ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Εξαγωγή Ονοματικών Οντοτήτων και Εμπλουτισμός Κειμένου με χρήση Σημασιολογικού Ιστού

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

των

Χρυσούλα Ζέρβα, Αλίκη Κοπανέλη

Επιβλέπων: Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2012

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Εξαγωγή Ονοματικών Οντοτήτων και Εμπλουτισμός Κειμένου με χρήση Σημασιολογικού Ιστού

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

των

Χρυσούλα Ζέρβα, Αλίκη Κοπανέλη

Επιβλέπων: Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 19η Ιουλίου 2012.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Ανδρέας-Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π.	Γεώργιος Στάμου Λέκτορας Ε.Μ.Π.	Στέφανος Κόλλιας Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2012

.....
Χρυσούλα Ζέρβα
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

.....
Αλίκη Κοπανέλη
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε κατά το ακαδημαικό έτος 2011-2012 στο Εθνικό Μετσόβιο Πολυτεχνείο. Θα θέλαμε να ευχαριστήσουμε τον επιβλέποντα καθηγητή κ. Ανδρέα-Γεώργιο Σταφυλοπάτη για την εμπιστοσύνη που μας έδειξε αναθέτοντάς μας την εργασία αυτή και για τη δυνατότητα που μας έδωσε να ασχοληθούμε με το συγκεκριμένο ενδιαφέρον θέμα. Επίσης, ευχαριστούμε θερμά τον κ. Γεώργιο Στάμου για την καθοδήγησή του καθ' όλη τη διάρκεια και την εξαιρετική συνεργασία που είχαμε. Ιδιαίτερες ευχαριστίες θα θέλαμε να απευθύνουμε προς τον δρ. Γεώργιο Σιόλα για τις υποδείξεις και την πολύτιμη συμβολή του σε όλα τα στάδια εκπόνησης της εργασίας. Τέλος, θα θέλαμε να ευχαριστήσουμε τις οικογένειες αλλά και όλους τους φίλους μας για την υπομονή και τη συμπαράστασή τους.

Περίληψη

Η παρούσα διπλωματική εργασία έχει ως αντικείμενο τη μελέτη και την ανάπτυξη δύο συστημάτων τα οποία επιδιώκουν τον εμπλουτισμό ακατέργαστων και αδόμητων κειμένων, γραμμένων σε φυσική γλώσσα, με χρήση Σημασιολογικού Ιστού και συγκεκριμένα των διασυνδεδεμένων δεδομένων της DBpedia. Καθοριστικής σημασίας κρίνεται ο εντοπισμός και η επιλογή μέσα από το κείμενο, μόνο εκείνων των φράσεων που αντιστοιχούν σε ονοματικές οντότητες της DBpedia και φέρουν την ανά περίπτωση επιθυμητή πληροφορία. Η εξαγωγή των οντοτήτων αυτών, δίνουν τη δυνατότητα άντλησης επιπρόσθετης πληροφορίας η οποία εμπλουτίζει το κείμενο με τον τρόπο που υπαγορεύει ο στόχος του κάθε συστήματος.

Το πρώτο σύστημα ονομάζεται "Σύστημα Σημασιολογικής Επισημείωσης και Εξαγωγής Συνοπτικής Αναπαράστασης Κειμένου" και προσανατολίζεται στην εξαγωγή των ονοματικών οντοτήτων από ένα δεδομένο κείμενο, το σύνολο των οποίων είναι ικανό να αποτελέσει μία επαρκή αναπαράστασή του. Συγκεκριμένα, μία αναπαράσταση θεωρείται αποδεκτή όταν συνοψίζει τις βασικές έννοιες του κειμένου και αρκεί για να το διαχωρίσει με σημασιολογικά κριτήρια από άλλα κείμενα. Μάλιστα, οι οντότητες που συνθέτουν την εν λόγω αναπαράσταση, παρέχονται από το σύστημα ταξινομημένες με βάση τη νοηματική βαρύτητα που θεωρείται πως έχει η κάθε μία για το εκάστοτε κείμενο. Για την ταξινόμηση και τη διαλογή των εντοπισμένων οντοτήτων χρησιμοποιούνται κριτήρια που βασίζονται σε δεδομένα αντλούμενα από τη Wikipedia και τη DBpedia. Η τελική αξιολόγηση των αποτελεσμάτων γίνεται με χρήση προσημειωμένων συνόλων κειμένων και των στατιστικών μεγεθών ακρίβειας και ανάκλησης.

Το δεύτερο σύστημα ονομάζεται "Σύστημα Ταυτοποίησης Προσώπων με χρήση Σημασιολογικού Ιστού" και αφορά τον εντοπισμό αναφορών σε πρόσωπα του πραγματικού κόσμου εντός ενός κειμένου. Στη συγκεκριμένη περίπτωση, γίνεται αναζήτηση στη γνωσιακή βάση της DBpedia προκειμένου να προσδιοριστεί ποιές από τις εντοπισμένες ονοματικές οντότητες πληρούν την παραπάνω συνθήκη με βάση τον τύπο δεδομένων που υποδηλώνει η σημασιολογία της κάθε μίας. Τα αποτελέσματα είναι ικανοποιητικά ως προς την ακρίβειά τους, σε σύγκριση και με υπάρχοντα συστήματα, ωστόσο περιορίζονται στον εντοπισμό οντοτήτων που είναι καταχωρημένες στη γνωσιακή βάση που χρησιμοποιήθηκε.

Λέξεις Κλειδιά: Σημασιολογικός Ιστός, γραμματική επισημείωση όρων, σημασιολογική επισημείωση όρων, διασυνδεδεμένα δεδομένα, γνωσιακή βάση δεδομένων, ταυτοποίηση προσώπων, ονοματική οντότητα, εξαγωγή ονοματικών οντοτήτων, RDF, DBpedia, Wikipedia

Abstract

The main object of the present thesis is the study and the development of two independent systems that attempt to enrich plain, unprocessed, natural language texts, using Semantic Web (Dbpedia Linked Data in particular). In the above mentioned procedure, the detection and extraction of the phrases that correspond to DBpedia's noun entities and "bear" the desired piece of information is of paramount importance. The extraction of these entities, facilitates the acquisition of extra related information, thus enriching the initial text according to the target of each system.

The first system, named "Condensed Representation Extraction and Semantic Text Annotation" - CRESTA, is oriented towards the extraction of a set of noun entities that can be considered an efficient representation of the input text. A representation is approved when succesful in summarising the fundamental text concepts and distinguishing its semantic context. In addition, CRESTA performs evaluation ranking over the above mentioned entities, based on their conceptual significance. The metrics necessary for the implementation of ranking and final selection procedures, are calculated using Wikipedia and DBpedia data. Pretagged corpora were used as an evaluation set for the observation of the CRESTA's performance, that was conducted based on precision and recall values.

The second system, named "Semantic Web based Person IDentification" - SWPID, aims at the detection of references to real world persons, within plain texts. For the purposes of this approach, the system queries Dbpedia knowledge database in order to identify the entities that fulfil the above mentioned condition, as indicated by their extracted semantics. The precision results are remarkably satisfactory compared to other tools performing the same task, however, the output results are obviously restricted to entities already included in the employed knowledge database.

Keywords: Semantic Web, Part of Speech tagging, semantic annotation, Linked Data, knowledge database, person identification, noun entity, noun entity extraction, RDF, DBpedia, Wikipedia

Περιεχόμενα

Κατάλογος πινάκων

Κατάλογος σχημάτων

Κεφάλαιο 1

Εισαγωγή

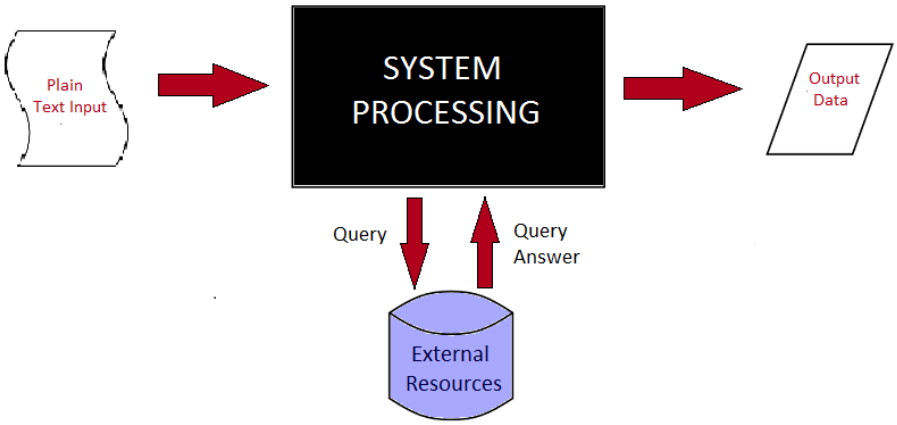
Τα τελευταία χρόνια, παρατηρείται μία διαρκώς αυξανόμενη συσσώρευση πληροφορίας στον Παγκόσμιο Ιστό. Στο μεγαλύτερο ποσοστό της, η πληροφορία αυτή, είναι αδόμητη και ανοργάνωτη τόσο στο περιεχόμενο όσο και στον τρόπο με τον οποίο αναρτάται στο διαδίκτυο. Επιπλέον, είναι προορισμένη για ανθρώπινη κατανάλωση (human consumption) καθιστώντας ανέφικτη την άμεση επεξεργασία της μέσω αυτοματοποιημένων διαδικασιών. Το αποτέλεσμα είναι να περιορίζεται σημαντικά η δυνατότητα εκμετάλλευσης της διαθέσιμης πληροφορίας, καθώς καθίσταται δύσκολη τόσο η αναζήτηση όσο και η διαχείρισή της. Το γεγονός αυτό, έστρεψε το ενδιαφέρον μεγάλου μέρους της επιστημονικής κοινότητας σε θεωρητικά μοντέλα, τεχνολογίες και εργαλεία που προσεγγίζουν τα ζητήματα αυτά συμβάλλοντας στην ανάπτυξη και επέκταση των επιστημονικών περιοχών της επεξεργασίας φυσικής γλώσσας (natural language processing-NLP), εξόρυξης κειμένου (text mining), εξαγωγή πληροφορίας (information extraction) και αναπαράστασης γνώσης (knowledge representation).

Η εξέλιξη των παραπάνω περιοχών καθώς και συνολικότερα της τεχνητής νοημοσύνης (artificial intelligence), έχει συμβάλλει στην ανάπτυξη μεθόδων αντιμετώπισης του προβλήματος της επεξεργασίας φυσικού ανθρώπινου λόγου από υπολογιστή και προσέγγισης του προβλήματος της κατανόησης φυσικού ανθρώπινου λόγου από τον υπολογιστή. Έτσι, δίνει τη δυνατότητα δημιουργίας συστημάτων τα οποία δέχονται ακατέργαστη πληροφορία και τη μετατρέπουν σε δομημένη (με τρόπο που υποδηλώνει τυπική σημασιολογία), διαχειρίσιμη από υπολογιστές. Παράλληλα, παρέχει τη δυνατότητα αυτόματης διασύνδεσης της πληροφορίας αυτής με άλλες αντίστοιχα δομημένες, οι οποίες είναι αναρτημένες στο διαδίκτυο. Μέσω τέτοιων διασυνδέσεων, όχι μόνο επιτυγχάνεται ο εμπλουτισμός της με δεδομένα εξωτερικών πηγών, αλλά παράλληλα εντάσσεται η πληροφορία αυτή σε ένα σύνολο οργανωμένων δεδομένων, συμβάλλοντας στην προσπάθεια μετατροπής του Παγκόσμιου Ιστού σε ένα οργανωμένο σύνολο που ονομάζεται ιστός δεδομένων (web of data). Στο πλαίσιο της διπλωματικής αυτής, αναπτύχθηκαν δύο συστήματα τα οποία προσεγγίζουν δύο διαφορετικές πτυχές του εν λόγω ζητήματος.

Βάση της λειτουργίας και των δύο συστημάτων, αποτελεί η χρήση του Σημασιολογικού Ιστού (Semantic Web), έννοια που διατυπώθηκε από τον Tim Berners-Lee το 1944 στο πρώτο

World Wide Web Conference, και έθεσε τις βάσεις για την πραγματοποίηση της ιδέας του Ιστού Δεδομένων, η οποία μέχρι τότε άγγιζε τα όρια της ουτοπίας. Ο Σημασιολογικός Ιστός αποτελεί ένα μοντέλο αναπαράστασης πληροφορίας, το οποίο διαχειρίζεται τα δεδομένα ως έννοιες (concepts) και επιχειρεί τη διασύνδεσή τους μέσω νοηματικών σχέσεων (ρόλων) και την οργάνωσή τους με ιεραρχικά κριτήρια. Η οργάνωση των εννοιών γίνεται με τη μορφή διασυνδεδεμένων μεταξύ τους γράφων και υπογράφων. Σήμερα, ο Σημασιολογικός Ιστός αποτελείται από εκατοντάδες υπογράφους. Τα τμήματά του που επιλέχθηκαν να χρησιμοποιηθούν, είναι όλα όσα συνδέονται άμεσα με τον γράφο της DBpedia. Η επιλογή αυτή έγινε καθώς ο συγκεκριμένος γράφος παρουσιάζει αφ' ενός το ευρύτερο φάσμα δεδομένων από πλευράς θεματολογίας, και αφ' ετέρου επαρκές πλήθος συνδέσμων προς άλλους γράφους. Τα πλεονεκτήματα της χρήσης του συγκεκριμένου μοντέλου στην ανάπτυξη των συστημάτων θα γίνουν εμφανή στη συνέχεια.

Πριν γίνει αναφορά, στο αντικείμενο κάθε συστήματος, στο παρακάτω διάγραμμα γίνεται μία γενική περιγραφή εισόδου εξόδου των δύο συστημάτων (Blackbox description).



Σχήμα 1.1: Γενική περιγραφή Συστημάτων

1.1 Σύστημα Σημασιολογικής Επισημείωσης και Εξαγωγής Συνοπτικής Αναπαράστασης Κειμένου

Το Σύστημα Σημασιολογικής Επισημείωσης και Εξαγωγής Συνοπτικής Αναπαράστασης Κειμένου (Condensed Representation Extraction and Semantic Text Annotation System - CRESTA), αποσκοπεί στην εξαγωγή μίας περιεκτικότερης αναπαράστασης ενός αδόμητου κειμένου (plain text), σημασιολογικά επισημειωμένης. Αναγκαία συνθήκη για να θεωρηθεί ένας όρος μέρος της αναπαράστασης, είναι η ύπαρξή του ως οντότητα (entity) στο Σημασιολογικό Ιστό. Με τον όρο περιεκτικότερη αναπαράσταση, εννοείται μία αναπαράσταση του κειμένου που θα διατηρεί

εκείνα μόνο τα δεδομένα τα οποία αρκούν για την περιγραφή της πληροφορίας που περιέχεται σε αυτό. Τα δεδομένα αυτά, είναι όροι (λέξεις ή φράσεις) του κειμένου, που θα πρέπει να παρουσιάζουν νοηματική συνοχή με το υπόλοιπο κείμενο, να είναι συγκεκριμένοι ως προς το θέμα του και να αρκούν για τη διάκρισή του από άλλα κείμενα ή δεδομένα με διαφορετικό περιεχόμενο.

Κατά την εξαγωγή των όρων με τα χαρακτηριστικά που περιγράφηκαν, θεωρείται καταλυτικής σημασίας η χρήση του Σημασιολογικού Ιστού, προκειμένου να έχουμε τα επιθυμητά αποτελέσματα. Χωρίς τη χρήση του, θα ήταν εφικτή η ανεύρεση όρων είτε εντός είτε εκτός του κειμένου που θα πληρούσαν αρκετές από τις προϋποθέσεις ένταξής τους σε μία περιεκτική αναπαράσταση του κειμένου ή του θέματός του, όπως έχει γίνει κατά το παρελθόν. Είναι, δηλαδή, δυνατή η χρήση διαφορετικών εξωτερικών πηγών πληροφορίας, όπως προσημειωμένα σύνολα δεδομένων (pre-tagged datasets), λεξικογραφικές βάσεις δεδομένων (lexical databases), οργανωμένες συλλογές δεδομένων όπως η Wikipedia κλπ. Με χρήση των παραπάνω, μέσω στατιστικών μεθόδων και τεχνικών μηχανικής μάθησης (machine learning), είναι δυνατός ο εντοπισμός και σε ορισμένες περιπτώσεις η κατάταξη όρων με νοηματική βαρύτητα ως προς το κείμενο. Με τον τρόπο αυτόν, παρέχεται μεν στο χρήστη σημαντική πληροφορία, την οποία, όμως, δεν θα μπορέσει να επεξεργαστεί, παρά μόνο εάν επιστρατεύσει τις απαιτούμενες γνώσεις και εμπειρία. Το γεγονός αυτό οφείλεται στο ότι κάθε ένας από τους προαναφερθέντες όρους αποτελεί μία απλή συμβολοσειρά (string). Ακόμα κι αν για κάθε επιλεγμένο όρο παρέχεται επιπρόσθετη πληροφορία, η πληροφορία αυτή θα συνεχίσει να είναι σύνολο συμβολοσειρών, αξιοποιήσιμη μόνο μέσω ανθρώπινης παρέμβασης.

Αντίθετα, διασφαλίζοντας πως κάθε όρος θα αντιστοιχεί σε μία οντότητα του Σημασιολογικού Ιστού και παρέχοντας στο χρήστη το σύνδεσμο σε αυτή γίνεται εφικτή η πρόσβαση σε ολόκληρο το σημασιολογικό πλαίσιο (semantic context) του όρου. Πλέον, για κάθε όρο, έχει γίνει αυτόματη σημασιολογική μετάφραση και από απλή συμβολοσειρά, έχει μετατραπεί σε οντότητα που αναπαριστά έννοια του κόσμου. Η παραπάνω μετάφραση είναι ιδιαίτερα σημαντική εξαιτίας της δυνατότητας που παρέχει αξιοποίησης του κάθε όρου αλλά και ολόκληρου του κειμένου από αυτοματοποιημένες διαδικασίες και εφαρμογές. Εύκολα, λοιπόν, ένα κείμενο που αναφέρεται σε κάποιο ειδικό ιατρικό ζήτημα, κατατάσσεται αυτόματα στην κατηγορία του ιατρικού κλάδου στον οποίο ανήκει, ακόμα και αν αυτός δεν αναφέρεται ρητά εντός του κειμένου. Έτσι, το αρχικό κείμενο μετατρέπεται σε αξιοποιήσιμη υπολογιστικά πηγή πληροφορίας. Επιπλέον, μέσω των συνδέσμων που παρέχονται για κάθε όρο, επιτυγχάνεται το πρώτο στάδιο για την ανάρτηση της πληροφορίας που περιέχεται στο κείμενο στον αντίστοιχο γράφο, η οποία συνεπάγεται την επέκταση του ίδιου του Σημασιολογικού Ιστού.

Βάσει των προαναφερθέντων, το σύστημα δέχεται στην είσοδό του ένα κείμενο (γραπτό ανθρώπινο λόγο), και αντλώντας πληροφορίες από εξωτερικές πηγές (Wikipedia και DBpedia), το επεξεργάζεται ώστε να εξάγει μία λίστα που θα περιλαμβάνει όρους του κειμένου ιεραρχημένους σύμφωνα με τη νοηματική τους βαρύτητα ως προς το κείμενο και τους συνδέσμους των όρων αυτών προς την DBpedia και προς άλλες εξωτερικές πηγές. Η επεξεργασία αναλύεται σε δύο διακριτά τμήματα σε σειρά μεταξύ τους. Το πρώτο αφορά όλες τις διαδικασίες εξαγωγής

από το κείμενο των όρων που πληρούν τις απαραίτητες προϋποθέσεις ένταξης στην αναπαράσταση της εξόδου. Στις διαδικασίες αυτές, περιλαμβάνονται γραμματική επισημείωση (part of speech tagging), εξαγωγή ονοματικών φράσεων (noun phrase extraction) και ταυτοποίηση της σημασιολογίας των φράσεων αυτών με οντότητες της DBpedia ώστε να ικανοποιείται η απαίτηση για διασύνδεση του κειμένου με το Σημασιολογικό Ιστό. Το δεύτερο τμήμα αφορά τις διαδικασίες αξιολόγησης των φράσεων αυτών, ώστε να εξασφαλισθεί ότι οι όροι που παρέχονται στο χρήστη παρουσιάζουν τη μέγιστη δυνατή νοηματική εγγύτητα με το θέμα του αρχικού κειμένου και να επιτευχθεί η ταξινόμησή τους με βάση αυτή. Η αξιολόγηση γίνεται με βάση τα στοιχεία που αντλούνται από το κείμενο, τη Wikipedia και την DBpedia.

1.2 Σύστημα Ταυτοποίησης Προσώπων με χρήση Σημασιολογικού Ιστού

Το Σύστημα Ταυτοποίησης Προσώπων με χρήση Σημασιολογικού Ιστού (Semantic Web based Person IDentification System - SWPID), στοχεύει στην αυτόματη αναγνώριση όρων ενός δεδομένου κειμένου που αντιστοιχούν σε οντότητες με μία συγκεκριμένη κοινή ιδιότητα. Αυτό σημαίνει πως κάθε τέτοιος όρος φέρει συγκεκριμένου τύπου πληροφορία. Χαρακτηριστικά παραδείγματα αποτελούν όροι που αντιπροσωπεύουν ημερομηνίες, τοποθεσίες, πρόσωπα του πραγματικού κόσμου, γνωστά έργα τέχνης, οργανισμούς κλπ.

Η αναγνώριση τέτοιων τύπων δεδομένων αποτελεί έναν από τους βασικότερους στόχους του κλάδου εξαγωγής πληροφορίας. Οι τεχνικές εξόρυξης κειμένου που επιστρατεύονται από το συγκεκριμένο κλάδο για την επίτευξη του παραπάνω στόχου, αποσκοπούν σε πρώτο επίπεδο στην αναγνώριση της μορφής των τύπων των δεδομένων που έχουν την επιθυμητή ιδιότητα και σε δεύτερο επίπεδο στην αναγνώριση της πληροφορίας που παρέχεται γι' αυτά, αποκλειστικά μέσα από το κείμενο. Για παράδειγμα, από την απλή πρόταση: "Ο Albert Einstein γεννήθηκε στην πόλη Ulm" οι υπάρχουσες τεχνικές εξόρυξης κειμένου δύνανται να επιτύχουν την αναγνώριση του όρου Albert Einstein ως αναπαράσταση προσώπου, τον όρο Ulm ως αναπαράσταση τοποθεσίας αλλά και τη μεταξύ τους σύνδεση μέσω του ρήματος "γεννήθηκε". Αν και τα παραπάνω συμπεράσματα είναι πολύ χρήσιμα, υστερούν στα εξής σημεία: Δεν παρέχουν καμία σημασιολογική πληροφορία για τους όρους που αναγνωρίστηκαν, πέραν του τύπου δεδομένων που αντιπροσωπεύουν (η λέξη "γεννήθηκε" είναι μία συμβολοσειρά αναγνωρισμένη ως ρήμα που συνδέει τους δύο όρους και απαιτείται η συμβολή ανθρώπινης γνώσης για την σημασιολογική μετάφρασή της). Επιπλέον, ο όρος Albert Einstein ταιριάζει στο πρότυπο ονοματεπώνυμο αλλά δεν υπάρχει δυνατότητα εξακρίβωσης της αντιστοίχισης του ονόματος αυτού σε πρόσωπο του πραγματικού κόσμου. Τέλος, η άντληση πληροφορίας για τους συγκεκριμένους όρους, περιορίζεται αποκλειστικά σε ό,τι αναφέρεται στο κείμενο.

Η προσέγγιση που υιοθετήθηκε στο πλαίσιο του παρόντος συστήματος επικεντρώνεται στην επίλυση των τριών παραπάνω προβλημάτων με τη χρήση Σημασιολογικού Ιστού. Σε αυτή την περίπτωση, η επιλογή των όρων εκείνων που εξάγονται από το κείμενο και η αναγνώριση

της ιδιότητάς τους γίνεται μέσω της αντιστοίχισης σε οντότητες της dbpedia. Συνεπώς, εξασφαλίζεται το γεγονός ότι ο κάθε όρος αναπαριστά έννοια του πραγματικού κόσμου που αντιστοιχεί στη ζητούμενη πληροφορία. Μάλιστα, παρέχοντας το σύνδεσμο προς την οντότητα, και δεδομένου ότι κάθε μία συνδέεται με τις υπόλοιπες μέσω νοηματικών σχέσεων, προσδίδεται σημασιολογική υπόσταση στη σχέση του όρου με οποιονδήποτε άλλον χωρίς ανθρώπινη παρέμβαση. Η σχέση αυτή δεν επιδιώκεται να αντληθεί μέσα από το κείμενο αλλά μόνο μέσα από το Σημασιολογικό Ιστό, εφόσον αναπαρίσταται σε αυτόν. Η προαναφερθείσα ιδιότητα του Σημασιολογικού Ιστού, παρέχει τη δυνατότητα άντλησης πληροφορίας επιπλέον του κειμένου για κάθε αναγνωρισμένο όρο.

Για το δεύτερο σύστημα, επιλέχθηκε προς αναγνώριση ο τύπος πληροφορίας: Πρόσωπα του πραγματικού κόσμου. Κατά τη λειτουργία του, λοιπόν, το σύστημα δέχεται στην είσοδό του ένα κείμενο και το επεξεργάζεται ανατρέχοντας στη dbpedia με στόχο την εξαγωγή εκείνων των όρων του κειμένου που αντιπροσωπεύουν πρόσωπα του πραγματικού κόσμου και του σύνδεσμού τους στην αντίστοιχη οντότητα της dbpedia. Πιο συγκεκριμένα, η επεξεργασία αναλύεται σε δύο τμήματα. Το πρώτο, αφορά όλες εκείνες τις διαδικασίες εξαγωγής των πιθανών όρων-στόχων και συμπίπτει σχεδόν εξ ολοκλήρου με το αντίστοιχο τμήμα του πρώτου συστήματος που περιγράφηκε παραπάνω. Το δεύτερο τμήμα, αφορά την απόπειρα αναζήτησης της ιδιότητας ”είναι άνθρωπος” για κάθε εντοπισμένη οντότητα και την προώθηση στην έξοδο εκείνων μόνο των όρων που αντιστοιχήθηκε επιτυχημένα στη συγκεκριμένη ιδιότητα.

Η χρήση του Σημασιολογικού Ιστού με τον προαναφερθέντα τρόπο επιβάλλει την κατά σύμβαση θεώρηση πως ο πραγματικός κόσμος μπορεί να αναπαρασταθεί πλήρως από τις οντότητες και τους ρόλους του τμήματος του ιστού που επιλέχθηκε. Η σύμβαση αυτή υπαγορεύει την επιλογή τμήματος που αναπαριστά μεγάλο όγκο δεδομένων, ώστε να είναι δυνατή η άντληση όσο το δυνατόν περισσότερων όρων, ανεξαρτήτως θεματολογίας. Κρίθηκε πως η DBpedia ανταποκρίνεται ικανοποιητικά στην απαίτηση αυτή. Ωστόσο, είναι προφανές πως ακόμα και σε αυτή την περίπτωση, αναμένεται ένα ποσοστό αποτυχίας στην αναγνώριση όρων που αν και πληρούν τις λοιπές προϋποθέσεις ένταξής τους στα αποτελέσματα εξόδου, δεν αντιστοιχούν σε οντότητα της DBpedia.

1.3 Διάρθρωση του κειμένου

Στα κεφάλαια που ακολουθούν, αναλύονται περαιτέρω τα παραπάνω συστήματα, καθώς και παρουσιάζεται το επιστημονικό και τεχνολογικό υπόβαθρου στο οποίο βασίστηκε η ανάπτυξή τους. Συγκεκριμένα:

- Στο δεύτερο κεφάλαιο με τίτλο ”Τεχνολογικό Υπόβαθρο και Σχετικές Εργασίες” αναλύονται θεωρίες, μέθοδοι και εργαλεία που έχουν αναπτυχθεί στους τομείς της επεξεργασίας φυσικής γλώσσας, της εξόρυξης κειμένου και της εξαγωγής πληροφορία. Επιπλέον, παρουσιάζονται τεχνολογίες και εφαρμογές που αφορούν το Σημασιολογικό Ιστό και την μέχρι σήμερα επέκτασή του, καθώς και επιλεγμένες από τη βιβλιογραφία εργασίες, που

θεωρήθηκε πως έχουν ιδιαίτερο ενδιαφέρον ως προς την κατανόηση, το σχεδιασμό και την υλοποίηση της παρούσας διπλωματικής εργασίας. Στόχος του κεφαλαίου είναι η ένταξη της εργασίας στο επιστημονικό και ερευνητικό της πλαίσιο και η παρουσίαση των ζητημάτων που συνέβαλλαν στην ανάπτυξη και ολοκλήρωσή της.

- Στο τρίτο κεφάλαιο με τίτλο ”Γενική Περιγραφή Συστήματος CRESTA”, γίνεται μία αναλυτική περιγραφή του σχεδιασμού της δομής του συγκεκριμένου συστήματος καθώς και των κανόνων και αρχών λειτουργίας που το διέπουν.
- Στο τέταρτο κεφάλαιο με τίτλο ”Γενική Περιγραφή Συστήματος SWPID”, γίνεται μία περιγραφή του σχεδιασμού της δομής του συγκεκριμένου συστήματος καθώς και των κανόνων και αρχών λειτουργίας που το διέπουν, η οποία επικεντρώνεται στα ζητήματα που διαφοροποιούνται σε σχέση με το πρώτο σύστημα.
- Στο πέμπτο κεφάλαιο με τίτλο ”Ζητήματα Υλοποίησης” περιγράφονται επιλογές και λεπτομέρειες υλοποίησης οι οποίες κρίνονται καθοριστικής σημασίας για τη λειτουργία και την απόδοση των συστημάτων, καθώς και για τη μορφή των αποτελεσμάτων τα οποία εξάγουν. Επιπλέον, γίνεται ανάλυση των εργαλείων που χρησιμοποιήθηκαν προκειμένου να γίνει κατανοητός ο τρόπος λειτουργίας τους καθώς και ο ρόλος τους στο πλαίσιο της διπλωματική εργασίας.
- Στο έβδομο κεφάλαιο με τίτλο ”Αξιολόγηση-Συμπεράσματα” παρουσιάζονται και σχολιάζονται τα αποτελέσματα των συστημάτων με τελικό στόχο την αξιολόγησή τους και την ανάδειξη ζητημάτων που επιδέχονται περαιτέρω μελέτης και βελτίωσης.

Κεφάλαιο 2

Τεχνολογικό Υπόβαθρο και Σχετικές Εργασίες

2.1 Επεξεργασία Φυσικής Γλώσσας

Η επεξεργασία φυσικής γλώσσας είναι ένας τομέας έρευνας και εφαρμογών που ασχολείται με την επεξεργασία και κατανόηση φυσικώς παραγόμενου ανθρώπινου κειμένου ή λόγου από υπολογιστές. Ερευνητές του τομέα αυτού συγκεντρώνουν πληροφορίες για το πώς επεξεργάζεται το ανθρώπινο μυαλό τη γλωσσική πληροφορία, με στόχο την ανάπτυξη εργαλείων και εφαρμογών βασισμένων σε τεχνικές που θα μιμούνται τις ανθρώπινες συμπεριφορές στο συγκεκριμένο ζήτημα [20]. Για το λόγο αυτό θα μπορούσε να πει κανείς ότι η επεξεργασία φυσικής γλώσσας είναι ένα πεδίο της περιοχής της τεχνητής νοημοσύνης παρότι σίγουρα χρησιμοποιεί γνώση και από πολλές άλλες επιστημονικές περιοχές όπως τη γλωσσολογία, τα μαθηματικά, την ηλεκτρονική, ακόμα και την ψυχολογία. Στόχος της επεξεργασίας φυσικής γλώσσας δεν είναι απλά η ανάλυση ενός κειμένου σε πρωτογενές επίπεδο αλλά η κατανόησή του από υπολογιστή[6]. Αυτός όμως ο στόχος είναι δύσκολο να επιτευχθεί και μέχρι τώρα έχουν γίνει προσπάθειες που επιλύουν ένα μέρος του προβλήματος. Αντίθετα σε αρχικά στάδια επεξεργασίας ενός κειμένου-κυρίως συντακτικής και γραμματικής ανάλυσης- έχουν γίνει πολύ σημαντικά βήματα προς την πλήρη επίλυση. Κάποια από τα απλά προβλήματα που έχουν επιλυθεί σε ικανοποιητικό βαθμό είναι[20]:

- κατάτμιση σε λεκτικές μονάδες, (tokenization)[18], διαδικασία κατά την οποία δεδομένου ενός κειμένου εξάγονται όροι κάθε ένας εκ των οποίων αποτελεί λεκτική μονάδα (token) όπως για παράδειγμα λέξεις.
- κατάτμιση σε προτάσεις (sentence segmentation), χωρισμός ενός κειμένου σε προτάσεις
- συντακτική ανάλυση (parsing), συντακτική ανάλυση ενός κειμένου με βάση μία δεδομένη τυπικά ορισμένη γραμματική.

- χαμηλού επιπέδου συντακτική ανάλυση (chunking), χαμηλού επιπέδου συντακτική ανάλυση κατά το οποίο αναγνωρίζονται ουσιαστικά, συγκροτήματα ουσιαστικών, ρήματα, ρηματικές φράσεις αλλά δεν προσδιορίζεται η συντακτική τους θέση σε μία πρόταση.

- γραμματική επισημείωση [22], γραμματικός προσδιορισμός κάθε λέξης του κειμένου.

- περιστολή λέξεων (stemming), η αναγωγή μιας λέξης στη ρίζα της. Στόχος αυτής της διαδικασίας δεν είναι απαραίτητα η αναγωγή στη γραμματική ρίζα της λέξης αλλά η ταυτοποίηση λέξεων που ενώ έχουν την ίδια ρίζα, εμφανίζονται σε διαφορετικές μορφές σε ένα κείμενο.

- εξαγωγή ονοματικών φράσεων [5], εξαγωγή από κείμενο, των φράσεων που έχουν ως σημείο αναφοράς ένα ή παραπάνω ουσιαστικά.

Οι παραπάνω εργασίες, συνήθως εμφανίζονται ως δευτερεύουσες εργασίες κάποιου γενικότερου στόχου η επίτευξη του οποίου δεν θα ήταν δυνατή χωρίς την επίλυση τέτοιων υποπροβλημάτων. Παράδειγμα ενός τέτοιου γενικού στόχου είναι η μετάφραση κειμένου από υπολογιστή, που αποτελεί και ένα από τα πιο δύσκολα προβλήματα στον τομέα της επεξεργασίας φυσικής γλώσσας αλλά και της τεχνητής νοημοσύνης γενικότερα. Αυτό συμβαίνει γιατί η μετάφραση ενός κειμένου απαιτεί τη συνολική γνώση που έχει ο άνθρωπος και χρησιμοποιεί όταν διαβάζει και μεταφράζει ένα κείμενο, όπως γνώση γραμματικής, σημασιολογικής ερμηνείας των λέξεων, γλωσσικών ιδιομορφιών, αντικειμένων του πραγματικού κόσμου κλπ. Ένα τέτοιο πρόβλημα βρίσκεται στην περιοχή της κατανόησης φυσικής γλώσσας (natural language understanding) καθώς απαιτεί και τη σημασιολογική ανάλυση του κειμένου. Άλλα προβλήματα με αντίστοιχη δυσκολία είναι [17]:

- αναγνώριση ονοματοδοτημένων οντοτήτων (named entity recognition), αναγνώριση των κύριων ονομάτων ενός κειμένου και του τύπου αυτών (πχ. όνομα ανθρώπου, τοποθεσία κλπ)

- εξαγωγή σχέσεων (relationship extraction), αναγνώριση των σχέσεων μεταξύ κυρίων ονομάτων ενός κειμένου.

- αυτόματη εξαγωγή περίληψης (automatic summarisation), διαδικασία κατά την οποία δεδομένου ενός κειμένου παράγεται η περίληψή του σε μορφή αναγνώσιμη από άνθρωπο.

- θεματική κατάτμιση (topic segmentation) και αναγνώριση θέματος (topic recognition), χωρισμός ενός κειμένου σε σημασιολογικώς συνεκτικά τμήματα και αναγνώριση του θέματος του κάθε τμήματος.

Στον τομέα της επεξεργασίας φυσικής γλώσσας μπορεί να θεωρηθεί και ότι ανήκουν τα πεδία της ανάκτησης πληροφορίας (information retrieval-IR) και της εξαγωγής πληροφορίας [6]. Οι δύο αυτές κατηγορίες εργασιών σχετίζονται άμεσα με την επεξεργασία φυσικής γλώσσας αλλά εκτείνονται και έξω από τα όρια αυτής. Η ανάκτηση πληροφορίας έχει να κάνει με

αναζήτηση, ανάκτηση και αποθήκευση πληροφορίας προερχόμενης από κείμενα, μεταδεδομένα από κείμενα, τον Παγκόσμιο Ιστό ή και συσχετιστικές βάσεις. Η εξαγωγή πληροφορίας ασχολείται με αναζήτηση σημασιολογικής πληροφορίας από κείμενο. Και ο δύο κατηγορίες, αν και αποτελούν ξεχωριστό τμήμα του τομέα "επιστήμη των υπολογιστών (computer science)", χρησιμοποιούν τεχνικές επεξεργασίας φυσικής γλώσσας προκειμένου να αναλύσουν ανθρώπινο γραπτό λόγο. Ένας άλλος σχετικός κλάδος είναι της εξόρυξης κειμένου. Ασχολείται με την άντληση χρήσιμης πληροφορίας από κείμενα, ανήκει στον τομέα τεχνολογίας ανθρώπινης γλώσσας και σχετίζεται άμεσα με την εξαγωγή πληροφορίας.

2.1.1 Γραμματική Επισημείωση

Η γραμματική επισημείωση σε ένα κείμενο είναι διαδικασία κατά την οποία αναγνωρίζεται κάθε λεκτική μονάδα του κειμένου ως μέρος του λόγου και σημειώνεται η ιδιότητα αυτή δίπλα στη λεκτική μονάδα[22]. Έτσι, η έξοδος ενός προγράμματος που υλοποιεί αυτή τη διαδικασία (Part of speech tagger) είναι οι λεκτικές μονάδες του κειμένου και από ένα tag για την κάθε μία που προσδιορίζει αυτή την ιδιότητά της. Με τον όρο tag γίνεται αναφορά σε μία ακολουθία χαρακτήρων που συμβολίζουν τα μέρη του λόγου. Υπάρχουν πολλά διαφορετικά συστήματα ορισμού των μερών του λόγου, άρα και συμβολισμού τους. Για παράδειγμα, υπάρχουν διαφορετικοί τρόποι αντιμετώπισης σημείων στίξης, ειδικών λέξεων κλπ. Πρέπει να σημειωθεί πως τα tags δεν προσδιορίζουν μέρος του λόγου όπως ο όρος αυτός αναφέρεται ως όρος της γλωσσολογίας. Αποσκοπούν σε μία πιο γενική αναγνώριση του συντακτικού και γραμματικού ρόλου που έχει η κάθε λεκτική μονάδα στην πρόταση, ώστε να είναι πιο εύκολα διαχειρίσιμη για περεταίρω ανάλυση του κειμένου.

Η γραμματική επισημείωση χρησιμοποιείται συνήθως ως ένα πρώτο επίπεδο επεξεργασίας κειμένου το οποίο θα χρησιμεύσει σε πιο σύνθετες εργασίες που ανήκουν στον τομέα της επεξεργασίας φυσικής γλώσσας, όπως αναλυτικότερη συντακτική ανάλυση, σημασιολογική ανάλυση, μεταφράσεις κλπ. Ένας σημαντικός παράγοντας που συνέβαλε στην πρόοδο για την γραμματική επισημείωση και στη δημιουργία αντίστοιχων εργαλείων επισημείωσης με αρκετά καλή ακρίβεια, είναι η ύπαρξη μεγάλων σωμάτων δεδομένων κάθε ένα από τα οποία ονομάζεται επισημειωμένο σύνολο δεδομένων (tagged corpus), δηλαδή μεγάλου όγκου κειμένων κάθε λέξη των οποίων ακολουθείται από μία ετικέτα (tag) γραμματικής επισημείωσης [22]. Η δημιουργία αυτών των σωμάτων έγινε κυρίως με μη αυτοματοποιημένες διαδικασίες με συμβολή ειδικών στον τομέα της γλωσσολογίας και με περιορισμένη χρήση κανόνων από στατικούς αυτόματους συντακτικούς αναλυτές.

Μέχρι σήμερα, η ακρίβεια των εργαλείων επισημείωσης έχει φτάσει μέχρι και το 96%, ποσοστό που αν και είναι αρκετά ικανοποιητικό, μπορεί να δημιουργήσει προβλήματα σε δεύτερο και τρίτο επίπεδο επεξεργασίας μετά το τέλος της επισημείωσης. Η δυσκολία επίτευξης μεγάλης ακρίβειας στη γραμματική επισημείωση οφείλεται σε δύο ανασταλτικούς παράγοντες [22]. Ο πρώτος έχει να κάνει με την ύπαρξη λέξεων με διφορούμενα νοήματα. Τέτοιες λέξεις που μπορούν να έχουν παραπάνω από μία σημασιολογικές ερμηνείες είναι πολύ δύσκολο να ανα-

γνωριστούν από τον υπολογιστή, καθώς θα μπορούσαν να εμφανιστούν με διαφορετικά tags σε διαφορετικές προτάσεις. Για τον άνθρωπο, γνωρίζοντας τη σημασιολογία της εκάστοτε λέξης αλλά και τα συμφραζόμενα, η επιλογή της σωστής ετικέτας είναι αρκετά πιο εύκολη.

Ο δεύτερος ανασταλτικός παράγοντας είναι άγνωστες για το εργαλείο γραμματικής επισημείωσης λέξεις. Πρόκειται για λέξεις τις οποίες δεν συνάντησε κατά την εκπαίδευσή του (δεν βρίσκονταν στο επισημειωμένο σύνολο δεδομένων που χρησιμοποιήθηκε) και κατ'επέκταση δεν ξέρει πώς θα τις αντιμετωπίσει.

Παρακάτω θα γίνει αναφορά σε κάποιες από τις σημαντικότερες μεθόδους που έχουν χρησιμοποιηθεί για γραμματική επισημείωση[22]:

- επισημείωση βασισμένη σε κανόνες (rule based tagging).
  Πρόκειται για μέθοδο που στηρίζεται σε κανόνες για την αναγνώριση του γραμματικού ρόλου των λέξεων. Οι κανόνες αυτοί απαιτούν τη βοήθεια ειδικών, αλλά και πολλές ώρες ανθρώπινης εργασίας ώστε να αναπτυχθούν και να γραφούν χειροκίνητα. Για το λόγο αυτό, συνολικά αυτές οι μέθοδοι, δεδομένης και της χαμηλής ακρίβειάς τους άρχισαν να αντικαθίστανται από μεθόδους μηχανικής μάθησης. Μία πολύ ενδιαφέρουσα εφαρμογή αυτής της προσέγγισης είναι η ανάπτυξη συστημάτων μάθησης βασισμένης σε κανόνες αλλά η μάθηση των κανόνων αυτών γίνεται αυτοματοποιημένα όπως θα περιγραφεί στη συνέχεια.

- Μάθηση βασισμένη σε μετασχηματισμούς.
  Πρόκειται για μέθοδο μάθησης κανόνων με χρήση "σωστά επισημειωμένου" σώματος δεδομένων. Η μέθοδος αυτή αποτελείται από δύο στάδια στα οποία χρησιμοποιούνται δύο διαφορετικά είδη δεδομένων. Το πρώτο είναι σώμα δεδομένων οι λεκτικές μονάδες του οποίου έχουν περάσει από διαδικασία επισημείωσης και είναι σωστά προσδιορισμένη η κάθε λέξη και το δεύτερο είναι σώμα δεδομένων στις λεκτικές μονάδες του οποίου αναθέτονται τυχαία tags κατά το πρώτο στάδιο της διαδικασίας μάθησης. Στη συνέχεια, στη δεύτερη φάση, χρησιμοποιείται ένα σύνολο από προκαθορισμένους κανόνες-μετασχηματισμούς. Οι μετασχηματισμοί αυτοί εφαρμόζονται στα τυχαίως επισημειωμένα δεδομένα και στη συνέχεια γίνεται σύγκριση των αποτελεσμάτων με βάση τα σωστά επισημειωμένα δεδομένα. Αφού εφαρμοστούν όλοι οι μετασχηματισμοί, επιλέγεται αυτός ο οποίος μειώνει περισσότερο το ποσοστό λάθους. Η διαδικασία επαναλαμβάνεται μέχρι να μην υπάρχει κανένας δυνατός μετασχηματισμός που μειώνει το λάθος κάτω από ένα συγκεκριμένο όριο που έχει επιλεχθεί. Έτσι, με το τέλος της εκπαίδευσης, το σύστημα έχει μάθει μετασχηματισμούς-κανόνες τους οποίους και θα εφαρμόσει στη συνέχεια για την γραμματική επισημείωση άλλων σωμάτων δεδομένων.

- Μέδοδος μαρκοβιανού μοντέλου.
  Τα εργαλεία που χρησιμοποιούν αυτή τη μέθοδο (με τις διάφορες παραλλαγές υλοποίησης) δεν βασίζονται πια σε κανόνες μετάβασης, αλλά σε στατιστικά δεδομένα που προκύπτουν από το επισημειωμένο σύνολο δεδομένων που χρησιμοποιείται. Σε γενικό πλαίσιο, στην προσέγγιση αυτή, θεωρούμε τυχαίεσ μεταβλητές X_1,...,X_T που παίρνουν τιμές από

ένα σύνολο καταστάσεων $S=\{s_1,...,s_N\}$. Οι τιμές της κάθε μεταβλητής δεν εξαρτώνται από τις τιμές των προηγούμενων μεταβλητών, παρά μόνο από την τιμή της αμέσως προηγούμενης. Επιπλέον, το σύστημα είναι στατικό-χρονικά αμετάβλητο. Επιπλέον, η πιθανότητα μετάβασης από την κατάσταση i στην κατάσταση j ορίζεται ως: $a_{ij}=P(X_{t+1}=s_j|X_t=s_i)$ Ισχύει $\Sigma a_{ij}=1$. Το σύνηθες μοντέλο εκπαίδευσης που χρησιμοποιείται σε αυτή την κατηγορία γραμματικής επισημείωσης είναι το Hidden Markov Model το οποίο αποτελεί μία όχι τόσο αυστηρή προσέγγιση του Markov Model καθώς επιτρέπεται πάνω από ένα μονοπάτι εξόδου από μία κατάσταση. Τυπικά, αυτό το μοντέλο αποτελείται από μία τετράδα στοιχείων: $<s_i, S, W, E>$
όπου s_i η αρχική κατάσταση, S το σύνολο των επιτρεπτών καταστάσεων, W το σύνολο των συμβόλων που χρησιμοποιούνται για κάθε μετάβαση, E το σύνολο των μεταβάσεων από κάθε κατάσταση στην επόμενη, κάθε στοιχείο του οποίου περιλαμβάνει και την πιθανότητα μετάβασης. Κατά την εφαρμογή ενός τέτοιου μοντέλου στην εργασία για γραμματική επισημείωση χρειάζεται να γίνει συσχετισμός ανάμεσα στις καταστάσεις και τα Part of Speech tags. Στην ενότητα 4.1.1 περιγράφεται αναλυτικά η λειτουργία του POS tagger που χρησιμοποιήθηκε στο πλαίσιο της διπλωματικής εργασίας.

- Άλλες στοχαστικές μέθοδοι που έχουν χρησιμοποιηθεί είναι οι μεγιστη εντροπία (maximum entropy), μηχανές διανυσμάτων υποστήριξης (support vector machines), νευρωνικά δίκτυα και στατιστικά δέντρα αποφάσεων (statistical decision trees).

Συνήθως, σε κάθε περίπτωση, η βασική αρχιτεκτονική είναι κοινή και περιλαμβάνει τα εξής στάδια [27]:

1. Κατάτμιση σε λεκτικές μονάδες.
2. Αναζήτηση όρων σε πηγές (λεξικά).
3. Αποσαφήνιση διφορούμενων όρων.

Συνολικά, στόχος όλων των συστημάτων που επιτελούν γραμματική επισημείωση είναι η επίτευξη του επιθυμητού αποτελέσματος με τη μικρότερη δυνατή ανθρώπινη, εξειδικευμένη εργασία. Γι' αυτό και η ανάπτυξη τέτοιων προγραμμάτων στηρίζεται πλέον μόνο σε μεθόδους μάθησης. Σε όλη αυτή την ερευνητική διαδικασία, είναι πολύ σημαντικό να αναφερθεί πως το σπουδαιότερο ρόλο επιτελούν τα κατασκευασμένα επισημειωμένα σώματα δεδομένων, τα οποία είναι και η βάση της εκπαίδευσης. Από τα σπουδαιότερα που χρησιμοποιούνται μέχρι σήμερα για την αγγλική γλώσσα είναι τα: Brown corpus [12] [1], Bank of English[2], και Penn Treebank[21].

2.1.2 Εξαγωγή Ονοματικών Φράσεων

Η εξαγωγή ονοματικών φράσεων από ένα κείμενο είναι μία πολύ μεγάλης σημασίας εργασία στον τομέα της επεξεργασίας φυσικής γλώσσας. Στις ονοματικές φράσεις βρίσκεται το

μεγαλύτερο μέρος της πληροφορίας που λαμβάνει ο άνθρωπος διαβάζοντας ένα κείμενο. Αυτό είναι προφανές αν σκεφτεί κανείς πως μία απλή, κύρια πρόταση έχει, συνήθως, τη δομή [45]:

Subject-Predicate

όπου το κατηγόρημα αποτελείται είτε από ένα ρήμα μόνο του είτε από ένα ρήμα και ένα ή παραπάνω αντικείμενα ή ακόμα και προσδιορισμούς διαφόρων τύπων.

Σε κάθε περίπτωση, η πρόταση χρησιμοποιείται για να δηλώσει την ενέργεια που εκτελεί ένα υποκείμενο (subject) σε ένα αντικείμενο (object), ή για να δηλώσει μία ενέργεια του υποκειμένου (αμετάβατο ρήμα), ή για να δηλώσει την κατάσταση στην οποία βρίσκεται ένα υποκείμενο. Η πρόταση, επομένως, προσδίδει ιδιότητα στο υποκείμενο η οποία μπορεί να εκφράζεται ώς ενέργεια, ως σχέση του υποκειμένου με το αντικείμενο κλπ. Δεδομένων των παραπάνω, το υποκείμενο (το οποίο δεν μπορεί παρά να είναι ονοματική φράση), από σημασιολογική σκοπιά, είναι το μεγαλύτερης σημασίας στοιχείο μιας πρότασης.

Φυσικά, όταν μιλάμε για επεξεργασία γλώσσας, τα προαναφερθέντα δεν μπορούν να αποτελέσουν κανόνα εξόρυξης σημασιολογικής πληροφορίας από κείμενα, αλλά μία απλή παρατήρηση, η οποία ίσως βοηθήσει στο πολυσύνθετο πρόβλημα της κατανόησης ανθρώπινης γλώσσας από υπολογιστή. Πρωτού γίνει αναφορά στις τεχνικές για εξαγωγή ονοματικών φράσεων, θα δοθεί ο ορισμός των ονοματικών φράσεων:

Ονοματική φράση (noun phrase) είναι μία φράση (ακολουθία λέξεων με συγκεκριμένα χαρακτηριστικά διάταξης) βασισμένη σε ένα κυρίαρχο ουσιαστικό (head noun) και αποτελείται από το ουσιαστικό αυτό και άλλες λέξεις που το χαρακτηρίζουν. Επεκτείνοντας τον ορισμό, σε ορισμένες ονοματικές φράσεις, η βάση της φράσης είναι όχι ουσιαστικό αλλά αντωνυμία ή οποιαδήποτε άλλη λέξη μπορεί να χρησιμοποιηθεί αυτόνομη ως υποκείμενο μιας πρότασης [45]. Οι υπόλοιπες λέξεις ή φράσεις της ονοματικής φράσης συνοψίζονται στις παρακάτω κατηγορίες:

- Ουσιαστικό. Για παράδειγμα: noun phrase extraction
  Η παραπάνω φράση αποτελείται από τρία ουσιαστικά. Το κυρίαρχο ουσιαστικό είναι η λέξη extraction ενώ τα υπόλοιπα δύο ουσιαστικά χρησιμοποιούνται για τον χαρακτηρισμό του ουσιαστικού.
- Επίθετο. Για παράδειγμα: the beautiful lady
- Άρθρο (πχ a, the), possessives a (πχ my, your), demonstratives (πχ that, this), numerals (πχ one, two), quantifiers (πχ many, much).
- Εμπρόθετη φράση. Για παράδειγμα: the President of the United States
  Η φράση of the United States αποτελεί prepositonal phrase και χαρακτηρίζει το κυρίαρχο ουσιαστικό (dog).
- Δευτερεύουσα αναφορική φράση. Για παράδειγμα: The building where I work

Η δεύτερη και τρίτη κατηγορία τοποθετούνται πριν το κυρίαρχο ουσιαστικό και το χαρακτηρίζουν. Οι δύο τελευταίες κατηγορίες τοποθετούνται πάντα μετά το κυρίαρχο ουσιαστικό.
Η εξαγωγή ονοματικών φράσεων, απαιτεί τον εντοπισμό και ταυτοποίηση όλων των παραπάνω κατηγοριών μέσα σε ένα κείμενο. Έτσι, έστω τα σύνολα:

$W=\{w_1, w_2..., w_n\}$,
όπου κάθε w_i με i=1..n αποτελεί λέξη του κειμένου,
$NP=\{np_1, np_2..., np_n\}$,
όπου $np_i= \{w_j, w_{j+1}..., w_{j+k}\}$, i=1..n, j,k>0 και j+k<n.

Η διαδικασία για εξαγωγή ονοματικών φράσεων μπορεί να θεωρηθεί ως μία σχεση από το σύνολο W στο σύνολο NP που ακολουθεί τους κανόνες που περιγράφηκαν παραπάνω.

Χρησιμοποιώντας έναν εργαλείο γραμματικής επισημείωσης σε πρώτο στάδιο και με την εφαρμογή κανόνων που προκύπτουν από τον ορισμό που δόθηκε σε μία ονοματική φράση, φαίνεται αρκετά εύκολη η εξαγωγή των ονοματικών φράσεων με δεδομένη τη γραμματική ετικέτα κάθε λέξης. Όμως, η βέλτιστη εξαγωγή ονοματικών φράσεων είναι αρεκτά πιο πολύπλοκη δεδομένου ότι μία λέξη μπορεί να ανήκει σε περισσότερες από μία φράσεις και πως μία ονοματική φράση μπορεί να περιλαμβάνει πάνω από μία ολοκληρωμένες ονοματικές φράσεις στο εσωτερικό της. Έτσι, το ίδιο ουσιαστικό μπορεί να αποτελεί κυρίαρχο ουσιαστικό για μία φράση και προσδιοριστικό ουσιαστικό για μια άλλη. Επιπλέον, οι γραμματικοί κανόνες στους οποίους υπόκειται η ονοματική φράση δεν είναι συγκεκριμένοι και καταγεγραμμένοι, με αποτέλεσμα να είναι δύσκολο να γίνει τόσο ο ορισμός τους όσο και η αναγνώρισή τους στο πλαίσιο του ορισμού αυτού[5].

Υπάρχουν δύο γενικές κατηγορίες μεθόδων εξαγωγής ονοματικών φράσεων. Η μία ανήκει στον τομέα της μηχανικής γνώσης (knowledge engineering) και η άλλη στον τομέα της μηχανικής μάθησης[5].

Στην πρώτη περίπτωση ανήκουν στατικοί επεξεργαστές ονοματικών φράσεων οι οποίοι χρησιμοποιούν έτοιμους γραμματικούς κανόνες προκειμένου να ορίσουν την δομή της ονοματικής φράσης. Οι κανόνες αυτοί, μπορούν να περιγραφούν με τη μορφή τυπικής γραμματικής ή με τη μορφή που έχει ένα αυτόματο πεπερασμένων καταστάσεων (finite state automaton). Ένας τέτοιος κανόνας είναι:

"Ένα άρθρο σηματοδοτεί την αρχή νέας ονοματικής φράσης."

Με βάση αυτό τον κανόνα έχουμε μετάβαση σε αρχική κατάσταση όταν εντοπιστεί άρθρο. Οι επεξεργαστές ονοματικών φράσεων αυτής της κατηγορίας είναι αρκετά αποδοτικοί και μικρού υπολογιστικού κόστους, αλλά παρουσιάζουν προβλήματα συντηρησιμότητας και προσαρμοστικότητας σε νέους γραμματικούς κανόνες. Επίσης, όταν πρόκειται για γλωσσικές εξαιρέσεις που παρουσιάζονται σε κάποιο κείμενο και δεν έχουν μοντελοποιηθεί ή δεν μπορούν να μοντε-

λοποιηθούν σε κανόνα σύστασης ονοματικής φράσης, υπάρχει αποτυχία εντοπισμού της φράσης η οποία θα μπορούσε να οδηγήσει σε αποτυχία εντοπισμού και άλλων φράσεων στην ίδια περίοδο. Σε αυτό το σημείο εγείρεται το ερώτημα για το κατά πόσο είναι δυνατόν να υπάρξει ένα τέτοιο σύνολο κανόνων το οποίο να καλύπτει όλες τις περιπτώσεις σύστασης ονοματικών φράσεων που μπορούν να εντοπιστούν σε ένα κείμενο. Δεδομένης της υποκειμενικότητας και της ποικιλίας μιας γλώσσας, ένα τέτοιο σύνολο είναι πολύ δύσκολο εώς αδύνατο να διαμορφωθεί. Στην περίπτωση, όμως, δημιουργίας ενός τέτοιου συνόλου, το αντίστοιχο εργαλείο θα μπορούσε να εξάγει όλες τις υπάρχουσες ονοματικές φράσεις ενός κειμένου.

Στην δεύτερη περίπτωση, ανήκουν οι επεξεργαστές ονοματικών φράσεων οποίοι χρησιμοποιούν παραδείγματα προκειμένου είτε να ενημερώσουν ήδη υπάρχοντες κανόνες, είτε να δημιουργήσουν νέους. Όσα περισσότερα παραδείγματα βλέπουν, τόσο πιο πιθανό είναι να διαπιστώσουν νέες γλωσσικές σχέσεις. Βασίζονται κυρίως στη στατιστική και τις πιθανότητες και βλέποντας πόσο καλά λειτουργεί ένας κανόνας σε συγκεκριμένο σώμα δεδομένων, προσδίδουν σε αυτόν βάρη. Υπάρχουν δύο διαφορετικές προσεγγίσεις στον τομέα της μάθησης: Η επιβλεπόμενη και η μη επιβλεπόμενη μάθηση. Στην επιβλεπόμενη μάθηση, χρησιμοποιείται ένα υποσημειωμένο σώμα δεδομένων και η προσαρμογή των κανόνων εξαγωγής ονοματικών φράσεων γίνεται με σύγκριση των αποτελεσμάτων του προγράμματος με τα σωστά αποτελέσματα του σώματος δεδομένων. Στη μη επιβλεπόμενη μάθηση δεν χρησιμοποιείται σώμα δεδομένων με γνωστά τα επιθυμητά αποτελέσματα της αναζήτησης ονοματικών φράσεων. Αντίθετα, γίνεται κάποιου είδους αξιολόγηση πάνω στην έξοδο του προγράμματος, και με βάση αυτή βελτιώνονται τα στατιστικά βάρη των κανόνων. Στον τομέα της επιβλεπόμενης μάθησης κάποιες από τις υλοποιημένες τεχνικές είναι:

- Μάθηση βασισμένη σε μετασχηματισμούς
- Μάθηση βασισμένη στη μνήμη
- Μέγιστη εντροπία
- Μέδοδος μαρκοβιανού μοντέλου
- Conditional Random Field
- Μηχανές διανυσμάτων υποστήριξης

2.2 Εξόρυξη Κειμένου

Ο όρος εξόρυξη κειμένου αναφέρεται στη διαδικασία εξαγωγής χρήσιμων πληροφοριών από κείμενα[53]. Κατά τη διαδικασία αυτή, χρησιμοποιούνται τεχνικές από τους τομείς ανάκτησης πληροφορίας, εξαγωγής πληροφορίας καθώς και επεξεργασίας φυσικλής γλώσσας [15]. Επίσης, συχνά απαιτείται η γνώση και η χρήση αλγορίθμων και μεθόδων για εξόρυξη δεδομένων (data mining), μηχανική μάθηση και στατιστική[53]. Στόχος είναι η ανάλυση κειμένων

κατανοητών από τον άνθρωπο και η αναγνώριση των σημαντικών πληροφοριών που περιέχονται σε αυτά. Η σημαντικότητα μιας πληροφορίας, ανάλογα με το πρόβλημα, καθορίζεται είτε σε γενικό πλαίσιο, είτε με βάση μία σημασιολογική απαίτηση. Ο προαναφερθείς στόχος καθιστά εμφανή και την εξάρτηση της εξόρυξης κειμένου από την επεξεργασία φυσικής γλώσσας, όπως ορίστηκε στην ενότητα 2.1 [15].

Για την επίτευξη του παραπάνω στόχου, τα πρώτα στάδια της διαδικασίας έχουν να κάνουν με μία προεπεξεργασία του κειμένου και την μετατροπή του σε μορφή δεδομένων καταλληλότερη για περαιτέρω επεξεργασία σε σχέση με το απλό κείμενο. Έτσι τεχνικές όπως κατάτμιση σε λεκτικές μονάδες και περιστολή λέξεων του τομέα της επεξεργασίας φυσικής γλώσσας, συνήθως είναι απαραίτητες. Επιπλέον, πρέπει να γίνει όσο το δυνατόν αποδοτικότερη αναπαράσταση ώστε να είναι εύκολη η προσπέλαση των λεκτικών μονάδων. Τελευταίο βήμα σε αυτά τα πρώτα στάδια είναι η γλωσσική επεξεργασία του κειμένου (πχ γραμματική επισημείωση, χαμηλού επιπέδου συντακτική ανάλυση, συντακτική ανάλυση και άλλες εργασίες του τομέα της επεξεργασίας φυσικής γλώσσας) [15].

Τα επόμενα στάδια επίλυσης τέτοιων προβλημάτων, εξαρτώνται από τη φύση του προβλήματος και έχουν αναπτυχθεί πολλές τεχνικές τα τελευταία χρόνια προς αυτή την κατεύθυνση για τα περισσότερα από αυτά. Κάποιες βασικές κατηγορίες προβλημάτων που ανήκουν στον τομέα αυτόν είναι[11]:

- ομαδοποίηση κειμένων (document clustering)
  Αναφέρεται στον χωρισμό κειμένων σε ομάδες σημασιολογικής συνάφειας.
- κατηγοριοποίηση κειμένων (document classification
  Αναφέρεται στην ανάθεση προκαθορισμένων κατηγοριών σε κείμενα. Για παράδειγμα, έχουν προκαθοριστεί οι θεματικές ενότητες τέχνη, αθλητισμός, οικονομία, πολιτική και, δεδομένου ενός συνόλου κειμένων, καθένα από αυτά ταξινομείται σε κάποια από τις κατηγορίες αυτές.
- εξαγωγή πληροφορίας
  Αναφέρεται στην διαδικασία εξαγωγής πληροφορίας (σημασιολογικής) από ένα κείμενο μέσα από κατανόηση στοιχείων του κειμένου. Κατά μία έννοια, πρόκειται για μία μορφή ολοκληρωμένης κατανόησης φυσικής γλώσσας. Η κατηγορία αυτή προβλημάτων, δεδομένου ότι βρίσκεται στην καρδιά της σημασιολογικής επεξεργασίας κειμένων, θα αναπτυχθεί εκτενέστερα στη συνέχεια.

Συνολικά, οι δύο γενικές προσεγγίσεις των τεχνικών που υπάρχουν για την επίλυση των παραπάνω προβλημάτων είναι, όπως ακριβώς αναφέρθηκε και στην περίπτωση του προβλήματος της εξαγωγής ονοματικών φράσεων, η μηχανική μάθηση και η μηχανική γνώσης. Ενώ γενικότερα στον τομέα της εξόρυξης κειμένου έχει επικρατήσει η μηχανική μάθηση (στατιστικές μέθοδοι, νευρωνικά δίκτυα (neural networks) κλπ)[33][53], στον τομέα της εξαγωγής πληροφορίας το δίλημμα παραμένει εύλογο, δεδομένου ότι το πρόβλημα του τομέα που ασχολείται με την κατανόηση φυσικής γλώσσας είναι ακόμα άλυτο.

2.3 Εξαγωγή Πληροφορίας

Δεδομένου ενός κειμένου ή ενός τμήματος κειμένου, μία από τις προκλήσεις στην περιοχή της τεχνολογίας λόγου είναι η αναγνώριση πληροφοριών με συγκεκριμένα χαρακτηριστικά μέσα από το κείμενο[28]. Δηλαδή, πληροφορίες που θα αποτελούν δεδομένα συγκεκριμένων τύπων και μάλιστα τύπων τους οποίους ο υπολογιστής θα μπορεί αυτόματα να διακρίνει κατηγοριοποιώντας κατά κάποιον τρόπο την πληροφορία και μάλιστα δίνοντάς της σημασιολογική υπόσταση. Η επίτευξη ενός τέτοιου στόχου αποτελεί αντικείμενο του τομέα της εξαγωγής πληροφορίας. Η άντληση τέτοιου είδους πληροφορίας και μάλιστα δομημένης είναι πολύ σημαντικό επίτευγμα ειδικά όταν γίνεται λόγος για μεγάλο όγκο δεδομένων σε μορφή κειμένου που είναι δύσκολο να προσπελαστούν και να υποστούν επεξεργασία σε όλο τους το εύρος από έναν άνθρωπο.

Ο πρώτος ορισμός που δόθηκε για την εξαγωγή πληροφορίας στο Message Understanding Conference το 1987 είναι "η εργασία εξαγωγής συγκεκριμένων, καλώς ορισμένων τύπων πληροφορίας από ομογενή σύνολα κειμένων, σε σαφώς οριοθετημένους τομείς και/ή συμπλήρωσης καθορισμένων προτύπων που ονομάζονται φόρμες (templates) με την πληροφορία που έχει εξαχθεί" [8].

Οι βασικοί τύποι πληροφορίας που επιθυμεί κάποιος να αναζητήσει σε πρώτο βαθμό, σε ένα κείμενο, είναι οι οντότητες το περιεχόμενο των οποίων είναι σημασιολογικά πλούσιο και οι μεταξύ τους σχέσεις. Σε δεύτερο επίπεδο και αναλύοντας την πληροφορία που θα μπορούσαν να δώσουν για τις οντότητες οι σχέσεις αυτές καταλήγουμε σε ένα σχήμα τεσσάρων ειδών πληροφορίας. Τα είδη αυτά είναι[8][13]:
-οντότητα, λέξη ή φράση που αντιπροσωπεύει ένα συγκεκριμένο αντικείμενο του κόσμου (είτε έχει υλική υπόσταση είτε όχι). Παραδείγματα οντοτήτων αποτελούν τα πρόσωπα, οι ημερομηνίες, τα έργα τέχνης κλπ.
-χαρακτηριστικό (attribute), ιδιότητα μιας οντότητας που την χαρακτηρίζει με κάποιο τρόπο. Για παράδειγμα, όνομα, περιγραφή, τύπος κλπ.
-γεγονός (fact), σχέση μεταξύ ιδιοτήτων, όπως για παράδειγμα η θέση ενός προσώπου σε μία επιχείρηση.
-συμβάν (event), ενέργεια στην οποία συμμετέχουν με κάποιον τρόπο πάνω από μία οντότητες.

Ένα σύστημα το οποίο είναι ικανό να αναγνωρίσει όλα αυτά τα είδη πληροφορίας, μπορεί να καλύψει σημασιολογικά ένα μεγάλο ποσοστό των εννοιών του κειμένου.

Συνήθως η εξαγωγή πληροφορίας γίνεται με τη βοήθεια καλών εργαλείων συντακτικής ανάλυσης και στη συνέχεια με χρήση της τεχνικής που ονομάζεται ταιριασμα προτύπων (pattern matching)[13] . Από τη γλωσσολογική προ-επεξεργασία του κειμένου, δίνονται πληροφορίες για το ρόλο κάθε λέξης η φράσης και γίνεται αναγνώριση κάποιων βασικών οντοτήτων όπως άνθρωποι και εταιρείες. Έπειτα, χρησιμοποιώντας τη γνώση του συντακτικού ρόλου των λέξεων, και με τη συμβολή έτοιμων προτύπων, γίνεται η αναγνώριση και των υπόλοιπων πληροφοριών που αντλούνται από το κείμενο και ιδιαίτερα των γεγονότων. Σε επόμενα βήματα της επεξερ-

γασίας, εξάγονται και πληροφορίες ακόμα λιγότερο εμφανείς όπως αυτές που προκύπτουν μέσω αποσαφήνισης της αναφορικής έννοιας αντωνυμιών.

Σημαντικό σχεδιαστικό ζήτημα σε αυτό το σημείο, είναι το κατά πόσο θα γίνει πλήρης συντακτική ανάλυση του κειμένου πριν το ταίριασμα προτύπων[7]. Παρ' ότι η πλήρης συντακτική ανάλυση θα οδηγούσε σε μεγαλύτερη ακρίβεια όσον αφορά τα πρότυπα, άρα και το ταίριασμά τους, τα υπάρχοντα εργαλεία έχουν δύο αρνητικά ζητήματα: Δεν έχουν αρκετά ακριβή αποτελέσματα και έχουν μεγάλο υπολογιστικό κόστος κατά τη διαδικασία αποσαφήνισης. Επομένως, αν και θεωρητικά η πλήρες συντακτική συντακτική ανάλυση θα διευκόλυνε πολύ την επίλυση του προβλήματος, τελικά πιθανώς να αποτελέσει αιτία για αστοχίες λόγω λάθος αποφάσεων.

Στην διαδικασία εξαγωγής πληροφορίας, μεγάλη βοήθεια μπορεί να προσφέρει η οντολογία (ontology)[7]. Η αναγνώριση οντοτήτων ως στιγμιότυπο (instance) η κάθε μία των κλάσεων μίας οντολογίας και η εφαρμογή των κανόνων της οντολογίας πάνω στις οντότητες και τις σχέσεις τους, βοηθάει στην εξαγωγή πληροφοριών πολλές φορές ακόμα και πέραν αυτών που δηλώνονται ρητά στο κείμενο. Για παράδειγμα, αν στο κείμενο δίνεται η πληροφορία: "Ο Χ απολύθηκε από την εταιρεία Υ" ως συμβάν τότε με την εφαρμογή του κανόνα:

απολίθηκε(Χ-άνθρωπος, Υ-εταιρεία)=>έχει_δουλέψει(Χ-άνθρωπος, Υ-εταιρεία)

εξάγεται η πληροφορία πως "Ο Χ υπήρξε υπάλληλος της εταιρείας Υ".
Σε πρώτο επίπεδο, η οντολογία μπορεί να χρησιμοποιηθεί όπως ήδη αναφέρθηκε για την αναγνώριση των βασικών εννοιών του κειμένου. Αυτό προϋποθέτει ότι η οντολογία είναι πλήρης σε σχέση με τον σαφώς ορισμένο τομέα πάνω στον οποίο γίνεται η εξαγωγή πληροφορίας. Επομένως, όλες οι έννοιες που χρειάζεται να αναγνωριστούν, πρέπει να υπάρχουν στην οντολογία ως στιγμιότυπα των κλάσεών της. Επιπλέον, πρέπει να υπάρχει μία σύνδεση ανάμεσα στα στιγμιότυπα της οντολογίας και την γραπτή τους αναπαράσταση ώστε να μπορεί να γίνει το ταίριασμα μεταξύ της έννοιας και της λέξης που την αναπαριστά. Αυτό προϋποθέτει πως έχει γίνει η γλωσσολογική επεξεργασία του κειμένου (και σίγουρα στην επεξεργασία αυτή πρέπει να περιλαμβάνεται περιστολή λέξεων με τρόπο που να υπάρχει συμφωνία μεταξύ των stemmed words του κειμένου και της γλωσσικής αναπαράστασης των στιγμιότυπων της οντολογίας). Με αυτό τον τρόπο, γίνεται κάποιου είδους σημειολογική επισημείωση (semantic tagging) στις λεκτικές μονάδες του κειμένου το οποίο βοηθάει πάρα πολύ στη μετέπειτα ανάλυση.

Σε δεύτερο επίπεδο, για να εξαχθούν πιο πλούσιες σημασιολογικώς πληροφορίες, χρειάζεται επιπλέον γνώση του τομέα η οποία πιθανότατα δεν υπάρχει στο κείμενο και πρέπει να αντληθεί από την οντολογία. Έτσι θα δοθεί η δυνατότητα να γίνουν κατανοητές σχέσεις μεταξύ εννοιών που αναφέρονται στο κείμενο και δεν μπορούν να έχουν σημασιολογική υπόσταση εκτός της οντολογίας.

2.4 Σημασιολογικός Ιστός

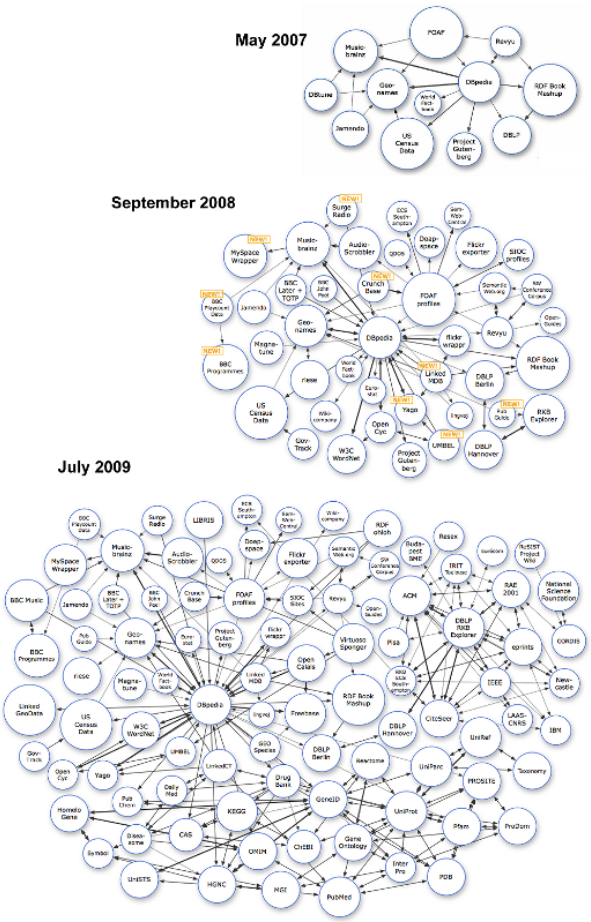
Ο Σημασιολογικός Ιστός (Semantic Web) είναι ένας ιστός δεδομένων, η ανάπτυξη του οποίου σηματοδοτεί την εξέλιξη του Παγκόσμιου Ιστού (World Wide Web) που αποτελούταν σε μεγάλο βαθμό από υλικό προορισμένο για ανθρώπινη κατανάλωση, σε έναν ιστό που θα περιλαμβάνει δεδομένα και πληροφορίες η διαχείριση των οποίων μπορεί να γίνει και από υπολογιστές[34]. Πρόκειται για έναν ιστό πληροφορίας που αντλείται από δεδομένα μέσα από διαδικασίες σημασιολογικής μετάφρασης συμβόλων. Μέσα από τις διαδικασίες αυτές, παρέχεται ένα είδος νοήματος κατά τη λογική σύνδεση των όρων που μπορεί να υποστηρίξει τη διαλειτουργικότητα μεταξύ διαφορετικών συστημάτων. Από την πρώτη φορά που αναφέρθηκε η ιδέα του μέχρι και σήμερα, έχουν γίνει πολλές προσπάθειες προς αυτή την κατεύθυνση, αλλά υπάρχει και ακόμα μεγαλύτερη ανάγκη ενσωμάτωσης δεδομένων από διαφορετικές επιστημονικές- και όχι μόνο- περιοχές σε ένα ενιαίο σύνολο με σημασιολογική σύνδεση[34].

Για να γίνει η μετάβαση από τη σύλληψη της ιδέας του Σημασιολογικού Ιστού, στην πραγματοποίησή της απαιτείται να υπάρχει μεγάλος όγκος δεδομένων στον Παγκόσμιο Ιστό σε μορφή καθορισμένη και κοινή για όλους, προσπελάσιμη, και εύκολα διαχειρίσιμη. Επιπλέον, θα πρέπει να είναι γνωστές και καθορισμένες οι σχέσεις αυτών των δεδομένων μεταξύ τους. Τα δεδομένα με τα συγκεκριμένα χαρακτηριστικά εμφανίζονται με τον όρο διασυνδεδεμένα δεδομένα (linked data) [19]. Τα διασυνδεδεμένα δεδομένα ως όρος ύπαρξης και ανάπτυξης του Σημασιολογικού Ιστού απαιτείται να είναι ενσωματωμένα στον Παγκόσμιο Ιστό ανεξάρτητα από την πηγή από την οποία προέρχονται. Για το λόγο αυτό υιοθετείται η αναπαράσταση αντικειμένων του κόσμου μέσω των URIs (Uniform Resource Identifier - Ενιαίο Αναγνωριστικό Πόρων) τα οποία αποτελούν χαρακτηριστικές περιγραφές τους. Η σχέση, η δομή και η σύνδεσή των περιγραφών αυτών απαιτείται να επιτρέπει την αυτόματη εξαγωγή συμπερασμάτων (reasoning) ώστε να χρειάζεται η λιγότερη δυνατή ανθρώπινη εργασία κατά το στάδιο της ενσωμάτωσης των δεδομένων[14].

Για την ικανοποίηση της τελευταίας απαίτησης, υιοθετήθηκε σε κάποιο βαθμό η χρήση οντολογιών[14]. Επιπλέον, υπήρχε και συνεχίζει να υπάρχει η ανάγκη για ανάπτυξη και χρήση γλωσσών που θα προσφέρουν σημασιολογική διαλειτουργικότητα. Σε αυτό το πλαίσιο αναπτύχθηκε το 1997 το RDF (Resource Description Framework - Πλαίσιο Περιγραφής Πόρων)από το World Wide Web Consortium[54]. Το RDF παρέχει μία απλή αλλά ισχυρή γλώσσα αναπαράστασης των σχέσεων μεταξύ URIs βασισμένη σε τριάδες[14]. Η σύνδεση των URIs που παρέχεται από το RDF μπορεί στη συνέχεια να χρησιμοποιηθεί από την OWL (Web Ontology Language - Γλώσσα Οντολογίας Διαδικτύου) ώστε να δοθούν οι σχέσεις μεταξύ των διάφορων λεξικών που χρησιμοποιούνται στο RDF και να εξυπηρετηθεί η σημασιολογική διαλειτουργικότητα που αναφέρθηκε παραπάνω. Δεδομένου ότι η OWL επιτρέπει να γίνει αναφορά σε έννοια μιας οντολογίας από μία άλλη, μπορεί να υπάρξει σύνδεση μεταξύ διαφορετικών οντολογιών (είτε ίδιων είτε ακόμα και διαφορετικών πεδίων ενδιαφέροντος)[50].

Η αρχή της προσπάθειας υλοποίησης ενός τέτοιου ιστού δεδομένων έγινε από την κοινότητα έρευνας για το Σημασιολογικό Ιστό και κυρίως από το W3C Linking Open Data (LOD)

project που ξεκίνησε τον Ιανουάριο του 2007. Ως αρχική προϋπόθεση, ήταν η εύρεση συνόλων δεδομένων με ανοιχτά δικαιώματα πρόσβασης, η μετατροπή τους σε RDF σύμφωνα με τις αρχές των LOD Linked Open Data - Διασυνδεδεμένα Δεδομένα με ανοιχτά δικαιώματα πρόσβασης όπως αυτές θα περιγραφούν στη συνέχεια και η δημοσίευσή τους στον ιστό[4]. Στο σχήμα 1 παρουσιάζεται η εξέλιξη του Ιστού Δεδομένων από το 2007 μέχρι το 2009[14].

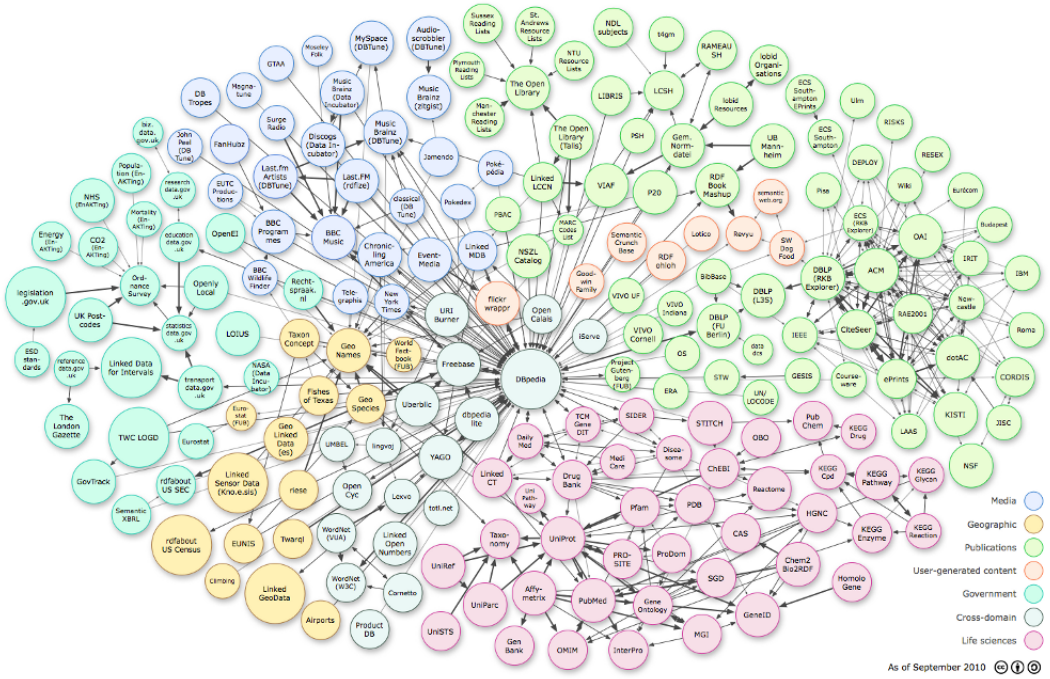


Σχήμα 2.1: Η εξέλιξη του Σημασιολογικού Ιστού μέχρι το 2009

Η δομή του σημασιολογικού ιστού όπως έχει επικρατήσει μέχρι σήμερα συνοψίζεται στα τρία επίπεδα που παρουσιάζονται στη συνέχεια. Εδώ πρέπει να σημειωθεί πως υπάρχουν κι άλλες διαφορετικές προσεγγίσεις για την οργάνωση και δημοσίευση των δεδομένων αλλά στο πλαίσιο της διπλωματικής αυτής θα παρουσιαστεί η αρχιτεκτονική του ιστού που έχει ήδη αναπτυχθεί και όχι οι προτάσεις για βελτίωση της δομής. Τα διασυνδεδεμένα δημοσιευμένα στον ιστό σύνολα δεδομένων φαίνονται στο σχήμα 2[14].

Τα τρία επίπεδα του ιστού είναι τα εξής[34]:

- RDF: επιτρέπει τον ισχυρισμό γεγονότων. Έτσι, για παράδειγμα μπορούμε να περιγράψουμε την πρόταση ”Ο Χ ονομάζεται George”



Σχήμα 2.2: Η εικόνα του σημασιολογικού ιστού το 2010

- RDFs (Resource Description Framework schema - Σχήμα Πλαισίου Περιγραφής Πόρων): επιτρέπει να περιγραφεί ένα λεξιλόγιο (vocabulary) και να χρησιμοποιηθεί στην αναπαράσταση αντικειμένων του κόσμου. Για παράδειγμα στην πρόταση ”Ο Χ είναι LivingPerson”, υπάρχει συγκεκριμένος παγκόσμιος ορισμός του LivingPerson ο οποίος έχει σημασιολογική υπόσταση.

- OWL: επιτρέπει την περιγραφή σχέσεων μεταξύ διαφορετικών λεξιλογίων. Με τον τρόπο αυτό, κλάσεις ή σχέσεις από διαφορετικά λεξιλόγια συνδέονται μεταξύ τους, επομένως και τα δεδομένα που περιγράφουν.

2.4.1 Διασυνδεδεμένα Δεδομένα

Τα δεδομένα που δημοσιεύονται στον Σημασιολογικό Ιστό, σύμφωνα με τον εμπνευστή της ιδέας Tim Berners-Lee, πρέπει να δομούνται με βάση τέσσερις αρχές [19] ώστε να μπορούν να θεωρηθούν LOD και να αποτελούν μέρος του Σημασιολογικού Ιστού.

Αρχή πρώτη:
Ως ονόματα των αντικειμένων πρέπει να χρησιμοποιούνται URIs και μόνο.
Αρχή δεύτερη:
Πρέπει να χρησιμοποιούνται HTTP HyperText Transfer Protocol - Πρωτόκολλο Μεταφοράς Υπερκειμένου URIs ώστε να μπορεί υπάρχει κάποια αναφορά σε αυτά τα ονόματα.

Αρχή τρίτη:
Αν κάποιος θελήσει να αναζητήσει πληροφορίες για κάποιο URI, πρέπει οι πληροφορίες αυτές να του παρέχονται μέσω RDF και SPARQL .
Αρχή τέταρτη:
Στις παραπάνω πληροφορίες, πρέπει να περιέχονται και links προς άλλα URIs.

Η πρώτη αρχή είναι ιδιαίτερα καθοριστικής σημασίας ώστε να μπορέσει κανείς να πει ότι ο παγκόσμιος ιστός μετατρέπεται όντως από ιστό αρχείων σε ιστό δεδομένων. Πλέον τα αντικείμενα τα οποία έχουν όνομα δεν είναι τα αρχεία που περιέχουν τις πληροφορίες, αλλά πραγματικά αντικείμενα του κόσμου (είτε αυτά έχουν υλική υπόσταση είτε όχι). Μπορεί να υπάρχει για παράδειγμα πλέον ένα URI το οποίο αποτελεί αναγνωριστικό για την έννοια ”άνθρωπος”, αλλά μπορεί να υπάρχει και URI που αποτελεί αναγνωριστικό για την περιγραφή της σχέσης ”γνωρίζω άνθρωπο”. Τα URIs, λοιπόν είναι αναγνωριστικά παγκόσμιας εμβέλειας. Η δεύτερη αρχή εξασφαλίζει τη δυνατότητα να μπορεί κάποιος να ανατρέξει σε κάποια περιγραφή του αντικειμένου αυτού η οποία να παρέχεται μέσω του πρωτοκόλλου HTTP το οποίο χρησιμοποιείται ευρέως για την μεταφορά δεδομένων στον παγκόσμιο ιστό. Η τρίτη αρχή προτείνει το πλαίσιο RDF για την περιγραφή των αντικειμένων ώστε πλέον η πληροφορία να μην παρέχεται μόνο για ανθρώπινη κατανάλωση μέσω πρωτοκόλλου HTTP αλλά να μπορεί να είναι προσπελάσιμη μέσω ερωτήσεων από υπολογιστή. Οι ερωτήσεις αυτές γίνονται κυρίως με χρήση της γλώσσας SPARQL (SPARQL Protocol and RDF Query Language - Πρωτόκολλο SPARQL και Γλώσσα Επερώτησης RDF). Η τέταρτη αρχή εξασφαλίζει τη διασύνδεση μεταξύ των δεδομένων που είναι και ο τελικός στόχος του Σημασιολογικού Ιστού.

Από τα πιο σημαντικά στοιχεία για τον ορισμό και την κατανόηση των LOD είναι η χρήση των HTTP URIs για την αναγνώριση αντικειμένων (εννοιών και σχέσεων) του πραγματικού κόσμου. Μέσα από την απαίτηση αυτή, προκύπτει το πρόβλημα της αναπαράστασης των αντικειμένων αυτών. Για παράδειγμα υπάρχει στον πραγματικό κόσμο η έννοια υπολογιστής και ταυτόχρονα υπάρχει ένας τεράστιος αριθμός αρχείων (τα οποία μάλιστα βρίσκονται στον παγκόσμιο ιστό) τα οποία περιγράφουν την έννοια αυτή. Όταν η αναφορά γίνεται σε ένα από αυτά τα αρχεία, τότε είναι εύκολο η αίτηση HTTP που αναζητάει την πληροφορία πίσω από το συγκεκριμένο URI να δώσει σαν απάντηση το ίδιο το αρχείο. Όταν όμως η αναφορά γίνεται στην έννοια του πραγματικού κόσμου, τότε- δεδομένου ότι η ίδια η έννοια είναι αδύνατο να μεταφερθεί σαν έννοια μέσω του παγκόσμιου ιστού- πρέπει με κάποιο τρόπο η απάντηση στην αίτηση να δίνει μία αναγνωριστική περιγραφή (resource description) του αντικειμένου αυτού η οποία βρίσκεται πίσω από το όνομα (URI). Η επίλυση του προβλήματος αυτού πρέπει να έχει το χαρακτηριστικό της μοναδικότητας του κάθε αντικειμένου και άρα του διαχωρισμού των αρχείων ως αντικείμενα του κόσμου, από τις έννοιες που αυτά περιγράφουν.

Η λύση στο πρόβλημα αυτό δίνεται από δύο στρατηγικές[14]:

- 303 URI
  Με τη στρατηγική αυτή, όταν κάποιος κάνει αίτηση HTTP αναζητώντας ένα αντικείμενο του πραγματικού κόσμου που δεν είναι αρχείο, επιστρέφεται ο κωδικός 303 και στη συνέχεια με μία δεύτερη αίτηση ζητείται και επιστρέφεται το αρχείο που έχει επιλεχθεί

για να περιγράψει το αντικείμενο αυτό. Το πρόβλημα αυτής της στρατηγικής είναι πως απαιτούνται δύο αιτήσεις HTTP προκειμένου να επιστραφεί το κατάλληλο αρχείο.

- Hash URIs
  Με τη στρατηγική αυτή, κάθε URI που αντιπροσωπεύει αντικείμενο του πραγματικού κόσμου χωρίζεται σε δύο τμήματα με διαχωριστικό στοιχείο το σύμβολο #. Το πρώτο τμήμα είναι το αρχείο που έχει την περιγραφή του αντικειμένου και το δεύτερο τμήμα είναι το προσδιοριστικό του ίδιου του αντικειμένου. Έτσι, γίνεται μία HTTP αίτηση, επιστρέφεται ολόκληρο το αρχείο (που πιθανότατα περιλαμβάνει περιγραφές πολλών αντικειμένων) και στη συνέχεια γίνεται αναζήτηση μέσα στο αρχείο ώστε να βρεθεί η περιγραφή του συγκεκριμένου αντικειμένου. Το πρόβλημα αυτής της στρατηγικής είναι πως στην περίπτωση που το εν λόγω αρχείο περιλαμβάνει μεγάλο αριθμό περιγραφών, τότε με μία αίτηση επιστρέφεται μεγάλος αριθμός δεδομένων από τα οποία θα χρησιμοποιηθεί μόνο ένα πολύ μικρό μέρος.

Πρακτικά χρησιμοποιούνται και οι δύο προσεγγίσεις ανάλογα με το πόσες περιγραφές έχει το κάθε αρχείο. Για παράδειγμα, στην περίπτωση των λεξιλογίων που θα παρουσιαστούν στη συνέχεια, χρησιμοποιούνται τα Hash URIs ενώ στην περίπτωση της DBpedia χρησιμοποιούνται τα 303 URIs. Επιπλέον χαρακτηριστικά των LOD θα αναλυθούν στη συνέχεια κατά την περιγραφή του RDF και της SPARQL.

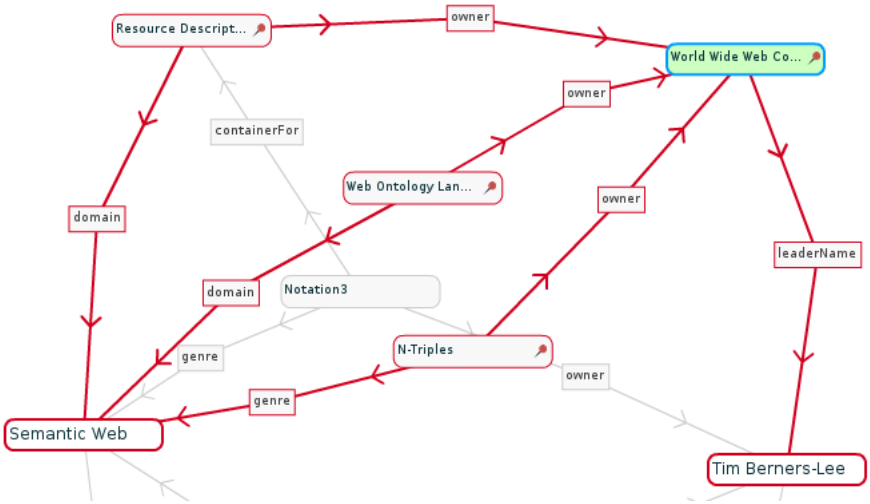
2.4.2 Πλαίσιο Περιγραφής Πόρων

Το RDF είναι ένα μοντέλο για την ανταλλαγή δεδομένων στον παγκόσμιο ιστό. Έχει στοιχεία τα οποία διευκολύνουν την διασύνδεση και τη συγχώνευσή τους ακόμα και όταν τα διάφορα σχήματα και λεξιλόγια που υποστηρίζονται κατά καιρούς διαφέρουν μεταξύ τους. Ιδιαίτερα, επιτρέπει και ενισχύει την εξέλιξη αυτών των σχημάτων στο χρόνο, χωρίς μάλιστα να απαιτείται να αλλάξουν οι καταναλωτές δεδομένων (data consumers) που χρησιμοποιούνται[34].

Το RDF [50] στηρίζεται στην απλή ιδέα της δημιουργίας ενός τύπου αναπαράστασης δεδομένων που ονομάζεται RDF τριάδα (triple) για κάθε σύνδεσμο μεταξύ δύο εννοιών. Στα δύο άκρα της τριάδας βρίσκονται τα URIs των αντικειμένων που συνδέονται μεταξύ τους, ενώ στο κέντρο της τριάδας βρίσκεται το URI του συνδέσμου. Έχουμε επομένως τη σύνδεση μεταξύ δύο αντικειμένων η οποία έχει τη δομή *υποκείμενο-κατηγόρημα-αντικείμενο*. Στην περίπτωση αυτή το κατηγόρημα δεν αναφέρεται στην τυπική γραμματική του ερμηνεία αλλά σε ένα ρόλο που συνδέει το υποκείμενο με το αντικείμενο. Τέτοιες συνδέσεις αντιπροσωπεύονται από τις ακμές του γράφου που σχηματίζει η δομή αυτή ενώ τα συνδεδεμένα αντικείμενα αντιπροσωπεύονται από τους κόμβους του. Πρόκειται για κατευθυνόμενο γράφο με φορά από το υποκείμενο της σύνδεσης προς το αντικείμενο (αν χρησιμοποιηθεί συντακτική ορολογία κατ'αντιστοιχία). Η γραφική αναπαράσταση του μοντέλου -μέσω τέτοιων γράφων- χρησιμοποιείται συχνά για την κατανόηση των συνδέσεων από τον άνθρωπο.

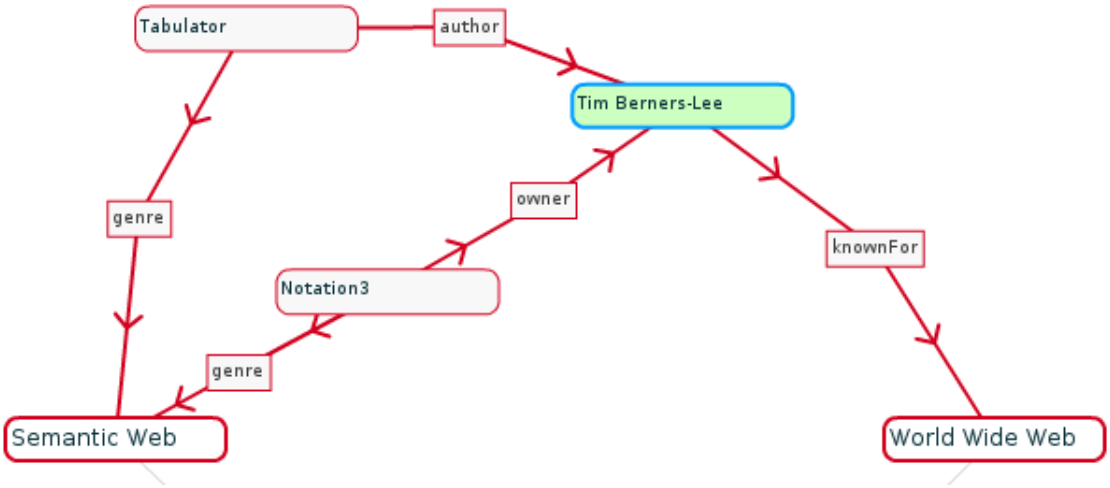
Ένα χαρακτηριστικό παράδειγμα γράφου δίνεται στο σχήμα 2.3. Ο γράφος αυτός προκύ-

πτει από την εφαρμογή της DBpedia, RelFinder [10]. Με την εφαρμογή αυτή, δεδομένων δύο εννοιών, εμφανίζονται όλοι οι υπογράφοι που διασυνδέουν με κάποιο τρόπο τη μία με την άλλη. Οι έννοιες του συγκεκριμένου παραδείγματος είναι οι: Tim Berners-Lee και Semantic Web. Με κόκκινο χρώμα φαίνεται ο υπογράφος που συσχετίζει τις δύο αυτές έννοιες με σημείο αναφοράς την έννοια World Wide Web Consortium. Συνολικότερα, τα ορθογώνια παραλληλόγραμμα δίνουν πληροφορία για το όνομα (label) των ακμών και αντιπροσωπεύουν το URI του δεύτερου από τα τρία στοιχεία της αντίστοιχης τριάδα. Τα άλλα δύο (αρχή και τέλος του διανύσματος) αντιπροσωπεύονται στο συγκεκριμένο γράφο από καμπυλοειδή παραλληλόγραμμα. Ένα



Σχήμα 2.3: Relationship Finder γράφος

δεύτερο, πιο συγκεκριμένο παράδειγμα γράφου δίνεται στο σχήμα 2.4.



Σχήμα 2.4: Relationship Finder γράφος

Σε αυτή την περίπτωση παρουσιάζεται η διασύνδεση των εννοιών World Wide Web και Semantic Web με σημείο αναφοράς την έννοια Tim Berners-Lee. Αναλύοντας τις πληροφορίες που μας δίνει ο γράφος αυτός, φαίνονται οι τριάδες οι οποίες χρησιμοποιήθηκαν για την εξαγωγή του. Έτσι, προκύπτει ο πίνακας 2.1:

Tabulator	genre	Semantic_Web
Notation3	genre	Semantic_Web
Tabulator	author	Tim_Berners-Lee
Notation3	author	Tim_Berners-Lee
Tim_Berners-Lee	author	World_Wide_Web

Πίνακας 2.1: Τριάδες σχήματος 2.4

Εδώ πρέπει να σημειωθεί πως οι παραπάνω τριάδες δεν υπάρχουν αυτούσιες σε RDF γράφο της Dbpedia. Τόσο στα παραπάνω σχήματα όσο και στο παράδειγμα των τριάδων γίνεται αναφορά στις έννοιες όχι με βάση το αναγνωριστικό τους (URI) αλλά με βάση το όνομα που χρησιμοποιεί ο άνθρωπος για να τις περιγράψει. Όπως, όμως, έχει αναφερθεί, στόχος του Σημασιολογικού Ιστού είναι η κατανάλωση δεδομένων από τον υπολογιστή, γι' αυτό και σύμφωνα με την πρώτη και δεύτερη αρχή του Tim Berners-Lee, στις τριάδες της DBpedia στη θέση των συγκεκριμένων ονομάτων, βρίσκονται τα URI ονόματα. Η σύνδεση μεταξύ του URI και της συμβολοσειράς που χρησιμοποιεί ο άνθρωπος γίνεται με βάση μία συγκεκριμένη κατηγορία τριάδων που εμφανίζουν ως ακμή-ρόλο το URI:
http://www.w3.org/2000/01/rdf-schema#label
Η σχέση αυτή συνδέει το αναγνωριστικό URI μίας έννοιας με τη συμβολοσειρά που αποτελεί όνομά της για ανθρώπινη κατανάλωση και θα αναλυθεί στη συνέχεια.

Έτσι, στην πραγματικότητα, για να προκύψει ο γράφος πίσω από την πρώτη τριάδα του παραδείγματος χρειάστηκαν οι εξής RDF τριάδες [9] :

<http://dbpedia.org/resource/ Semantic_Web>	<http://www.w3.org/2000/01/rdf-schema#label>	”Semantic Web”@en
<http://dbpedia.org/resource/ Tabulator>	<http://www.w3.org/2000/01/rdf-schema#label>	”Tabulator”@en
<http://dbpedia.org/property/ genre>	<http://www.w3.org/2000/01/rdf-schema#label>	”genre”@en
<http://dbpedia.org/resource/ Tabulator>	<http://dbpedia.org/property/ genre>	<http://dbpedia.org/resource/ Semantic_Web>

Πίνακας 2.2: RDF τριάδες σχήματος 2.4

Από εδώ και στο εξής, θα ονομάζουμε υποκείμενο τον κόμβο-αρχή κάθε διανύσματος του κατευθυνόμενου γράφου, κατηγόρημα (predicate) την ακμή του και αντικείμενο τον κόμβο-τέλος του σε αντιστοιχία με τη συντακτική ανάλυση των όρων της πρότασης που σχηματίζεται με χρήση ανθρώπινου λόγου από κάθε τριάδα.

Έχοντας συνοψίσει τα βασικά χαρακτηριστικά του τρόπου αντιμετώπισης των αντικειμένων στο RDF, θα προσδιοριστούν οι τύποι συνδέσεων που γίνονται μεταξύ των δεδομένων στο πλαίσιο αυτό. Στο εξής, κάθε τμήμα πληροφορίας που διαθέτει αναγνωριστικό URI θα αναφέρεται ως οντότητα και θα θεωρείται πως σχετίζεται σημασιολογικά με την έννοια που αντιπροσωπεύει. Οι κατηγορίες τριάδων που εμφανίζονται στην RDF περιγραφή κάθε οντότητας είναι οι εξής[50]:

1. Τριάδες που περιγράφουν την οντότητα με γράμματα-συμβολοσειρές. Πρόκειται για την περιγραφή της οντότητας από την οπτική του ανθρώπου.
2. Τριάδες που περιγράφουν την οντότητα μέσω συνδέσεων ΠΡΟΣ άλλες έννοιες. Τέτοιοι τύποι συνδέσεων μπορούν να περιγράφουν για παράδειγμα το δημιουργό, τον κάτοχο, την χώρα γέννησης κλπ. Τέτοιοι τύποι συνδέσμων ονομάζονται εξερχόμενες ακμές (outgoing links) για την αντίστοιχη οντότητα.
3. Τριάδες που περιγράφουν την οντότητα μέσω συνδέσεων ΑΠΟ άλλες έννοιες. Τέτοιοι τύποι συνδέσεων μπορούν να περιγράφουν για παράδειγμα το έργο του οποίου δημιουργός είναι η περιγραφόμενη οντότητα , το αντικείμενο του οποίου κάτοχος είναι η περιγραφόμενη οντότητα, τον άνθρωπο του οποίου χώρα γέννησης είναι η περιγραφόμενη οντότητα. Τέτοιοι τύποι συνδέσμων ονομάζονται εισερχόμενες ακμές (incoming links).
4. Τριάδες που περιγράφουν οντότητες σχετικές με τη συγκεκριμένη οντότητα.
5. Τριάδες που περιγράφουν δεδομένα για τις περιγραφές, δηλαδή μεταδεδομένα όπως η τελευταία ενημέρωση της περιγραφής κλπ.
6. Τριάδες που περιγράφουν μεγαλύτερα σύνολα δεδομένων στα οποία ανήκει η περιγραφή της οντότητας.

Οι κατηγορίες 1 και 2 είναι οι πιο σημαντικές καθώς περικλείουν πολύ σημαντική πληροφορία για κάθε οντότητα. Η πρώτη κατηγορία επιτρέπει τη σύνδεση των οντοτήτων του σημασιολογικού ιστού με την αναφορά τους σε αρχεία που περιέχουν ανθρώπινο λόγο. Έτσι το RDF μπορεί να ενσωματωθεί σε εξυπηρετητή πελάτη (client server) και να αποτελέσει βάση για πολλές εφαρμογές. Η δεύτερη κατηγορία δίνει όλα τα σημαντικά χαρακτηριστικά της οντότητας που προκύπτουν από συνδέσεις. Για τη συγκεκριμένη κατηγορία τριάδων η εν λόγω οντότητα είναι υποκείμενο και όλες οι συνδέσεις έχουν σαν στόχο να την περιγράψουν. Η κατηγορία 3 είναι επίσης αρκετά σημαντική αλλά στοχεύει πιο πολύ στο να χρησιμοποιηθεί η οντότητα (που αποτελεί αντικείμενο σε αυτή την περίπτωση) στην περιγραφή του υποκειμένου της τριάδας. Παρόλα αυτά αν υπάρχει μία ακμή από το a στο b είναι αρκετά σημαντικό στην περιγραφή του b να περιλαμβάνεται η συγκεκριμένη ακμή ως ρόλος, παρ'ότι το a δεν αποτελεί αντικείμενο σε κάποια από τις τριάδες της περιγραφής αυτής. Αυτό μπορεί να διευκολύνει για αντίστροφες αναζητήσεις και δίνει μεγαλύτερη συνοχή στο Σημασιολογικό Ιστό. Οι υπόλοιπες κατηγορίες αν και όχι τόσο βασικές και άμεσα αναγκαίες για την περιγραφή μιας οντότητας χρειάζονται και διευκολύνουν την περιήγηση στον Σημασιολογικό Ιστό και επεκτείνουν την διαθέσιμη πληροφορία σχετικά με την ονότητα αυτή.

2.4.3 Σχήμα Πλαισίου Περιγραφής Πόρων

RDFs (Resource Description Framework schema - Σχήμα Πλαισίου Περιγραφής Πόρων)[50] ονομάζεται η γλώσσα περιγραφής RDF λεξιλογίου (RDF Vocabulary Description Language). Πρακτικά, η χρήση της αποσκοπεί στη σημασιολογική επέκταση του RDF ώστε να μπορούμε να μιλάμε για τις έννοιες κλάση (class) και ιδιότητα (property). Οι ιδιότητες που υπάρχουν στο RDF εμφανίζονται σαν χαρακτηριστικά των οντοτήτων ή αντιπροσωπεύουν μεταξύ τους σχέσεις. Όμως, δεν υπάρχει κάποιος τρόπος να περιγραφούν οι ίδιες οι ιδιότητες ούτε οι σχέσεις τους με άλλες ιδιότητες ή με άλλες οντότητες. Στον Πίνακα 1 παρουσιάζονται οι κλάσεις και οι ιδιότητες που ορίζονται από το RDFs [50]:

Class name	**Comment**
rdfs:Resource	The class resource, everything.
rdfs:Literal	The class of literal values, e.g. textual strings and integers.
rdf:XMLLiteral	The class of XML literals values.
rdfs:Class	The class of classes.
rdf:Property	The class of RDF properties.
rdfs:Datatype	The class of RDF datatypes.
rdf:Statement	The class of RDF statements.
rdf:Bag	The class of unordered containers.
rdf:Seq	The class of ordered containers.
rdf:Alt	The class of containers of alternatives.
rdfs:Container	The class of RDF containers.
rdfs:ContainerMembershipProperty	The class of container membership properties, rdf:_1, rdf:_2, ..., all of which are sub-properties of 'member'.
rdf:List	The class of RDF Lists.

Πίνακας 2.3: RDF(s) classes

Property name	Comment	Domain	Range
rdf:type	The subject is an instance of a class.	rdfs:Resource	rdfs:Class
rdfs:subClassOf	The subject is a subclass of a class.	rdfs:Class	rdfs:Class
rdfs:subPropertyOf	The subject is a subproperty of a property.	rdf:Property	rdf:Property
rdfs:domain	A domain of the subject property.	rdf:Property	rdfs:Class
rdfs:range	A range of the subject property.	rdf:Property	rdfs:Class
rdfs:label	A human-readable name for the subject.	rdfs:Resource	rdfs:Literal
rdfs:comment	A description of the subject resource.	rdfs:Resource	rdfs:Literal
rdfs:member	A member of the subject resource.	rdfs:Resource	rdfs:Resource
rdf:first	The first item in the subject RDF list.	rdf:List	rdfs:Resource
rdf:rest	The rest of the subject RDF list after the first item.	rdf:List	rdf:List
rdfs:seeAlso	Further information about the subject resource.	rdfs:Resource	rdfs:Resource
rdfs:isDefinedBy	The definition of the subject resource.	rdfs:Resource	rdfs:Resource
rdf:value	Idiomatic property used for structured values (see the RDF Primer for an example of its usage).	rdfs:Resource	rdfs:Resource
rdf:subject	The subject of the subject RDF statement.	rdf:Statement	rdfs:Resource
rdf:predicate	The predicate of the subject RDF statement.	rdf:Statement	rdfs:Resource
rdf:object	The object of the subject RDF statement.	rdf:Statement	rdfs:Resource

Πίνακας 2.4: RDF(s) properties

Όπως φαίνεται, αυτό που προσπαθεί να γίνει με αυτό το λεξιλόγιο-βάση είναι κυρίως να κατηγοριοποιηθούν τα URIs σε κλάσεις και να καθοριστούν οι μεταξύ τους σχέσεις (και κυρίως οι ιεραρχικές σχέσεις).

2.4.4 Γλώσσα Οντολογίας Διαδικτύου

Τα RDF και RDFs επιτρέπουν την αναπαράσταση μόνο ενός τμήματος της οντολογικής γνώσης που μπορούμε να έχουμε. Έτσι, υπάρχουν κάποιες σχέσεις κλάσεων και ιδιοτήτων που δεν μπορούν να περιγραφούν. Για παράδειγμα με χρήση RDF/RDFs δεν μπορούμε να περιγράψουμε ξένες κλάσεις (disjoint classes). Συνεπώς, αν έχουμε ορίσει δύο κλάσεις Άντρας, Γυναίκα, δεν

υπάρχει τρόπος να πούμε πως κάποια έννοια δεν μπορεί να είναι στιγμιότυπο και των δύο κλάσεων ταυτόχρονα. Επίσης, δεν μπορούμε να έχουμε αριθμητικούς περιορισμούς. Έτσι, αν έχουμε την ιδιότητα έχει_γονείς, δεν υπάρχει τρόπος να δηλώσουμε πως μία συγκεκριμένη έννοια δεν μπορεί συνδεθεί με την ιδιότητα αυτή με πάνω από δύο έννοιες. Χρειαζόμαστε, λοιπόν, μία γλώσσα περιγραφής οντολογίας που θα είναι πλουσιότερη από την RDF Schema και θα προσφέρει τις παραπάνω εκφραστικές δυνατότητες και άλλες που δεν έχουν αναφερθεί. Κατά το σχεδιασμό μίας τέτοιας γλώσσας, πρέπει να βρεθεί η χρυσή τομή ανάμεσα στην εκφραστικότητα και τον αποδοτικό μηχανισμό αυτόματης εξαγωγής συμπερασμάτων. Γενικά είναι γνωστό πως όσο πιο πλούσια εκφραστικά είναι μία γλώσσα, τόσο πιο πολύπλοκη γίνεται η διαδικασία αυτόματης εξαγωγής συμπερασμάτων. Επομένως, χρειάζεται ένας συμβιβασμός ώστε να δημιουργηθεί μία γλώσσα αρκετά εκφραστική αλλά όχι τόσο ώστε να καταλήγει μη αποδοτική.

Η OWL[49] είναι μία γλώσσα η οποία χρησιμοποιεί τον ευέλικτο τρόπο με τον οποίο το RDF (βασισμένο σε XML σύνταξη) διαχειρίζεται τα δεδομένα και τους δίνει σημασιολογική υπόσταση έτσι ώστε να ξεπερνάει τη βασική σημασιολογία του RDF schema. Η ικανότητα αυτή της OWL είναι πολύ μεγάλης σημασίας, κυρίως για εφαρμογές που θέλουν να κάνουν αναζήτηση και συγχώνευση πληροφορίας από διαφορετικές πηγές και κοινότητες. Αν και τα διάφορα λεξιλόγια που υπάρχουν είναι επαρκή για την ανταλλαγή δεδομένων ανάμεσα σε δύο μέρη που έχουν προσυμφωνήσει σε ορισμούς, η έλλειψη σημασιολογίας εμποδίζει τις εφαρμογές από το να εκτελέσουν αυτή την εργασία με αξιοπιστία, όταν πρόκειται για νέα λεξιλόγια. Στην έκταση του σημασιολογικού ιστού, ο ίδιος όρος πιθανώς να εμφανίζεται με διαφορετική σημασία, σε διαφορετικό νοηματικό πλαίσιο και ταυτόχρονα διαφορετικοί όροι πιθανώς να εμφανίζονται ως αναφορά σε αντικείμενα που έχουν ακριβώς την ίδια σημασία. Με τη χρήση της OWL, τα διαφορετικά λεξιλόγια συνδέονται σημασιολογικά μεταξύ τους και οι διαφορετικοί RDF γράφοι ενώνονται σε ένα ενιαίο ”σύννεφο” δεδομένων που επιτρέπει την ανταλλαγή και αναζήτηση από τη μία άκρη ως την άλλη. Έχει, λοιπόν, επιτευχθεί η διαλειτουργικότητα ανάμεσα σε πολυάριθμα και με αυτόνομη ανάπτυξη και διαχείριση λεξιλόγια/συστήματα περιγραφής λεξιλογίων. Παρακάτω, παρουσιάζονται κάποιες από τις ιδιότητες της OWL που θεωρούνται πολύ χρήσιμα στον σημασιολογικό ιστό.

owl: *equivalentClass*
owl: *equivalentProperty*
Οι δύο αυτές ιδιότητες επιτρέπουν στην OWL να λειτουργεί συνδετικά ανάμεσα σε διαφορετικά λεξιλόγια και συνδέει μεταξύ τους ξεχωριστούς RDF γράφους από διαφορετικές πηγές. Η πρώτη ορίζει κλάσεις που ενώ έχουν διαφορετικό όνομα έχουν την ίδια σημασία και η δεύτερη κάνει ακριβώς την ίδια δουλειά για τις ιδιότητες.

owl: *inverseOf*
Η ιδιότητα αυτή, είναι ιδιότητα ιδιοτήτων. Όταν δύο ιδιότητες συνδέονται με την ιδιότητα *inverseOf* τότε για κάθε τριάδα που έχει μία από τις δύο ιδιότητες ως κατηγόρημα, υπονοείται άλλη μία τριάδα που έχει αντεστραμμένους τους ρόλους υποκειμένου-αντικειμένου και

την άλλη ιδιότητα ως κατηγόρημα. Για παράδειγμα αν ορίσουμε τις ιδιότητες hasDirector και directed τότε για κάθε τριάδα της μορφής *<production> hasDirector <director>* θα υπονοεί και μία τριάδα της μορφής *<director> directed <production>*.

2.4.5 Λεξιλόγια

Συνολικά, στο σημασιολογικό ιστό και σε επίπεδα υψηλότερα από το RDF ορίζονται αντικείμενα και σχέσεις, σε βάθος που καθορίζεται από τα εργαλεία που χρησιμοποιούνται. Οι ορισμοί αυτοί γίνονται μέσα σε ταξινομίες, λεξιλόγια και οντολογίες τα οποία εκφράζονται με χρήση των: SKOS(Simple Knowledge Organization System)[48]-για απλούς ιεραρχικούς ορισμούς, RDFs και OWL. Από σύστημα σε σύστημα, αυξάνεται η εκφραστικότητα και χτίζονται τα στρώματα του ιστού, με χαμηλότερο το RDF.

Η χρήση της OWL και η ελευθερία που προσφέρει στο να αναπτυχθεί ένα νέο λεξιλόγιο, αφήνει το περιθώριο σε όποιον προσθέτει δεδομένα να ορίζει αυθαίρετα λεξιλόγια και να τα επισυνάπτει στον ιστό, γεγονός που δημιουργεί άσκοπες καθυστερήσεις και αυξάνει τον όγκο του ιστού χωρίς να αυξάνει ανάλογα την πληροφορία που αυτός περιλαμβάνει. Για το λόγο αυτό, εμφανίζεται η ανάγκη επαναχρησιμοποίησης υπαρχόντων λεξιλογίων. Γενικότερα η επαναχρησιμοποίηση ορισμών που έχουν οριστεί ήδη θα διευκολύνει στη συνοχή του ιστού αλλά και στην αποδοτικότητα της αυτόματης εξαγωγής συμπερασμάτων. Η παρακάτω λίστα παρουσιάζει μερικά από τα ήδη υπάρχοντα λεξιλόγια, που καλύπτουν συνηθισμένους τύπους δεδομένων και χρησιμοποιούνται ευρέως:

- Το Dublin Core Metadata Initiative (DCMI) Metadata Terms vocabulary ορίζει γενικά χαρακτηριστικά μεταδεδομένων όπως τίτλο, δημιουργό, ημερομηνία και θέμα.
- Το Friend-of-a-Friend (FOAF) vocabulary ορίζει όρους για την περιγραφή ανθρώπων, τις δραστηριότητές τους και τις σχέσεις τους με άλλους ανθρώπους και αντικείμενα.
- Το Semantically-Interlinked Online Communities (SIOC) vocabulary είναι σχεδιασμένο για να περιγράφει χαρακτηριστικά από διδικτυακές κοινότητες όπως χρήστες, δημοσιεύσεις και χώροι δημόσιας συζήτησης.
- Το Description of a Project (DOAP) vocabulary ορίζει όρους για την περιγραφή projects λογισμικού, κυρίως αυτών που είναι ανοιχτού κώδικα.
- Η Music Ontology ορίζει όρους για την περιγραφή διαφόρων εννοιών σχετικών με μουσική όπως καλλιτέχνες, μουσικά κομμάτια και ερμηνείες.
- Η Programmes Ontology ορίζει όρους για την περιγραφή προγραμμάτων όπως εκπομπές τηλεόρασης και ραδιοφώνου.
- Η Good Relations Ontology ορίζει όρους για την περιγραφή προϊόντων, υπηρεσιών και άλλων εννοιών σχετικών με εφαρμογιές ηλεκτρονικού εμπορίου.

- Το Creative Commons (CC) schema ορίζει όρους για την περιγραφή αδειών πνευματικών δικαιωμάτων σε RDF.
- Η Bibliographic Ontology (BIBO) παρέχει έννοιες και ιδιότητες για την περιγραφή βιβλιογραφικών αναφορών.
- Το OAI Object Reuse and Exchange vocabulary χρησιμοποιείται από πολλές πηγές εκδόσεων για να αναπαραστήσει χαρακτηριστικά όπως διαφορετικές εκδόσεις ενός αρχείου ή στοιχεία για την εσωτερική δομή του.
- Το Review Vocabulary παρέχει ένα λεξιλόγια για την αναπαράσταση κριτικών και αξιολογήσεων σε προϊόντα και υπηρεσίες.
- Το Basic Geo (WGS84) vocabulary ορίζει όρους όπως για την περιγραφή γεωγραφικά τοποθετημένων εννοιών.

Ολοκληρώνοντας το κεφάλαιο του σημασιολογικού ιστού, πρέπει να αναφερθεί πως δεν πρόκειται για μία οντολογία καλώς ορισμένη αλλά για ένα σύνολο γράφων με σημασιολογικές προεκτάσεις. Για το λόγο αυτό, τονίζεται πως κάθε ισχυρισμός (με την έννοια *assertion* που χρησιμοποιείται στις οντολογίες), δεν αντιπροσωπεύει γεγονός αλλά απλό ισχυρισμό. Έτσι, είναι πιθανό να υπάρχουν αλληλοσυγκρουόμενοι ισχυρισμοί, όπως αυτοί δίνονται από τις διαφορετικές ανεξάρτητες υποκειμενικές πηγές. Η παραδοχή αυτή είναι που δίνει στο Σημασιολογικό Ιστό τη δυνατότητα υλοποίησης με πραγματικά δεδομένα και επέκτασής του σε τέτοια έκταση ώστε να μπορεί να παρακολουθήσει τον όγκο πληροφορίας που υπάρχει στον παγκόσμιο ιστό. Η σημασιολογία που δίνεται, αντιστοιχεί στον πραγματικό κόσμο και είναι αναπόφευκτο να ακολουθεί την υποκειμενικότητά του κατά την αναπαράστασή του. Αυτή είναι μία σημαντική διαφοροποίηση του παγκόσμιου ιστού από την απλή δημιουργία μίας και μοναδικής οντολογίας καλώς ορισμένης και πρέπει πάντα να λαμβάνεται υπ'όψιν κατά τις αναζητήσεις σε αυτόν.

2.4.6 SPARQL

Η SPARQL (SPARQL Protocol and RDF Query Language) είναι μία γλώσσα που κατασκευάστηκε ωστε να παρέχει τη δυνατότητα ερωτήσεων κ αναζητήσεων σε σε βασεις δεδομένων RDF. Καθιερώθηκε από το RDF Data Access Working Group (DAWG) του World Wide Web Consortium, και θεωρείται μία από τις θεμελιώδεις τεχνολογίες του Σημασιολογικού Ιστού.[51]

Όσον αφορά στην αρχή λειτουργίας της, η SPARQL είναι μία γλώσσα που στηρίζεται στο ταιριασμα γράφων (graph matching). Συγκεκριμένα, έχοντας μία πηγή δεδομένων, έστω D, μια αναζήτηση βασίζεται σε κάποιο συγκεκριμένο πρότυπο, το οποίο κατά την υλοποίηση της αναζήτησης επιχειρείται να ταιριάξει με κάποιο από τα δεδομένα της πηγής D. Οι τιμές που λαμβάνονται από το ταίριασμα αυτό, υπόκεινται σε περαιτέρω επεξεργσία προκειμένου να προκύψει η απάντηση στην αναζήτηση του χρήστη. Έτσι μία SPARQL αναζήτηση αποτελείται από τρία

μέρη. Το πρώτο μέρος είναι το κομμάτι που αφορά το ταίριασμα προτύπων, δηλαδή το ταίριασμα προτύπων γράφων, και περιλαμβάνει χαρακτηριστικά όπως λογικές πράξεις, φώλιασμα (nesting) και φιλτράρισμα (filtering) ταιριάσματος, ορισμένα από τα οποία, θα περιγραφούν παρακάτω. Το δεύτερο μέρος είναι οι τροποποιητές της λύσης, οι οποίοι επεξεργάζονται την έξοδο του ταιριάσματος προτύπου (η οποία δίνεται σε μορφή πίνακα που θα παρουσιαστεί παρακάτω), και την τροποποιούν εφαρμόζοντας τελεστές που επιτυγχάνουν διάκριση, ταξινόμηση, περιορισμό κλπ των στοιχείων της εξόδου. Το τελευταίο μέρος αφορά τη μορφή της εξόδου του αποτελέσματος στο χρήστη που μπορεί να είναι διαφόρων τύπων (επιλογή συγκεκριμένων τιμών μεταβλητών, απαντήσεις τύπου ναι/όχι κλπ). [30]

Συντακτικά η γλώσσα παρουσιάζει σημαντικές ομοιότητες με την SQL, αλλά είναι προσαρμοσμένη ώστε να διαχειρίζεται πληροφορία σε μορφή τριάδων (RDF δεδομένων). Συγκεκριμένα η SPARQL υποστηρίζει σύζευξη και διάζευξη προτάσεων (λογική ένωση και λογική τομή), φιλτράρισμα αποτελεσμάτων, δηλαδή φιλτράρισμα ως προς την τιμη του αποτελέσματος, καθώς και δυνατότητα καθορισμού προαιρετικών απαιτήσεων (optional patterns). Επιπλέον υποστηρίζει και άλλες πιο σύνθετες εντολές, ορισμένες από τις οποίες οι οποίες χρησιμοποιήθηκαν και στη διπλωματική αναλύονται παρακάτω. Για τη χρήση της SPARQL από κάποιο χρήστη (άνθρωπο ή μηχανή) απαιτείται ένα SPARQL σημείο πρόσβασης (endpoint) που θα παρέχει σύνδεση με βάσεις δεδομένων RDF.
Η SPARQL διαθέτει τέσσερις διαφορετικές μορφές αναζήτησης ανάλογα με το σκοπό του χρήστη:[51] [38]

- SELECT query
  Χρησιμοποιείται για την εξαγωγή τιμών από ένα σημείο πρόσβασης SPARQL και επιστρέφει τα αποτελέσματα σε μορφή πίνακα.
- CONSTRUCT query
  Χρησιμοποιείται για την εξαγωγή πληροφορίας από ένα SPARQL endpoint και μετατρέπει τα αποτελέσματα σε έγκυρη RDF μορφή.
- ASK query
  Χρησιμοποιείται σε αναζητήσεις που επιστρέφουν True/False αποτέλεσμα.
- DESCRIBE query
  Χρησιμοποιείται για την εξαγωγή ενός RDF γράφου από το σημείο πρόσβασης SPARQL, η οποία γίνεται με βάση τις πληροφορίες που δίνει/ζητάει ο χρήστης στην αναζήτηση.

Κάθε μία από τις παραπάνω μορφές ερώτησης ξεκινάει με τη χαρακτηριστική λέξη κλειδί (SELECT, CONSTRUCT, ASK, DESCRIBE) και μία μεταβλητή και ακολουθεί ένα μπλοκ που αρχίζει με τη λέξη WHERE που προσδιορίζει και περιορίζει την αναζήτηση. Σημειώνεται πως στην περίπτωση του DESCRIBE, η χρήση του WHERE είναι προαιρετική. Στο WHERE μπλοκ ο χρήστης περιορίζει την αναζήτηση προσδιορίζοντας την τιμή ενός ή περισσότερων στοιχείων από τις τριάδες στις οποίες θα γίνει αναζήτηση. Επίσης στο συγκεκριμένο μπλοκ εφαρμόζονται και τυχόν φίλτρα ή λοιπές σύνθετες εντολές. Τα στοιχεία των τριάδων που δεν προσδιορίζονται θεωρούνται μεταβλητές. Η SPARQL κάνει ενοποίηση των μεταβλητών και των τριάδων

κατα τη διαδικασία υλοποίησης της αναζήτησης.
Στην περαιτέρω συντακτική ανάλυση θα χρησιμοποιηθεί μόνο SELECT καθώς είναι η πιο συνηθισμένη μορφή που χρησιμοποιείται για εξαγωγή πληροφορίας και αυτή που χρησιμοποιείται στην παρούσα διπλωματική εργασία.
Παρακάτω παρουσιάζονται κάποια χαρακτηριστικά για την χρήση των ερωτήσεων.[51] [30]

1. Τα δεδομένα εισόδου σε κάθε αναζήτηση δίνονται με τη μορφή συμβολοσειρών και IRIs (Internationalized Resource Identifier) που αποτελεί γενικευμένη μορφή του URI έτσι ωστε να υποστηρίζεται καλύτερα η χρήση URL. Επίσης ενώ τα URIs περιορίζονται σε ένα υποσύνολο μόνο του ASCII character set, τα IRIs μπορεί να περιέχουν χαρακτήρες από το Universal Character Set (Unicode/ISO 10646), συμπεριλαμβάνοντας Ελληνικά, Κινέζικα, Ιαπωνέζικα, Κορεάτικα, Κυριλλικούς χαρακτήρες κτλ. Για παράδειγμα τα:
<http://www.w3.org/2001/XMLSchema#>, <xml2> και <http://www.παράδειγμα> είναι εξίσου αποδεκτά IRIs.

2. Μία μεταβλητή ορίζεται με τη χρήση του ”?” ή του ”$”. Για παράδειγμα το -?uri- και το -$uri- θεωρείται μεταβλητή. Στη συνέχεια για λόγους συνέπειας θα χρησιμοποιούμε μόνο το ”?”.

3. Τα IRIs ορίζονται πάντα μέσα σε <> ενώ οι συμβολοσειρές σε ’ ’. Επίσης στο τέλος της κάθε συμβολοσειράς μπορεί να δίνεται και η γλώσσα στην οποία είναι γραμμένο ως εξής: ’some_string’@en. Τέλος οι αριθμητικές τιμές δίνονται κανονικά χωρίς την ανάγκη χρήσης συμβόλου για να αναγνωριστούν.

4. Τα αποτελέσματα στην έξοδο μίας αναζήτησης SPARQL μπορεί να είναι σύνολα δεδομένων (results sets) ή RDF γράφοι.

Για την κατανόηση της λειτουργίας των queries καθώς και της λειτουργίας ορισμένων επιπλέον συναρτήσεων/λέξεων κλειδιών που χρησιμοποιήθηκαν παραθέτουμε ορισμένα χαρακτηριστικά παραδείγματα χρησιμοποιώντας τα δεδομένα και τις οντολογίες της dbpedia.

ΠΑΡΑΔΕΙΓΜΑ 1 : Εύρεση υποκειμένων

```
SELECT ?uris
WHERE {
        ?uris <http://www.w3.org/2004/02/skos/core#broader>
        <http://dbpedia.org/resource/Category:Computer_science>
};
```

Τα αποτελέσματα είναι της μορφής:
http://dbpedia.org/resource/Category:Algorithms
http://dbpedia.org/resource/Category:Theoretical_computer_science
http://dbpedia.org/resource/Category:Artificial_intelligence
. . .
Στο παραπάνω παράδειγμα εντοπίζονται τα υποκείμενα (το πρώτο στοιχείο) όλων των τριάδων

που έχουν σαν αντικείμενο το <http://dbpedia.org/resource/Category:Computer_science> και κατηγόρημα το <http://www.w3.org/2004/02/skos/core#broader>. Τα αντικείμενα τυπώνονται σε μία στήλη πίνακα με τίτλο uris (λόγω του ονοματος της αντίστοιχης μεταβλητής). Η λέξη κλειδί VARCHAR δηλώνει τον τύπο της μεταβλητής.

ΠΑΡΑΔΕΙΓΜΑ 2: Εύρεση περισσότερων μεταβλητών

```
SELECT ?uris ?broader
WHERE {
        ?uris <http://www.w3.org/2004/02/skos/core#broader> ?broader };
```

Τα αποτελέσματα είναι της μορφής:

http://dbpedia.org/resource/Category:Radiometry	http://dbpedia.org/resource/Category:Electromagnetic_radiation
http://dbpedia.org/resource/Category:Optics	http://dbpedia.org/resource/Category:Electromagnetic_radiation
http://dbpedia.org/resource/Category:Polarization	http://dbpedia.org/resource/Category:Electromagnetic_radiation
http://dbpedia.org/resource/Category:Light	http://dbpedia.org/resource/Category:Electromagnetic_radiation

Στο παραπάνω παράδειγμα βλέπουμε πως η SELECT μπορεί να επιστρέφει και αποτελέσματα για περισσότερες από μία μεταβλητές, όπως στη συγκεκριμένη περίπτωση όπου και το υποκείμενο και το αντικείμενο είναι μεταβλητά. (παρουσιάζεται τμήμα του πίνακα γιατί ήταν ιδιαίτερα εκτενής).

ΠΑΡΑΔΕΙΓΜΑ 3: Χρήση COUNT

```
SELECT count (?uris)
WHERE {
        ?uris <http://www.w3.org/2004/02/skos/core#broader> ?broader
};
```

Τα αποτελέσματα είναι της μορφής:
INTEGER
1462164

Στην παραπάνω περίπτωση χρησιμοποιούμε το ίδιο query με το προηγούμενο παράδειγμα αλλά με την προσθήκη της εντολής count με αποτέλεσμα να επιστρέφεται ο αριθμός όλων των τριάδων με τα χαρακτηριστικά που προσδιορίσαμε. Παρατηρούμε ότι στο αποτέλεσμα αναφέρεται η μεταβλητή ως callret-0 και δίνεται ο χαρακτηρισμός INTEGER.

Σε περίπτωση που το query γίνεται μέσω jena, θα πρέπει να ονοματοδοτηθεί η μεταβλητή count ως εξής:

```
SELECT (count (?uris) as ?counter)
WHERE {
        ?uris <http://www.w3.org/2004/02/skos/core#broader> ?broader
};
```

Επίσης η καταμέτρηση μπορεί να γίνει παράλληλα με την εκτύπωση των αποτελεσμάτων ως εξής:

```
SELECT ?uris (count (?uris) as ?counter)
WHERE {
        ?uris <http://www.w3.org/2004/02/skos/core#broader> ?broader
};
```

ΠΑΡΑΔΕΙΓΜΑ 4: Χρήση DISTINCT
Σε περίπτωση που κάποιο URI προκύπτει περισσότερες από μία φορές μπορεί να χρησιμοποιηθεί το distinct προκειμένου να τυπώνεται κάθε αποτέλεσμα μία φορά. Συνεπώς με το query:

```
SELECT distinct ?broader
WHERE {
        ?uris <http://www.w3.org/2004/02/skos/core#broader> ?broader
};
```

Ένα μέρος τον απαντήσεων που παίρνουμε είναι οι εξής:
http://dbpedia.org/resource/Category:Electromagnetic_radiation
http://dbpedia.org/resource/Category:Climatology
http://dbpedia.org/resource/Category:Climate_forcing
http://dbpedia.org/resource/Category:Latin_letters
http://dbpedia.org/resource/Category:Autism
(Για καλύτερη κατανόηση μπορεί να γίνει σύγκριση με το παράδειγμα 2)

ΠΑΡΑΔΕΙΓΜΑ 5: Χρήση πολλών προσδιοριστικών προτάσεων και μεταβλητών
Είναι δυνατόν να δίνονται περισσότερες από μία προσδιοριστικές προτάσεις όπως για παράδειγμα :

```
SELECT ?uri ?cat ?broader
WHERE {
        ?uri <http://www.w3.org/2000/01/rdf-schema#label> ”Computer”@en.
        ?uri <http://purl.org/dc/terms/subject>                 ?cat.
        ?cat<http://www.w3.org/2004/02/skos/core#broader>?broader
}
```

Τα αποτελέσματα είναι της μορφής:

http://dbpedia.org/resource/Computer http://dbpedia.org/resource/Category:Computing
http://dbpedia.org/resource/Category:Digital_technology
http://dbpedia.org/resource/Computer http://dbpedia.org/resource/Category:Computers

Εδω τυπωνονται ολες οι categories και broader categories του resource uri που αντιστοιχεί στη συμβολοσειρά ”Computer”. Η αναζήτηση και η ευρεση των αποτελεσμάτων γίνεται με αναδρομική ενοποίηση των μεταβλητών και αντιστοίχισή τους σε URIs.

ΠΑΡΑΔΕΙΓΜΑ 6: Χρήση FILTER
Με χρήση φίλτρων μπορεί να περιοριστεί ακόμα περισσότερο η αναζήτηση. Τα φίλτρα εφαρμόζονται συνήθως για περιορισμό αριθμητικών αποτελεσμάτων, ή για ταίριασμα συμβολοσειρών. Ακολουθούν παραδείγματα φίλτρων:

```
SELECT ?resource
WHERE {
        ?resource <http://somewhere/peopleInfo#age> ?age .
        FILTER (?age >= 24)
}
```

Integer constraint: κρατούνται μόνο τα αποτελέσματα με τιμή μεγαλύτερη του 24

```
SELECT ?g
WHERE {
        ?y <http://www.w3.org/2001/vcard-rdf/3.0#Given> ?g .
FILTER regex(?g, ”rt”, ”i”)
}
```

Ταίριασμα συμβολοσειρών: κρατούνται μόνο τα αποτελέσματα που περιέχουν το φθόγγο rt. Το ”i” υποδηλώνει ότι είναι αποδεκτά και πεζά και κεφαλαία γράμματα.

ΠΑΡΑΔΕΙΓΜΑ 7: Χρήση PREFIX
Ακόμα παρέχεται η δυνατότητα προσδιορισμού και χρήσης συντομεύσεων ώστε να μη χρειάζεται η επανάληψη εκτενών IRI στο σώμα της ερώτησης. Η χρήση του prefix επιτρέπει να μην επαναλαμβάνεται όλο το σώμα ενός IRI, όταν αυτό αποτελεί ορισμένο τμήμα μίας τυπικής οντολογίας:
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
Στο κυρίως σώμα της ερώτησης μπορούμε να γράψουμε:

```
SELECT ?x
WHERE{
        ?x foaf:name ?y };
```

αντί για

```
SELECT ?x
WHERE{
        ?x <http://xmlns.com/foaf/0.1/name> ?y
};
```

Η χρήση prefixes είναι ιδιαίτερα χρήσιμη σε σύνθετες αναζητήσεις.
Τέλος πρέπει να σημειωθεί ότι εκτός από πληροφορία για τριάδες η sparql μπορεί να διαχειριστεί και πληροφορία για γράφους. Δίνεται δηλαδή η δυνατότητα να περιοριστεί η αναζήτηση σε ένα συγκεκριμένο γράφο, καθώς και να αναζητηθεί ο γράφος ή οι γράφοι που περιέχουν συγκεκριμένα δεδομένα ή έχουν συγκεκριμένα χαρακτηριστικά.

ΠΑΡΑΔΕΙΓΜΑΤΑ 8-10: Χρήση Γράφων

ΠΑΡΑΔΕΙΓΜΑ 8

```
SELECT distinct ?g
WHERE {
graph ?g {?uri <http://www.w3.org/2004/02/skos/core#broader> ?uri2}
};
```

Η απάντηση που δίνεται είναι η:
VARCHAR
categories

Πρόκειται για τον γράφο που περιλαμβάνει τριάδες με το κατηγόρημα
<http://www.w3.org/2004/02/skos/core#broader>

ΠΑΡΑΔΕΙΓΜΑ 9

```
SELECT distinct ?g
WHERE {
        graph ?g {?uri ?r ?uri2}
};
```

Τα αποτελέσματα είναι της μορφής:
http://www.openlinksw.com/schemas/virtrdf#
http://localhost:8890/DAV

http://dbpedia4.org

Όλοι οι γράφοι του σημείου πρόσβασης SPARQL. Περιλαμβάνονται και default γράφοι του συστήματος όπως: http://www.openlinksw.com/schemas/virtrdf# http://localhost:8890/DAV

ΠΑΡΑΔΕΙΓΜΑ 10

```
SELECT distinct ?uri
WHERE {
        graph ?categories {?uri ?x ?y}.
        FILTER regex (?uri, "super", "i")
};
```

Τα αποτελέσματα είναι της μορφής:
http://www.openlinksw.com/virtrdf-data-formats#default-iid-SuperFormats
http://www.openlinksw.com/virtrdf-data-formats#default-iid-nullable-SuperFormats
http://www.openlinksw.com/virtrdf-data-formats#sql-numeric-literal-SuperFormats
http://www.openlinksw.com/virtrdf-data-formats#sql-numeric-literal-nullable-SuperFormats

Πρόκειται για υποκείμενα που περιέχουν τη συμβολοσειρά super στο γράφο categories. (Παρουσιάζεται μέρος του συνολικού πίνακα των αποτελεσμάτων καθώς ήταν ιδιαίτερα εκτενής)

2.4.7 DBpedia

Η DBpedia είναι ένας τόπος πρόσβασης σε σημασιολογικά οργανωμένη πληροφορία και δεδομένα που έχουν εξαχθεί από τη Wikipedia καθώς και σε εφαρμογές διαχείρησης και αξιοποίησης της πληροφορίας αυτής. Συνολικά το έργο υλοποίησης της DBpedia έχει στόχο να αποσπάσει δομημένη πληροφορία από τα δεδομένα της Wikipedia και να διαθέσει τη σημασιολογικά οργανωμένη αυτή πληροφορία στο διαδίκτυο, σε μορφή Resource Description Framework - Πλαίσιο Περιγραφής Πόρων (RDF). Παρέχει τη δύνατότητα σύνθετων αναζητήσεων στο συνολικό ογκο των δεδομένων της καθώς και δυνατότητα σύνδεσης άλλων συνόλων δεδομένων με τα δεδομένα της DBpedia. Η απόπειρα αυτή ξεκίνησε από το Ανοιχτό Πανεπιστήμιο του Βερολίνου και το πανεπιστήμιο του Leipzig, σε συνεργασία με την Openlink Software. Το πρώτο συνολο δεδομένων διαθέσιμο για δημόσια χρήση, δημοσιεύτηκε το 2007. Τα σύνολα δεδομένων που δημοσιεύονται έκτοτε, διατίθενται με ανοιχτά δικαιώματα, επιτρέποντας την αξιοποίηση τους από οποιονδήποτε χρήστη. [3]

Για τη δημιουργία όλων των συνόλων δεδομένων και των μεταξύ τους σχέσεων (ρόλων) που τα συνδέουν, έχουν χρησιμοποιηθεί άρθρα της Wikipedia, καθώς αποτελούν μια από τις μεγαλύτερες, σημαντικότερες και εγκυρότερες πηγές δεδομένων στο διαδίκτυο, η οποία επεκτείνεται και ελέγχεται διαρκώς. Συνεπώς, αποτελεί μία πολύ καλή πηγή πληροφορίας για την εξαγωγή

δεδομένων της DBpedia. [40] Τα αρθρα της Wikipedia αποτελούνται κυρίως από απλό κείμενο, αλλά περιέχουν και κάποιου τύπου δομημένη πληροφορία όπως εικόνες, γεωγραφικές συντεταγμένες, εξωτερικούς συνδέσμους, πίνακες ”infobox”, οργάνωση σε κατηγορίες, περιλήψεις (abstracts), καθως και εσωτερικούς συνδέσμους σε άλλα άρθρα της Wikipedia. Αυτή η πληροφορία οργανώνεται κατάλληλα σε RDF μορφή ωστε να είναι προσβάσιμη για τυχον αναζητήσεις των χρηστών. Το τρέχον σύνολο δεδομένων στη γνωσιακή βάση περιλαμβάνει περίπου τρία εκατομμύρια ”οντότητες”, συμπεριλαμβανομένων εικόνων, ατόμων, εταιριών, περιοχών κτλ.[9] Βεβαίως πρέπει να σημειωθεί πως ο τρόπος δημιουργίας αλλά και ανανέωσης των άρθρων της Wikipedia, καθώς και ελέγχου της συνέπειάς τους, η οποία απαιτεί ανθρώπινη παρέμβαση και γίνεται χειροκίνητα, συνεπάγεται κάποιο ποσοστό νοηματικών συγκρούσεων και ανακριβειών το οποίο αναπόφευκτα κληροδοτείται και στα δεδομένα της DBpedia [40]

Για κάθε μία από αυτές τις οντότητες, η DBpedia ορίζει ένα καθολικά μοναδικό αναγνωριστικό (μέσω συστήματος ταυτοποίησης), στο οποίο μπορεί να αναχθεί μέσω Διαδικτύου σε μια πλούσια περιγραφή RDF. Στην περιγραφή μπορεί να συμπεριλαμβάνονται ορισμοί (σε 30 γλώσσες), σχέσεις με άλλους πόρους, ταξινομήσεις σε τέσσερις εννοιολογικές ιεραρχικές δομές, συνδέσεις σε επίπεδο δεδομένων με άλλες πηγές δεδομένων στον Παγκόσμιο Ιστό, οι οποίες περιγράφουν επίσης την οντότητα. Όσο η DBpedia εξελίσεται ένας διαρκώς αυξανόμενος αριθμός εκδοτών δημόσιων δεδομένων (public data publishers) αρχισαν να κατασκευάζουν συνδέσμους σε δεδομένα της DBpedia. Έτσι πλέον η DBpedia αποτελεί κεντρικό κόμβο διασύνδεσης για τον υπό εξέλιξη ιστό δεδομένων. [3][40]

Για την κατασκευή του αναγνωριστικού κάθε οντότητας, χρησιμοποιείται ο αντίστοιχος τίτλος του αγγλικού άρθρου της dbpedia. Οι πληροφορίες που προέρχονται από άρθρα σε άλλες γλώσσες αντιστοιχίζονται αμφίδρομα στα αναγνωριστικά αυτά. Σε κάθε οντότητα ανατίθεται ένα Uniform Resource Identifier - Ενιαίο Αναγνωριστικό Πόρων (URI) σύμφωνα με το σχήμα: *http://dbpedia.org/resource/Ονομα*, όπου το Όνομα προέρχεται από το αντίστοιχο URL (Uniform Resource Locator -Ενιαίος Εντοπιστής Πόρων) του άρθρου της Wikipedia, το οποίο έχει τη μορφή *http://en.wikipedia.org/wiki/Ονομα.* Η διαδικασία αυτή κατασκευής URI, προσδίδει τις εξής ιδιότητες [3] :

- Τα URIs της DBpedia καλύπτουν μια ευρεία περιοχή εγκυκλοπεδικών θεμάτων, θεμάτων επικαιρότητας, γενικού και ειδικού ενδιαφέροντος.
- Διαθέτουν έγκυρους ορισμούς που προέρχονται από ομόφωνη απόφαση της κοινότητας.
- Η διαχείρηση τους υπόκειται σε σαφείς και επίσημες πολιτικές διαχείρισης.
- Κάθε οντότητα διαθέτει μία εκτενή και αναλυτική γραπτή περιγραφή σε τουλάχιστον ένα γνωστό διαδικτυακό τόπο (Wikipedia), γεγονός ιδαίτερα χρήσιμο για παραπομπές.

Τα δεδομένα της DBpedia είναι οργανωμένα σε RDF τριάδες, ωστόσο δεν υπάρχει τυπική οντολογία που να καλύπτει το σύνολο των δεδομένων τόσο λόγω του όγκου όσο και λόγω του τύπου τους. Παρ’ όλα αυτά, μέρος των δεδομένων διαθέτει κάποιας μορφής οργανωμένη ταξι-

νόμηση αντίστοιχη με αυτή μίας τυπικής οντολογίας. Συγκεκριμένα εντοπίζονται τα παρακάτω τέσσερα σχήματα ταξινόμησης, τα οποία ικανοποιούν διαφορετικές απαιτήσεις εφαρμογών.

1. Wikipedia Categories
   Η dbpedia περιέχει μια αναπαράσταση του συστήματος των κατηγοριών Wikipedia βασισμένη στο σύστημα SKOS. [1] Συνολικά υπάρχουν πάνω από 415.000 κατηγορίες. Το κύριο πλεονέκτημα του συστήματος των κατηγοριών είναι ότι επεκτείνεται και ενημερώνεται διαρκώς από χιλιάδες συντάκτες της Wikipedia. Ωστόσο, παρά το μεγάλο εύρος των κατηγοριών, εντοπίζεται ένα βασικό μειονέκτημα που σχετίζεται με την μη επαρκή ιεραρχική δομή μεταξύ των κατηγοριών. Συγκεκριμένα, ενώ χρησιμοποιούνται εκτενώς ορισμένοι ιεραρχικοί ρόλοι (κατηγορήματα) που ορίζονται από το SKOS (πχ broader categories , related categories), η ιεραρχία που προκύπτει δεν είναι σωστή, καθώς υπάρχουν κύκλοι στο σύστημα των κατηγοριών, και οι κατηγορίες συχνά υποδεικνύουν μία μάλλον χαλαρή συγγένεια μεταξύ των άρθρων αποτυγχάνοντας να απεικονίσουν πλήρως τη μεταξύ τους σχέση.

2. YAGO
   Το σχήμα ταξινόμησης Yago αποτελείται από 286.000 κλάσεις που αποτελούν μια βαθιά ιεραρχία υπαγωγής. Το σχήμα δημιουργήθηκε χαρτογραφώντας τις κατηγορίες-φύλλα της Wikipedia, δηλαδή τις κατηγορίες που δεν έχουν υποκατηγορίες, με τη χρήση των WordNet synsets. Για τη χαρτογράφηση χρησιμοποιήθηκε συγκεκριμένος αλγόριθμος κατάταξης που περιγράφεται στο [39]. Χαρακτηριστικά της ιεραρχίας Yago είναι το βάθος της και η κωδικοποίηση πολλών πληροφοριών σε μια κατηγορία (π.χ. υπάρχει η κατηγορία *MultinationalCompaniesHeadquarteredInTheNetherlands*). Ενώ Yago επιτυγχάνει, σε γενικές γραμμές, υψηλή ακρίβεια, υπάρχουν ορισμένα εμφανή λάθη και παραλείψεις (π.χ. η κατηγορία που αναφέρεται δεν είναι μια υποκατηγορία του *MultinationalCompanies* όπως θα ήταν επιθυμητό) λόγω της αυτόματης παραγωγής της ταξινόμησης. Από κοινού με τη dbpedia έχει αναπτυχθεί ένα script που αναθέτει αυτόματα Yago κλάσεις σε DBpedia οντότητες.

3. UMBEL
   Η οντολογία της UMBEL (Upper Mapping and Binding Exchange Layer) είναι μία ”ελαφριά” οντολογία που έχει δημιουργηθεί για διασύνδεση διαδικτυακού περιεχομένου και δεδομένων. Η συγκεκριμένη οντολογία προέρχεται από την OpenCyc και αποτελείται από 20.000 κλάσεις. Οι κλάσεις της OpenCyc με τη σειρά τους προέρχονται εν μέρει από Cyc συλλογές δεδομένων (collections), οι οποίες βασίζονται σε WordNet synsets. Δεδομένου οτι η Yago χρησιμοποιεί επίσης WordNet synsets και βασίζεται στην Wikipedia, μια χαρτογράφηση από OpenCyc κλάσεις στη DBpedia μπορεί να προκύψει μέσω UMBEL. Η ταξινό-

[1] Το SKOS είναι ένα έργο ανάπτυξης προδιαγραφών και προτύπων για την υποστήριξη της χρήσης των συστημάτων οργάνωσης γνώσης (KOS : Knowledge Organisation Systems), όπως οι θησαυροί, τα συστήματα ταξινόμησης, συστήματα κλάσης αντικειμένων και ταξινομήσεις, στο πλαίσιο του Σημασιολογικού Ιστού.Το SKOS παρέχει ένα πρότυπο τρόπο να συστηματοποιούνται τα συστήματα οργάνωσης γνώσης με χρήση του πλαισίου RDF. Η κωδικοποίηση της πληροφορίας αυτής σε RDF επιτρέπει τη χρήση της σε ποικιλία εφαρμογών πληροφορικής με διαλειτουργικό τρόπο.

μηση συντηρείται από την UMBEL και οι λεπτομέρειες για τη διαδικασία παραγωγής της οντολογίας είναι προσβάσιμες από τον αντίστοιχο ιστότοπο.

4. DBpedia Ontology
   Η οντολογία της DBpedia αποτελείται από 170 κλάσεις που σχηματίζουν ιεραρχία υπαγωγής με σχετικά μικρό βάθος. Περιλαμβάνει 720 ιδιότητες με προσδιορισμούς εύρους και τομέα. Η οντολογία αυτή δημιουργήθηκε αυτόματα από τα συνηθέστερα χρησιμοποιούμενα πρότυπα που εντοπίζονται στο infobox στην αγγλική έκδοση της Wikipedia. Η οντολογία χρησιμοποιείται ως άξονας για την εξαγωγή συμπαγούς πληροφορίας από το infobox όπως αναφέρεται και στην περιγραφή των δεδομένων παρακάτω.

Τα υπόλοιπα δεδομένα, τα οποία έχουν κωδικοποιηθεί σε RDF μορφή αλλά δεν ανήκουν σε καμία από τις παραπάνω ταξινομημένες ομάδες, διαχωρίζονται ανάλογα με το κατηγόρημα που χαρακτηρίζει κάθε τριάδα.

Τα δεδομένα της DBpedia είναι προσβάσιμα μέσω τεσσάρων διαφορετικών μηχανισμών ανάλογα με τις ανάγκες που έχει ο χρήστης ή ο πράκτορας (agent). Οι μηχανισμοί αυτοί αναλύονται σύντομα παρακάτω:

1. Διασυνδεδεμένα δεδομένα
   Πρόκειται για μία μέθοδο δημοσίευσης δεδομένων RDF στον Ιστό και διασύνδεσης τους με άλλους σχετικούς διαδικτυακούς πόρους. Η μέθοδος αυτή, βασίζεται σε HTTP URIs ως αναγνωριστικά πόρων και το πρωτόκολλο HTTP για την ανάκτηση των αντίστοιχων περιγραφών. Τα αναγνωριστικά της DBpedia (όπως *http://dbpedia.org/resource/Berlin)* είναι κατασκευασμένα ώστε να επιστρέφουν (α) RDF περιγραφές όταν η αναζήτηση στα δεδομένα γίνεται από πράκτορες του Σημασιολογικού Ιστού (όπως δεδομένα ή προγράμματα περιήγησης ή μηχανές αναζήτησης του Σημασιολογικού Ιστού), και (β) μια απλή HTML μορφή που περιέχει την ίδια πληροφορία και απευθύνεται σε παραδοσιακά προγράμματα περιήγησης στον Ιστό.

2. Σημείο πρόσβασης SPARQL
   Διατίθεται ένα σημείο πρόσβασης SPARQL, το οποίο καθιστά εφικτή την αναζήτηση στη γνωσιακή βάση της DBpedia. Συγκεκριμένα, αυτοματοποιημένες εφαρμογές μπορούν να στείλουν ερωτήματα με χρήση του πρωτοκόλλου SPARQL στο σημείο πρόσβασης, κάνοντας χρήση του αντίστοιχου URL (*http://dbpedia.org/sparql*). Εκτός από τη standard SPARQL, το endpoint υποστηρίζει διάφορες επεκτάσεις της γλώσσας που είναι χρήσιμες για την ανάπτυξη εφαρμογών πελάτη, όπως η αναζήτηση πλήρους κειμένου σε επιλεγμένα κατηγορήματα RDF και συναρτήσεις συγκεντρωτικών αποτελεσμάτων, όπως η COUNT(). Καθώς πρόκειται για διαδικτυακή εφαρμογή η οποία έχει αυστηρές απαιτήσεις και περιορισμούς σε χρόνο απόκρισης και αποφανσιμότητα υπάρχουν όρια στην πολυπλοκότητα και το βάθος των ερωτημάτων/αναζητήσεων που είναι αποδεκτά καθώς και στον όγκο των δεδομένων εξόδου που παρέχονται ως αποτέλεσμα (ορισμένες πολύ γενικές ερωτήσεις μπορεί να δίνουν ως αποτέλεσμα στην έξοδο πίνακες με εκατομμύρια URIs που έχουν ανασυρθεί από τη γνωσιακή βάση δεδομένων). Για τους ίδιους λόγους,

δεν είναι φορτωμένα στη βάση όλα τα RDF δεδομένα που διαθέτει η DBpedia αλλά επιλέγονται μόνο ορισμένοι θεμελιώδεις τύποι κατηγορημάτων οι οποίοι φορτώνονται, με αποτέλεσμα να μην είναι ποτέ διαθέσιμα όλα τα σύνολα δεδομένων στη βάση.

3. RDF Dumps
   Οι τριάδες RDF που έχουν εξαχθεί από τη Wikipedia έχουν οργανωθεί σε αρχεία με βάση το κατηγόρημα κάθε τριάδας, και στη συνέχεια τα κατηγορήματα οργανώνονται σε ομάδες με σημασιολογικά κριτήρια. Οι τριάδες που ανήκουν σε κάθε τέτοια ομάδα αποτελούν ένα αρχείο τύπου N-triple. Τα αρχεία αυτά είναι διαθέσιμα για λήψη από τους επισκέπτες του ιστότοπου της DBpedia. Συνεπώς το σύνολο των δεδομένων της βάσης είναι προσβάσιμο για χρήση και μπορεί να αξιοποιηθεί σε πληθώρα εφαρμογών και απομακρυσμένους εξυπηρετητές. Σημειώνεται πως για τις ανάγκες της παρούσας διπλωματικής χρησιμοποιήθηκε αυτή η μέθοδος καθώς υπήρχε η ανάγκη για πολλαπλές και σύνθετες αναζητήσεις καθώς και χρήσης δεδομένων που δεν ήταν διαθέσιμα στο διαδικτυακό σημείο πρόσβασης SPARQL. Συνεπώς όπως περιγράφεται και στο Παράρτημα 1, εγκαταστάθηκε ο εξυπηρετητής virtuoso, στον οποίο φορτώθηκαν τα σύνολα δεδομένων τα οποία θεωρήθηκαν απαραίτητα για την υλοποίηση της εφαρμογής.

4. Lookup Index
   Η υπηρεσία αυτή σχετίζεται με τη δημοσίευση LOD (Linked Open Data - Διασυνδεδεμένα Δεδομένα με ανοιχτά δικαιώματα πρόσβασης) δεδομένων. Διατίθενται κατάλληλες υπηρεσίες αναζήτησης οι οποίες διευκολύνουν τη διασύνδεση εξωτερικών δεδομένων σε πόρους της DBpedia. Συγκεκριμένα, υπάρχει υπηρεσία που διευκολύνει την επιλογή κατάλληλων URi και ετικεττών για κάθε νέα σύνδεση δεδομένων. Βασίζεται σε ευρετήριο Lucene δείκτη παρέχοντας σταθμισμένη αναζήτηση ετικέτας, η οποία συνδυάζει την σύγκριση συμβολοσειρών και συσχέτιση κατάταξης (παρόμοια με PageRank) για να βρει τα πιο πιθανά αποτελέσματα για ένα συγκεκριμένο όρο. Το DBpedia lookup είναι διαθέσιμο ως διαδικτυακή υπηρεσία στη http://lookup.dbpedia.org/api/search.asmx.

Η γνωσιακή βάση της DBpedia είναι συνδεδεμένη με διάφορες άλλες πηγές δεδομένων στο Διαδίκτυο σύμφωνα με τις αρχές των LOD. Σήμερα, περιέχει 4,9 εκατομμύρια εξερχόμενες συνδέσεις RDF που οδηγούν σε συμπληρωματικές πληροφορίες σχετικά με οντότητες της DBpedia, καθώς και μεταπληροφορίες (meta-data) σχετικά με στοιχεία πολυμέσων που απεικονίζουν κάθε οντότητα. Δεδομένου ότι τελευταία ένας αυξανόμενος αριθμός εκδοτών δημόσιων δεδομένων έχουν αρχίσει να RDF συνδέσεις με DBpedia οντότητες, οι εισερχόμενες συνδέσεις, μαζί με τις εξερχόμενες συνδέσεις που δημοσιεύθηκαν μέσω του DBpedia, καθιστούν τη DBpedia έναν από τους κεντρικούς κόμβους διασύνδεσης του Παγκοσμίου Ιστού των δεδομένων.
Το σύνολο των RDF dumps που διαθέτει η DBpedia παρουσιάζεται παρακάτω:

- Titles
  Πρόκειται για τις θεμελιώδεις RDF τριάδες που προσδιορίζουν το αναγνωριστικό κάθε οντότητας. Ως οντότητα θεωρείται κάθε έννοια που έχει άρθρο με το όνομα της στη

Wikipedia και ο τίτλος του άρθρου δίνεται ως αντικείμενο της τριάδας. Είναι η μόνη περίπτωση που εμφανίζεται η συμβολοσειρά του τίτλου στην τριάδα καθώς σε κάθε άλλη τριάδα που αναφέρεται στην συγκεκριμένη έννοια χρησιμοποιείται το αναγνωριστικό που εντοπίζεται στο υποκείμενο της συγκεκριμένης τριάδας

- Short Abstracts
  Περιέχει τις πρώτες 500 λέξεις από το abstract του αντίστοιχου άρθρου της Wikipedia.

- Extended Abstracts
  Περιέχει ολόκληρο το abstract του αντίστοιχου άρθρου της Wikipedia.

- Images
  Το dump αυτο περιέχει το σύνολο των πληροφοριών που σχετίζονται με τις εικόνες στα άρθρα της wikipedia. Αφ'ενός, αντιστοιχίζει κάθε εικόνα σε κάθε άρθρο της Wikipedia στο αναγνωριστικό URI που αντιστοιχεί στο συγκεκριμένο άρθρο. το κατηγόρημα περιγράφει τον τύπο της εικόνας *<http://xmlns.com/foaf/0.1/thumbnail>, <http://xmlns.com/foaf/0.1/depiction>* Με τη χρήση του *<http://xmlns.com/foaf/0.1/thumbnail>* προσδιορίζει τη θέση της εικόνας στο αρχείο εικόνων της Wikipedia και με τη χρήση του <http://purl.org/dc/elements/1.1/rights> προσδιορίζει τα δικαιώματα κάθε εικόνας στο αρχείο αυτό.

- Geographic Coordinates
  Παρουσιάζει για κάθε περιοχή που περιγράφεται στη Wikipedia τις εξαγόμενες πληροφορίες σχετικά με τις συντεταγμένες της. Σε κάθε περιοχή αντιστοιχούν τέσσερις τριάδες προκειμένου να προσδιοριστούν πλήρως οι συντεταγμένες, να διαχωριστεί το latitude από το longitude, και να είναι εμφανής η πληροφορία πως το αναγνωριστικό URI της οντότητας έχει πληροφορία συντεταγμένων.

- Raw Infobox Properties
  Περιέχει το σύνολο των εξαγόμενων πληροφοριών από το πλαίσιο infobox κάθε άρθρου. Για κάθε τύπο πληροφορίας χρησιμοποιείται το: *<http://dbpedia.org/property/something>* ως κατηγόρημα, με τη χρήση της κατάλληλης λέξης που προσδιορίζει την ιδιότητα αυτή.

- Raw Infobox Property Definitions
  Προσδιορίζουν τη συμβολοσειρά και τον τύπο που αντιστοιχεί σε κάθε αναγνωριστικό URI των ιδιοτήτων που χρησιμοποιούνται για το infobox.

- Homepages
  Δίνει το σύνδεσμο στην κεντρική σελίδα του ιστότοπου της ανίστοιχης οντότητας αν αυτός υπάρχει.

- PersonData
  Για όλες τις οντότητες που αναφέρονται σε πρόσωπα δίνει το σύνολο των σχετικών με άτομα πληροφοριών που έχουν εξαχθεί από την Wikipedia και μπορούν να οργανωθούν είτε στην οντολογία της DBpedia είτε στη foaf, την purl, και την τυπική οντολογία για RDF. Οι αντίστοιχες τριάδες για τη γερμανική γλώσσα βρίσκονται στο dump PND.

- Articles Categories
  Αντιστοιχίζει κάθε οντότητα (το χαρακτηριστικό της URI) με τα URI's των κατηγοριών της Wikipedia στις οποίες ανήκει.

- Categories (Labels)
  Προσδιορίζει τη συμβολοσειρά που αντιστοιχεί στο αναγνωριστικό URI κάθε κατηγορίας.

- Categories (Skos)
  Περιέχει πληροφορίες για τις κατηγορίες (categories) που χρησιμοποιούνται. Τα κατηγορήματα που χρησιμοποιούνται σε αυτό το σύνολο δεδομένων είναι δύο τύπων. Είτε ανήκουν στην οντολογία SKOS και συνδέουν κατηγορίες μεταξύ τους με σχέσεις γενίκευσης ή σημασιολογικής σχετικότητας, είτε ανήκουν στην τυπική οντολογία του w3-RDF και χρησιμοποιούνται για να προσδιορίσουν τον τύπο της κάθε κατηγορίας. (πχ Consept)

- External Links
  Συνδέει το αναγνωριστικό κάθε οντότητας με εξωτερικούς συνδέσμους που παρέχονται στο αντίστοιχο άρθρο της wikipedia.

- Links to Wikipedia Article
  Περιέχει πληροφορίες για τη σελίδα στην Wikipedia που αντιστοιχεί στο αναγνωριστικό κάθε οντότητας. Συγκεκριμένα με το κατηγόρημα *<http://xmlns.com/foaf/0.1/page>* δίνει το σύνδεσμο στη σελίδα, με το *<http://purl.org/dc/elements/1.1/language>* τη γλώσσα του άρθρου και με το *<http://xmlns.com/foaf/0.1/primaryTopic>* κάνει την αντίστροφη σύνδεση συνδέοντας το URL κάθε σελίδας με το αντίστοιχο αναγνωριστικό URI της οντότητας που περιγράφει.

- Wikipedia Pagelinks
  Περιέχει όλες τις τριάδες που αφορούν εσωτερικούς συνδέσμους ανάμεσα στα άρθρα της wikipedia. Το υποκείμενο κάθε τριάδες είναι η οντότητα στης οποίας το άρθρο εντοπίζουμε σύνδεσμο σε άλλο άρθρο, και το αντικείμενο είναι το αναγνωριστικό URI της οντότητας στην οποία δείχνει ο σύνδεσμος.

- Redirects
  Περιέχει όλες τις τριάδες που αφορούν στις οντότητες που αν και είναι καταχωρημένες στη Wikipedia δεν έχουν ξεχωριστό άρθρο αλλά ανακατευθύνουν το χρήστη σε κάποιο άλλο άρθρο. Κάθε τριάδα υποδεικνύει σε ποιά οντότητα γίνεται η ανακατεύθυνση.

- Disambiguation Links
  Περιέχει όλες τις τριάδες που αφορούν στις οντότητες που είναι καταχωρημένες στη Wikipedia αλλά υπάρχουν περισσότερα του ενός άρθρα που μπορεί να αντιστοιχούν σε αυτές. Για κάθε τέτοια οντότητα υπάρχουν τόσες τριάδες όσες και τα άρθρα που μπορεί να αντιστοιχούν σε αυτές. Τα αρθρα δηλαδή που θα εμφανίζονταν στη σελίδα αποσαφήνισης της Wikipedia αν γινόταν αναζήτηση της συγκεκριμένης οντότητας.

- Page IDs
  Περιέχει όλες τις τριάδες που συσχετίζουν το URL κάθε αρθρου στη Wikipedia με το

χαρακτηριστικό του ID. Υπάρχει επίσης και το αντίστοιχο Revision ID κάθε έννοιας το οποίο περιέχεται στο αντίστοιχο dump.

- Ontology Infobox Types
  Περιέχει όλες τις τριάδες της μορφής: $object rdf:type $class οι οποίες αντλήθηκαν σαν πληροφορία από τα infoboxes της wikipedia.

- Ontology Infobox Properties
  Περιέχει τριάδες που φέρουν πληροφορία από τα infoboxes της Wikipedia και έχουν εξαχθεί με τη μέθοδο αυστηρής εξαγωγής, βασισμένης σε οντολογίες (strict ontology based extraction)[3]. Είναι πολύ σημαντικό το γεγονός πως το συγκεκριμένο σύνολο τριάδων, περιέχει πολύ υψηλού επιπέδου πληροφορία καθώς όλα τα κατηγορήματα είναι της μορφής: *http://dbpedia.org/ontology/NAME* και υπάρχει ένα κατηγόρημα και μόνο για κάθε υπονοούμενη σχέση. Αυτό σημαίνει ότι αν σε δύο διαφορετικά λεξιλόγια χρησιμοποιούνται κατηγορήματα που αντιπροσωπεύουν τη σχέση "τόπος_γέννησης", όλα αυτά τα διαφορετικά κατηγορήματα έχουν ενοποιηθεί σε ένα το οποίο βρίσκεται στο χώρο ονομάτων που προαναφέρθηκε.

- Ontology Infobox Properties Specific Περιέχει τριάδες που φέρουν πληροφορία από τα infoboxes της Wikipedia και έχουν εξαχθεί με τη μέθοδο χαλαρής εξαγωγής, βασισμένης σε οντολογίες (loose ontology based extraction) [3]

Επιπλέον των παραπάνω συνόλων δεδομένων τα οποία βασίζονται στην πληροφορία της Wikipedia, γίνεται αοπόπειρα διάθεσης των διασυνδεδεμένων δεδομένων άλλων γνωσιακών βάσεων, οι οποίες περιγράφουν επίσης οντότητες της DBpedia. Συγκεκριμένα είναι διαθέσιμα δεδομένα από τις εξής γνωσιακές βάσεις: *RDF Bookmashup, Bricklink, DailyMed, DBLP, Diseasome, DrugBank, EUNIS, Eurostat, CIA, Factbook, flickr, wrappr, Freebase, Geonames, GeoSpecies, GADM, Project, Gutenberg, Italian, Public, Schools, LinkedMDB, MusicBrainz, New, York, Times, Cyc, Revyu, SIDER, TCMGeneDIT, Umbel, US, Census, WikiCompany, WordNet, YAGO2*. Τα δεδομένα αυτά περιέχουν RDF τριάδες της μορφής:
<http://dbpedia.org/resource/Οντότητα1> <http://www.w3.org/2002/07/owl#sameAs> <http://www4.wiwiss.fu-berlin.de/diseasome/resource/Αντίστοιχη_Οντότητα1>,
που ουσιαστικά συνδέουν την οντότητα της DBpedia με τον υπόλοιπο Σημασιολογικό Ιστό. Τα δεδομένα αυτά, αν και περιορισμένα προς το παρόν, είναι ιδιαίτερης σημασίας καθως θέτουν τις βάσεις για την ενοποίηση των αναγνωριστικών URIs με βάση τους κανόνες του LOD και τη δυνατότητα άμεσης πρόσβασης σε κάθε δεδομένο του Ιστού με την κατοχή ενός μόνο αναγνωριστικού URI που προσδιορίζει μία συγκεκριμένη έννοια.

2.5 Σχετικές Εργασίες

Έχει καταστεί σαφές από τα προηγούμενα κεφάλαια, οτι στο πλαίσιο της διπλωματικής αυτής γίνεται μια απόπειρα σύνθεσης, της θεωρίας και των τεχνολογιών δύο διαφορετικών

πεδίων της πληροφορικής. Συγκεκριμένα, χρησιμοποιούνται τεχνικές από το πεδίο της εξόρυξης κειμένου και της επεξεργασίας φυσικής γλώσσας, προκειμένου να γίνει η αρχική επιλογή των προς εξέταση όρων του αρχικού κειμένου. Παράλληλα χρησιμοποιείται ο Σημασιολογικός Ιστός και τεχνολογίες που σχετίζονται με διασυνδεδεμένα δεδομένα, προκειμένου να γίνει η τελική επιλογή, η σημασιολογική επισημείωση αλλά και η κατάταξη των αποτελεσμάτων που προωθούνται στο χρήστη. Προκειμένου λοιπόν να γίνει επιλογή των τεχνικών που θα χρησιμοποιηθούν και από τα δύο πεδία, μελετήθηκαν πρακτικές και θεωρητικές εργασίες που ανήκουν είτε στην μία είτε στην άλλη κατηγορία. Ιδιαίτερη σημασία δόθηκε σε εργασίες που χρησιμοποιούν ίδιες ή αντίστοιχες πηγές δεδομένων, όπως η DBpedia και η Wikipedia. Στη συνέχεια παρουσιάζονται οι εργασίες αυτές, οι οποίες θεωρήθηκε ότι σχετίζονται με την παρούσα διπλωματική (είτε ως προς τον στόχο, είτε ως προς τις χρησιμοποιούμενες τεχνικές), οργανωμένες ανάλογα με το επιστημονικό πεδίο στο οποίο εντάσσονται.

2.5.1 Εξαγωγή σημαντικών όρων με χρήση συλλογών κειμένων

Η αυτοματοποιημένη εξαγωγή σημαντικών όρων, είτε με τη μορφή φράσεων και λέξεων κλειδιών, είτε με τη μορφή όρων ευρετηρίου, είναι μια εργασία η οποία έχει απασχολήσει μεγάλο μέρος της επιστημονικής κοινότητας που ασχολείται με την Εξόρυξη Κειμένου. Μάλιστα, πρόκειται για μία διαδικασία που εκτός από τη σημασία που φέρει ούτως ή άλλως η επιτυχημένη αυτοματοποίηση της, είναι ιδιαίτερα σημαντική και ως ενδιάμεσο στάδιο για την επίτευξη άλλων σχετικών εργασιών όπως η κατηγοριοποίηση κειμένων, η ομαδοποίηση κειμένων, η περίληψη κειμένων κλπ. Στη διεθνή βιβλιογραφία υπάρχει πληθώρα προτεινόμενων μεθόδων οι οποίες χρησιμοποιούν ως πηγή δεδομένων συλλογές κειμένων που διαθέτουν σύνολα από φράσεις-κλειδιά προσημειωμένες από ειδικούς. Ωστόσο, τα τελευταία χρόνια έχουν γίνει απόπειρες χρήσης διαφορετικών πηγών δεδομένων, όπως είναι η Wikipedia,[52] λόγω των ιδιαίτερων χαρακτηριστικών που παρουσιάζουν. Συγκεκριμένα, τέτοιες πηγές, αν και δε διαθέτουν σύνολα προσημειωμένων όρων, διαθέτουν δομικά στοιχεία εμφανώς διαχωρισμένα από το κυρίως σώμα του κειμένου και εύκολα επεξεργάσιμα από εφαρμογές. Ακόμα διαθέτουν χωρισμό κειμένων σε κατηγορίες και δομές διασύνδεσης των κειμένων μεταξύ τους (υπερσύνδεσμοι - hyperlinks), καθώς και δεδομένα για αποσαφήνιση και συσχέτιση όρων. Τέλος, το πλήθος και το εύρος της θεματολογίας των κειμένων τους είναι ιδιαίτερα θετικοί παράγοντες που καθιστούν τέτοιου τύπου διαδικτυακές εγκυκλοπαίδειες κατάλληλες για χρήση.

Παρακάτω παρουσιάζονται δύο ερευνητικές εργασίες, οι οποίες χρησιμοποιούν μεθόδους εξόρυξης πληροφορίας από τη Wikipedia προκειμένου να εξάγουν σημαντικές φράσεις από μη επεξεργασμένα κείμενα. Στην πρώτη εργασία που παρουσιάζεται, η εξαγωγή των φράσεων αξιοποιείται για αυτόματη κατηγοριοποίηση κειμένων, ενώ στη δεύτερη για καλύτερη ομαδοποίηση κειμένων.

Human-competitive automatic topic indexing

Αναλυτικότερα, στη διδακτορική της διατριβή, με τίτλο ”Human-competitive automatic topic indexing” [23], η Olena Medelyan χρησιμοποιεί την εξόρυξη κειμένου από τη Wikipedia σε συν-

δυασμό με την τεχνική ανάθεσης όρων (term assignment) από λεξικό και την τεχνική αυτοματοποιημένης ετικετοποίησης (automatic tagging), προκειμένου να καταλήξει σε φράσεις που επαρκούν για την ανάθεση κατηγοριών και υπο-κατηγοριών σε κείμενα.

Κατά την υλοποίηση και των τριών τεχνικών, κρίνεται απαραίτητη η εξαγωγή από το αρχικό κείμενο όλων των όρων που είναι υποψήφιοι προς εξέταση. Η προσέγγιση που επιλέγεται είναι η εξαντλητική εξόρυξη του κειμένου ώστε να εξαχθεί κάθε πιθανός συνδυασμός n-grams, φράσεων που αποτελούνται από το πολύ n-λέξεις. Στη συγκεκριμένη προσέγγιση δε λαμβάνεται υπόψη στην αρχική εξόρυξη κανένας γραμματικός ή συντακτικός κανόνας. Στη συνέχεια το διάνυσμα των n-grams κανονικοποιείται (ανάλογα με την τεχνική στην οποία θα αξιοποιηθεί) με τη χρήση μεθόδων όπως η αφαίρεση τερματικών λέξεων (stopwords), μετατροπή κεφαλαίων χαρακτήρων σε πεζούς, περιστολή λέξεων και αναδιάταξη λέξεων. Οι επιλεγμένοι όροι, στη συνέχεια, θα αξιοποιηθούν με χρήση των τριών τεχνικών που αναφέρονται παραπάνω. Από τις τεχνικές αυτές, αξίζει να αναλυθεί η χρήση της Wikipedia, καθώς σχετίζεται περισσότερο με το αντικείμενο της παρούσας διπλωματικής εργασίας.

Για την αρχική ανάδειξη των n-grams που θα διατηρηθούν στην έξοδο, γίνεται απόπειρα αντιστοίχησης τους με κάποιο τίτλο άρθρου της Wikipedia, καθώς ο τίτλος κάθε άρθρου θεωρείται πως αντιπροσωπεύει μία αυτούσια έννοια/οντότητα. Σύμφωνα με την Medelyan, η χρήση της Wikipedia με την παραπάνω μέθοδο συνιστά σημαντική πρόκληση, καθώς το σύνολο δεδομένων της είναι ιδιαίτερα εκτενές. Αυτό συνεπάγεται, οτι το σύνολο των όρων που εντοπίζονται ως τίτλοι άρθρων της Wikipedia είναι μη ικανοποιητικό καθώς περιέχει και αρκετά γενικούς όρους οι οποίοι δεν παρουσιάζουν σημασιολογική συγγένεια με το θέμα του κειμένου. Για το σκοπό αυτό απαιτείται περαιτέρω επεξεργασία και διαχωρισμός των υποψήφιων φράσεων. Ο διαχωρισμός αυτός υλοποιείται με χρήση της μετρικής keyphraseness, η οποία δίνεται από τον τύπο:

$$keyphraseness = \frac{L}{C},$$

όπου L ο αριθμός των συνολικών εμφανίσεων ενός όρου στα άρθρα της Wikipedia ως σύνδεσμος και C ο αριθμός των συνολικών εμφανίσεων ενός όρου στη Wikipedia ανεξάρτητα αν είναι σύνδεσμος ή όχι. Με τον τρόπο αυτό, κατατάσσονται οι φράσεις με βάση την "ειδικότητα τους" ως προς το θέμα του κειμένου. Χρησιμοποιώντας κάποιο κατώφλι και κρατώντας μόνο τα n-grams με keyphraseness μεγαλύτερο του κατωφλίου αυτού, αποκλείεται μεγάλο μέρος των όρων, οι οποίοι δε φέρουν σημαντική πληροφορία ως προς το νόημα του κειμένου (όροι που μπορούν να αλλάξουν χωρίς να επηρεάζεται σημαντικά το θέμα και το νόημα του αρχικού κειμένου).

Στη συνέχεια γίνεται επιπρόσθετη επεξεργασία των επιλεγμένων όρων, με σκοπό να αντιστοιχηθούν καλύτερα (ώστε να φέρουν το ορθό νοηματικό περιεχόμενο) στους τίτλους των άρθρων της Wikipedia. Συγκεκριμένα, αν κάποιο n-gram αντιστοιχεί σε καταχώρηση της Wikipedia που ανακατευθύνει (redirects) σε άρθρο με διαφορετικό τίτλο, γίνεται αυτόματη ανανέωση του όρου και αλλαγή του n-gram με το σωστό τίτλο του άρθρου. Επιπλέον, σε περίπτωση που υπάρχουν αμφισημίες λόγω του πολλαπλού νοηματικού περιεχομένου ενός όρου, χρησιμοποιούνται

οι σελίδες αποσαφήνισης (disambiguation pages) της Wikpedia με σκοπό να γίνει επιλογή του σωστού νοήματος του όρου. Για το σκοπό αυτό, υπολογίζεται η σημασιολογική συγγένεια του αρχικού κειμένου με το κάθε άρθρο της Wikipedia το οποίο φέρει ως τίτλο τον υπο-εξέταση όρο και σε συνάρτηση με τη γενικότητα ή μή του όρου επιλέγεται ο τίτλος του οποίου το άρθρο φέρει τη μέγιστη σημασιολογική συγγένεια (σημειώνεται πως οι μέθοδοι υπολογισμού της συγγένειας βασίζονται στο [26] και σχετίζονται με το πλήθος των όρων του κειμένου που ειναι κοινοί με όρους-συνδέσμους του κάθε άρθρου).

Η αξιολόγηση του αποτελέσματος της παραπάνω μεθόδου έγινε με χρήση συνόλων προσημειωμένων κειμένων, τα οποία μάλιστα είχαν επισημειωθεί περισσότερες από μία φορές από διαφορετικές ομάδες ειδικών. Σύμφωνα με τα αποτελέσματα οι όροι που εντοπίστηκαν με την παραπάνω μέθοδο παρουσιάζουν recall (ανάκληση) πάνω από 50% σε κάθε περίπτωση. Μάλιστα, σημειώνεται πως όσον αφορά τους όρους οι οποίοι είχαν προταθεί από τουλάχιστον τρεις ομάδες ειδικών (δηλαδή τους όρους αυτούς που συνοψίζουν καλύτερα το θέμα του κειμένου), τα ποσοστά ανάκλησης ήταν μεγαλύτερα του 70%, γεγονός που υποδεικνύει πως η απόδοση της συγκεκριμένης μεθόδου αυξάνεται όσο μειώνεται το πλήθος των όρων που επιζητούνται ως αποτέλεσμα και όσο γίνεται πιο συγκεκριμένο το θέμα και το περιεχόμενο των όρων.

Exploiting Wikipedia Knowledgefor Conceptual Hierarchical Clustering of Documents

Στην εργασία με τίτλο "*Exploiting Wikipedia Knowledge for Conceptual Hierarchical Clustering of Documents*" [36], χρησιμοποιούνται αντίστοιχες τεχνικές με αυτές που παρουσιάσαμε παραπάνω, σε συνδυασμό με την χρήση μεθόδων εξόρυξης κειμένου, προκειμένου να επιτύχει ομαδοποίηση κειμένων. Υποστηρίζεται, μάλιστα, οτι η προσέγγιση αυτή επιτυγχάνει να αντιμετωπίσει τα μειονεκτήματα της πιο δημοφιλούς μεθόδου που χρησιμοποιείται για το συγκεκριμένο σκοπό, του σάκου με λέξεις (Bag-of-Words). Συγκεκριμένα η χρήση της Wikipedia επιλέχθηκε γιατί δίνει τη δυνατότητα αξιοποίησης του σημασιολογικού περιεχομένου κάθε όρου για την κατάταξη και επιλογή των όρων αλλά και δεν απαιτεί μεγάλο διανυσματικό χώρο για την υλοποίηση της.

Προκειμένου να επιλεγεί το σύνολο των υπό εξέταση όρων, αντί για n-grams, προτιμάται η γραμματική επισημείωση κάθε λέξης, και η εξαγωγή ονοματικών φράσεων με χρήση απλών κανόνων που βασίζονται στην επισημείωση ώστε να επιλέξουν τους όρους με την κατάλληλη αλληλουχία λέξεων. Από το διάνυσμα αυτό οι φράσεις ελέγχονται στη συλλογή κειμένων της dbpedia και διατηρούνται μόνο οι όροι οι οποίοι εντοπίζονται ώς τίτλοι άρθρου. Ακολουθούνται και πάλι επεξεργασίες όσον αφορά την αποσαφήνιση λέξεων, αντίστοιχες με αυτές που χρησιμοποιούνται στην [23].

Για την περαιτέρω διαλογή των όρων που εντοπίστηκαν, αξιοποιούνται οι εξής μετρικές:

- Σταθμισμένη συχνότητα (weighted frequency): Γινόμενο του αριθμού εμφανίσεων μίας φράσης στο κείμενο και του αριθμού λέξεων της φράσης
- Κατάταξη Συνδέσμων (LinkRank) : Μέτρική που αντιπροσωπεύει το πλήθος των κοινών συνδέσμων μεταξύ του αρχικού κειμένου και του άρθρου της Wikipedia που αντιπροσωπεύει μια φράση του αρχικού κειμένου.

- Εννοιολογική συγγένεια (Concept similarity): Συσχέτιση μεταξύ των διανυσμάτων tf-idf του αρχικού κειμένου και άρθρου της Wikipedia που αντιπροσωπεύει μια φράση του αρχικού κειμένου. Εξάγεται με βάση το cos() ο βαθμός συσχέτισης για κάθε φράση.
- Σειρά Εμφάνισης (Order Rank) : Αντιπροσωπεύει ένα μέτρο της θέσης της πρώτης εμφάνισης μίας φράσης στο κείμενο, δίνοντας μεγαλύτερο βάρος σε αυτές που εντοπίζονται στην αρχή του κειμένου θεωρώντας ότι είναι σημαντικές για το κείμενο.
- Εξειδίκευση ως προς το θέμα του κειμένου (Keyphraseness) : Πρόκειται για το ίδιο χαρακτηριστικό που περιγράφεται και στην [23]

Οι παραπάνω μετρικές συνυπολογίζονται προκειμένου να εξαχθεί ένα τελικό διάνυσμα φράσεων μαζί με το αντίστοιχο βάρος της κάθε μίας το οποίο αντιπροσωπεύει τη σημασία της ως προς το κείμενο. Σύμφωνα με τη συγκεκριμένη μελέτη με την χρήση κατάλληλων βαρών, επιτυγχάνεται ιεράρχηση των φράσεων ώστε αυτές με το μεγαλύτερο συνολικό βάρος να μπορούν να χρησιμοποιηθούν ως έννοιες (concepts) κατάλληλες για χρήση σε τεχνικές ομαδοποίησης κειμένου.

Συμπερασματικά, και δεδομένων και άλλων εργασιών όπως [52] [26] οι οποίες επιδιώκουν τη χρήση της Wikipedia για την αξιολόγηση φράσεων και την εξαγωγή σημασιολογικών συμπερασμάτων για φράσεις και κείμενα, μπορούμε να ισχυριστούμε πως υπάρχει μία στροφή στην κατεύθυνση των διευρυμένων διαδικτυακών πηγών δεδομένων αλλά και της σημασιολογικής επισημείωσης φράσεων κατά τη διάρκεια της εξόρυξης κειμένου με σκοπό τη βελτίωση της εξαγωγής πληροφορίας από κείμενο.

2.5.2 Εξαγωγή Πληροφορίας σχετιζόμενη με το Σημασιολογικό Ιστό

Στην ενότητα αυτή, θα παρουσιαστούν εργασίες που έχουν γίνει πάνω στο αντικείμενο της εξαγωγής πληροφορίας και σχετίζονται άμεσα με το σημασιολογικό ιστό, είτε χρησιμοποιώντας τμήμα του για εξαγωγή συμπερασμάτων, είτε ενισχύοντας τις προσπάθειες εμπλουτισμού του. Φαίνεται, πως καθώς έχουν αρχίσει να διαφαίνονται οι δυνατότητες που προσφέρει κυρίως για την εξαγωγή και ανάκτηση πληροφορίας, το ενδιαφέρον στρέφεται όλο και περισσότερο στην δημιουργία εφαρμογών που θα τις αξιοποιούν ή που θα συμβάλλουν στην επέκτασή τους. Οι εργασίες που ακολουθούν, ασχολούνται κυρίως με αυτά τα δύο ζητήματα και σε κάποιο βαθμό χρησιμοποιούν ως προστάδια επεξεργασίας ιδέες και μεθόδους που περιγράφηκαν στις σχετικές εργασίες της προηγούμενης ενότητας. Κατά την περιγραφή τους, στο σημείο αυτό, θα δοθεί έμφαση στα τμήματα που αφορούν τη διπλωματική αυτή και αξιοποιήθηκαν κατά την υλοποίηση των συστημάτων CRESTA και SWPID.

Ontology Based Information Extraction

Στη διπλωματική εργασία με τίτλο "*Ontology Based Information Extraction*" [46] του Carlos Vicient Monllao , περιγράφεται μία προσπάθεια σημασιολογικής επισημείωσης ενός τυχαίου κειμένου με χρήση οντολογίας. Η προσπάθεια αυτή, έγινε στο πλαίσιο του προγράμματος DAMASK (Data Mining Algorithms with Semantic Knowledge-Αλγοριθμοι Εξόρυξης Δεδομένων με Σημασιολογική

Πληροφορία) το οποίο επιδιώκει την άντληση πληροφορίας από κείμενα που προέρχονται από διαφορετικές πηγές με χρήση υπαρχουσών οντολογιών, την κατηγοριοποίησή τους, και την τελική σημασιολογική μετάφραση των κατηγοριών του προηγούμενου βήματος. Ολόκληρο το πρόγραμμα αυτό, μπορεί να ενταχθεί στην συνολικότερη προσπάθεια που γίνεται ανάπτυξης μεθόδων και τεχνολογιών επέκτασης του Σημασιολογικού Ιστού. Η εργασία που ενδιαφέρει την παρούσα διπλωματική αφορά το πρώτο τμήμα του προγράμματος.

Στην εν λόγω εργασία, η εξαγωγή πληροφορίας γίνεται από κείμενα τα οποία περιγράφουν μία συγκεκριμένη έννοια του πραγματικού κόσμου και χωρίζεται σε τρία στάδια. Στο πρώτο στάδιο γίνεται η εξαγωγή ονοματικών οντοτήτων οι οποίες θεωρείται ότι σχετίζονται άμεσα με την περιγραφόμενη έννοια. Στη συνέχεια, επιχειρείται η ανεύρεση γενικότερων εννοιών στις οποίες ανήκει κάθε έννοια. Τέλος, για κάθε μία από τις γενικές αυτές έννοιες, γίνεται μια προσπάθεια αντιστοίχισής της με κλάσεις οντολογίας, η οποία περιγράφει επαρκώς τον τομέα του θέματος του κειμένου.

Η εξαγωγή ονοματικών οντοτήτων στο πρώτο επίπεδο της επεξεργασίας του κειμένου γίνεται με χρήση κανόνων της αγγλικής γλώσσας, που λαμβάνουν υπ'όψιν την χρήση κεφαλαίων γραμμάτων και τα πρότυπα με βάση τα οποία συντίθεται στη γλώσσα αυτή τέτοιου είδους φράσεις (φράσεις που περιγράφουν έννοιες του πραγματικού κόσμου). Οι διαδικασίες που επιτελούν την εργασία αυτή περιγράφονται και στο [29]. Οι κανόνες που αναφέρθηκαν, χρησιμοποιούνται κυρίως όταν πρόκειται για αδόμητο, ακατέργαστο κείμενο, ενώ στην περίπτωση που χρησιμοποιούνται άρθρα της Wikipedia είναι δυνατό να γίνει αξιοποίηση των υπερσυνδέσμων που υπάρχουν. Το βασικό πρόβλημα, πέραν της αναγνώρισης όλων των ονοματικών οντοτήτων από το κείμενο, είναι η επιλογή αυτών που σχετίζονται με την περιγραφόμενη από το κείμενο έννοια. Όλες οι υπόλοιπες, θεωρούνται θόρυβος και δεν πρέπει να ληφθούν υπ'όψιν κατά την περαιτέρω επεξεργασία. Ο διαχωρισμός αυτός γίνεται με τη χρήση του παγκόσμιου ιστού και ενός μέτρου συσχετισμού μεταξύ της περιγραφόμενης έννοιας και της υποψήφιας ονοματικής οντότητας, ο υπολογισμός του οποίου γίνεται με αναζήτηση της συχνότητας των κοινών εμφανίσεων των δύο φράσεων σε κείμενα του ιστού, σε σχέση με την εμφάνιση αποκλειστικά της υποψήφιας ονοματικής οντότητας. Το συγκεκριμένο μέτρο θεωρείται αξιόπιστο και δίνει μία λίστα ονοματικών οντοτήτων πάνω στις οποίες βασίζεται η κυρίως επεξεργασία.

Στο επόμενο στάδιο, γίνεται προσπάθεια σημασιολογικής επισημείωσης, σε δύο επίπεδα. Στο πρώτο επίπεδο, γίνεται η ανεύρεση γενικών εννοιών που μπορούν να θεωρηθούν ως κατηγορίες στις οποίες εντάσσονται κάθε μία από τις επικρατούσες οντότητες του προηγούμενου σταδίου, ενώ σε δεύτερο επίπεδο, χρησιμοποιείται οντολογία για να γίνει ταίριασμα μεταξύ των παραπάνω κατηγοριών και κλάσεων με τις οποίες συνδέονται. Η αναζήτηση των γενικών εννοιών (οι οποίες μπορούν να θεωρηθούν και ως κατηγορίες), γίνεται με χρήση αναζήτησης προτύπων στο διαδίκτυο, τα οποία υποννοούν κάποιο βαθμό ιεραρχίας. Όταν βρεθούν οι γενικές έννοιες κάθε οντότητας, γίνεται χρήση της επιλεγμένης για το εκάστοτε κείμενο οντολογίας, με στόχο να αντιστοιχηθεί κάθε έννοια-κατηγορία με μία κλάση της οντολογίας ώστε να γίνει η σημασιολογική επισημείωση. Η αντιστοίχιση αυτή μπορεί να γίνεται αποκλειστικά με σύγκριση των ονομάτων της κλάσης και της έννοιας-κατηγορίας. Αν αποτύχει, οι κατηγο-

ρίες επεκτείνονται σε υπο/υπερ-κατηγορίες (και με χρήση του Wordnet) ώστε να γίνει εκ νέου προσπάθεια ταιριάσματος.

Από τα πολύ σημαντικά στοιχεία της εργασίας που μόλις περιγράφηκε είναι η χρήση του Παγκόσμιου Ιστού συνολικά σαν πηγή πληροφορίας τόσο για την επιλογή των σχετικών με το κείμενο εννοιών, όσο και για την ανεύρεση και επιλογή των κατηγοριών των ονοματικών οντοτήτων. Επιπλέον, η χρήση οντολογιών, μετατρέπει την εξαγόμενη περιγραφή, από περιγραφή στιγμιοτύπων, σε περιγραφή κλάσεων.

Semantic Text Mining with Linked Data

Στα πρακτικά του πέμπτου Διεθνούς Συνεδρίου για τα INC, IMS και IDC, εντοπίζεται μία ιδιαίτερα αξιόλογη εργασία, με τίτλο "*Semantic Text Mining with Linked Data*"[16] η οποία συνδυάζει τις τεχνικές εξόρυξης ιστού και εξόρυξης κειμένου με τεχνικές που σχετίζονται με το Σημασιολογικό Ιστό και τα διασυνδεδεμένα δεδομένα, με σκοπό να οδηγήσει στη δημιουργία νέων σημασιολογικών γράφων από απλά κείμενα του διαδικτυακού ιστού και στη συνέχεια να συνδέσει τους γράφους αυτούς στο Σημασιολογικό Ιστό, επιτυγχάνοντας τη μετατροπή μίας ιστοσελίδας απλού κειμένου, σε εμπλουτισμένη και σημασιολογικά επισημειωμένη πληροφορία, και επεκτείνοντας παράλληλα το Σημασιολογικό Ιστό.

Στη συγκεκριμένη εργασία εισάγεται και χρησιμοποιείται η έννοια της ονοματικής οντότητας (name entity) αντί για ονοματική φράση ή όρος. Μάλιστα, η υλοποίηση περιορίζεται σε ένα πολύ συγκεκριμένο πλαίσιο αποδεκτών οντοτήτων που συνίστανται σε πρόσωπα, περιοχές, εταιρίες και οργανισμούς, καθώς θεωρείται δύσκολος ο ακριβής εντοπισμός άλλων οντοτήτων με αυτοματοποιημένο τρόπο. Για τον εντοπισμό των παραπάνω οντοτήτων, χρησιμοποιείται ο "*Stanford Named Entity Recognizer(SNER)*" στα ανακτημένα τμήματα που συνιστούν το κυρίως κείμενο κάθε ιστοσελίδας. Στη συνέχεια γίνεται απόπειρα δημιουργίας γράφου που συνδέει τις ονοματικές οντότητες που έχουν εξαχθεί για κάθε πρόταση του κειμένου. Οι γράφοι που εξάγονται συντίθενται σε ένα τελικό γράφο.

Οι γράφοι δημιουργούνται με τη μορφή RDF τριπλετών οι οποίες ωστόσο έχουν προς το παρόν κενό κατηγόρημα. Το κατηγόρημα, δηλαδή ο ρόλος (σχέση) που συνδέει κάθε δυάδα ονοματικών οντοτήτων προσδιορίζεται με χρήση βάσεων δεδομένων του Σημασιολογικού Ιστού (όπως η DBpedia) που χρησιμοποιούν επίσης τριάδες RDF. Ένα σημαντικό πρόβλημα στην περίπτωση αυτή είναι η ανάγκη να διατηρηθεί μόνο η χρήσιμη πληροφορία του αρχικού γράφου. Για το σκοπό αυτό, χρησιμοποιείται ο αλγόριθμος εξαγωγής συχνών υπογράφων (frequent subgraph mining). Στη συνέχεια, κάθε τριάδα των επιλεγμένων υπογράφων αντιπαραβάλλεται στην βάση δεδομένων της DBpedia, με σκοπό μέσω ταιρίασματος προτύπων να εντοπισθεί το κατάλληλο κατηγόρημα που συνδέει τις δύο τριάδες μεταξύ τους. Μάλιστα, για τις δυάδες οντοτήτων για τις οποίες δεν εντοπίζεται άμεσα κατηγόρημα χρησιμοποιείται κατάλληλος αναδρομικός αλγόριθμος ο οποίος αναζητά εξαντλητικά με αναζήτηση κατά βάθος σχέση μεταξύ των δύο οντοτήτων με βάση το εξής σκεπτικό: "Αν έχουμε δύο διαφορετικές οντότητες σε δύο διαφορετικές τριάδες ως εξής : "<οντότητα1> <κατηγόρημα1> <οντότητα2>" "<οντότητα2> <κατηγόρημα2> <οντότητα3>" και το αντικείμενο της μίας τριάδας ταυτίζεται με το υποκείμενο της άλλης τότε θεωρείται πως οι δύο οντότητες μπορούν να συσχετιστούν σε μία νέα τριάδα

με κατηγόρημα ”<οντότητα1> <κατηγόρημα1 (οντότητα2) κατηγόρημα2> <οντότητα3>”.

Η εργασία αυτή είναι ιδιαίτερα ενδιαφέρουσα καθώς υποδεικνύει έναν τρόπο επέκτασης του υπάρχοντος Σημασιολογικού Ιστού που ανάγεται σε πλήρως αυτοματοποιημένες διαδικασίες και δεν απαιτεί ανθρώπινη παρέμβαση.

Clustering of Rough Set Related Documents with Use of Knowledge from DBpedia

Στην εργασία με τίτλο ”*Clustering of Rough Set Related Documents with Use of Knowledge from DBpedia*” [41], γίνεται μία προσπάθεια ομαδοποίησης ενός συνόλου κειμένων με κοινή θεματολογία, βασισμένη σε πληροφορίες αντλούμενες από τη βάση γνώσης της DBpedia. Η προσπάθεια αυτή, εντάσσεται στο πλαίσιο του προγράμματος SYNAT, το οποίο έχει σαν στόχο τη δημιουργία μίας ενοποιημένης πλατφόρμας για την αποθήκευση και εξυπηρέτηση πληροφορία αναφερόμενη σε τομείς της επιστήμης και της τεχνολογίας. Η συγκεκριμένη εργασία, ασχολείται με εκείνα τα ζητήματα που αφορούν την ανάπτυξη εργαλείων και μεθόδων τα οποία θα διευκολύνουν τη δημιουργία ευρετηρίου, την αναζήτηση και την ανάκτηση κειμένων εντός μεγάλων συλλογών, με βάση το σημασιολογικό τους περιεχόμενο.

Η προσέγγιση του παραπάνω ζητήματος έγινε με βάση την DBpedia. Στόχος είναι η ομαδοποίηση των κειμένων με σημασιολογικά κριτήρια, στηριζόμενη στις σχέσεις που υπάρχουν ανάμεσα σε κάθε κείμενο και την DBpedia. Συγκεκριμένα, σε πρώτο στάδιο το κείμενο αναλύθηκε σε λέξεις και χρησιμοποιήθηκε το κριτήριο tf-idf για την αξιολόγησή τους. Στη συνέχεια, έγινε προσπάθεια συσχέτισης κάθε λέξης με έννοιες της DBpedia. Για κάθε κείμενο, υπολογίζεται ένας βαθμός συσχέτισης ανάμεσα στο ίδιο το κείμενο και τις έννοιες οι οποίες περιέχονται σε αυτό. Ο υπολογισμός, στηρίζεται στις περιλήψεις (abstracts) που είναι αποθηκευμένα στη βάση γνώσης της DBpedia, για κάθε έννοια. Έτσι, τα κείμενα, αναπαρίστανται πλέον από ένα σύνολο εννοιών (που ανταιστοιχεί περίπου στο 1% των εννοιών που εντοπίστηκαν αρχικά), και με βάση τα σύνολα αυτά, γίνεται η ομαδοποίησή τους.

Σημαντικό στοιχείο στη συγκεκριμένη εργασία είναι η αξιοποίηση της DBpedia ως εξωτερική πηγή πληροφορίας (όπως συνήθως γίνεται με άλλα σώματα δεδομένων του διαδικτύου) και ο εντοπισμός του βαθμού συσχέτισης της κάθε έννοιας με το κείμενο με βάση τις κοινές εμφανίσεις εννοιών σε σχέση με τα abstracts.

Κεφάλαιο 3

Γενική Περιγραφή Συστημάτων

3.1 Εισαγωγή

Στο κεφάλαιο αυτό, θα γίνει παρουσίαση του σχεδιασμού των δύο συστημάτων, της γενικής λειτουργίας τους, των συγκεκριμένων σκοπών που εξυπηρετούν και της σημασίας εκπλήρωσής τους. Επιπλέον θα δοθεί η ανάλυση των κρίσιμων αποφάσεων σχετικά με τα μέσα που χρησιμοποιήθηκαν, τις αρχιτεκτονικές επιλογές και τον τρόπο χειρισμού επιμέρους εργασιών. Λόγω των διαφορετικών απαιτήσεων που υπάρχουν, κρίθηκε απαραίτητο η περιγραφή να χωριστεί σε δύο υποενότητες, μία για κάθε σύστημα. Πρέπει να σημειωθεί, πως τα ζητήματα που αφορούν την καθ' αυτή υλοποίηση, δεν θα συμπεριληφθούν στο κεφάλαιο αυτό.

Κάθε ένα από τα δύο συστήματα, δέχεται στην είσοδό του ένα κείμενο και παρέχει στην έξοδό του ένα σύνολο από όρους που εμφανίζονται αυτούσιοι στο κείμενο και οι οποίοι ικανοποιούν τον στόχο που έχει τεθεί. Επιπλέον, και τα δύο συστήματα αποτελούνται από δύο διακριτά τμήματα μεταξύ τους. Το πρώτο τμήμα αφορά το σχηματισμό του συνόλου P όλων εκείνων των όρων που αποτελούν πιθανές φράσεις-στόχους, ενώ το δεύτερο τμήμα αφορά την επιλογή στοιχείων του συνόλου αυτού για το σχηματισμό του τελικού συνόλου T των φράσεων-στόχων που θα προωθηθούν στην έξοδο. Για τα δύο αυτά σύνολα ισχύει η σχέση:

$P \subseteq T$

Όσον αφορά το πρώτο τμήμα και την επιλογή των υποψήφιων φράσεων-στόχων, βασική προϋπόθεση επιλογής μιας φράσης αποτελεί η δυνατότητα αντιστοίχισής της με μία από τις οντότητες της DBpedia. Η προϋπόθεση εξυπηρετεί δύο σκοπούς. Ο πρώτος έχει να κάνει με το γεγονός πως ο υπολογιστής δεν έχει γνώση του πραγματικού κόσμου και κατ'επέκταση του νοήματος τόσο του κειμένου όσο και κάθε έννοιας που βρίσκεται μέσα σε αυτό και αναπαρίσταται από μία φράση του. Επομένως, είναι ανάγκη προκειμένου να γίνει κάποιου είδους σημασιολογική και όχι μόνο επεξεργασία, να υπάρχουν μία ή περισσότερες πηγές πληροφορίας οι οποίες θα συνθέτουν μία κατά σύμβαση πλήρη αναπαράσταση του κόσμου και στην οποία θα έχει τη δυνατότητα να ανατρέχει το κάθε σύστημα προκειμένου να πάρει τις πληροφορίες που χρειά-

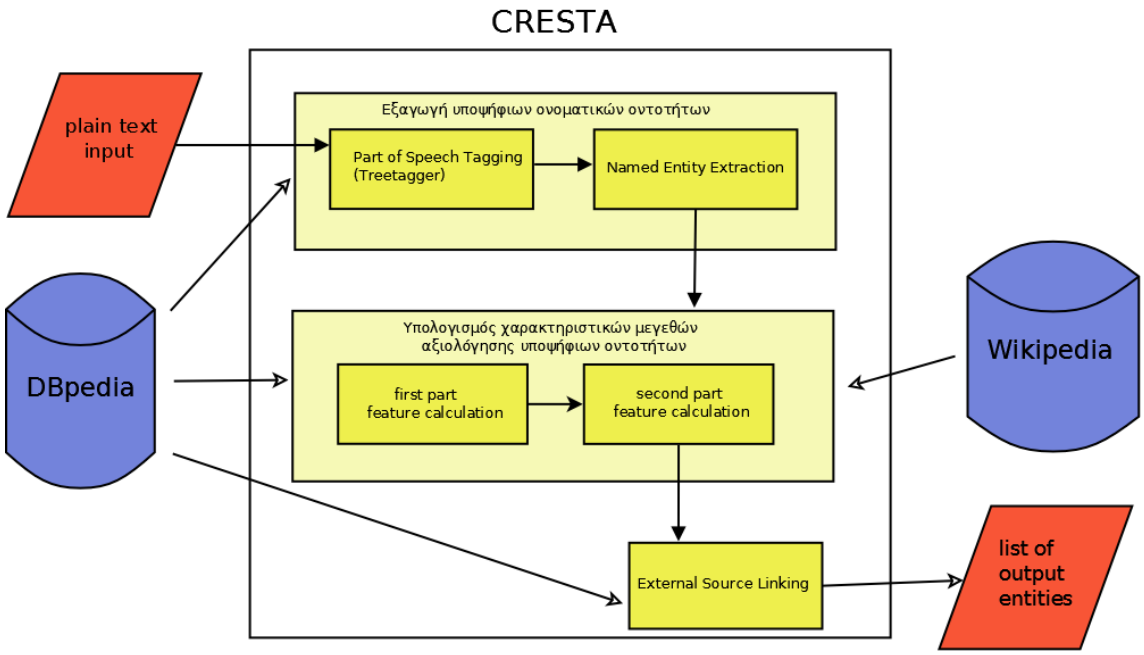
ζεται. Ο δεύτερος, έχει να κάνει με τα πλεονεκτήματα που μπορεί να προσφέρει η σύνδεση του κειμένου με το Σημασιολογικό Ιστό. Η σύνδεση αυτή, αφ'ενός διευκολύνει κατά πολύ την εργασία που γίνεται στο δεύτερο τμήμα των συστημάτων (διαδικασία επιλογής του τελικού συνόλου των όρων που δίνονται σαν έξοδος), αφ' ετέρου εμπλουτίζει τους όρους που εξάγονται μέσω της σημασιολογικής τους επισημείωσης και τους προσδίδει χαρακτηριστικά σημασιολογικής οντότητας. Για το λόγο αυτό, στην έξοδο των δύο συστημάτων δεν δίνονται μόνο οι επιλεγμένες φράσεις-στόχοι όπως αυτές αναφέρονται στο κείμενο, αλλά και ο σύνδεσμος της κάθε μίας με την DBpedia.

Όσον αφορά το δεύτερο τμήμα και την τελική απόφαση θεώρησης μιας φράσης (οντότητας) ως φράση-στόχο, η διαδικασία που ακολουθείται σε κάθε σύστημα διαφοροποιείται σε σημαντικό βαθμό καθώς διαφοροποιείται και ο στόχος εξόδου. Μία βασική διαφοροποίηση είναι το γεγονός πως η έξοδος του συστήματος CRESTA περιλαμβάνει κατάταξη των επιλεγμένων όρων με χρήση ποσοτικοποιημένων κριτηρίων που επιδιώκεται να αναπαραστήσουν με τον καλύτερο δυνατό τρόπο το βαθμό συνοχής καθενός από αυτούς με το θέμα του κειμένου. Αντίθετα, η έξοδος του συστήματος SWPID, δεν περιλαμβάνει ταξινομημένους όρους, ενώ η πληροφορία που επιδιώκεται να δοθεί γι' αυτούς με την είσοδό τους στο τελικό σύνολο αποτελεσμάτων είναι η σύνδεσή τους με οντότητα της DBpedia που αναπαριστά πρόσωπο του πραγματικού κόσμου.

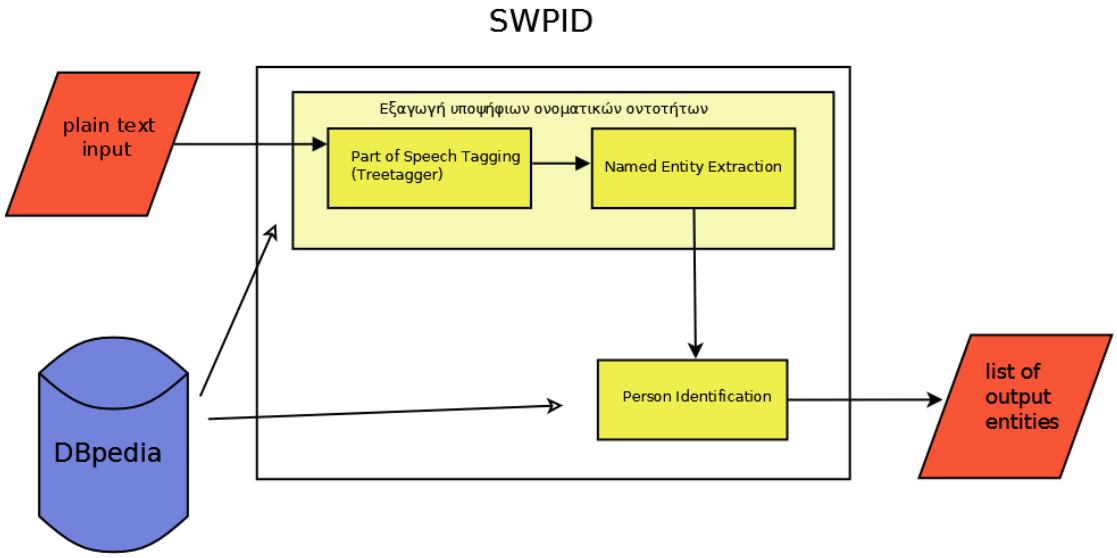
Στο σημείο αυτό, αξίζει να σημειωθεί πως τόσο τα κείμενα εισόδου όσο και τα αποτελέσματα εξόδου των συστημάτων, περιλαμβάνουν δεδομένα διατυπωμένα στην αγγλική γλώσσα. Η επιλογή της γλώσσας αυτής έγκειται στο γεγονός πως υπάρχει πληθώρα εργαλείων επεξεργασίας αγγλικής γλώσσας σε σχέση με οποιαδήποτε άλλη. Επιπλέον αν και η DBpedia έχει καταχωρήσεις σε 17 γλώσσες επιπλέον της αγγλικής, οι καταχωρήσεις αυτές είναι πολύ περιορισμένες σε πλήθος δεδομένων και κρίνεται πως δε μπορούν να θεωρηθούν επαρκείς για την περιγραφή του κόσμου. Επομένως, οι γραμματικοί, συντακτικοί και λεξιλογικοί κανόνες που χρησιμοποιούνται και αναλύονται στη συνέχεια, αναφέρονται εξ ολοκλήρου, στην αγγλική γλώσσα και γραμματική. Παρ' όλα αυτά, με τροποποίηση των κανόνων και προσαρμογή τους στην εκάστοτε γραμματική και σύνταξη, η διαδικασία που περιγράφεται παρακάτω μπορεί να αξιοποιηθεί και για κείμενα σε άλλες γλώσσες.

Η αρχιτεκτονική των δύο συστημάτων είναι συμβατή με την τεχνοτροπία ροής δεδομένων και ακολουθεί το μοντέλο αυλού/φίλτρου (pipes and filters). Δημιουργείται, επομένως, μία ροή δεδομένων, από την είσοδο ως την έξοδο. Η ροή αυτή έχει ως ενδιάμεσα στάδια (φίλτρα) αυτόνομα μέρη προγράμματος, τα οποία δέχονται σαν είσοδο τα δεδομένα του προηγούμενου σταδίου, εκτελούν μια σειρά από υπολογισμούς και προωθούν τα δεδομένα που παράγουν στο επόμενο στάδιο.

Συνολικά η αρχιτεκτονική των δύο συστημάτων φαίνεται στα παρακάτω σχήματα:



Σχήμα 3.1: Αρχιτεκτονική πρώτου συστήματος



Σχήμα 3.2: Αρχιτεκτονική δεύτερου συστήματος

Στις επόμενες ενότητες, παρουσιάζεται αναλυτικότερα η λειτουργία κάθε συστήματος.

3.2 Σύστημα CRESTA

Στόχος του πρώτου συστήματος είναι η αναζήτηση μέσα από το κείμενο, εκείνων των φράσεων που συνοψίζουν το θέμα του και το περιεχόμενό του με τον καλύτερο δυνατό τρόπο και άρα αποτελούν μία ικανοποιητική αναπαράστασή του, καθώς και η σύνδεσή τους με οντότητες του σημασιολογικού ιστού της DBpedia. Πρέπει να σημειωθεί πως το σύστημα προσανατολίζεται σε φράσεις οι οποίες εντοπίζονται αυτούσιες στο κείμενο και δεν κάνει σύνθεση φράσεων (σε αντίθεση με τα περισσότερα συστήματα εξαγωγής φράσεων κλειδιών (keyphrase generators)). Η ανάπτυξή του, προσεγγίζει την επίλυση του προβλήματος με βάση κάποιες παρατηρήσεις σχετικά με τα χαρακτηριστικά που θα έχουν οι φράσεις αυτές. Αρχικά, πρόκειται για φράσεις που δεν θα περιέχουν ρήματα και ρηματικούς προσδιορισμούς, καθώς δεν είναι ιδιαίτερα πλούσιες σε πληροφορία οι ενέργειες που γίνονται από αντικείμενα του κόσμου ή σε αντικείμενα του κόσμου αλλά τα ίδια τα αντικείμενα του κόσμου. Αυτό συμβαίνει γιατί η ίδια η ύπαρξη ενός αντικειμένου ή μίας έννοιας γενικότερα, υπονοεί τις ενέργειες τις οποίες η έννοια αυτή μπορεί να εκτελέσει ή στις οποίες μπορεί να συμμετάσχει παθητικά και επομένως θα θεωρείται σημασιολογικά ισχυρότερη. Η επιλογή αυτή, συμβαδίζει και με την απαίτηση ταύτισης κάθε εξαγόμενου όρου με οντότητα της DBpedia, καθώς οι οντότητες αυτές (εκτός από συγκεκριμένες εξαιρέσεις όπως ορισμένοι τίτλοι έργων τέχνης) δεν διαθέτουν ρηματικούς τύπους κατά τη λεκτική αναπαράστασή τους. Τέλος, οι φράσεις θα πρέπει να είναι ”σπάνιες φράσεις” (φράσεις με όσο γίνεται πιο ειδικό νόημα) και ταυτόχρονα να αναφέρονται όσο πιο συχνά γίνεται στο υπό εξέταση κείμενο. Επίσης, θα πρέπει να έχουν συνοχή με το υπόλοιπο κείμενο, δηλαδή σημασιολογική συγγένεια με τις υπόλοιπες φράσεις, ώστε να έχουν και μεγαλύτερη πιθανότητα να προσεγγίζουν τη θεματολογία του.

Αξίζει να σημειωθεί πως η χρήση της DBpedia δίνει κάποια επιπλέον πλεονεκτήματα πέραν αυτών που αναφέρθηκαν στην εισαγωγή του κεφαλαίου. Το πρώτο και βασικότερο, είναι πως αντιστοιχίζοντας τους όρους του κειμένου σε οντότητες της DBpedia, δόθηκε η δυνατότητα χρήσης του σημασιολογικού πλαισίου στο οποίο εντάχθηκε ο κάθε όρος ώστε να εξακριβωθεί η θεματική συνοχή του με τους υπόλοιπους και κατ’επέκταση με το ίδιο το κείμενο. Επιπλέον κατά το δεύτερο τμήμα επεξεργασίας του παρόντος συστήματος, δεδομένου του ότι οι οντότητες της DBpedia έχουν περιγραφές στο σύνολο δεδομένων της Wikipedia, δόθηκε η δυνατότητα να χρησιμοποιηθεί η Wikipedia για εξαγωγή στατιστικών συμπερασμάτων,ιδιαίτερα χρήσιμων κατά για τη διάκριση των φράσεων με μεγάλη σημασιολογική βαρύτητα σε σχέση με τις υπόλοιπες. Όταν οι φράσεις αυτές, που αποτελούν οντότητες-στόχους επιλεχθούν, τότε επιδιώκεται η σύνδεσή τους με εξωτερικές πηγές πληροφορίας, όπως αυτές παρέχονται από την βάση της DBpedia.

Παρακάτω θα γίνει ανάλυση του κάθε σταδίου-φίλτρου χωριστά.

3.2.1 Γραμματική Επισημείωση

Το αρχικό κείμενο που παίρνουμε στην είσοδο είναι ένα απλό αρχείο κειμένου, (σε μορφή txt), το οποίο δε φέρει καμία επισημείωση ή μεταδεδομένα προς αξιοποίηση. Πρόκειται δηλαδή, για ένα διάνυσμα συμβολοσειρών. Προκειμένου να γίνει μια αρχική διάκριση των υποψήφιων φράσεων-στόχων είναι απαραίτητο να γίνει διάκριση λέξεων και προτάσεων και ταυτόχρονα γραμματική αναγνώριση και ετικετοποίηση κάθε λέξης του κειμένου. Είναι απαραίτητη δηλαδή η γραμματική επισημείωση, η οποία ακολουθεί τους κανόνες που περιγράφηκαν στην ενότητα 2.1.1. Για το σκοπό αυτό χρησιμοποιήθηκε ο Tree-tagger[44]. Το εργαλείο αυτό, δέχεται το μη προσημειωμένο κείμενο στην είσοδο του ενώ η έξοδός του είναι ένας πίνακας τριών στηλών, όπου κάθε γραμμή αντιστοιχεί σε μία λέξη ή σημείο στίξης του κειμένου. Η σειρά των λέξεων διατηρείται όπως στο κείμενο. Η πληροφορία που παρέχεται από την έξοδο του tagger είναι η μορφή κάθε λέξης και η ετικέτα της με σημειωμένο το μέρος του λόγου στο οποίο ανήκει. Σημειώνεται πως η stemmed μορφή που παρέχεται από τον tagger δεν είναι η πλήρως κανονικοποιημένη μορφή της λέξης (δηλαδή η ρίζα της). Είναι απλά η ίδια λέξη στον ενικό και σε πρώτο πρόσωπο.

3.2.2 Εξαγωγή Ονοματικών Οντοτήτων

Ο πίνακας-έξοδος του προηγούμενου σταδίου αντιστοιχεί ουσιαστικά σε μία πρώτη επισημείωση του κειμένου και χρησιμοποιείται για την περαιτέρω ανάλυση και εντοπισμό των πιθανών φράσεων-στόχων. Η διαδικασία αυτή είναι γνωστή ως εξαγωγή ονοματικών φράσεων. Σημειώνεται πως οι οντότητες της DBpedia, στη συντριπτική τους πλειοψηφία, δεν περιέχουν ρηματικές φράσεις και αποτελούν μία μορφή ονοματικών φράσεων, σημαντικά πιο απλοποιημένη σε σχέση με τον τυπικό ορισμό της ονοματικής φράσης που δόθηκε στην ενότητα 2.1.2. Συγκεκριμένα πρόκειται για ονοματικές φράσεις οι οποίες αποτελούνται από σχετικά μικρό αριθμό λέξεων, δεν εμπεριέχουν δευτερεύουσες προτάσεις και διαθέτουν στην πλειοψηφία τους περιορισμένο ή και μηδενικό αριθμό επιθετικών ή άλλων προσδιορισμών. Καθώς λοιπόν οι φράσεις-στόχοι θα πρέπει να πληρούν αυτά τα χαρακτηριστικά, κρίθηκε αναποτελεσματική η χρήση έτοιμων εργαλείων για χαμηλού επιπέδου συντακτική ανάλυση ή εξαγωγή ονοματικών φράσεων καθώς η εξαγωγή ονοματικών φράσεων που παράγουν ακολουθεί τον τυπικό γραμματικό ορισμό. Συνεπώς, ήταν αναγκαία η διατύπωση και χρήση συγκεκριμένων κανόνων κατάτμησης των προτάσεων και κριτηρίων απόφασης σχετικά με το χαρακτηρισμό λέξεων ή ομάδων λέξεων.

Η πλήρης υλοποίηση της κατάτμησης του κειμένου που γίνεται στο πρώτο στάδιο μπορεί να χωριστεί σε περισσότερα υπο-στάδια. Στο πρώτο γίνεται ένας αρχικός διαχωρισμός και εντοπισμός ονοματικών φράσεων με σχετική ελαστικότητα ως προς τα κριτήρια που χρησιμοποιούνται ενώ στα υπόλοιπα, γίνεται περαιτέρω επεξεργασία και κατάτμηση των φράσεων αυτών ώστε να ταιριάζουν στο πρότυπο των φράσεων της DBpedia. Οι κανόνες και τα κριτήρια που αφορούν την αρχική κατάτμηση του κειμένου, χρησιμοποιούν ως δεδομένα μόνο την γραμματική ετικέτα κάθε λέξης και την αλληλουχία των λέξεων (άρα και των ετικετών) στο κείμενο.

Οι κανόνες επεξεργασίας των ήδη διαχωρισμένων ονοματικών φράσεων χρησιμοποιούν πληροφορίες από την DBpedia, προκειμένου να αξιολογηθούν και να αποφασιστεί το κατά πόσο θα υποστούν περαιτέρω επεξεργασία.

Ακολουθούν οι κανόνες που αφορούν στην αρχική κατάτμηση και στη συνέχεια οι λεπτομέρειες υλοποίησης τους.

1. Η επεξεργασία των λέξεων γίνεται με τη σειρά εμφάνισης τους στο κείμενο (η οποία συμπίπτει με τη σειρά εμφάνισης στον πίνακα)

2. Κάθε λέξη ανήκει το πολύ σε μία ονοματική φράση. Συνεπώς οι λέξεις μπορούν να διαχωριστούν με κριτήριο το αν ανήκουν ή όχι σε μία ονοματική φράση. Προφανώς οι λέξεις που ανήκουν στην ίδια ονοματική φράση βρίσκονται σε διαδοχικές θέσεις στο κείμενο.

3. Με την αρχή της εξέτασης κάθε λεκτικής μονάδας, αυξάνεται ένας μετρητής λέξεων του κειμένου, εκτός αν πρόκειται για σημείο στίξης.

4. Κατά την επεξεργασία του κειμένου, υπάρχουν δύο διακριτές καταστάσεις, οι οποίες κατά σύμβαση θα αποκαλούνται ”εντός ονοματικής φράσης” ή ”Εντός ΟΦ” και ”εκτός ονοματικής φράσης” ή ”Εκτός ΟΦ”. ”Εντός ονοματικής φράσης” θεωρούμε ότι βρίσκεται το σύστημα όταν η λέξη που ελέγχθηκε τελευταία δεν άνηκε σε ονοματική φράση, ενώ ”εκτός ονοματικής φράσης”, όταν η λέξη που ελέγχθηκε τελευταία, άνηκε σε ονοματική φράση.

5. Στην αρχή της επεξεργασίας (πριν την επεξεργασία της πρώτης λέξης του κειμένου) το σύστημα βρίσκεται σε κατάσταση ”εκτός ονοματικής φράσης”.

6. Οταν αποφασιστεί οτι κάποια λέξη δεν ανήκει σε ονομαστική φράση απορρίπτεται και δεν ελέγχεται εκ νέου από το σύστημα.

Οι μεταβάσεις τους συστήματος, σηματοδοτούν την εκκίνηση και την ολοκλήρωση της κατασκευής κάθε ονοματικής φράσης με τον εξής τρόπο: Μετάβαση από την κατάσταση ”Εκτός ΟΦ” στην κατάσταση ”Εντός ΟΦ” δηλώνει την εκκίνηση της κατασκευής της ονοματικής φράσης. Μετάβαση από την κατάσταση ”Εντός ΟΦ” στην κατάσταση ”Εκτός ΟΦ” δηλώνει την ολοκλήρωση της κατασκευής της ονοματικής φράσης.

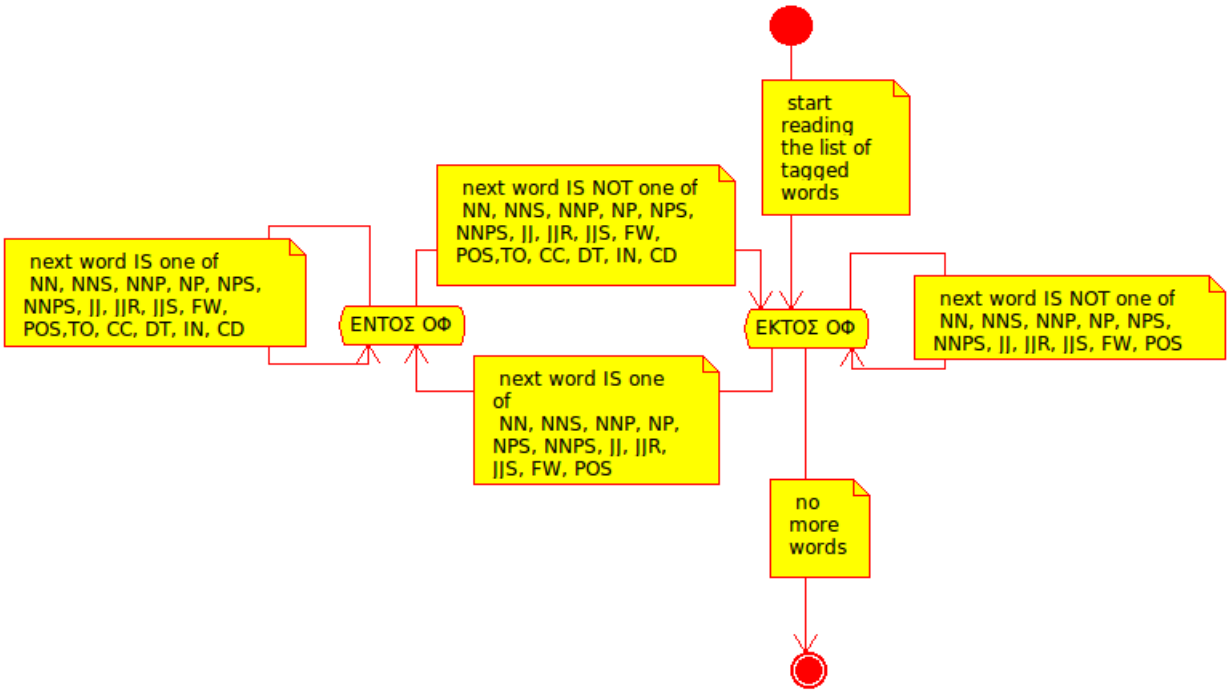
Οι υπόλοιποι κανόνες εξαρτώνται από την κατάσταση του συστήματος πριν την επεξεργασία της επόμενης λέξης:

- Το σύστημα βρίσκεται σε κατάσταση εκτός ονοματικής φράσης (εκτος ΟΦ)
  -Αν η υπο εξέταση λέξη έχει μία από τις επόμενες ετικέτες: ”**NN**”, ”**NNS**”, ”**NNP**”, ”**NP**”, ”**NPS**”, ”**NNPS**”, ”**JJ**”, ”**JJR**”, ”**JJS**”, ”**FW**”, ”**POS**”, δηλαδή είναι είτε ουσιαστικό, είτε επιθετικός προσδιορισμός, είτε ξένη λέξη, τότε η κατάσταση του συστήματος αλλάζει και γίνεται ”εντός ονοματικής φράσης”, και ξεκινάει η κατασκευή μίας νέας ονοματικής φράσης.

-Αν η υπό εξέταση λέξη δεν έχει μία από τις παραπάνω ετικέτες, τότε η κατάσταση του συστήματος παραμένει ”εκτός ονοματικής φράσης”. Η υπό επεξεργασία λέξη απορρίπτεται, και το συστημα προχωράει στην επεξεργασία της επόμενης λέξης.

- Το σύστημα βρίσκεται σε κατάσταση εντός ονοματικής φράσης (εντός ΟΦ)
  -Αν η υπό εξέταση λέξη έχει μία από τις επόμενες ετικέτες: ”**NN**”, ”**NNS**”, ”**NNP**”, ”**NP**”, ”**NPS**”, ”**NNPS**”, ”**JJ**”, ”**JJR**”, ”**JJS**”, ”**FW**”, ”**POS**”,”**TO**”, ”**CC**”, ”**DT**”, ”**IN**”, ”**CD**”, δηλαδή είναι ένα από τα παρακάτω: ουσιαστικό, επιθετικός προσδιορισμός, ξένη λέξη, αριθμός, άρθρο, πρόθεση ή σύνδεσμος, τότε η κατάσταση του συστήματος παραμένει ”εντός ονοματικής φράσης”, η υπό εξέταση λέξη ανήκει στην ονοματική φράση που είναι ”υπό κατασκευή”.
  -Αν η υπό εξέταση λέξη δεν έχει μία από τις παραπάνω ετικέτες, τότε η κατάσταση του συστήματος αλλάζει και γίνεται ”εκτός ονοματικής φράσης”. Η υπό εξέταση λέξη απορρίπτεται, και το σύστημα προχωράει στην επεξεργασία της επόμενης λέξης.

Η παραπάνω διαδικασία συνοψίζεται στο διάγραμμα καταστάσεων του σχήματος 3.3



Σχήμα 3.3: Διάγραμμα καταστάσεων κατά τον σχηματισμό Ονοματικών Φράσεων

Στην υλοποίηση του συστήματος, κάθε νέα ονοματική φράση αποτελεί ένα αντικείμενο με συγκεκριμένα χαρακτηριστικά. Ειδικότερα, τα χαρακτηριστικά κάθε φράσης χωρίζονται σε δυο κατηγορίες. Αφ’ ενός τα χαρακτηριστικά τα οποία είναι γνωστά άμεσα με τα την κατασκευή του αντικειμένου, όπως ο αριθμός λέξεων, η συμβολοσειρά, η θέση στο κείμενο κλπ, και αφ’ ετέρου τα χαρακτηριστικά τα οποία αφορούν τη σημασιολογική βαρύτητα της φράσης και τη συνοχή της σε σχέση με τις υπόλοιπες φράσεις, τα οποία υπολογίζονται σε δεύτερο επίπεδο επεξεργασίας.

Η ολοκλήρωση του πρώτου ελέγχου του κειμένου, η οποία περιγράφεται παραπάνω, καταλήγει στην εξαγωγή μίας λίστας από αντικείμενα που αντιπροσωπεύουν τις πιθανές φράσεις-στόχους και συνεπώς ολόκληρο το κείμενο. Στο εξής δε χρειάζεται να ανατρέξει κανείς εκ νέου στο κείμενο καθώς κάθε νέα επεξεργασία μπορεί να γίνει στη λίστα αυτή των φράσεων. Καθώς οι φράσεις αυτές έχουν προκύψει με αυτοματοποιημένο τρόπο, απαιτείται μετά το στάδιο αυτό, αλλά και μετά από τα περισσότερα από τα επόμενα στάδια επεξεργασίας, ένα στάδιο βελτιστοποίησης των φράσεων, όπου αφαιρούνται τυχόν περιττές λέξεις (άρθρα, σύνδεσμοι, προθέσεις) από το τέλος κάθε φράσης, αλλά και μονολεκτικές φράσεις που δεν περιέχουν ουσιαστικό. Στη συνέχεια, οι φράσεις ταξινομούνται αλφαβητικά και οι εκείνες που εντοπίζονται περισσότερες από μία φορές ενοποιούνται σε ένα αντικείμενο με την κατάλληλη τιμή αριθμού εμφανίσεων στο κείμενο.

Το επόμενο βήμα αφορά στον έλεγχο των υποψήφιων φράσεων ώστε να εντοπισθούν εκείνες οι οποίες αντιστοιχούν σε κάποια καταχωρημένη οντότητα στη βάση της DBpedia. Συγκεκριμένα, ελέγχονται διαδοχικά όλες οι συμβολοσειρές για να εντοπιστούν αυτές που διαθέτουν ομώνυμο αναγνωριστικό URI (resource description URI) . Δηλαδή, γίνεται αναζήτηση για τον εξής τύπο τριάδας:

<http://dbpedia.org/resource/symboloseira> <http://www.w3.org/2000/01/rdf-schema#label> ”sumboloseira”@en.

Οι φράσεις που εντοπίστηκαν στη DBpedia κρατούνται ώστε να αξιοποιηθούν στο δεύτερο τμήμα του προγράμματος. Οι υπόλοιπες περνάνε από διαδοχικά στάδια περαιτέρω επεξεργασίας με σκοπό να ελεγχθεί η πιθανότητα να εντοπιστούν στη βάση σε πιο απλοποιημένη μορφή.

Η διαδικασία βελτιστοποίησης των φράσεων, αλφαβητικής ταξινόμησης, ενοποίησης πολλαπλών εμφανίσεων και αναζήτησης στη DBpedia που περιγράφηκε παραπάνω παρεμβάλλονται μετά από κάθε νέο κύκλο επεξεργασίας των μη εντοπισμένων φράσεων, ώστε να ενημερώνεται η λίστα με τις φράσεις που έχουν εντοπιστεί στη DBpedia και άρα αποτελούν υποψήφιες φράσεις-στόχους. Το σύνολο των διαδικασιών αυτών θα αναφέρεται στο εξής ως διαδικασία ενημέρωσης της λίστας.

Το πρώτο στάδιο μετά-επεξεργασίας αφορά στην κατάτμηση των φράσεων που δεν εντοπίστηκαν, σε μικρότερες επιμέρους φράσεις. Για να επιτευχθεί ορθολογική και βέλτιστη κατάτμηση, η διαδικασία ”σπάει” τις φράσεις στα σημεία που εντοπίζει συνδέσμους (CC), προθέσεις (IN), χωροχρονικά επιρρήματα (WRB), απαρέμφατο (TO) και κτητικές αντωνυμίες (POS), καθώς οι λέξεις αυτές συχνά υποδηλώνουν έναρξη νέας ονοματικής φράσης. Ακολουθεί η διαδικασία ενημέρωσης της λίστας.

Οι μη-εντοπισμένες φράσεις που έχουν απομείνει αποτελούνται πλέον κυρίως από επιθετικούς προσδιορισμούς και ουσιαστικά. Σύμφωνα με την αγγλική γραμματική οι επιθετικοί προσδιορισμοί συνήθως προηγούνται του ουσιαστικού το οποίο προσδιορίζουν. Συνεπώς στο επόμενο στάδιο, αφαιρούνται σταδιακά οι επιθετικοί προσδιορισμοί (JJ, JJR, JJS) αλλά και αριθμητικά (CD) από την αρχή κάθε ονοματικής φράσης. Ακολουθεί η διαδικασία ενημέρωσης της λίστας και εαν υπάρχουν ακόμα μή εντοπισμένες φράσεις που δεν ξεκινάνε με ουσιαστικό,

η διαδικασία επαναλαμβάνεται.

Στο τελευταίο στάδιο επεξεργασίας, σε αντίθεση με τα προηγούμενα, οι εναπομείναντες φράσεις δεν ελέγχονται με βάση κάποιο γραμματικό κανόνα. Συγκεκριμένα, για κάθε φράση χρησιμοποιείται ένα μεταβλητού μήκους ”παράθυρο”, το οποίο επιλέγει και ελέγχει διαδοχικά όλες τις πιθανές αλληλουχίες λέξεων στη φράση ώστε να διαπιστωθεί αν κάποια από αυτές θα μπορούσε να αποτελεί υποψήφια φράση-στόχο. Ο έλεγχος αυτός, βασίζεται στην παραδοχή οτι άν κάποια αλληλουχία λέξεων εντός της υπό-έλεγχο φράσης (η οποία δεν έχει εντοπιστεί στη βάση της DBpedia) εντοπιστεί στη βάση της DBpedia και υπάρχει αναγνωριστικό URI που της αντιστοιχεί, τότε αποτελεί μία υποψήφια φράση στόχο. Η συγκεκριμένη διαδικασία ελέγχου είναι εξαντλητική, δηλαδή ξεκινάει με ένα παράθυρο ελέγχου μήκους (n-1) λέξεων, (όπου n ο αριθμός των λέξεων της υπό εξέτασης φράσης) και αφού διατρέξει όλη τη φράση, μειώνει το μήκος του παραθύρου κατά ένα και τη διατρέχει από την αρχή. Σε κάθε βήμα ελέγχεται η αλληλουχία λέξεων που βρίσκεται εντός του παραθύρου. Σε περίπτωση που η υπό-έλεγχο αλληλουχία εντοπιστεί στη DBpedia, θεωρείται νέα υποψήφια φράση-στόχος, αφαιρείται από τη φράση που την περιείχε και τοποθετείται στη λίστα με τις σωστές. Οι αλληλουχίες λέξεων πριν και μετά τη φράση που εντοπίστηκε, θεωρούνται νέες πιθανές φράσεις και ελέγχονται ξανά με την ίδια διαδικασία, με σκοπό τον εντοπισμό περαιτέρω υποψήφιων φράσεων-στόχων.

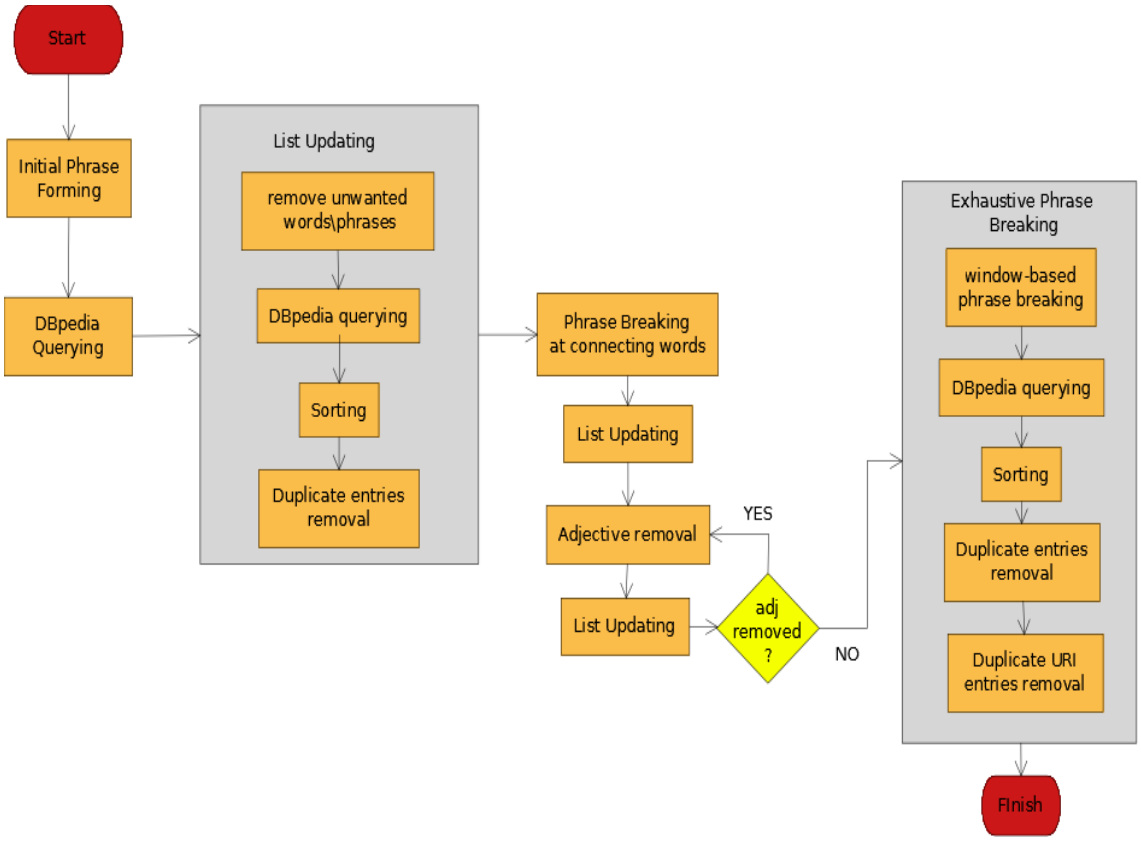
Η ολοκλήρωση της παραπάνω διαδικασίας συνεπάγεται και την ολοκλήρωση του πρώτου τμήματος επεξεργασίας. Πριν την έξοδο υλοποιείται για μία τελευταία φορά η διαδικασία ενημέρωσης της λίστας, καθώς η λίστα με τις υποψήφιες φράσεις στόχους πρέπει να ελεγχθεί για την αφαίρεση και ενοποίηση των διπλών εμφανίσεων της ίδιας φράσης. Στη συνέχεια η λίστα διοχετεύεται στο δεύτερο τμήμα επεξεργασίας.

Μετά το πέρας της διαδικασίας που περιγράφηκε, οι όροι που έχουν εξαχθεί δεν είναι απλές ονοματικές φράσεις αλλά κάθε μία από αυτές αντιστοιχεί σε μία οντότητα της DBpedia. Για το λόγο αυτό, η διαδικασία κρίθηκε ορθότερο να ονομαστεί εξαγωγή ονοματικών οντοτήτων (noun entity extraction) και όχι εξαγωγή ονοματικών φράσεων παρ’ότι η δεύτερη διαδικασία αποτελεί τμήμα της πρώτης. Οι υποεργασίες του σταδίου αυτού, όπως περιγράφηκαν παραπάνω, φαίνονται αναλυτικά στο διάγραμμα ροής που ακολουθεί (σχήμα 3.4).

3.2.3 Ποσοτικοποίηση πληροφορίας από κείμενο και από Wikipedia

Στο στάδιο αυτό, η λίστα με τις υποψήφιες οντότητες εμπλουτίζεται με ποσοτικοποιημένη πληροφορία που αντιπροσωπεύει την ”δυνατότητα” της κάθε οντότητας να βρίσκεται ανάμεσα στις επικρατέστερες για την ικανοποίηση του στόχου του συστήματος. Στο πλαίσιο αυτό υπολογίζονται τρία αριθμητικά χαρακτηριστικά τα οποία συνοψίζονται ως εξής:

1. Πρώτη εμφάνιση
   Το χαρακτηριστικό αυτό έχει να κάνει με την θέση του κειμένου στην οποία εμφανίζεται για πρώτη φορά η πρώτη λέξη της φράσης που αντιπροσωπεύει την κάθε οντότητα στο κείμενο. Μάλιστα, δεδομένου ότι στην συγκεκριμένη περίπτωση, όσο πιο μικρή είναι η



Σχήμα 3.4: Διάγραμμα Ροής εξαγωγής ονοματικών οντοτήτων

θέση της πρώτης εμφάνισης της πρώτης λέξης, αυξάνεται η πιθανότητα η οντότητα να είναι σημαντική για το κείμενο, η τιμή που υπολογίζεται ως θέση αντιστρέφεται, ώστε να επιδιώκεται η μεγιστοποίηση του μέτρου αυτού και όχι η ελαχιστοποίησή του.

2. KeyPhraseness

   Το χαρακτηριστικό αυτό έχει να κάνει με το πόσο ειδικού περιεχομένου είναι μία λέξη. Για τον υπολογισμό της τιμής του, χρησιμοποιούνται τα κείμενα της Wikipedia και η πληροφορία που μπορεί να αντληθεί από αυτά μέσω του mediawiki API[24]. Η πληροφορία αυτή αναλύεται σε δύο στοιχεία της υπό εξέταση οντότητας. Το πρώτο είναι ο αριθμός των links που δείχνουν προς το άρθρο της Wikipedia που αποτελεί περιγραφή της οντότητας αυτής και θα αναφέρεται στο εξής με τον όρο backlinks. Το δεύτερο είναι όλες οι εμφανίσεις της φράσης που αντιπροσωπεύει την οντότητα σε όλα συνολικά τα άρθρα της wikipedia. Η τελική τιμή προκύπτει από τον εξής τύπο:

$$keyphraseness = \frac{backlinks}{totallinks}$$

   Και για το χαρακτηριστικό αυτό επιδιώκεται η μεγιστοποίησή του, καθώς όσο μεγαλύτερος είναι ο παρονομαστής τόσο πιο συχνή είναι η φράση και άρα έχει μικρότερη πιθανό-

τητα να είναι σημαντική, ενώ όσο μεγαλύτερος είναι ο αριθμητής, η φράση χρησιμοποιείται περισσότερες φορές με την ιδιότητα του συνδέσμου και επομένως είναι σημαντική για περισσότερα άρθρα της wikipedia. Άρα αυξάνεται και η πιθανότητά της να είναι σημαντική και για το κείμενο που εξετάζεται. Ένα χαρακτηριστικό παράδειγμα είναι η λέξη "and". Παρότι εμφανίζεται πάρα πολλές φορές στα άρθρα της wikipedia (άρα θα έχει μεγάλο παρονομαστή), ελάχιστες από αυτές αποτελεί link προς το άρθρο "And" (άρα έχει μικρό αριθμητή). Τελικά το keyphraseness θα είναι αρκετά χαμηλό, γεγονός που αποτελεί σημαντική πληροφορία για την πιθανότητα η φράση αυτή να αποτελεί σημαντικό όρο του κειμένου.
Στο σημείο αυτό θα πρέπει να σημειωθεί πως το χαρακτηριστικό αυτό επιχειρήθηκε να εξαχθεί εν μέρη από τη DBpedia, καθώς στο σύνολο RDF δεδομένων "Wikipedia Pagelinks" που παρουσιάστηκε στην ενότητα 2.4.7, υπάρχει καταχωρημένη η πληροφορία των εσωτερικών συνδέσμων της Wikipedia και μάλιστα είναι εύκολο να ανακτηθεί με μία SPARQL αναζήτηση ο αριθμός των backLinks σε σημαντικά συντομότερο χρονικό διάστημα. Ωστόσο, λόγω προβήματος στα δεδομένα της DBpedia το οποίο δεν επιλύθηκε από τους υπεύθυνους συντήρησης του συστήματος στο χρονικό διάστημα εκπόνησης της συγκεκριμένης διπλωματικής, δεν ήταν δυνατό να μελετηθεί η προσέγγιση αυτή.

3. Term Frequency
   Το χαρακτηριστικό αυτό έχει να κάνει με την συχνότητα εμφάνισης κάθε φράσης στο κείμενο, και μάλιστα σε συνδυασμό με το μέγεθός της. Προφανώς, μία φράση που εμφανίζεται πολλές φορές στο κείμενο, θα έχει μεγαλύτερη πιθανότητα να είναι και σημαντικός όρος του. Ταυτόχρονα, μία φράση που αποτελείται από μεγάλο αριθμό λέξεων, φαίνεται πως είναι πιο σημαντική σαν όρος του κειμένου, σε σχέση με μία φράση που έχει ίδιο αριθμό εμφανίσεων αλλά αποτελείται από μικρότερο αριθμό λέξεων. Οι παρατηρήσεις αυτές ποσοτικοποιούνται ως εξής:

$$termfrequency = frequency * numberofwords$$

   Και για αυτό το χαρακτηριστικό επιδιώκεται η μεγιστοποίησή του.

Τα τρία αυτά χαρακτηριστικά κανονικοποιούνται και στη συνέχεια υπολογίζεται ο γραμμικός τους συνδυασμός με βάρη που υπολογίστηκαν πειραματικά. Οι διαδικασίες κανονικοποίησης και υπολογισμού βαρών θα αναλυθούν κατά την περιγραφή της υλοποίησης στην ενότητα 4.4. Με βάση το γραμμικό συνδυασμό των τριών αυτών μεγεθών, διατηρείται μόνο ένας αριθμός από τις αρχικές οντότητες που δόθηκαν ως είσοδος στο στάδιο αυτό και αυτές μόνο διοχετεύονται στο επόμενο στάδιο ως υποψήφιες οντότητες.

3.2.4 Εξαγωγή και Ποσοτικοποίηση πληροφορίας από DBpedia

Κατά το στάδιο αυτό, υπολογίζονται δύο ακόμα χαρακτηριστικά τα οποία θα ενισχύσουν τη διαδικασία αξιολόγησης των υποψήφιων οντοτήτων. Τα χαρακτηριστικά αυτά αποσκοπούν στο

να ενισχύσουν τις φράσεις που φαίνονται να παρουσιάζουν σημασιολογική συγγένεια μεταξύ τους, θεωρώντας πως σε ένα κείμενο με συγκεκριμένη θεματολογία η πλειοψηφία των σημαντικών φράσεων αναμένεται να συνδέονται σημασιολογικά. Είναι σημαντικό το γεγονός ότι το βήμα αυτό έπεται του υπολογισμού των υπολοίπων χαρακτηριστικών καθώς και της πρώτης επιλογής των σημαντικότερων φράσεων, καθώς έτσι αποφεύγονται σφάλματα και αποκλίσεις που οφείλονται στο θόρυβο που προκαλείται από περιττές φράσεις που δε σχετίζονται με το θέμα του κειμένου αλλά μπορεί να σχηματίζουν εννοιολογικές ομάδες μεταξύ τους και να επηρεάζουν αρνητικά τα αποτελέσματα.

Για τον υπολογισμό των τιμών των χαρακτηριστικών αυτών απαιτείται η άντληση πληροφορίας από τη βάση της DBpedia ως εξής:

1. **Αξιοποίηση Κατηγοριών (categories) της DBpedia**
   Στη συγκεκριμένη περίπτωση, αξιοποιούνται τα URIs και τα κατηγορήματα που σχετίζονται με την κατηγοριοποίηση των εννοιών-οντοτήτων (DBpedia categories). Ουσιαστικά αναζητούνται οι κοινές κατηγορίες (και ευρύτερες κατηγορίες) με τις οποίες σχετίζονται οι επιλεγμένες οντότητες και αποδίδεται ένα βάρος στην κάθε οντότητα ανάλογα με τον αριθμό των κατηγοριών αυτών στις οποίες ανήκει. Το βάρος αυτό αποτελεί ένα βαθμό συσχέτισης με το θέμα (ή θέματα) του κειμένου. Όπως έχει αναλυθεί και στην ενότητα (2.4.7) μία οντότητα συνδέεται με τις κατηγορίες στις οποίες ανήκει με τριάδες του τύπου:
   entity <http://purl.org/dc/terms/subject> category
   Όπου entity είναι μία από της υποψήφιες οντότητες του κειμένου, ενώ category ειναι μία από τις κατηγορίες στις οποίες ανήκει. Αφού διατρέχεται μία φορά η λίστα με τις υπό εξέταση φράσεις του κειμένου, δημιουργείται μία λίστα με τις κατηγορίες που εντοπίζονται. Κάθε κατηγορία αποτελεί ένα αντικείμενο το οποίο περιέχει πληροφορία σχετικά με το ποιες φράσεις-οντότητες ανήκουν σε αυτό. Στη συνέχεια η λίστα με τις κατηγορίες που έχει δημιουργηθεί υφίσταται επιπλέον επεξεργασία. Στη δεύτερη επεξεργασία αξιοποιούνται τα κατηγορήματα *<http://www.w3.org/2004/02/skos/core#broader>* και *<http://www.w3.org/2004/02/skos/core#related>*, τα οποία συνδέουν τις υπάρχουσες κατηγορίες μεταξύ τους. Αναζητούνται τριάδες της μορφής:
   category <http://www.w3.org/2004/02/skos/core#broader> category2
   category <http://www.w3.org/2004/02/skos/core#related> category3
   όπου category η κατηγορία που εντοπίζεται στη λίστα και category2 η κατηγορία που συνδέεται με την αρχική με σχέση γενίκευσης (η 2 ειναι γενικότερη κατηγορία της 1). Η category3 συνδέεται με την αρχική με απλή σημασιολογική σχέση. Οι νέες κατηγορίες που εντοπίζονται προστίθενται και αυτές στη λίστα ως νέα αντικείμενα ενώ κληρονομούν τις οντότητες με τις οποίες σχετίζονται οι κατηγορίες που συνδέονται με αυτές. Πρέπει να σημειωθεί πως αυτή η διαδικασία εμπλουτισμού της λίστας των κατηγοριών θα μπορούσε να διεξάγεται επαναληπτικά σε βρόχο καθώς οι νέες κατηγορίες που προστίθενται ενδεχομένως συνδέονται με άλλες κτλ. Ωστόσο, επιλέχθηκε η επανάληψη να μην έχει βάθος μεγαλύτερο από 1 καθώς όπως έχει ήδη αναφερθεί το κομμάτι της οντολογίας που αφορά στις κατηγορίες της DBpedia δεν είναι σωστά ιεραρχημένο με αποτέλεσμα να

εντοπίζονται κύκλοι του τύπου
category1 <http://www.w3.org/2004/02/skos/core#broader> category2
category2 <http://www.w3.org/2004/02/skos/core#broader> category1
γεγονός το οποίο αλλοιώνει την ορθότητα των αποτελεσμάτων. Επιπλέον, κρίθηκε πως με μία επανάληψη δημιουργείται ένα ικανοποιητικό σε μέγεθος σύνολο κατηγοριών οι οποίες μένουν κοντά στο θέμα του κειμένου. Το γεγονός αυτό σταδιακά παύει να ισχύει όσο προστίθενται στη λίστα γενικότερες κατηγορίες, οι οποίες συνδέονται με τις αρχικές οντότητες με σχέσεις βάθους μεγαλύτερου του 3.

Στη συνέχεια, επιλέγονται τα categories τα οποία σχετίζονται με περισσότερες από μία οντότητες. Ακολουθεί η αντίστροφη διαδικασία, κατά την οποία για κάθε υποψήφια οντότητα, υπολογίζεται με πόσα από τα προαναφερθέντα υποψήφια categories (μαζί με τα related και broader) σχετίζεται. Η τιμή αυτή αποτελεί άλλο ένα χαρακτηριστικό προς μεγιστοποίηση καθώς, με όσα περισσότερα από τα κοινά categories σχετίζεται μία οντότητα, τόσο μεγαλύτερη συνοχή έχει με τις υπόλοιπες οντότητες του κειμένου, άρα έχει και μεγαλύτερη πιθανότητα να έχει σημασιολογική συγγένεια με το θέμα.

2. **Αξιοποίηση περιλήψεων (abstracts) της DBpedia**
Αυτή τη φορά, η πληροφορία που αντλείται από τη βάση της DBpedia είναι σε μορφή κειμένου. Πρόκειται για το κείμενο που αποτελεί την περίληψη του άρθρου της Wikipedia που αντιστοιχεί σε κάθε μία από τις υποψήφιες οντότητες. Μάλιστα, αυτή η πληροφορία αναζητάται τόσο για τις υποψήφιες οντότητες που έχουν επιλεχθεί από το προηγούμενο στάδιο όσο και για τις υποψήφιες οντότητες που έχουν αποκλειστεί από το προηγούμενο στάδιο. Το τμήμα του κειμένου που αναζητείται περιέχεται σε RDF τριάδες της μορφής:
entity <http://dbpedia.org/ontology/abstract> text
όπου entity είναι μία από τις οντότητες που έχουν αντληθεί από το αρχικό κείμενο, και Κείμενο είναι το abstract της. Στη συνέχεια, καθένα από αυτά τα κείμενα, υπόκειται στη διαδικασία εξαγωγής ονοματικών οντοτήτων που έχει περιγραφεί παραπάνω. Τελικά, λοιπόν, προκύπτει για κάθε οντότητα του κειμένου μία λίστα από οντότητες οι οποίες αναφέρονται στην περίληψή της που θα ονομάζεται από εδώ και στο εξής ”abstract list” Η παραπάνω πληροφορία που αντλήθηκε από τη βάση της DBpedia χρησιμοποιείται ως εξής:

Για κάθε μία από τις επιλεγμένες υποψήφιες οντότητες που δίνονται ως έξοδος από το τρίτο στάδιο, υπολογίζεται πόσα στοιχεία της abstract list της είναι κοινά με τα στοιχεία όλων των abstract lists των υποψηφίων οντοτήτων του κειμένου συνολικά (τόσο των αποκλεισμένων όσο και των επιλεγμένων). Είναι προφανές ότι οι επιλεγμένες υποψήφιες οντότητες που έχουν περισσότερες κοινές abstract οντότητες με το υπόλοιπο κείμενο, αντιπροσωπεύουν καλύτερα το κείμενο θεματικά και σημασιολογικά. Το χαρακτηριστικό μέγεθος αυτό που θα αποκαλείται ”abstract” υπολογίζεται αναλυτικά από τον παρακάτω τύπο:

$$abstract_{\text{i}} = \frac{abstract\_list_{\text{i}} \bigcap \sum abstract\_lists}{\sum abstract\_lists}$$

Όπου 0<i<αριθμός των επιλεγμένων υποψήφιων οντοτήτων

Έχοντας ολοκληρώσει την παραπάνω διαδικασία, κάθε μία από τις επιλεγμένες υποψήφιες οντότητες έχει δύο επιπλέον χαρακτηριστικά, το category feature και το abstract feature. Ο γραμμικός συνδυασμός των χαρακτηριστικών αυτών και του γραμμικού συνδυασμού των χαρακτηριστικών του 3ου σταδίου δίνει έναν βαθμό σε κάθε επιλεγμένη υποψήφια οντότητα που κατατάσσει τις υποψήφιες οντότητες με βάση την εκτιμώμενη πιθανότητα που έχουν να αποτελούν οντότητες στόχους του συστήματος. Πριν διοχετευτούν οι οντότητες αυτές στο επόμενο στάδιο, γίνεται επιλογή των πιο σημαντικών από αυτών με βάση των βαθμό τους. Το ποσοστό των οντοτήτων που θα επιλεχθούν σε αυτό το σημείο, αποτελεί επιλογή του χρήστη.

3.2.5 Σύνδεση με εξωτερικές πηγές πληροφορίας

Στο στάδιο αυτό, παρέχονται ως είσοδος οι επιλεγμένες οντότητες του προηγούμενου σταδίου και για κάθε μία από αυτές αναζητείται η σύνδεσή τους με εξωτερικές πηγές. Οι πηγές αυτές, παρέχονται από την DBpedia και η αναζήτησή τους γίνεται σε RDF τριάδες της μορφής:
Οντότητα <http://xmlns.com/foaf/0.05/page> Wiki_link
Οντότητα <http://dbpedia.org/ontology/wikiPageExternalLink> External_link
όπου, Οντότητα είναι μία από τις τελικώς επιλεγμένες οντότητες και Wiki_link, External_link, τα URIs που αποτελούν σύνδεσμο της κάθε οντότητας με εξωτερικές πηγές. Πιο συγκεκριμένα, τα Wiki_links είναι σύνδεσμοι προς σελίδες της wikipedia, ενώ τα External_links σύνδεσμοι προς άλλες εξωτερικές πηγές.
Ακόμα, υλοποιούνται και δομές σύνδεσης με γνωσιακές βάσεις δεδομένων εκτός της DBpedia, δηλαδή ανάκτησης για κάθε επιλεγμένη οντότητα του αναγνωριστικού URI που της αντιστοιχεί σε άλλες βάσεις γνώσης. Αυτό μπορεί να επιτευχθεί με αναζήτηση σε RDF τριάδες της μορφής:
Οντότητα <http://www.w3.org/2002/07/owl#sameAs> External_Uri
στα δεδομένα της DBpedia που περιέχουν πληροφορία εξωτερικών πηγών. Βέβαια καθώς η προσπάθεια διασύνδεσης με εξωτερικές πηγές είναι πολύ πρόσφατη, τα διαθέσιμα στους χρήστες δεδομένα είναι σημαντικά περιορισμένα (κάτω από 1000 τριάδες για κάθε γνωσιακή βάση), και συνεπώς δε μπορούν να εντοπιστούν αναγνωριστικά URI για τις περισσότερες οντότητες. Ωστόσο, είναι προφανές πως η διαδικασία αυτή έχει σημαντικές προοπτικές εξέλιξης στο μέλλον.

3.3 Σύστημα SWPID

Στο δεύτερο σύστημα, οι οντότητες που αναζητούνται είναι οντότητες που αντιστοιχούν σε πρόσωπα του πραγματικού κόσμου. Επομένως γίνεται αναζήτηση των φράσεων οι οποίες μπορούν να θεωρηθούν ανθρώπινα ονόματα και μάλιστα ονόματα υπαρκτών προσώπων (είτε ιστορικών, είτε της επικαιρότητας). Ο στόχος αυτός εντάσσεται στο πλαίσιο του γενικότερου προβλήματος της αναγνώρισης πληροφορίας συγκεκριμένου τύπου. Για την επίλυση του, επιλέχθηκε η αξιοποίηση της πληροφορίας της DBpedia καθώς στα δεδομένα της περιέχονται

ρόλοι που προσδιορίζουν τον τύπο κάθε οντότητας και τη συνδέουν με τον τύπο αυτόν. Έτσι, με μία απλή αναζήτηση είναι εύκολο να εξακριβωθεί κατά πόσο ένας από τους επιλεγμένους όρους του πρώτου τμήματος του συστήματος αντιστοιχεί σε μία οντότητα που αναπαριστά πρόσωπο. Επίσης, εφόσον η αναζήτηση έχει θετική έκβαση, έχει, ουσιαστικά, βρεθεί η διασύνδεση του εν λόγω όρου με οντότητα της DBpedia άρα έχει αυτομάτως βρεθεί το σημασιολογικό πλαίσιο του όρου αυτού.

Η παραπάνω επιλογή επίλυσης του προβλήματος με το οποίο ασχολείται το σύστημα αυτό, έχει πολλαπλά πλεονεκτήματα πέραν του βασικότερου της σύνδεσης του κειμένου με το Σημασιολογικό Ιστό. Πρώτον, καθώς δεν χρησιμοποιήθηκαν πρότυπα ονομάτων ή λεξικά όρων κατά την αναζήτηση, δόθηκε η πολύ σημαντική δυνατότητα εντοπισμού όρων που αποτελούν αναγνωριστικά ανθρώπινων προσώπων και δεν είναι εξ ολοκλήρου ονόματα. Έτσι μπορεί να θεωρηθεί ως όρος που αντιστοιχεί σε πρόσωπο ο όρος "Pope Julius II". Δεύτερον, το γεγονός πως δεν χρησιμοποιήθηκε κάποιου είδους συντακτική ανάλυση για την αναγνώριση προσώπων, δεν υπάρχει περίπτωση να αναγνωριστεί ως άνθρωπος η αναφορά σε κάποιο φανταστικό χαρακτήρα (πχ κινουμένων σχεδίων) ή σε κάποιο ζώο στο οποίο αποδίδονται ανθρώπινες ενέργειες. Είναι σαφές πως οι όροι που θα επιλεχθούν θα είναι αναγνωριστικά συγκεκριμένων ανθρώπων και μάλιστα, ανατρέχοντας κάποιος στον σύνδεσμο της DBpedia που παρέχεται στην έξοδο, έχει πρόσβαση σε ένα ικανοποιητικό σύνολο πληροφοριών για το αναπαριστώμενο πρόσωπο.

Τα παραπάνω πλεονεκτήματα, είναι, βέβαια, αξιοποιήσιμα μόνο στο βαθμό που το αναφερόμενο στο κείμενο πρόσωπο περιγράφεται μέσω URI περιγραφής στην DBpedia. Αν δεν υπάρχει η αντίστοιχη καταχώρηση, το σύστημα δεν είναι σε θέση να εξάγει οποιουδήποτε είδους πληροφορία για το αναφερόμενο πρόσωπο. Το γεγονός αυτό οδηγεί στο συμπέρασμα, πως αναμένεται τα αποτελέσματα να περιλαμβάνουν μόνο πρόσωπα της ιστορίας ή της επικαιρότητας, γνωστά στο ευρύ κοινό και εμπίπτει στις απώλειες που θα υπάρξουν εξαιτίας της παραδοχής που έχει γίνει σχετικά με την κατά σύμβαση πλήρη περιγραφή του κόσμου από τον ιστό της DBpedia.

3.3.1 Γραμματική Επισημείωση

Το πρώτο τμήμα επεξεργασίας όπως ήταν αναμενόμενο είναι κοινό με αυτό του πρώτου συστήματος, αφού και για αυτό το σύστημα, ο σχηματισμός των φράσεων που θα αποτελέσουν τις υποψήφιες οντότητες στηρίζεται στις ετικέτες γραμματικής επισημείωσης. Έτσι, το τμήμα αυτό, θα έχει ως είσοδο το υπό-εξέταση κείμενο, και θα προωθεί στην έξοδό του όλες τις λέξεις του κειμένου επισημειωμένες με ετικέτες που αντιστοιχούν στο μέρος του λόγου που ανήκει κάθε λέξη, σύμφωνα με την απόφαση του TreeTagger.

3.3.2 Εξαγωγή Ονοματικών Οντοτήτων

Το δεύτερο τμήμα επεξεργασίας, εκτελεί την ίδια διεργασία με το αντίστοιχο του πρώτου συστήματος. Βέβαια στο συγκεκριμένο σύστημα, οι στόχοι είναι πολύ πιο συγκεκριμένοι και

προσανατολίζονται σε κύρια ονόματα και μάλιστα ονόματα τα οποία αντιστοιχούν σε υπαρκτά πρόσωπα. Για το λόγο αυτό, έχει τροποποιηθεί η τελευταία επεξεργασία της λίστας όπου οι φράσεις ελέγχονται εξαντλητικά με χρήση παραθύρου. Στο στάδιο αυτό έχει εισαχθεί ένας επιπλέον κανόνας, με βάση τον οποίο, όταν ο έλεγχος συναντά αλληλουχία δυο ή και περισσότερων λέξεων που αποτελούν όλες κύρια ονόματα τα οποία δεν εντοπίζει στη βάση της DBpedia, να τα αφαιρεί από την υπό εξέταση φράση ώστε να μην γίνεται έλεγχος για πιθανές μικρότερες φράσεις που υπάρχουν εντός της και εντοπίζονται στη DBpedia. Ο λόγος για τον οποίο κρίθηκε απαραίτητη η προσθήκη του κανόνα αυτού στη συγκεκριμένη περίπτωση είναι πως αν μια φράση που αποτελείται από κύρια ονόματα και αντιστοιχεί για παράδειγμα στο ονοματεπώνυμο ενός υπαρκτού ατόμου, σπάσει σε μικρότερες είναι πολύ πιθανό αν κάποιο τμήμα του ονοματεπώνυμου (πχ μόνο το όνομα ή μόνο το επώνυμο) εντοπιστεί στα δεδομένα της DBpedia να γίνει λάθος αναγνώριση του ατόμου. Για καλύτερη κατανόηση του προβλήματος ακολουθεί συγκεκριμένο παράδειγμα:

Στη φράση ”His ancestry was Ashkenazi Jewish, with his paternal line having supplied the rabbis of Trier since 1723, a role that had been taken up by his own grandfather, Meier Halevi Marx” γινεται αναφορά στον ” Meier Halevi Marx”. Καθώς δεν υπάρχει κάποια καταχώρηση για το συγκεκριμένο πρόσωπο η αναζήτηση ολόκληρου του ονοματεπώνυμου δεν επιστρέφει κάποιο αποτέλεσμα. Αν στη συνέχεια διαχωρίζουμε όνομα και επώνυμο και τα αναζητήσουμε ξεχωριστά παίρνουμε το εξής αποτέλεσμα: ”Marx” που προφανώς αντιστοιχεί σε διαφορετικό πρόσωπο από αυτό της αρχικής φράσης. Συνεπώς ο παραπάνω κανόνας αποσκοπεί στην αποφυγή της αλλοίωσης του τελικού αποτελέσματος.

Ακόμα για τον ίδιο λόγο, αλλά και για να επιταχυνθεί η λειτουργία του συστήματος, στη συγκεκριμένη διαδικασία ελέγχεται κατά πόσον οι φράσεις που αναζητούνται στη DBpedia είναι αλληλουχία κύριων ονομάτων, ώστε να μην κρατούνται στη λίστα ονοματικές φράσεις που δεν είναι πιθανό να αντιστοιχούν σε ονοματεπώνυμα.

Στο τέλος του τμήματος αυτού όπως και στο προηγούμενο σύστημα έχουμε στην έξοδο μία λίστα με υποψήφιες οντότητες-στόχους οι οποίες τροφοδοτούνται στο 2ο τμήμα επεξεργασίας του συστήματος.

3.3.3 Ταυτοποίηση Προσώπων

Οι υποψήφιες φράσεις-στόχοι που δέχεται σαν είσοδο το σύστημα έχουν όλες καταχώριση στη DBpedia. Συγκεκριμένα, αντιστοιχούν σε ένα συγκεκριμένο αναγνωριστικό URI. Μέσω του URI αυτού κάθε φράση-οντότητα συνδέεται με άλλα δεδομένα που την προσδιορίζουν. Προκειμένου λοιπόν να αναγνωρίστουν οι φράσεις οι οποίες περιγράφουν υπαρκτά πρόσωπα, είναι απαραίτητη η γνώση του τρόπου με τον οποίο καταχωρούνται στην RDF περιγραφή της DBpedia, πληροφορίες που σχετίζονται με την ανθρώπινη ιδιότητα μιας οντότητας. Σημειώνεται ότι η παραπάνω διαδικασία ονομάζεται ταυτοποίηση προσώπων (person identification).

Για κάθε καταχωρημένο άτομο υπάρχει ποικιλία τριάδων που σχετίζεται με την ανθρώ-

πινη ιδιότητα του. Συγκεκριμένα, η πληροφορία ”Person” εντοπίζεται τόσο στην οντολογία της DBpedia (http://dbpedia.org/ontology/Person) στο λεξικό της foaf που χρησιμοποιείται και αυτό κατά τη διαδικασία χαρτογράφησης της DBpedia (http://xmlns.com/foaf/0.05/Person) καθώς και από το λεξικό schema.org (http://schema.org/Person) που χρησιμοποιείται με αντίστοιχο τρόπο. Έτσι τα περισσότερα resource URI's ατόμων, συνδέονται και με τα τρία αυτά URI's με τριάδες στο ίδιο μοτίβο με το παράδειγμα που ακολουθεί:

<http://dbpedia.org/resource/Albert_Einstein> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dbpedia.org/ontology/Person>
<http://dbpedia.org/resource/Albert_Einstein> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://schema.org/Person>
<http://dbpedia.org/resource/Albert_Einstein> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://xmlns.com/foaf/0.05/Person>

Ωστόσο, δεν είναι καταχωρημένοι στα λεξικά της foaf και της schema όλα τα εν ζωή πρόσωπα τα οποία διαθέτουν καταχώρηση στη DBpedia. Επίσης η οντολογία της DBpedia, όπως έχει αναφερθεί και στην αντίστοιχη ενότητα (2.4.7) βασίζεται εν μέρη στα παραπάνω λεξικά για την υλοποίηση της και επιπρόσθετα, καθώς κατασκευάζεται με χειροκίνητες μεθόδους, η κατασκευή της εξελίσσεται με πολύ αργότερους ρυθμούς σε σχέση με τους ρυθμούς ανανέωσης και ενημέρωσης του περιεχομένου της Wikipedia και της DBpedia.

Για το σκοπό αυτό αναζητήθηκαν εναλλακτικές τριάδες που να μπορούν να εξασφαλίσουν την ανθρώπινη ιδιότητα σε περίπτωση που κάποιο αναγνωριστικό URI δεν αντιστοιχεί σε κάποια από τις παραπάνω τριάδες αλλά παρ όλα αυτά περιγράφει συγκεκριμένο άτομο.

Προέκυψε ότι σχεδόν στο σύνολό τους τα αναγνωριστικά URI των παραπάνω ατόμων συνδέονται μέσω του κατηγορήματος *<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>* με το αντικείμενο *<http://dbpedia.org/resource/Category:Living_people>*. Μάλιστα η χρήση της τριάδας αυτής είναι αρκετά διαδεδομένη αλλά και ασφαλής καθώς προέρχεται από την κατηγοριοποίηση της wikipedia και περιλαμβάνει πάνω από 500000 καταχωρήσεις ατόμων.

Μάλιστα η αξιοποίηση των κατηγοριών ”Living people” και ”Dead People” οδήγησε στο συμπέρασμα ότι υπάρχουν μόνο εν ζωή πρόσωπα τα οποία δεν είναι καταχωρημένα στα λεξικά foaf και schema.

Συγκεκριμένα η αναζήτηση:

```
SELECT count (?x) WHERE {
        ?x ?z <http://dbpedia.org/resource/Category:Living_people>
        OPTIONAL {
                ?y <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>  <http://xmlns.com/foaf/0.05/Person> .
                FILTER (?x = ?y) .
        }
        FILTER ( !BOUND(?y) )
```

```
}
```

όπου αναζητάμε το πλήθος των ατόμων που δεν είναι καταχωρημένα στο foaf αλλά ανήκουν στην κατηγορία Living_people έδωσε αποτέλεσμα: 53022 άτομα, ενώ αντίστοιχα η αναζήτηση:

```
SELECT count (?x) WHERE {
        ?x ?z <http://dbpedia.org/resource/Category:Dead_people>
        OPTIONAL {
            ?y <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>  <http://xmlns.com/foaf/0.05/Person> .
            FILTER (?x = ?y) .
        }
        FILTER ( !BOUND(?y) )
}
```

όπου αναζητάμε το πλήθος των ατόμων που δεν είναι καταχωρημένα στο foaf αλλά ανήκουν στην κατηγορία Dead_people έδωσε αποτέλεσμα: 0 άτομα.

Συνεπώς προκειμένου να καθοριστεί αν κάποιο αναγνωριστικό URI αντιστοιχεί σε πρόσωπο ελέγχεται η ύπαρξή του ώς υποκείμενο στις τρεις τριάδες που αναφέρθηκαν παραπάνω καθώς και στην τριάδα:

<http://dbpedia.org/resource/Name_Surname> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dbpedia.org/resource/Category:Living_people>

Οι φράσεις της λίστας λοιπόν ελέγχονται με τη σειρά και αν το resource κάποιας ανήκει σε τουλάχιστον μία από τις παραπάνω τριάδες τότε η αντίστοιχη φράση σημειώνεται ως πρόσωπο (Person), και ο έλεγχος συνεχίζει στην επόμενη λέξη. Στο τέλος στην έξοδο του προγράμματος δίνεται αρχείο το οποίο εκτυπώνει το σύνολο των υπαρκτών προσώπων που εντοπίστηκαν στο κείμενο καθώς και το αναγνωριστικό URI που αντιστοιχεί σε κάθε πρόσωπο ώστε να είναι εφικτή η σύνδεση με το Σημασιολογικό Ιστό μέσω DBpedia και η ανάκτηση επιπρόσθετης πληροφορίας για το συγκεκριμένο πρόσωπο.

Κεφάλαιο 4

Ζητήματα Υλοποίησης

Στο κεφάλαιο αυτό θα αναλυθούν κάποια ζητήματα υλοποίησης και των δύο συστημάτων που κρίνονται απαραίτητα για την κατανόηση της λειτουργίας τους. Για το μεγαλύτερο μέρος της υλοποίησης χρησιμοποιήθηκε η γλώσσα προγραμματισμού JAVA 6 σε συνδυασμό με SPARQL για το τμήμα της άντλησης πληροφορίας από την DBpedia. Το σημαντικότερο τμήμα της υλοποίησης παρουσιάζεται στην δεύτερη ενότητα του κεφαλαίου αυτού και έχει να κάνει με την απόφαση για τον τρόπο αναπαράστασης του κειμένου και τη δομή των δεδομένων που θα διοχετεύεται σε κάθε στάδιο του προγράμματος. Η αναπαράσταση αυτή σχετίζεται άμεσα με την θεώρηση των φράσεων-στόχων ως ενιαία αντικείμενα τα οποία με την ολοκλήρωση των πρώτων σταδίων αποτελούν και πρακτικά αυθύπαρκτες οντότητες. Στην τρίτη ενότητα, παρουσιάζονται βασικές αποφάσεις υλοποίησης που θεωρήθηκαν σημαντικές για την κατανόηση των διαδικασιών που ακολουθήθηκαν. Η τελευταία ενότητα του κεφαλαίου εξηγεί ζητήματα που σχετίζονται με την κανονικοποίηση χαρακτηριστικών μεγεθών αξιολόγησης και την επιλογή βαρών για τον γραμμικό συνδυασμό αυτών, τα οποία αποτελούν και το τελικό κριτήριο διαμόρφωσης της τελικής κατάταξης των υποψήφιων οντοτήτων-στόχων. Προφανώς, η ενότητα αυτή αναφέρεται μόνο στο σύστημα CRESTA καθώς το SWPID δεν περιλαμβάνει τέτοιου είδους αξιολόγηση των αντλούμενων από το κείμενο οντοτήτων. Στην ενότητα που ακολουθεί, παρουσιάζονται τα εργαλεία που χρησιμοποιήθηκαν αυτούσια, ώστε να γίνει κατανοητή η λειτουργία και ο ρόλος τους στο πλαίσιο της υλοποίησης.

4.1 Εργαλεία

4.1.1 TreeTagger

Για τη γραμματική επισημείωση χρησιμοποιήθηκε ο πιθανοτικός TreeTagger που αναπτύχθηκε από τον Helmut Schmid στο πλαίσιο του TC project στο Institute for Computational Linguistics, University of Stuttgart[44]. Πρόκειται για έναν Hidden Marcov Model tagger που χρησιμοποιεί

στατιστικούς κανόνες και δέντρα αποφάσεων.

Ο treeTagger δέχεται σαν είσοδο απλό κείμενο και δίνει σαν έξοδο ένα διάνυσμα για κάθε λέξη του κειμένου, το οποίο περιλαμβάνει την αρχική μορφή της λέξης όπως εμφανίζεται στο κείμενο, την επικρατέστερη ετικέτα γραμματικής επισημείωσης της λέξης και την κανονικοποιημένη (stemmed) μορφή της. Το σύνολο των ετικετών t, το οποίο περιλαμβάνει όλες τις πιθανές αποκρίσεις του Treetagger για κάθε λέξη παρουσιάζεται συνοπτικά στον πίνακα του Παραρτήματος 3[43]

Η λειτουργία του TreeTagger καθώς και η διαδικασία εκπαίδευσής του παρουσιάζεται συνοπτικά στη συνέχεια[32]:
Έστω $P(w_1w_2...w_n, t_1t_2...t_n)$ η πιθανότητα μίας ακολουθίας λέξεων $w_1w_2...w_n$ που έχουν προσημειωθεί με ετικέτα γραματικής επισημείωσης. Επίσης θα γίνουν δύο παραδοχές. Η πρώτη είναι πως η πιθανότητα μίας λέξης w_n εξαρτάται μόνο από τη γραμματική της επισημείωση t_n, και η δεύτερη πως η πιθανότητα μιας επιλογής γραμματικής ετικέτας εξαρτάται μόνο από τις αντίστοιχες ετικέτες των k προηγούμενων λέξεων. Οι παραπάνω παραδοχές σε συνδυασμό με το θεώρημα του Bayes δίνουν την παρακάτω σχέση:

$$P(w_1w_2...w_n, t_1t_2...t_n)= P(w_i|t_i)P(t_i|t_{i-k}...t_{i-1})P(w_1w_2...w_{n-1}, t_1t_2...t_{n-1})$$

Πιθανοτικά μοντέλα αυτού του τύπου, όπου το επόμενο στάδιο εξαρτάται μόνο από τα k προηγούμενα, ονομάζονται Markov Models k-οστής τάξης. Σε αντίθεση με τον TreeTagger, οι Ngram Taggers, για τον υπολογισμό των πιθανοτήτων μετάβασης χρησιμοποιούν την αρχή της εκτίμηση μέγιστης πιθανοφάνειας (Maximum Likelihood Estimation). Σύμφωνα με την αρχή αυτή, η πιθανότητα μετάβασης υπολογίζεται με βάση την συχνότητα εμφάνισης της ακολουθίας $t_it_{i-1}...t_{i-k}$ σε σχέση με τη συχνότητα εμφάνισης της ακολουθίας $t_{i-1}...t_{i-k}$. Η μέθοδος αυτή παρουσιάζει αρκετά προβλήματα, κυρίως στην περίπτωση που η πιθανότητα μετάβασης είναι πολύ μικρή ή ακόμα και μηδενική. Τότε, ο υπολογισμός δεν είναι αξιόπιστος. Ένας τρόπος αντιμετώπισης του προβλήματος είναι η αντικατάσταση των μηδενικών τιμών με πολύ μικρές θετικές τιμές, και η επανακανονικοποίηση των πιθανοτήτων ώστε να έχουν άθροισμα 1. Για να αποφευχθούν τα παραπάνω προβλήματα, ο TreeTagger χρησιμοποιεί δέντρο αποφάσεων για των υπολογισμό των πιθανοτήτων μετάβασης. Το δέντρο αποφάσεων χτίζεται αναδρομικά με τη χρήση ενός συνόλου δεδομένων για εκπαίδευση, αποτελούμενο από ακολουθίες ετικετών, με την εφαρμογή του αλγορίθμου ID3[31] προσαρμοσμένου.Σε κάθε βήμα επεκτείνεται η δημιουργία ενός δέντρου το μέγεθος του οποίου αυξάνεται αναδρομικά για κάθε ένα από τα υποσύνολα που δημιουργούνται από τις δοκιμές του προηγούμενου βήματος. Μετά την αρχική δημιουργία του δέντρου, πρέπει να γίνει κλάδεμα (pruning). Αυτό επιτυγχάνεται με τη χρήση του σταθμισμένου κέρδους πληροφορίας(weighted information gain). Πρόκειται για ένα μέγεθος που έχει να κάνει με την πληροφορία που δίνει ένας κόμβος μετά την ολοκλήρωση της κατασκευής του δέντρου. Οι κόμβοι που δεν ξεπερνούν το δεδομένο όριο για το G, αποκόπτονται από το δέντρο. Τέλος, ο υπολογισμός της καλύτερης ακολουθίας tags δεδομένης μίας ακολουθίας λέξεων γίνεται αποδοτικά με χρήση του αλγορίθμου Viterby[47].

Το λεξικό που χρησιμοποιείται για την αναζήτηση των πιθανών tags μιας λέξης αποτελείται

από 3 μέρη, το ολοκληρωμένο λεξικό (fullform lexicon), το λεξικό καταλήξεων (suffix lexicon) και το προεπιλεγμένο λεξικό (default lexicon). Για την κάθε λέξη γίνεται αναζήτηση σε κάθε ένα από τα τρία επίπεδα-μέρη του λεξικού με τη σειρά και σταματάει αν σε κάποιο επίπεδο της αναζήτησης βρεθεί αποτέλεσμα. Ενδιαφέρον παρουσιάζει το suffix lexicon το οποίο λειτουργεί με τη μορφή δέντρου κατασκευασμένου με τη χρήση συνόλου εκπαίδευσης προσημειωμένων λέξεων. Μετά τη δημιουργία του δέντρου, γίνεται και σε αυτό το σημείο κλάδεμα κόμβων με τη χρήση κριτηρίου ποσότητας πληροφορίας κόμβου.

Για τις ανάγκες της συγκεκριμένης εργασίας εξετάστηκαν τόσο ο TreeTagger, όσο και ο Stanford log-lenear POS Tagger[42]. Αν και ο Stanford POS tagger ήταν λίγο πιο ακριβής, τελικά επιλέχθηκε ο TreeTagger, για λόγους ταχύτητας, καθώς η διαφορά των δύο εργαλείων στην ταχύτητα ήταν αναντίστοιχα μεγάλη σε σχέση με τη διαφορά της ακρίβειάς τους.

4.1.2 Mediawiki API

Το MediaWiki API είναι μία διαδικτυακή υπηρεσία η οποία παρέχει άμεση και υψηλού επιπέδου πρόσβαση στα δεδομένα που βρίσκονται στις βάσεις του MediaWiki (ένα Wiki το οποίο χρησιμοποιείται κυρίως για τις ανάγκες της Wikipedia). Τα προγράμματα πελατών μπορούν να συνδέονται, να βλέπουν τα δεδομένα και να κάνουν αλλαγές απλά στέλνοντας HTTP αιτήσεις προς τον εξυπηρετητή. Η χρήση του API γίνεται απλά στέλνοντας μία ερώτηση HTTP για κάποιο URL (με συγκεκριμένη μορφή που θα προσδιοριστεί στη συνέχεια), μέσω οποιασδήποτε γλώσσας προγραμματισμού ή απλά μέσω κάποιου web browser. Παρακάτω θα δοθεί ένα χαρακτηριστικό παράδειγμα τέτοιου URL και θα προσδιοριστούν τα επιμέρους τμήματά του και η λειτουργία που το καθένα από αυτά επιτελεί:

http://en.wikipedia.org/w/api.php?format=xml&action=query&list=search&srredirects

&srprop=snippet&srsearch=Athens

Το πρώτο τμήμα που διακρίνεται είναι το http://en.wikipedia.org/w/api.php? το οποίο αποτελεί και το endpoint του API. Είναι η βάση για το URL του API του αγγλικού Wikipedia και με αυτό ξεκινάνε όλα τα URLs που χρησιμοποιούν το API. Η HTTP αίτηση που αντιστοιχεί στο συγκεκριμένο URL, επιστρέφει το αρχείο που περιλαμβάνει την περιγραφή του API και τις διάφορες δυνατότητες σχηματισμού URL για τη χρήση του. Το υπόλοιπο κομμάτι του URL περιλαμβάνει τις παραμέτρους που προσδιορίζουν την αναζήτηση και τα αποτελέσματά της. Το τμήμα: format=xml ζητάει τα αποτελέσματα να δοθούν σε μορφή XML (extensible markup language). Οι υπόλοιπες επιλογές για τη μορφή του αποτελέσματος είναι json, jsonfm, php, phpfm, wddx, wddxfm, xml, xmlfm, yaml, yamlfm, rawfm, txt, txtfm, dbg, dbgfm, dump, dumpfm. Το τελευταίο κομμάτι προσδιορίζει την ενέργεια που θα πραγματοποιηθεί πάνω στα δεδομένα. Σε αυτό το παράδειγμα, η ενέργεια είναι μία ερώτηση (action=query), και πιο συγκεκριμένα, το ζητούμενο στην ερώτηση είναι να γίνει αναζήτηση μιας συγκεκριμένης φράσης (list=search). Η φράση αυτή είναι η ”Athens” (srsearch=Athens). Επιπλέον, κατά την αναζήτηση ζητείται να ληφθούν υπόψιν και οι redirect pages (srredirects). Για κάθε φορά που αναφέρεται η φράση

Athens, επιστρέφεται ένα μικρό απόσπασμα (srprop=snippet), με περιορισμό τα 10 αποτελέσματα που είναι και ο περιορισμός που τίθεται αυτόματα (by default) κατά την ερώτηση, από τη στιγμή που δεν έχει προσδιοριστεί κάποιος άλλος αριθμός ως άνω φράγμα. Φυσικά, στην απάντηση θα επιστραφεί και ο αριθμός των συνολικών φορών που βρέθηκε η φράση αυτή. Ο όρος που χρησιμοποιείται από το API για την πληροφορία αυτή είναι: totalHits.

Ένα άλλο παράδειγμα είναι το εξής:

http://en.wikipedia.org/w/api.php?format=xml&action=query&list=backlinks

&bllimit=1000&redirects&blredirect&blnamespace=0&bltitle=Athens.

Εδώ, όπως και πριν, τα πρώτα δύο τμήματα προσδιορίζουν το αγγλικό API και τη μορφή του αποτελέσματος. Στη συνέχεια προσδιορίζεται πως θα γίνει ερώτηση αλλά αυτή τη φορά δεν θα έχει τη μορφή αναζήτησης φράσης αλλά την αναζήτησης συνδέσμων από άλλες σελίδες του wikipedia προς τη συγκεκριμένη σελίδα που αντιστοιχεί στο άρθρο με τίτλτο Athens (bltitle=Athens). Αυτή τη φορά, επιστρέφεται ένα xml αρχείο που περιλαμβάνει τη λίστα των backlinks. Επειδή τα αποτελέσματα συνήθως είναι περισσότερα από το όριο (1000), χρειάζεται να ελέγξουμε το κατά πόσο υπάρχει συνέχεια στην απάντηση. Αν υπάρχει συνέχεια, τότε στο τέλος του αρχείου δίνεται ένας κωδικός, ο οποίος προστίθεται στο ίδιο URL και στέλνεται εκ νέου HTTP ερώτηση η απάντηση στο οποίο είναι το αρχείο με τη συνέχεια των αποτελεσμάτων. Η διαδικασία των HTTP ερωτήσεων συνεχίζεται μέχρι να μην υπάρχουν άλλα αποτελέσματα.

Το API έχει πολλές ακόμα δυνατότητες όπως αναζήτηση των links κάθε σελίδας προς άλλες σελίδες της Wikipedia, ή αναζήτηση των κατηγοριών της Wikipedia στις οποίες ανήκει κάθε άρθρο.

4.1.3 Virtuoso

Ο Virtuoso είναι ένας επεκτάσιμος, ανεξαρτητος πλατφόρμας εξυπηρετητής (server), ο οποίος συνδυάζει διαχείριση σχεσιακών, διαγραμματικών και αναλυτικών δεδομένων, με λειτουργίες διαδικτιακού server και πλατφόρμας παροχής υπηρεσιών. Στην παρούσα εργασία χρησιμοποιήθηκε ως τοπικός εξυπηρετητής γνωσιακής βάσης, ώστε να υπάρχει η δυνατότητα φόρτωσης και διαχείρισης των δεδομένων της DBpedia, δηλαδή δεδομένων σε μορφή RDF.

Η χρήση του κρίθηκε αναγκαία, και προτιμήθηκε έναντι της χρήσης των ήδη φορτωμένων και διαθέσιμων online δεδομένων της DBpedia, για δύο λόγους. Πρώτον, το σύνολο των δεδομένων που είναι φορτωμένο στον αντίστοιχο διαδικτυακό εξυπηρετητή, ανανεώνεται συχνά χωρίς να διατίθεται εύκολα προσβάσιμη λίστα των αλλαγών, γεγονός που μπορεί να επηρεάσει με μη προβλέψιμο τρόπο τα αποτελέσματα των μετρήσεων. Δεύτερον, δεν είναι όλα τα διαθέσιμα σύνολα δεδομένων προσβάσιμα μέσω διαδικτύου, με αποτέλεσμα να μη μπορεί να αξιοποιηθεί το συνολο της πληροφορίας που εχει εξαχθεί από τη Wikipedia.

Ακολουθούν οι βασικές αρχές λειτουργίας του Virtuoso, ως προς την αποθήκευση και τη χρηση των δεδομένων.

Ο virtuoso αποθηκεύει δεδομένα με τη μορφή τριάδων RDF. Συγκεκριμένα κάθε νέα τριάδα, ή σύνολο τριάδων, φορτώνεται σε ένα γράφο (προκαθορισμένο, ή καθορισμένο από το χρήστη). Παράλληλα δίνεται και η δυνατότητα σύνδεσης με περισσότερους από έναν γράφους. Η τρέχουσα κατάσταση της βάσης αποθηκεύεται σε ένα πίνακα δεδομένων RDF_QUAD. Κάθε τριάδα αντιπροσωπεύεται από μία γραμμή στον πίνακα, με τέσσερα στοιχεία: τα στοιχεία της τριάδας (υποκείμενο-κατηγόρημα-αντικείμενο) και το γράφο στον οποίο ανήκει.

Η αναζήτηση δεδομένων και η εξαγωγή πληροφοριών από τα δεδομένα του virtuoso, γίνεται μέσω SQL και SPARQL. Συγκεκριμένα, υπάρχει ενσωματωμένο πρόγραμμα με sql και παρέχεται η δυνατότητα χρήσης SPARQL στο εσωτερικό του SQL. Ένα υποερώτημα SPARQL ή ένας παραγόμενος πίνακας γίνεται αποδεκτό είτε ως δήλωση SQL είτε ως ένα υποερώτημα ή πίνακας κάποιου άλλου αποδεκτού τύπου. Έτσι η SPARQL κληρονομεί όλες τις συγκεντρωτικές λειτουργίες και λειτουργίες ομαδοποίησης της SQL, καθώς και οποιαδήποτε ενσωματωμένη ή οριζόμενη από το χρήστη λειτουργία.

Ακόμα υπάρχουν κατάλληλα προργάμματα οδήγησης ώστε ο virtuoso να λειτουργεί με εφαρμογές RDF πλαισίων όπως τα Jena, Sesame and Redland. Έτσι εφαρμογή γραμμένη σε αυτά μπορεί να χρησιμοποιήσει το virtuoso ως επεξεργαστή αποθήκευσης και αναζήτησης. Εσωτερικά, η SPARQL μεταφράζεται σε SQL κατά το χρόνο της μετάφρασης του κώδικα αναζήτησης.

4.2 Δομές-Αντικείμενα

Όπως έχει αναφερθεί σε προηγούμενα κεφάλαια, η πληροφορία που δίνεται ως έξοδος από το σύστημα είναι οντότητες, οι αναγνωριστικές συμβολοσειρές των οποίων υπάρχουν στο αρχικό υπό-εξέταση κείμενο. Για το λόγο αυτό επιλέχθηκαν ως δομές δεδομένων αντικείμενα που αντιπροσωπεύουν γλωσσικές οντότητες του κειμένου και στη συνέχεια αποκτούν σημασιολογική υπόσταση. Οι δομές αυτές είναι οι εξής:

- **Λέξη (Word)** Πρόκειται για λέξεις του κειμένου, όπως αυτές δίνονται ως έξοδος από το στάδιο της γραμματικής επισημείωσης (μαζί με τη γραμματική ετικέτα της κάθε μίας και την semi-stemmed μορφή της που αποθηκεύονται ως χαρακτηριστικά των αντικειμένων της κλάσης αυτής μαζί με άλλες πληροφορίες που θα αναλυθούν στη συνέχεια).

- **Φράση (Phrase) και η υποκλάση Ονοματική Φράση (NounPhrase)** Τα αντικείμενα της κλάσης αυτής αποτελούν συλλογές λέξεων και μάλιστα με συγκεκριμένη διάταξη (που είναι πάντα αυτή με την οποία η κάθε ακολουθία λέξεων εμφανίζεται στο κείμενο). Πρόκειται λοιπόν για μία λίστα από λέξεις, οι οποίες θα αποτελέσουν τις υποψήφιες φράσεις-στόχους και στη συνέχεια τις φράσεις στόχους των δύο συστημάτων. Η κλάση Phrase υλοποιείται ως αφηρημένος τύπος κλάσης (abstract class) (abstract class) ενώ η κλάση NounPhrase ως υποκλάση της Phrase. Ο διαχωρισμός αυτός έγινε για λόγους επεκτασιμότητας, καθώς πιθανές φράσεις-στόχοι, θα μπορούσαν να αποτελέσουν εκτός από τις ονοματικές φράσεις όπως έχουν οριστεί στο πλαίσιο της διπλωματικής αυτής, και άλλου

τύπου φράσεις. Ένα τέτοιο παράδειγμα διαφορετικού τύπου φράσης που είχε προταθεί κατά το σχεδιασμό των απαιτήσεων ήταν οι φράσεις εντός εισαγωγικών. Η πρόταση αυτή σχετίζεται με την παρατήρηση πως υπάρχει ένας περιορισμένος αριθμός οντοτήτων της DBpedia που αναπαριστώνται από φράσεις που περιέχουν ρήματα ή γενικότερα στοιχεία, τα οποία δεν περιλαμβάνονται στα στοιχεία που λαμβάνει υπ'όψιν το σύστημα κατά την παρούσα υλοποίηση. Χαρακτηριστικά παραδείγματα οι τίτλοι ταινιών ή έργων τέχνης γενικότερα. Προκειμένου, λοιπόν, να μην απορριφθούν τέτοιες φράσεις, έγινε προσπάθεια να αναζητούνται στη βάση δεδομένων της DBpedia και όλες οι φράσεις που βρίσκονται σε εισαγωγικά. Όμως, ο διαφορετικός τρόπος αντιμετώπισης σε επίπεδο αναπαράστασης κειμένου από πηγή σε πηγή, αλλά και από τον ίδιο τον TreeTagger, δημιούργησε προβλήματα κατά την αναγνώριση των χαρακτήρων που αντιπροσωπεύουν εισαγωγικά αρχής και τέλους μίας τέτοιας φράσης. Η ιδέα τελικώς απορρίφθηκε. Παρ' όλα αυτά, ο διαχωρισμός μεταξύ της κλάσης Phrase και της υποκλάσης της NounPhrase επιτρέπει την επέκταση του συστήματος ώστε να είναι εύκολα προσαρμόσιμο σε εφαρμογές διαφορετικών απαιτήσεων. Για κάθε ονοματική φράση (στιγμιότυπο της κλάσης NounPhrase) αποθηκεύονται ως χαρακτηριστικά κάποια χαρακτηριστικά της τα οποία είτε αντλούνται από το ίδιο το κείμενο κατά το σχηματισμό των αντικειμένων των κλάσεων αυτών, είτε συμπληρώνονται κατά τη εξέλιξη των σταδίων με πληροφορίες από την Wikipedia και την DBpedia.

- **Λίστα ονοματικών φράσεων (MainList)** Πρόκειται για την κλάση των αντικειμένων που αποτελούν λίστες των ονοματικών φράσεων που κατασκευάζονται από το κείμενο, και είναι αυτή η τελική δομή δεδομένων που διοχετεύεται από κάθε προηγούμενο στάδιο του προγράμματος στο επόμενο. Οι μέθοδοι της κλάσης αυτής είναι αυτές που ελέγχουν την ταξινόμηση των NounPhrases και επεξεργάζονται τις πληροφορίες που είναι αποθηκευμένες σε αυτά, έχοντας έτσι τη γενική εποπτεία του συστήματος.

Αναλυτικότερα, κάθε μία από τις τρεις αυτές δομές-κλάσεις παρουσιάζεται στις παρακάτω τρεις υποενότητες.

4.2.1 Word

Κάθε αντικείμενο της κλάσης αυτής δημιουργείται όταν μία λέξη του κειμένου θεωρείται ότι πληροί τις προϋποθέσεις ώστε να ανήκει σε κάποια NounPhrase. Όταν συμβαίνει αυτό, τη NounPhrase είναι υπεύθυνη να καλέσει τον κατασκευαστή της κλάσης Word και να αρχικοποιήσει τα παρακάτω χαρακτηριστικά ως εξής:

- String text: Πρόκειται για τη συμβολοσειρά της κάθε λέξης όπως αυτό εμφανίζεται στο κείμενο

- String stemmed: Πρόκειται για τη semi-stemmed μορφή της κάθε λέξης όπως αυτή δίνεται ως έξοδος από το στάδιο γραμματικής επισημείωσης.

- String pos: Πρόκειται για τη γραμματική ετικέτα της κάθε λέξης όπως αυτό δίνεται ως έξοδος από το στάδιο γραμματικής επισημείωσης.
- Integer wc: Πρόκειται για τον μοναδικό ακέραιο που αποτελεί ταυτότητα της κάθε λέξης και αντιπροσωπεύει την θέση της στο κείμενο.
- Integer id: Πρόκειται για τον ακέραιο που αποτελεί μοναδικό αναγνωριστικό της κάθε λέξης μέσα σε μία NounPhrase και αντιπροσωπεύει την θέση της μέσα σε αυτή. Προφανώς δύο λέξεις επιτρέπεται να έχουν το ίδιο id αρκεί να μην βρίσκονται στην ίδια ονοματική φράση.

Τα χαρακτηριστικά αυτά, περιλαμβάνουν πληροφορία που αντλείται για κάθε λέξη από το ίδιο το κείμενο και από τη γραμματική επισημείωσή του, επομένως, οι τιμές τους είναι γνωστές όταν έρχεται η ώρα για την κατασκευή τέτοιων αντικειμένων. Ο κατασκευαστής της κλάσης Word καλείται από μέθοδο της κλάσης Phrase, όταν αποφασιστεί ότι η εν λόγω λέξη πρέπει να ανήκει στη συγκεκριμένη φράση-απόφαση που παίρνεται κατά την κατασκευή της φράσης αυτής.

4.2.2 Phrase-NounPhrase

Τα αντικείμενα της κλάσης αυτής, κατασκευάζονται στο στάδιο της ανάλυσης της εξόδου του σταδίου γραμματικής επισημείωσης κατά το στάδιο εξαγωγής ονοματικών οντοτήτων. Πρόκειται για αντικείμενα που αντιπροσωπεύουν τις υποψήφιες φράσεις-στόχους και, ουσιαστικά, συγκρατούν ως πληροφορία τις λέξεις από τις οποίες αποτελούνται και κάποια μετρήσιμα χαρακτηριστικά που ταυτοποιούν ή αξιολογούν την κάθε φράση. Όλα τα χαρακτηριστικά της κλάσης NounPhrase κληρονομούνται απευθείας από την κλάση Phrase αλλά αρχικοποιούνται κατά την κλήση του κατασκευαστή NounPhrase, όταν κατά την ανάλυση του γραμματικώς επισημειωμένου κειμένου, θεωρηθεί πως πληρούνται οι προϋποθέσεις για την εκκίνηση της κατασκευής μίας ονοματικής φράσης. Τα χαρακτηριστικά της κλάσης Phrase είναι τα εξής:

- Integer id: Το μοναδικό id κάθε φράσης και προσδιορίζει τη θέση της στην λίστα των φράσεων.
- Integer wn: Ο αριθμός των λέξεων που περιλαμβάνονται σε κάθε φράση
- Integer type: Ένας ακέραιος που προσδιορίζει τι τύπου είναι η συγκεκριμένη φράση. (Χρησιμοποιείται στην περίπτωση που υπάρχουν και διαφορετικού είδους φράσεις πέρα από NounPhrase, για την παρούσα υλοποίηση αρχικοποιείται πάντα στο 0 και δεν ξαναχρησιμοποιείται κατά την εκτέλεση του προγράμματος)
- Integer appearCounter: Ο αριθμός των επαναλήψεων της φράσης αυτής στο κείμενο.
- Integer Found: Ένας ακέραιος που προσδιορίζει το κατά πόσο έχει βρεθεί μια φράση ως οντότητα στην DBpedia ή όχι. Η τιμή του είναι 0 ή 1. Προφανώς όταν για μία φράση

ισχύει Found=0, τότε η φράση αυτή απορρίπτεται ή επανεξετάζονται τμήματά της. Στην αντίθετη περίπτωση, μένει η ίδια καθόλη τη διάρκεια της εκτέλεσης του προγράμματος μέχρι και την τελική έξοδό του.

- double termFrequency: Δεδαδικός αριθμός που προσδιορίζει την συχνότητα όρου όπως αυτή έχει οριστεί ως κριτήριο αξιολόγησης της φράσης στο προηγούμενο κεφάλαιο.
- double firstAppearence: Δεκαδικός αριθμός που αντιπροσωπεύει την πρώτη εμφάνιση της πρώτης λέξης κάθε φράσης στο κείμενο, όπως έχει οριστεί ως κριτήριο αξιολόγησης στο προηγούμενο κεφάλαιο.
- double totalHits: Δεκαδικός αριθμός που αντιπροσωπεύει τον αριθμό εμφανίσεων της φράσης στην Wikipedia.
- double metricSum: Δεκαδικός αριθμός που αντιπροσωπεύει το γραμμικό συνδυασμό των μεγεθών term frequency, first appearence, keyPhraseness.
- double backLinks: Δεκαδικός αριθμός που προσδιορίζει τον αριθμό εμφανίσεων της φράσης ως link προς άλλα άρθρα της Wikipedia.
- double categ: Δεκαδικός αριθμός που αντιπροσωπεύει το χαρακτηριστικό μέγεθος αξιολόγησης που σχετίζεται με τις κατηγορίες τις DBpedia, όπως αυτό αναλύθηκε στο προηγούμενο κεφάλαιο.
- double keyPhraseness: Δεκαδικός αριθμός που αντιπροσωπεύει το χαρακτηριστικό μέγεθος αξιολόγησης keyPhraseness όπως αυτό αναλύθηκε στο προηγούμενο κεφάλαιο.
- String phraseString: Η συμβολοσειρά της φράσης που βρίσκεται αυτούσια μέσα στο κείμενο.
- String stemmedString: Η συμβολοσειρά της φράσης με κάθε λέξη της να βρίσκεται στη semi-stemmed μορφή της.
- String underscoreString: Η συμβολοσειρά της φράσης με κάθε λέξη της να βρίσκεται στη semi-stemmed μορφή της και με τη σύνδεση των λέξεων να γίνεται με χρήση underscore (κάτω παύλας).
- String uriString: Το URI της οντότητας της DBpedia που αντιπροσωπεύει η φράση ως συμβολοσειρά.
- String wikiLink: Το URL του αντίστοιχου άρθρου της wikipedia που αναφέρεται στην οντότητα που αντιπροσωπεύει η φράση, με τη μορφή συμβολοσειράς.
- ArrayList <String> extLinks: Η λίστα με τις συμβολοσειρές που αναπαριστούν τα URLs όλων των συνδέσμων της οντότητας αυτής με εξωτερικές πηγές, όπως αυτές προτείνονται από την DBpedia.
- ArrayList Word: Η λίστα με τα αντικείμενα της κλάσης Word που περιλαμβάνει κάθε φράση.

Για τα παραπάνω χαρακτηριστικά και τη χρησιμότητά τους, αξίζει να σχολιαστούν τα εξής:

1. Τα χαρακτηριστικά word number, appearCounter, backLinks, totalHits χρησιμοποιούνται μόνο για τον υπολογισμό των χαρακτηριστικών μεγεθών αξιολόγησης. Τα πρώτα δύο αντλούνται απευθείας από το κείμενο (παρότι δεν είναι γνωστά κατά την κατασκευή της φράσης) ενώ τα δύο επόμενα αντλούνται από την Wikipedia μέσω του mediawiki API.

2. Οι τόσες διαφορετικές μορφές String που διατηρούνται για την κάθε φράση, έχουν να κάνουν με διαφορετικές προσεγγίσεις της φράσης ως συμβολοσειράς κατά τη διάρκεια των φάσεων του προγράμματος. Πιο συγκεκριμένα, το χαρακτηριστικό phraseString δίνει πληροφορία για το πώς εντοπίστηκε την πρώτη φορά η κάθε φράση στο κείμενο, ώστε να γνωρίζουμε την αρχική μορφή της. Το χαρακτηριστικό stemmed String δίνει την φράση με κάθε λέξη της να έχει υποστεί μερικό stemming. Έτσι μπορούμε να ενοποιήσουμε φράσεις που ενώ βρίσκονται σε διαφορετική μορφή στο κείμενο, είναι ακριβώς οι ίδιες και άρα να εντοπίσουμε την τιμή του χαρακτηριστικού appearCounter. Το χαρακτηριστικό uriString είναι το uri της DBpedia, και χρησιμοποιείται για τα queries στην βάση δεδομένων της DBpedia. Τέλος, το χαρακτηριστικό underscoreString δημιουργήθηκε με πρόθεση να βοηθήσει στην αναζήτηση οντότητας στην DBpedia, καθώς συχνά τα αναγνωριστικά των οντοτήτων έχουν τη μορφή: Λέξη1_Λέξη2_Λέξη3. Παρόλα αυτά, κατά την διεκπεραίωση των queries, αναγνωρίζονται τόσο τα stemmedStrings όσο και τα underscoreStrings επομένως, πρακτικά, το χαρακτηριστικό underscoreString δεν χρησιμοποιείται και μπορεί να καταργηθεί. Παραμένει στον κώδικα για λόγους επεκτασιμότητας.

4.2.3 MainList

Η MainList ως λίστα των αντικειμένων τύπου NounPhrase που έχουν εντοπιστεί, είναι πολύ σημαντική κυρίως για τις μεθόδους της, που διατρέχουν ολόκληρη τη λίστα και εφαρμόζουν τις ίδιες ενέργειες πάνω σε όλα τα αντικείμενά της. Ως χαρακτηριστικά υπάρχουν μόνο κάποια χαρακτηριστικά μεγέθη ολόκληρης της λίστας που έχουν να κάνουν με ελάχιστες/μέγιστες τιμές των χαρακτηριστικών μεγεθών αξιολόγησης των φράσεών της. Πιο συγκεκριμένα τα χαρακτηριστικά της κλάσης MainList παρουσιάζονται στη συνέχεια:

- double minTot: Η ελάχιστη τιμή του χαρακτηριστικού μεγέθους keyPhraseness που παρουσιάζεται.

- double maxTot: Η μέγιστη τιμή του χαρακτηριστικού μεγέθους keyPhraseness που παρουσιάζεται.

- double minLeF: Η ελάχιστη τιμή του χαρακτηριστικού μεγέθους term frequency που παρουσιάζεται.

- double maxLeF: Η μέγιστη τιμή του χαρακτηριστικού μεγέθους term frequency που παρουσιάζεται.

- double minFap: Η ελάχιστη τιμή του χαρακτηριστικού μεγέθους firstAppearence που παρουσιάζεται.

- double maxFap: Η μέγιστη τιμή του χαρακτηριστικού μεγέθους term frequency που παρουσιάζεται.

- ArrayList phrases: Η λίστα με όλες τις φράσεις που περιλαμβάνονται στη MainList.

Τα παραπάνω χαρακτηριστικά δεν παρουσιάζουν κάποιο ιδιαίτερο ενδιαφέρον ως προς το ρόλο που επιτελούν καθώς είναι καθαρά διαδικαστικός. Αντίθετα, οι μέθοδοι της κλάσης αυτής, είναι μέθοδοι που ελέγχουν τη λίστα με τα δεδομένα που διοχετεύονται από το ένα στάδιο στο άλλο και επιτελούν όλες τις λειτουργίες που έχουν να κάνουν με τη διαχείριση των φράσεων (διαγραφή φράσης, εισαγωγή φράσης, αναδιάταξη της σειράς των φράσεων, διαδικασία αφαίρεσης διπλών φράσεων, κλπ). Ακόμη, στη συγκεκριμένη κλάση βρίσκονται οι μέθοδοι για την εκτύπωση διαφόρων χαρακτηριστικών των φράσεων, οι οποίες είναι από τις σημαντικότερες διεργασίες δεδομένης της δυνατότητας που δίνει στον άνθρωπο να έχει την εποπτεία των αποτελεσμάτων κάθε σταδίου του συστήματος αλλά και των τελικών αποτελεσμάτων που αποτελούν την τελική έξοδο του συστήματος προς το χρήστη (και είναι το ζητούμενο από την αρχή του σχεδιασμού του). Τέλος, στην ίδια κλάση βρίσκονται και οι μέθοδοι που αφορούν στη διόρθωση προβλημάτων ροής που ενδέχεται να εμφανιστούν όταν τα δεδομένα της MainList επηρεάζονται από μεθόδους εξωτερικής κλάσης. Έτσι, πέρα από τον διαδικαστικό χαρακτήρα των μεθόδων αυτών, εξασφαλίζεται και η δυνατότητα ορθής λειτουργίας του συστήματος ακόμα και σε περιπτώσεις προβληματικής επέκτασης ή συντήρησής του.

Δύο επιπλέον δευτερεύουσες δομές δεδομένων που χρησιμοποιούνται στο 4ο Στάδιο για την αποθήκευση των αποτελεσμάτων που παίρνουν από την DBpedia είναι στη γενική τους μορφή οι εξής:

- **VirtuosoUri**
  Είναι η κλάση τα στιμγιότυπα της οποίας αντιπροσωπεύουν ένα αποτέλεσμα (URI) που δόθηκε σαν απάντηση από το Virtuoso σε μία SPARQL ερώτηση. Για κάθε ένα από τα αποτελέσματα αυτά, αποθηκεύεται στη συγκεκριμένη δομή η συμβολοσειρά με το URI (uriString), ο αριθμός των εμφανίσεων του URI αυτού ως απάντηση σε SPARQL query (tf) και το σύνολο των ids των οντοτήτων που ανήκουν στην τρέχουσα μορφή του αντικειμένου της κλάσης MainList με τα οποία συνδέεται μέσω RDF τριάδων το URI αυτό (resourceSet). Είναι σημαντικό να τονιστεί πως πρόκειται για abstract κλάση, η οποία επεκτείνεται από κλάσεις που συγκεκριμενοποιούν τον τύπο του URI που αναζητείται. Κατά τις δοκιμαστικές εκτελέσεις του συστήματος, μέχρι να βρεθούν ικανοποιητικά αποτελέσματα, υλοποιήθηκαν αρκετές τέτοιες κλάσεις που επεκτείνουν την VirtuosoUri, για URIs που είχαν να κάνουν με τον τύπο των οντοτήτων, τις ιδιότητες των οντοτήτων κοκ. Τελικώς διατηρήθηκε μόνο δύο κλάσεις που χρησιμοποιούν και επεκτείνουν το την κλάση VirtuosoUri. Οι κλαεις αυτές είναι οι Category και Person.

- **VirtuosoList**
  Είναι η κλάση που αποτελεί τη δομή του συνόλου των αντικειμένων τύπου VirtuosoUri. Πρόκειται για μία κλάση που παρέχει μία λίστα τέτοιων αντικειμένων και μεθόδους για την διαχείριση της λίστας αυτής. Όπως συμβαίνει και με τις κλάσεις VirtuosoUri, έτσι και εδώ, η κλάση VirtuosoList είναι abstract και επεκτείνεται από κλάσεις που συγκεκριμενοποιούν τον τύπο των αντικειμένων που αποθηκεύουν στη λίστα τους. Τελικώς, δεδομένου πως αντικείμενα VirtuosoUri είναι μόνο τα αντικείμενα τύπου Category και Person, οι κλάσεις που επεκτείνουν την VirtuosoList είναι η CategoryList και η PersonList.

4.3 Περαιτέρω ανάλυση ζητημάτων υλοποίησης Συστήματος CRESTA

4.3.1 Θέματα υλοποίησης Γραμματικής Επισημείωσης

Το 1ο στάδιο όπως έχει ήδη αναφερθεί, απλώς καλεί τον TreeTagger ώστε να διενεργηθεί γραμματική επισημείωση σε όλες τις λέξεις του κειμένου που δίνεται ως είσοδος. Το μόνο που αξίζει να σημειωθεί στο σημείο αυτό είναι πως το κείμενο που δίνεται ως είσοδος πρέπει να είναι της μορφής txt. Επιπλέον, προκειμένου να υπάρχει η δυνατότητα να εκτελεστεί το πρόγραμμα για πάνω από ένα κείμενο, το κάθε κείμενο πρέπει να έχει το όνομα ”input<Integer>.txt”. Κατά την έναρξη λειτουργίας του συστήματος ζητείται από το χρήστη να προσδιορίσει τον ακέραιο που αντιστοιχεί στο πρώτο και στο τελευταίο κείμενο. Για παράδειγμα αν ο χρήστης εισάγει τους αριθμούς 10 12, θα εκτελεστεί τρεις φορές, μία για καθένα από τα κείμενα:
input10.txt, input11.txt, input12.txt.

4.3.2 Θέματα υλοποίησης Εξαγωγής Ονοματικών Οντοτήτων

Το στάδιο αυτό, τόσο για το σύστημα CRESTA όσο και για το σύστημα SWPID, υλοποιείται μέσω της κλάσης NounPhraseExtractor(), η οποία δέχεται ως είσοδο τον πίνακα της εξόδου του δεύτερου σταδίου, και στη συνέχεια καλώντας μεθόδους της MainList αλλά και άλλων κλάσεων, υλοποιεί τις διαδικασίες επεξεργασίας και κατασκευής ονοματικών φράσεων που περιγράφηκαν στην ενότητα 3.2.2. Στη συνέχεια παρουσιάζονται τα βήματα υλοποίησης μαζί με τις κλάσεις και τις μεθόδους που συμμετέχουν σε κάθε βήμα.

1. Κατάτμηση του κειμένου σε διευρυμένες πιθανές ονοματικές φράσεις. Για το σκοπό αυτό καλείται ο κατασκευαστής της κλάσης FirstProc, ο οποίος με τη σειρά του καλεί τη μέθοδο processLine(). Μέσω της μεθόδου check() που καλείται από την processLine για κάθε γραμμή του πίνακα διαδοχικά, υλοποιείται ο έλεγχος για την εφαρμογή των γραμματικών και συντακτικών κανόνων που παρατέθηκαν στην αντίστοιχη ενότητα του κεφαλαίου 3. Η check() είναι θεμελιώδης μέθοδος καθώς όταν ο έλεγχος το υποδεικνύει, δημιουργεί

τα στιγμιότυπα των ονοματικών φράσεων και τα προσθέτει στη MainList αλλά και των λέξεων (Word) και τα προσθέτει στην εκάστοτε υπό κατασκευή NounPhrase.

2. Διαδικασία βελτιστοποίησης φράσεων, η οποία περιλαμβάνει τις εξής διαδικασίες:

   - Αφαίρεση περιττών λέξεων από τις φράσεις της MainList αλλά και περιττών μονολεκτικών φράσεων από τη MainList, η οποία υλοποιείται από τη μέθοδο: FixPhrase()
   - Αφαίρεση διπλών εμφανίσεων των φράσεων και ενοποίηση των χαρακτηριστικών τους. Για το συγκεκριμένο βήμα γίνεται πρώτα αλφαβητική ταξινόμηση της MainList από τη μέθοδο getSorted() της κλάσης MergeSort και στη συνέχεια καλείται η μέθοδος removeDupl() η οποία είναι μέθοδος της κλάσης MainList και αφαιρεί τις πολλαπλές εμφανίσεις μίας φράσης κρατώντας αυτή η οποία εμφανίζεται πρώτη. Επίσης η ίδια μέθοδος ενημερώνει το μετρητή εμφανίσεων κάθε φράσης.
   - Αναζήτηση κάθε φράσης στη βάση της DBpedia μέσω της κλάσης GetResourceUri(), η οποία ενημερώνει τα χαρακτηριστικά Found και UriString καθε αντικειμένου τύπου Phrase που βρίσκεται στην τρέχουσα μορφή της MainList, αναλογα με το αν το SPARQL query επιστρέφει ή οχι το αναγνωριστικό URI που αντιστοιχεί στη φράση. Σημειώνεται ότι στην περίπτωση που επιστραφεί URI που είναι υποκείμενο σε τριάδα η οποία δηλώνει ανακατεύθυνση (redirect), τότε το αναγνωριστικό URI της οντότητας αντικαθίσταται με το αντικείμενο της τριάδας και σε κάθε άλλη αναζήτηση θα χρησιμοποιείται αυτό.
     Υπενθυμίζεται πως οι διαδικασίες του βήματος αυτού επαναλαμβάνονται αυτούσιες μετά από κάθε ένα από τα βήματα περαιτέρω επεξεργασίας που περιγράφοντα στη συνέχεια.

3. Κατάτμηση σύνθετων φράσεων εκεί που εντοπίζονται συνδετικές λέξεις. Αυτό το στάδιο μετα-επεξεργασίας υλοποιείται από τις μεθόδους της MainList:
   -breakPhrases()(Ελέγχει κάθε φράση με Found=0 για το κατά πόσο πρέπει να διασπαστεί σε μικρότερες φράσεις.)
   -breakThis() (Χρησιμοποιεί τη μέθοδο breakMe() της κλάσης Phrase ωστε να εντοπίσει τα σημεία στα οποία θα γίνει η κατάτμηση. Στη συνέχεια για κάθε τμήμα της κατατμημένης φράσης, δημιουργεί και προσθέτει στη λίστα ένα νέο NounPhrase στιγμιότυπο ενώ αφαιρεί το παλιό.)

4. Αφαίρεση (επιθετικών) προσδιορισμών από την αρχή NounPhrases που δεν έχουν ακόμα αναγνωριστεί ως οντότητες της DBpedia. Η επεξεργασία αυτή υλοποιείται μέσω των μεθόδων της MainList:
   -removeAdj()
   -removeThis()
   (Ελέγχουν κάθε φράση και σε περίπτωση που εντοπίσουν επιθετικό προσδιορισμό το αφαιρούν.)
   Σημειώνεται πως το βήμα αυτό καθώς και η διαδικασία βελτιστοποίησης που το ακολουθεί επαναλαμβάνονται έως ότου να μην υπάρχει μη-εντοπισμένη στη DBpedia φράση η οποία να έχει επιθετικούς προσδιορισμούς στην αρχή της.

5. Διαχωρισμός σε δύο λίστες. Η MainList που περιέχει το σύνολο των φράσεων, χωρίζεται σε δύο νέα στιγμιότυπα, από τα οποία το ένα περιέχει τις εντοπισμένες και το άλλο τις μη εντοπισμένες φράσεις. Η διαδικασία αυτή διεκπεραιώνεται από την κλάση SeparateLists.

6. Τελικός έλεγχος και κατάτμηση μη εντοπισμένων φράσεων με χρήση μεταβλητού μήκους παραθύρου. Η διαδικασία αυτή υλοποιείται από την κλάση FifthProc(). Η κλάση αυτή εξετάζει κατά πόσο τμήματα των φράσεων που δεν έχουν εντοπιστεί, υπάρχουν στην DBpedia. Την ευθύνη της διάσπασης σε τμήματα που αναζητούνται εκ νέου αναλαμβάνει η μέθοδος της FifthProc():
   breakPhrase()

4.3.3 Θέματα υλοποίησης Ποσοτικοποίησης Πληροφορίας από Wikipedia και κείμενο

Το τρίτο στάδιο, όπως περιγράφηκε στην αντίστοιχη ενότητα 3.2.3, υλοποιείται με τα εξής βήματα με τελικό στόχο τον υπολογισμό των χαρακτηριστικών της MainList totalHits, backLinks, firstAppearence, termFrequency, keyPhraseness και metricSum και την αποκοπή μέρος των οντοτήτων που ανήκουν σε αυτήν και θεωρούνται απορριπτέες:

1. Ανάκτηση ποσοτικοποιημένων πληροφοριών από κείμενο. Καλείται η κλάση OrderRank η οποία υπολογίζει και αποθηκεύει την τιμή του χαρακτηριστικού firstAppearence.

2. Ανάκτηση ποσοτικοποιημένων πληροφοριών από wikipedia. Καλούνται οι κλάσεις WikiAPI, WikiBackLinks οι οποίες ζητούν, ανακτούν και αποθηκεύουν από το mediawiki API τις τιμές των χαρακτηριστικών των αντικειμένων τύπου NounPhrase, totalHits και backLinks αντίστοιχα.

3. Υπολογισμός των χαρακτηριστικών term frequency και keyPhraseness. Καλούνται από την κλάση Start οι μέθοδοι της κλάσης MainList:

   -setTermFrequency()
   -setKeyPhraseness()

4. Κανονικοποίηση firstAppearence, termFrequency, keyPhraseness. Καλούνται από την κλάση Start οι μέθοδοι της κλάσης MainList:

   -minMax()(Υπολογίζει μέγιστα και ελάχιστα)
   -convertLengthEpiF()(Πρώτη εκδοχή χωρίς όρισμα/λογαρίθμιση)
   -convertKeyP()
   -convertFirstAp()
   -IFminusMin() (Αφαιρεί το μικρότερο αριθμό ώστε να μετατραπούν όλες οι τιμές σε θετικές)
   -keyPminusMin()
   -firstMinusMin()
   -averageLF() (Υπολογίζει το μέσο όρο του μεγέθους)

-averageFirstAp()
-averageKeyP()
-convertLengthEpiF(avLF) (Δεύτερη εκδοχή που δέχεται όρισμα το μέσο όρο-
Διαιρεί όλες τις τιμές με το μέσο όρο)
-convertFirstAp(avFAp)
-convertKeyP(avKeyP)

5. Υπολογισμός metricSum. Καλείται από την κλάση Start η μέθοδος της κλάσης MainList:
   makeSum()

6. Ταξινόμηση με βάση το metricSum. Καλείται η κλάση SortMetric η οποία ταξινομεί τα αντικείμενα NounPhrase που βρίσκονται στη λίστα της MainList με κριτήριο το metricSum και χρησιμοποιεί για την ταξινόμηση αυτή τον αλγόριθμό mergesort.

7. Αποκοπή του 50% των αποτελεσμάτων. Καλείται από την κλάση Start η μέθοδος της κλάσεις της MainList:
   keepTheMostImportant()

Ας σημειωθεί για λόγους κατανόησης των δεδομένων που διοχετεύονται στο επόμενο στάδιο, πως η τελευατία μέθοδος επιστρέφει δύο καινούρια αντικείμενα MainList. Το πρώτο είναι η MainList που περιλαμβάνει ταξινομημένες τις οντότητες που βρισκόντουσαν στις πρώτες 50% των θέσεων της μέχρι τώρα κατάταξής τους, ενώ το δεύτερο είναι η MainList που περιλαμβάνει ταξινομημένες τις υπόλοιπες οντότητες, οι οποίες και έχουν απορριφθεί.

4.3.4 Θέματα υλοποίησης Ποσοτικοποίησης Πληροφορίας από DBpedia

Με τα δεδομένα που περιγράφηκαν παραπάνω και θεωρούνται ως είσοδος για το 4ο στάδιο, η υλοποίηση της διαδικασίας που περιγράφηκε στην ενότητα 3.2.4 για το στόχο αυτό περιλαμβάνει τα εξής βήματα και υποβήματα:

1. Υπολογισμός του κριτηρίου αξιολόγησης των οντοτήτων category.

   - Αναζήτηση όλων των κατηγοριών στις οποίες ανήκουν οι οντότητες που βρίσκονται στη MainList με τις επιλεγμένες οντότητες. Η αναζήτηση αυτή γίνεται μέσω της κλάσης GetVirtUri και της υποκλάσης της GetCategory.
   - Δημιουργία αντικειμένου Category για κάθε μία από τις κατηγορίες που βρέθηκαν και αποθήκευσή της στην CategoryList με χρήση της μεθόδου της CategoryList:
     addVirtuosoUri()
   - Προσαρμογή των χαρακτηριστικών του αντικειμένου Category που συναντάται για δεύτερη ή πλέον φορά με χρήση της μεθόδου της CategoryList:
     addExistingVirt()
   - Γίνεται αναζήτηση των broader και related με τα οποία σχετίζονται τα ήδη υπάρχοντα categories. Για κάθε URI που δίνεται σαν αποτέλεσμα της αναζήτησης αυτής,

επαναλαμβάνονται τα δύο τελευταία υποβήματα. Η αναζήτηση γίνεται με χρήση των κλάσεων GetBroader και GetRelated αντίστοιχα που είναι υποκλάσεις της GetVirtUri.

- ”Φιλτράρισμα” της CategoryList που έχει προκύψει (συμπεριλαμβανομένων των αποτελεσμάτων broader και related) ώστε να κρατηθούν στη λίστα μόνο εκείνα τα αντικείμενα τύπου Category που έχουν χαρακτηριστικό tf>=2. Τη διαδικασία αυτή αναλαμβάνει η μέθοδος της CategoryList:
  filterVirt()
- Αντίστροφη διαδικασία κατά την οποία κάθε οντότητα αποθηκευμένη στη MainList με τις επιλεγμένες οντότητες αποθηκεύει στο χαρακτηριστικό της category τον αριθμό των αντικειμένων της CategoryList με τα οποία συνδέεται. Αυτή η σύνδεση επιβεβαιώνεται μέσω του χαρακτηριστικού της Category resourceSet όπως έχει περιγραφεί παραπάνω. Υπεύθυνη για τη διαδικασία αυτή είναι η μέθοδος:
  uriCount() (Δέχεται σαν όρισμα τη mainList και είναι υπεύθυνη για την αποθήκευση σε όλα τα αντικείμενα NounPhrase της λίστας της της αντίστοιχης τιμής του χαρακτηριστικού category)
- Κανονικοποίηση του χαρακτηριστικού μεγέθους category, σύμφωνα με τις μεθόδους της κλάσης MainList:

  -minCateg() (Υπολογίζει την ελάχιστη τιμή)
  -convertCateg() (Λογαρίθμηση)
  -categMinusMin()(Αφαιρείται από όλες τις τιμές η ελάχιστη ώστε να είναι θετικά τα μεγέθη)
  -averageCat() (Υπολογισμός του μέσου όρου)
  -convertCateg() (Δεύτερη εκδοχή με όρισμα το μέσο όρο των τιμών-
  Διαίρεση όλων των τιμών με το μέσο όρο)

2. Υπολογισμός του κριτηρίου αξιολόγησης των οντοτήτων abstract.

   - Για κάθε αντικείμενο τύπου NounPhrase που βρίσκεται αποθηκευμένο στη MainList με τις επιλεγμένες οντότητες αναζητείται το αντίστοιχο abstract URI. Τη διαδικασία της ερώτησης αναλαμβάνει η κλάση ProcessAbstracts.
   - Κάθε abstract αποθηκεύεται σε ένα txt αρχείο στο οποίο διενεργείται γραμματική επισημείωση και Entity Extraction (1ο και 2ο στάδιο του συστήματος). Τις διαδικασίες αυτές τις εκκινεί επίσης η κλάση ProcessAbstracts.
   - Όλες οι οντότητες που προέρχονται από κείμενα abstracts μετατρέπονται σε Treeset μέσω της μεθόδου της MainList:
     convertToTreeSet()

   - Κάθε ένα από τα παραπάνω TreeSets αποθηκεύονται σε μία δομή τύπου ArrayList <TreeSet>. Κάθε στοιχείο της ArrayList αυτής αντιστοιχεί στο TreeSet μίας εκ των υποψηφίων οντοτήτων-στόχων. Επαναλαμβάνεται ότι κάθε TreeSet περιλαμβάνει αντικείμενα τύπου NounPhrase που εξήχθησαν από ένα abstract. Το υποβήμα αυτό εκτελείται από την κλάση ProcessAbstracts και μάλιστα η ArrayList που διατηρεί τα TreeSets αποτελεί χαρακτηριστικό της κλάσης αυτής (abstractsList).

- Για κάθε ένα από τα TreeSets ελέγχεται πόσα κοινά NounPhrases περιλαμβάνει με τα NounPhrases των δύο MainList που δοθήκανε σαν είσοδος στο στάδιο αυτό. Πρακτικά, αυτό σημαίνει πως υπολογίζεται πόσες κοινές οντότητες βρέθηκαν στο abstract κάθε οντότητας και στο αρχικό υπό-εξέταση κείμενο. Ο κάθε αριθμός των κοινών στοιχείων, αποθηκεύεται σε μία θέση μίας νέας ArrayList που αποτελεί και χαρακτηριστικό της ProcessAbstracts (foundOthers). Κάθε μία από τις θέσεις της ArrayList αντιστοιχεί σε ένα από τα NounPhrases της MainList με τις επιλεγμένες οντότητες στόχους. Το υποβήμα αυτό του ελέγχου και της αποθήκευσης εκτελείται ολόκληρο από τη μέθοδο της ProcessAbstracts:
  findCommons()

- Τα στοιχεία της λίστας foundOthers υπόκεινται σε κανονικοποίηση με χρήση της μεθόδου της κλάσης ProcessAbstracts:
  normalize()

3. Υπολογισμός του γραμμικού συνδυασμού των εξής μεγεθών: metricSum (υπολογισμένο από το προηγούμενο στάδιο), category (υπολογισμένο στο πρώτο βήμα του παρόντος σταδίου) και abstract (υπολογισμένο στο δεύτερο βήμα του παρόντος σταδίου). Ο γραμμικός συνδυασμός αυτός αποθηκεύεται και πάλι στο χαρακτηριστικό metricSum των NounPhrases.

4.3.5 Θέματα υλοποίησης Σύνδεσης με Εξωτερικές Πηγές

Για το στάδιο αυτό, υλοποιούνται τα παρακάτω βήματα:

1. Ζητούνται από το Virtuoso Server οι πληροφορίες της μορφής wikiLinks για κάθε NounPhrase που βρίσκεται στη MainList που περιλαμβάνει τις επιλεγμένες οντότητες. Οι απαντήσεις αποθηκεύονται στο αντίστοιχο χαρακτηριστικό κάθε NounPhrase. Τη διαδικασία αυτή υλοποιεί η κλάση GetWikiLinks που είναι υποκλάση της κλάσης GetVirtUri.
2. Ζητούνται από το Virtuoso Server οι πληροφορίες της μορφής external links για κάθε NounPhrase που βρίσκεται στη MainList που περιλαμβάνει τις επιλεγμένες οντότητες. Οι απαντήσεις αποθηκεύονται στο αντίστοιχο χαρακτηριστικό (που ειναι τύπου ArrayList) κάθε NounPhrase. Τη διαδικασία αυτή υλοποιεί η κλάση GetExtLinks που είναι υποκλάση της κλάσης GetVirtUri.

4.4 Υπολογισμός Κανονικοποιήσεων και Βαρών γραμμικών συνδυασμών Χαρακτηριστικών μεγεθών

Η ενότητα αυτή, θα ασχοληθεί με στατιστικά ζητήματα των χαρακτηριστικών μεγεθών αξιολόγησης που χρησιμοποιούνται. Τα ζητήματα αυτά αφορούν στην κανονικοποίηση των μεγεθών

αυτών ώστε να είναι συγκρίσιμα και στην επιλογή βαρών για το γραμμικό συνδυασμό τους ώστε να βρεθεί το τελικό κριτήριο αξιολόγησης που θα είναι και το κλειδί της τελικής ταξινόμησης των στοιχείων της MainList.

4.4.1 Κανονικοποιήσεις Μεγεθών

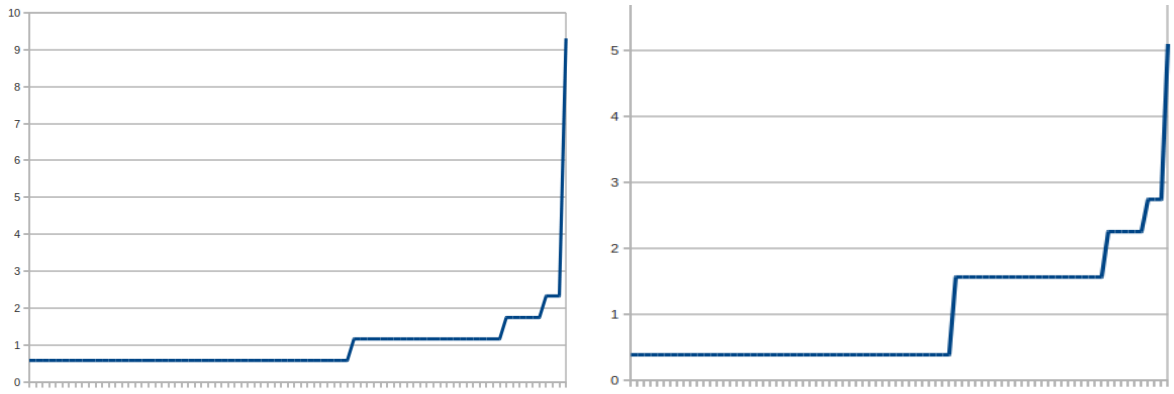
Τα χαρακτηριστικά μεγέθη όπως υπολογίζονται πριν την κανονικοποίησή τους έχουν εντελώς διαφορετική μορφή μεταξύ τους. Για το λόγο αυτό, προκειμένου να είναι εφικτή η σύγκρισή τους μέσω γραμμικού συνδυασμού, απαιτήθηκε η κανονικοποίησή τους. Επιλέχθηκε η κανονικοποίηση αυτή να γίνει μεταφέροντας τις τιμές των μεγεθών γύρω από τη μονάδα (να έχουν μέσο όρο ίσο με 1). Πρακτικά αυτό επιτυγχάνεται με τον εξής τρόπο:
$k_i = \frac{nonk_i}{avevage}$
όπου k_i η i-οστή κανονικοποιημένη τιμή του μεγέθους, $nonk_i$ η i-οστή μη κανονικοποιημένη τιμή του μεγέθους και average ο μέσος όρος όλων των μη κανονικοποιημένων τιμών του μεγέθους.

Αν και αναμενόταν η κανονικοποίηση αυτή να είναι αρκετή ώστε να επιτρέψει τον υπολογισμό του γραμμικού συνδυασμού, εμφανίστηκε το εξής πρόβλημα: Σε πρώτες πειραματικές μετρήσεις που εκτελέστηκαν εμφανίστηκαν τα μεγέθη να έχουν τιμές με λογαριθμική κατανομή. Υπήρχαν, δηλαδή, κάποιες πολύ μεγάλες τιμές, οι οποίες ανέβαζαν πολύ το μέσο όρο με αποτέλεσμα κατά την κανονικοποίηση να ελαττώνονται σε οριακά μηδενικές τιμές οι διαφορές των υπόλοιπων τιμών. Επιλέχθηκε λοιπόν, για όλα τα μεγέθη, πριν γίνει η κανονικοποίηση γύρω από τη μονάδα, να υποστούν λογαρίθμηση. Στο σημείο αυτό υπήρχε ακόμη ένα πρόβλημα, καθώς σχεδόν σε όλα τα μεγέθη, υπήρχαν και κάποιες μηδενικές τιμές, οι οποίες δεν μπορούσαν να υποστούν λογαρίθμηση. Οι τιμές αυτές εξισώθηκαν με την ελάχιστη μη μηδενική τιμή του κάθε μεγέθους. Εδώ πρέπει να σημειωθεί πως κατά τη λογαρίθμηση, ένα μεγάλο ποσοστό των τιμών των μεγεθών προέκυψε αρνητικό, με αποτέλεσμα να γίνει μετάθεση των τιμών πριν τον υπολογισμό των μέσων όρων και την κανονικοποίηση. Στα σχήματα 4.1, 4.2, 4.3, παρουσιάζονται έξι διαγράμματα. Τα τρία από αυτά είναι οι τιμές τριών μεγεθών κανονικοποιημένων χωρίς λογαρίθμηση. Οι υπόλοιπες είναι οι τιμές των ίδιων μεγεθών κανονικοποιημένων με λογαρίθμηση.

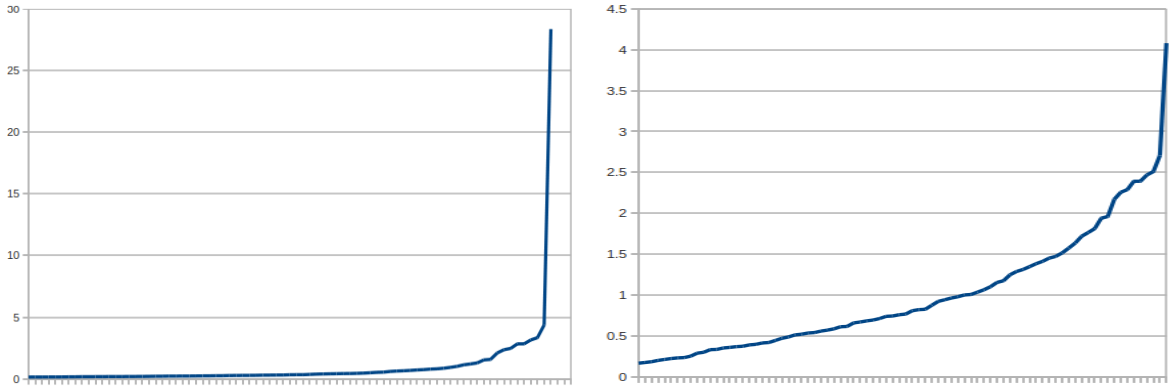
4.4.2 Υπολογισμός Βαρών γραμμικών συνδυασμών

Για το σύστημα, υπολογίζεται δύο φορές ο γραμμικός συνδυασμός μεγεθών.

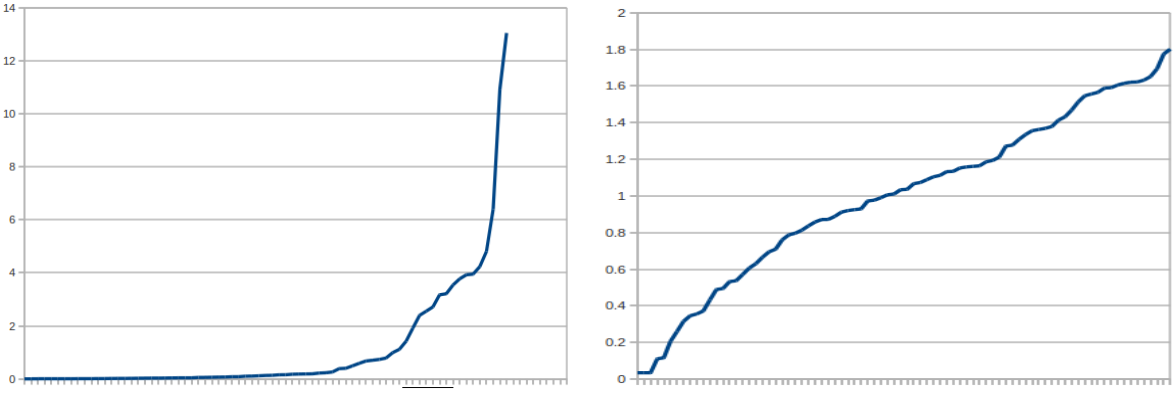
- Γραμμικός συνδυασμός των term frequency, keyPhraseness, first appearence:
  $metricSum = tfW * termFrequency + kpW * keyPhraseness + faW * firstAppearence$
  όπου $tfW + kpW + faW = 1$
- Γραμμικός συνδυασμός των metricSum, abstract, category:
  $totalSum = msW * meticSum + abW * abstract + catW * category$
  όπου $msW + abW + catW = 1$



Σχήμα 4.1: Διάγραμμα τιμών term frequency. Αριστερά: Χωρίς λογαρίθμηση. Δεξιά: Με λογαρίθμηση



Σχήμα 4.2: Διάγραμμα τιμών first appearence. Αριστερά: Χωρίς λογαρίθμηση. Δεξιά: Με λογαρίθμηση



Σχήμα 4.3: Διάγραμμα τιμών keyPhraseness. Αριστερά: Χωρίς λογαρίθμηση. Δεξιά: Με λογαρίθμηση

Υπάρχουν, λοιπόν, έξι τιμές βαρών που πρέπει να υπολογιστούν. Ο υπολογισμός έγινε πειραματικά σε δύο στάδια. Στο πρώτο στάδιο, χρησιμοποιήθηκαν featured articles της wikipedia.

Πρόκειται για άρθρα της wikipedia των οποίων τα links έχουν σημειωθεί από άνθρωπο ώστε να είναι σχετικά με το θέμα του κειμένου. Η διαδικασία που ακολουθήθηκε είναι η εξής:
Με δείγμα είκοσι τέτοιων κειμένων, υπολογίστηκε ο γραμμικός συνδυασμός των χαρακτηριστικών μεγεθών term frequency, keyPhraseness, first appearence με όλα τα βάρη να παίρνουν τιμές στο διάστημα [0,1] με βήμα 0.2.Προφανώς τηρήθηκε η απαίτηση:
$metricSum = tfW * termFrequency + kpW * keyPhraseness + faW * firstAppearence$

Για τα αποτελέσματα που προέκυψαν υπολογίστηκαν οι δείκτες ακρίβεια (precision) και ανάκληση (recall) ως εξής:

- Θεωρούμε σωστές φράσεις-στόχους τις φράσεις εκείνες που είναι links σε κάθε άρθρο.

- Θεωρούμε λάθος φράσεις-στόχους τις φράσεις εκείνες που υπάρχουν σε ένα άρθρο αλλά δεν είναι links.

$$precision = \frac{correct\_phrases\_found}{total\_phrases\_found}$$

$$recall = \frac{correct\_phrases\_found}{total\_correct\_phrases}$$

Μετά από τον υπολογισμό αυτό για όλα τα κείμενα, υπολογίστηκαν οι μέσοι όροι precision και recall για κάθε τριάδα βαρών, και υπολογίστηκε το fmeasure της κάθε τριάδας βαρών ως εξής:

$$fmeasure = \frac{2*precision*recall}{precision+recall}$$

Τα αποτελέσματα φαίνονται στον πίνακα 4.1.

Συνεπώς, φαίνεται πως οι καλύτερες τιμές fmeasure βρίσκονται για το βάρος του keyPhraseness στο διάστημα [0.2, 0.6] και το βάρος του first appearence στο διάστημα [0.4, 0.8]. Έτσι υπολογίστηκε ξανά ο γραμμικός συνδυασμός των τριών μεγεθών, αυτή τη φορά για βάρη: keyPhraseness στο διάστημα [0.2, 0.45] και term frequency στο διάστημα [0.3, 0.75] με βήμα 0.05.
Τα αποτελέσματα εμφανίζονται στον πίνακα 4.2.

Τα βάρη που βρίσκονται στην πρώτη γραμμή του σχήματος 9.5 είναι και η τελική επιλογή.

Στο δεύτερο στάδιο υπολογισμού βαρών, εφαρμόζεται ένα διαφορετικό κριτήριο επιλογής και χρησιμοποιείται διαφορετικό σύνολο δεδομένων. Το κριτήριο έχει να κάνει με τη βελτίωση της τελικής κατάταξης των αποτελεσμάτων που τελικώς διατηρήθηκαν ως τα επιθυμητά αποτελέσματα. Η διαδικασία που ακολουθήθηκε είναι η εξής:

Πίνακας 4.1: Πειραματικές τιμές fmeasure για διάφορους συνδυασμούς βαρών

kfW	fapW	tfW	fmeasure
0.4	0	0.6	0.403
0.4	0.2	0.4	0.396
0.2	0.2	0.6	0.395
0.2	0	0.8	0.394
0.6	0	0.4	0.389
0	0.2	0.8	0.385
0.8	0	0.2	0.378
0	0.4	0.6	0.375
0.2	0.4	0.4	0.371
0.6	0.2	0.2	0.365
0.4	0.4	0.2	0.360
0.8	0.2	0	0.357
1	0	0	0.3568
0.6	0.4	0	0.3453
0	0	1	0.339
0	1	0	0.339
0	0.6	0.4	0.329
0.4	0.6	0	0.322
0.2	0.6	0.2	0.320
0	0.8	0.2	0.280
0.2	0.8	0	0.277

Κάθε ένα από τα κείμενα του συνόλου δεδομένων, δόθηκε ως είσοδος στο σύστημα για διαφορετικές τιμές των βαρών του γραμμικού συνδυασμού:
$finalMetricSum = msW * metricSum + catW * category + abW * abstract$

Κάθε ένα από τα τρία παραπάνω βάρη, πήρε τιμές στο διάστημα [0,1] με βήμα 0.2. Τα αποτελέσματα που δόθηκαν ως έξοδος αξιολογήθηκαν ως εξής:

1. Για κάθε μία από τις ν φράσεις της λίστας με τα επιλεγμένα keyphrases δόθηκε τυχαία ένας μοναδικός ακέραιος κ ώστε 1<κ<ν. Το άθροισμα των αριθμών αυτών που αποτελεί και το άθροισμα των ν πρώτων όρων της αριθμητικής προόδου:
   $a_n=a_{n-1}+1$, $a_1=1$ν Θα είναι το μέτρο σύγκρισης για την αξιολόγηση των αποτελεσμάτων και θα ονομάζεται στο εξής άθροισα σωστής κατάταξης.

2. Για κάθε λίστα που δίνεται ως έξοδος σε κάθε εκτέλεση του προγράμματος σε κάθε μία από τις μ οντότητες-στόχους δίνεται ένας μοναδικός ακέραιος λ, 1<λ<μ ξεκινώντας από την αρχή της λίστας και μέχρι το τέλος. Πρακτικά, όσο ψηλότερα στην κατάταξη της λίστας βρίσκεται μία οντότητα, τόσο μικρότερο αριθμό θα έχει.

Πίνακας 4.2: Πειραματικές τιμές fmeasure για τιμές βαρών στη βέλτιστη περιοχή

kpW	faW	tfW	fmeasure
0.4	0.1	0.5	0.418
0.45	0.05	0.5	0.417
0.4	0.05	0.55	0.417
0.25	0.05	0.7	0.415
0.3	0.1	0.6	0.415
0.35	0.05	0.6	0.415
0.3	0.15	0.55	0.415
0.3	0.05	0.65	0.414
0.25	0.1	0.65	0.414
0.45	0.1	0.45	0.414
0.4	0.15	0.45	0.413
0.2	0.05	0.75	0.413
0.2	0.1	0.7	0.4130
0.2	0.15	0.65	0.412
0.25	0.2	0.55	0.4112
0.25	0.15	0.6	0.4109
0.2	0.2	0.6	0.410
0.3	0.2	0.5	0.409
0.35	0.2	0.45	0.408
0.25	0.25	0.5	0.408
0.45	0.15	0.4	0.408
0.2	0.25	0.55	0.407
0.4	0.2	0.4	0.407
0.3	0.25	0.45	0.406
0.35	0.25	0.4	0.403
0.45	0.2	0.35	0.398
0.4	0.25	0.35	0.398
0.35	0.15	0.5	0.391
0.35	0.1	0.55	0.385
0.45	0.25	0.3	0.374

3. Υπολογίζεται το άθροισμα των ακεραίων που δόθηκαν σε οντότητες που υπάρχουν ως keyphrases στο επισημειωμένο dataset, το οποίο θα ονομάζεται πειραματικό άθροισμα. Στην ιδανική περίπτωση, το άθροισμα αυτό είναι ίσο με το άθροισμα σωστής κατάταξης.

4. Υπολογίζεται η τιμή του μεγέθους:
   $ranking\_proximity = target\_sum - experimental_sum$
   Από τον παραπάνω τύπο, είναι εμφανές ότι στην ιδανική περίπτωση
   $ranking\_proximity \to 0$

Από τα πειραματικά αποτελέσματα που προέκυψαν κατά τη διαδικασία αυτή, φαίνεται πως η μικρότερη απόκλιση σε σχέση με την επιθυμητή κατάταξη παρουσιάζεται για τις τιμές των βαρών στα διαστήματα:

$msW \in [0.6, 0.85]$

$catW \in [0.05, 0.3]$

Προφανώς το τρίτο βάρος υπολογίζεται από τα προηγούμενα δύο καθώς το άθροισμα των τριών πρέπει να είναι ίσο με τη μονάδα. Η διαδικασία των παραπάνω τριών βημάτων επαναλαμβάνεται για βάρη στα προαναφερθέντα διαστήματα, αυτή τη φορά με βήμα 0.05. Τα αποτελέσματα που προέκυψαν έδειξαν ότι η καλύτερη δυνατή κατάταξη επιτυγχάνεται για βάρη:

$msW = 0.6$

$catW = 0.15$

$abW = 0.2$

και φαίνονται στον πίνακα 4.3. Ο πίνακας αυτός, παρουσιάζει τους κανονικοποιημένους μέσους όρους του μεγέθους ranking_proximity όπως ορίστηκε παραπάνω για κάθε τριάδα βαρών που εξετάστηκε.

Πίνακας 4.3: Βάρη τελικού γραμμικού συνδυασμού

msW	catW	abW	ranking proximity
0.65	0.15	0.2	1.103
0.7	0.15	0.15	1.084
0.75	0.1	0.15	1.073
0.65	0.1	0.25	1.030
0.6	0.25	0.15	1.016
0.7	0.25	0.05	1.001
0.75	0.15	0.1	1.001
0.8	0.05	0.15	0.998
0.85	0.05	0.1	0.998
0.8	0.1	0.1	0.997
0.75	0.05	0.1	0.994
0.6	0.3	0.1	0.993
0.7	0.05	0.25	0.992
0.6	0.2	0.2	0.991
0.8	0.15	0.05	0.982
0.65	0.25	0.1	0.975
0.6	0.15	0.25	0.973
0.6	0.1	0.3	0.972
0.65	0.2	0.15	0.972
0.75	0.2	0.05	0.971
0.7	0.2	0.1	0.971
0.7	0.1	0.2	0.954
0.65	0.3	0.05	0.951

Η αξιολόγηση του συστήματος θα γίνει με βάση τα βάρη αυτά τόσο για τον πρώτο, όσο και για τον δεύτερο γραμμικό συνδυασμό.

4.5 Περαιτέρω ζητήματα υλοποίησης Συστήματος SWPID

Όπως έχει προαναφερθεί, η υλοποίηση του δεύτερου συστήματος παρουσιάζει πολλά κοινά σημεία με την υλοποίηση του πρώτου, κυρίως στα δύο πρώτα στάδια. Για το λόγο αυτό, σε αυτή την ενότητα θα γίνει αναφορά μόνο στις λεπτομέρειες της υλοποίησης που συνιστούν ουσιαστικές διαφορές μεταξύ των δύο συστημάτων. Τα τμήματα της υλοποίησης των διαδικασιών του δεύτερου συστήματος που παρουσιάστηκαν στο κεφάλαιο 8 αλλά δεν αναφέρονται στην ενότητα αυτή είναι ταυτόσημα με τα αντίστοιχα τμήματα υλοποίησης του πρώτου συστήματος.

Σε αντιστοιχία με το πρώτο σύστημα ο έλεγχος της επεξεργασίας και ροής δεδομένων του δεύτερου συστήματος γίνεται από μία κεντρική κλάση, την StartPersons. Η κλάση αυτή ουσιαστικά καλεί τις μεθόδους και τους κατασκευαστές άλλων κλάσεων που υλοποιούν κάθε στάδιο επεξεργασίας του συστήματος.

Όσον αφορά στα στάδια εντοπισμού υποψήφιων φράσεων στόχων, η μόνη διαφορά στην υλοποίηση του δεύτερου συστήματος εντοπίζεται στην τελική διαδικασία του δεύτερου συστήματος, δηλαδή τη διαδικασία κατάτμησης μη εντοπισμένων φράσεων με χρήση μεταβλητού μήκους παραθύρου. Συγκεκριμένα, στο σύστημα αναγνώρισης ατόμων χρησιμοποιήθηκε διαφορετική κλάση, με την ονομασία PFifthProc(), η οποία καλείται από την StartPersons. Η κλάση αυτή εκτελεί την ίδια λειτουργία με τη FifthProc του πρώτου συστήματος, με τη μόνη διαφορά να εντοπίζεται στον κώδικα υλοποίησης της μεθόδου breakPhrase() η οποία εμπεριέχει τον εξής επιπλέον κανόνα ελέγχου:

```
for (int k=j; k<j+window-1; k++){
    if (!(helpPh.getWord(k).pos.equals(”NP”) || helpPh.getWord(k).pos.equals(”NPS”))){
      np=false;
    }
}
```

Με τον κανόνα αυτό εντοπίζει μόνο τις φράσεις που εμπεριέχουν κύρια ονόματα και στη συνέχεια ελέγχει μόνο αυτές ως προς την ύπαρξη τους στη DBpedia, σε αντίθεση με την αντίστοιχη διαδικασία του συστήματος CRESTA το οποίο ελέγχει όλες τις φράσεις.

Το τρίτο (και τελευταίο γι αυτό το σύστημα) στάδιο, είναι αυτό στο οποίο τα δύο συστήματα διαφοροποιούνται ριζικά ως προς τη λειτουργία τους και την έξοδό τους στο χρήστη. Το στάδιο αυτό υλοποιείται ουσιαστικά από τις κλάσεις GetPerson, Person και PersonList οι οποίες είναι υπο-κλάσεις των αφηρημένων κλάσεων GetVirtUri, VirtuosoUri και VirtuosoList αντίστοιχα. Ο κατασκευαστής της GetPerson καλείται από την StartPersons και ελέγχει κάθε μία από τις εντοπισμένες στη DBpedia οντότητες ώστε να διαπιστωθεί αν ανήκουν σε τριάδες που δηλώνουν την ιδιότητα person (όπως αυτές περιγράφηκαν στο κεφάλαιο 8).

Για κάθε οντότητα που αναγνωρίζεται ως υπαρκτό πρόσωπο δημιουργείται ένα νέο στιγμιότυπο της κλάσης Person (υπο-κλάση της VirtuosoUri) το οποίο εντάσσεται στο στιγμιότυπο της PersonList που έχει δημιουργηθεί. Στη διαδικασία αυτή συμμετέχουν οι εξής μέθοδοι:

- StoreQueryResult() της κλάσης GetPerson
- addPerson() της κλάσης PersonList

Η τελευταία διαδικασία του σταδίου αυτού αφορά στην έξοδο στο χρήστη και υλοποιείται μέσω της κλήσης της μεθόδου PrintPersons() η οποία ανήκει στην κλάση PersonList. Μέσω της μεθόδου αυτής τυπώνεται σε αρχείο txt λίστα η οποία περιέχει το ονοματεπώνυμο του ατόμου έτσι όπως αναφέρεται στο κείμενο και το resource uri που του αντιστοιχεί στη DBpedia.

Κεφάλαιο 5

Αξιολόγηση Συστημάτων

Στο κεφάλαιο αυτό θα παρουσιαστούν τα αποτελέσματα από την εκτέλεση των δύο συστημάτων και θα αξιολογηθούν ποσοτικά και ποιοτικά. Συγκεκριμένα, η παρουσίαση αυτή θα γίνει σε δύο ξεχωριστές ενότητες, μία για κάθε σύστημα. Κάθε ενότητα, θα έχει την εξής δομή: Θα παρουσιαστούν με τη σειρά η μέθοδος αξιολόγησης, τα αποτελέσματα, εκτενής σχολιασμός τους και προτάσεις για βελτίωση και επέκταση που προκύπτουν συμπερασματικά ερμηνεύοντας τα αποτελέσματα.

5.1 Σύστημα CRESTA

Το σύστημα CRESTA, αξιολογήθηκε με χρήση κειμένων των συνόλων δεδομένων με βάση τα οποία έγινε ο υπολογισμός των βαρών των γραμμικών συνδυασμών. Ωστόσο, χρησιμοποιήθηκε διαφορετικό τμήμα των συνόλων δεδομένων αυτών. Η αξιολόγηση των κειμένων, γίνεται μέσω των μεγεθών ακρίβεια και ανάκληση, όπως αυτά ορίστηκαν στο κεφάλαιο 3. Πιο συγκεκριμένα, η διαδικασία υπολογισμού των μεγεθών αξιολόγησης συνοψίζεται στα παρακάτω βήματα:

1. Εκτέλεση προγράμματος με είσοδο τυχαίως επιλεγμένα κείμενα του εκάστοτε συνόλου δεδομένων (Nguyen2007 και Wikipedia).

2. Σύγκριση οντοτήτων που αποτελούν έξοδο του προγράμματος και προσημειωμένων φράσεων κάθε κειμένου.

3. Ποσοτικοποίηση του αποτελέσματος της σύγκρισης μέσω υπολογισμού precision και recall. Ο υπολογισμός των δύο αυτών μεγεθών γίνεται για κάθε κείμενο χωριστά. Επιπλέον υπολογίζεται πολλές φορές για κάθε κείμενο, για διαφορετικό ποσοστό επί των αποτελεσμάτων κάθε φορά. Συγκεκριμένα, υπολογίζονται precision και recall για τα ποσοστά:
2%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%.

4. Υπολογισμός μέσου όρου των τιμών precision και recall για κάθε ποσοστό ξεχωριστά.

5. Κατασκευή διαγράμματος για τα ζεύγη (precision, recall) κάθε ποσοστού.

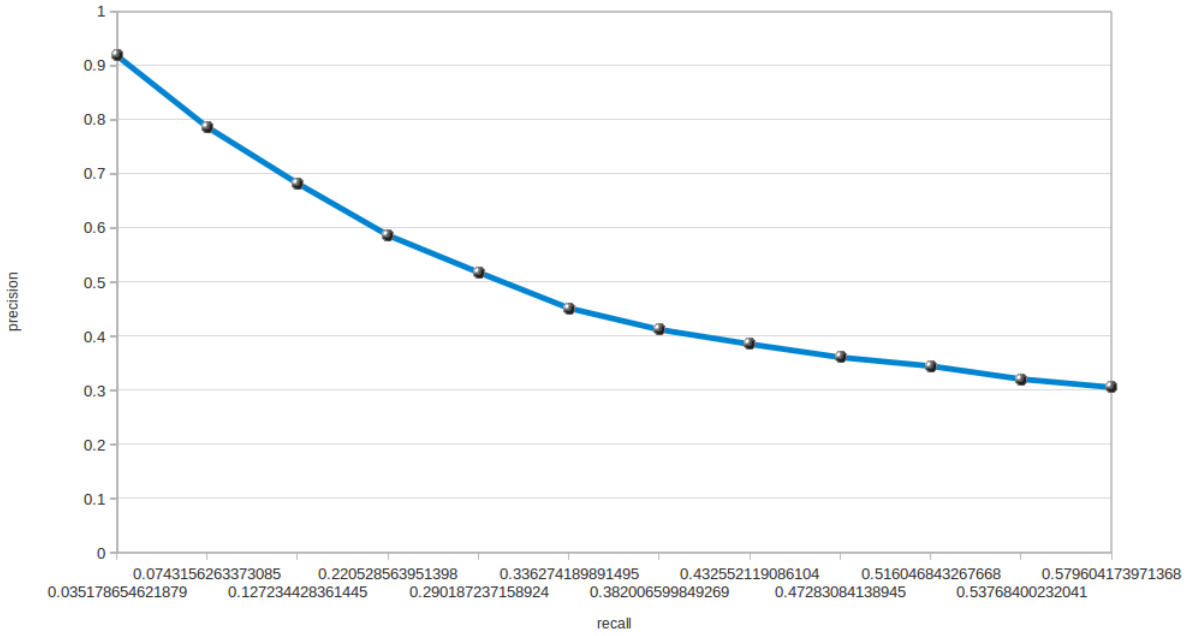
Στις δύο υποενότητες που ακολουθούν, παρουσιάζονται τα αποτελέσματα των βημάτων 4 και 5 που αφορούν συγκεντρωτικά αποτελέσματα όλων των κειμένων του κάθε συνόλου δεδομένων.

5.1.1 Αξιολόγηση με χρήση κειμένων της Wikipedia

Για τα κείμενα του συνόλου δεδομένων αυτού όταν δίνονται ως είσοδοι στο σύστημα, παράγονται ως έξοδοι κατά μέσο όρο 360.8 ονοματικές οντότητες και συγκρίνονται με κατά μέσο όρο 211.5 προσημειωμένους συνδέσμους ανά κείμενο. Παρακάτω θα παρουσιαστεί ο πίνακας με τα συγκεντρωτικά αποτελέσματα των μέσων όρων των τιμών precision και recall για τα είκοσι κείμενα που εξετάστηκαν καθώς και το αντίστοιχο διάγραμμα.

Πίνακας 5.1: Precision & Recall για το σύνολο δεδομένων των featured articles της DBpedia

percentage	precision	recall
2%	0.918596681096681	0.035178654621879
5%	0.786556384261936	0.0743156263373085
10%	0.682326989438402	0.127234428361445
20%	0.586490881624642	0.220528563951398
30%	0.51805542360722	0.290187237158924
40%	0.451935747871328	0.336274189891495
50%	0.412350852297142	0.382006599849269
60%	0.385534736850841	0.432552119086104
70%	0.361108149054015	0.47283084138945
80%	0.345014324876288	0.516046843267668
90%	0.320857498941784	0.53768400232041
100%	0.305698782498506	0.579604173971368



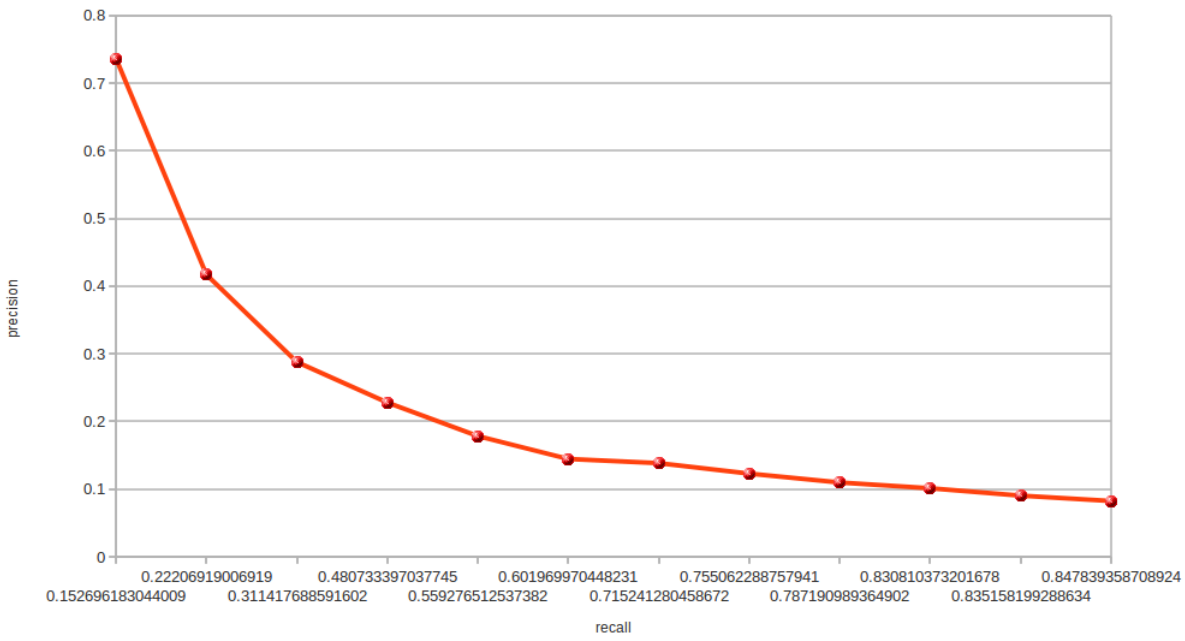
Σχήμα 5.1: precision-recall για το σύνολο δεδομένων των featured articles της Wikipedia

5.1.2 Αξιολόγηση με χρήση κειμένων του συνόλου δεδομένων Nguyen2007

Για τα κείμενα του συνόλου δεδομένων αυτού, το σύστημα παράγει ως έξοδο κατά μέσο όρο 251.4 ονοματικές οντότητες οι οποίες συγκρίνονται με κατά μέσο όρο 21.3 προσημειωμένες φράσεις. Οι συγκεντρωτικές τιμές precision και recall για κάθε ποσοστό επί του συνόλου των αποτελεσμάτων κάθε κειμένου φαίνονται στον παρακάτω πίνακα. Επιπλέον έχει κατασκευαστεί το διάγραμμα που απεικονίζει τη σχέση precision-recall για τις προαναφερθείσες τιμές.

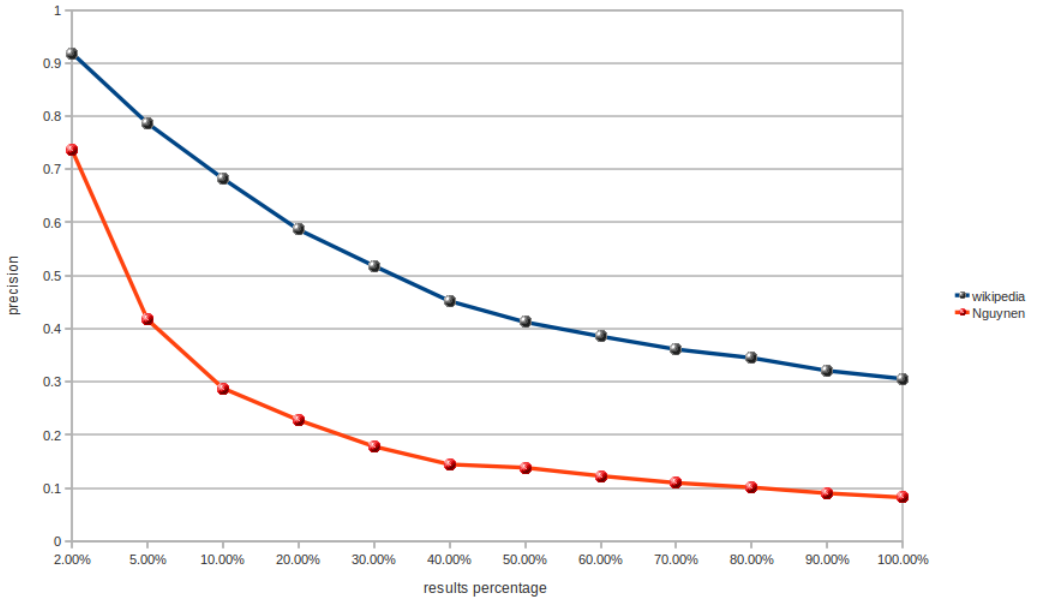
Πίνακας 5.2: Precision & Recall για το σύνολο δεδομένων Nguyen2007

percentage	precision	recall
2%	0.736428571428571	0.152696183044009
5%	0.41697705802969	0.22206919006919
10%	0.287804804804805	0.311417688591602
20%	0.227919877086011	0.480733397037745
30%	0.178368533991806	0.559276512537382
40%	0.1445471907666	0.601969970448231
50%	0.138384988225143	0.715241280458672
60%	0.122884657381361	0.755062288757941
70%	0.109762645485814	0.787190989364902
80%	0.101258944130135	0.830810373201678
90%	0.0903176771338847	0.835158199288634
100%	0.0824124576629651	0.847839358708924

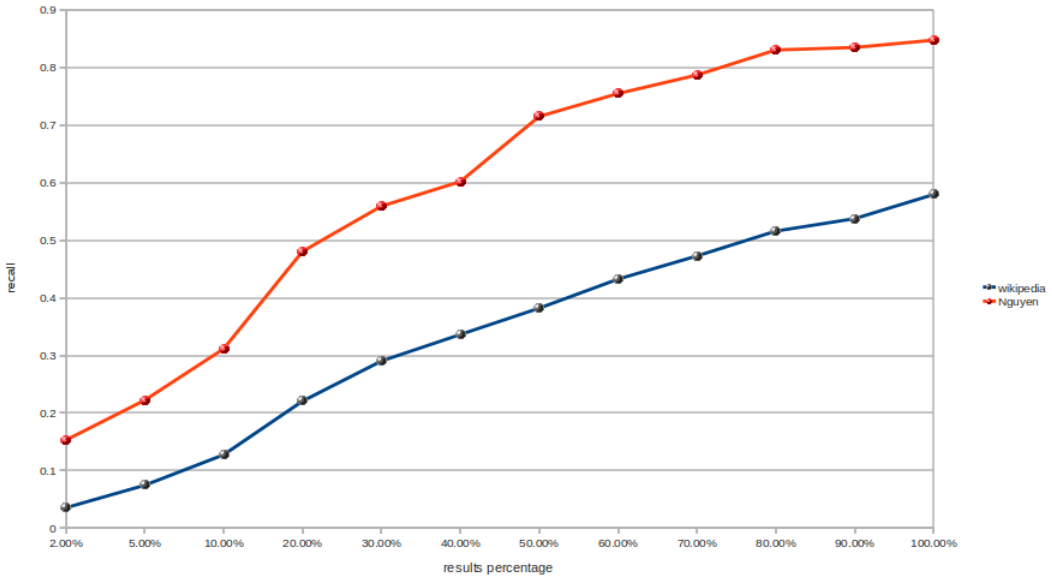


Σχήμα 5.2: precision-recall για το σύνολο δεδομένων Nguyen2007

Στη συνέχεια, στο πλαίσιο του σχολιασμού των αποτελεσμάτων, θα γίνει, εκτός των άλλων και σύγκριση μεταξύ των αποτελεσμάτων των δύο σύνολα δεδομένων. Για το λόγο αυτό θεωρήθηκε χρήσιμη η παρουσίαση συγκριτικών διαγραμμάτων για precision, recall, καθώς και το συνδυασμό τους. Τα διαγράμματα αυτά παραθέτονται σε αυτό το σημείο:



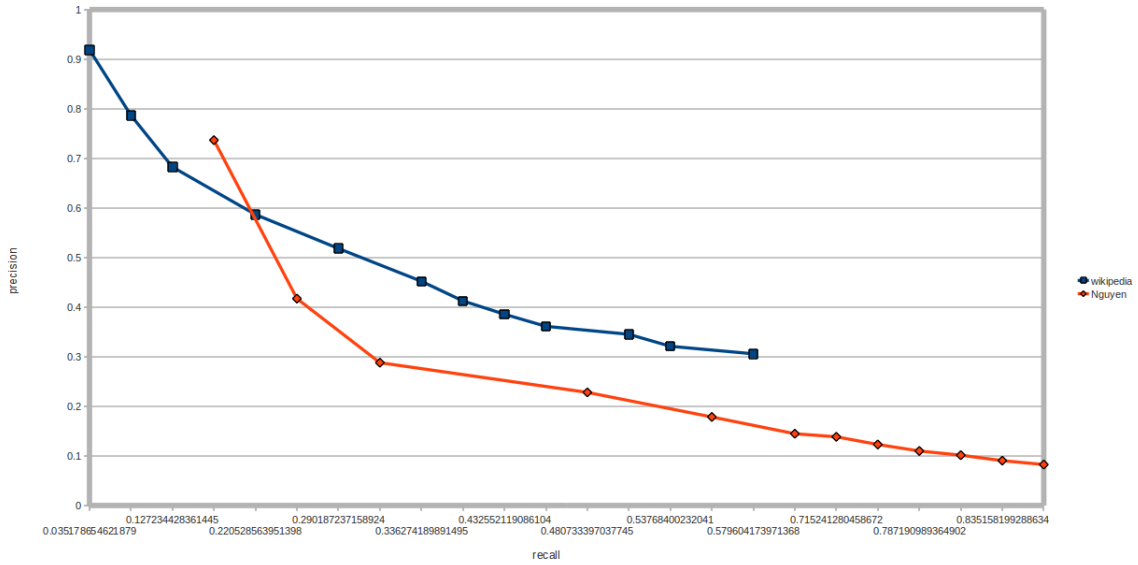
Σχήμα 5.3: Precisions Comparison



Σχήμα 5.4: Recalls Comparison

5.1.3 Σχολιασμός αποτελεσμάτων

Από τα αποτελέσματα μπορούν να εξαχθούν δύο ειδών συμπεράσματα.Το πρώτο έχει να κάνει με τη σχέση που προκύπτει μεταξύ precision και recall καθώς μεταβάλλονται τα ποσοστά και το δεύτερο με την διαφορά που παρουσιάζεται μεταξύ των αποτελεσμάτων των δύο



Σχήμα 5.5: Precision-Recall Comparison

συνόλων δεδομένων που χρησιμοποιήθηκαν.

Οσον αφορά τις γραφικές παραστάσεις των σχημάτων 10.1 και 10.2 φαίνεται πως καθώς αυξάνεται ο αριθμός των αποτελεσμάτων που εξετάζονται, μειώνεται ο μέσος όρος των τιμών του precision και αυξάνεται ο μέσος όρος των τιμών του recall. Η παρατήρηση αυτή ήταν αναμενόμενη δεδομένης της ποιοτικής σημασίας των δύο μεγεθών. Το μέγεθος precision εκφράζει την ακρίβεια των αποτελεσμάτων, δηλαδή το κατά πόσο στο σύνολο των υπό εξέταση αποτελεσμάτων περιλαμβάνονται επιθυμητοί όροι. Η λέξη ”επιθυμητός” αναφέρεται προφανώς στο στόχο που πρέπει να πληρούν τα αποτελέσματα ώστε να θεωρούνται αποδεκτά, όπως αυτός έχει οριστεί στα προηγούμενα κεφάλαια. Το μέγεθος recall αντίθετα, εκφράζει την πληρότητα των αποτελεσμάτων, δηλαδή το κατά πόσο στο σύνολο των υπό εξέταση αποτελεσμάτων περιέχονται όλοι οι όροι που θεωρούνται αποδεκτοί.

Με βάση την παραπάνω ερμηνεία των μεγεθών είναι απόλυτα λογικό το γεγονός πως όσο αυξάνεται ο όγκος των αποτελεσμάτων που γίνονται δεκτά ως έξοδος, μειώνεται η ακρίβεια και επομένως, προστίθενται στα αποτελέσματα και μη αποδεκτοί όροι. Αντίστοιχα λογικό είναι το γεγονός πως κατά τη μεταβολή αυτή, αυξάνονται οι αποδεκτοί όροι μεταξύ των αποτελεσμάτων. Το σημαντικό στοιχείο που πρέπει να σημειωθεί στην παρατήρηση αυτή είναι πως κατά τη συγκεκριμένη διαδικασία αξιολόγησης, το τμήμα των αποτελεσμάτων που εξεταζόταν σε κάθε ποσοστό δεν επιλεγόταν τυχαία από το σύνολο. Αντίθετα, επιλέγονταν πάντα τα πρώτα n στοιχεία εξόδου που αντιστοιχούσαν στο εκάστοτε ποσοστό επί του συνόλου. Δεδομένου ότι τα στοιχεία αυτά είναι ταξινομημένα με βάση το μέγεθος finalMetricSum (έχει οριστεί στο κεφάλαιο 4), τα πρώτα n στοιχεία είναι αυτά που επιλέγονται ως τα πιο πιθανά να αποτελούν αποδεκτές έξοδοι του συστήματος. Επομένως, οι αυξημένες τιμές precision στους πρώτους όρους της εξόδου αποτελούν πολύ θετικό στοιχείο κατά την αξιολόγηση της κατάταξης των

αποτελεσμάτων που αποτελούσαν πιθανές οντότητες-στόχους και δίνονται ως απόκριση στο χρήστη.

Όσον αφορά την σύγκριση μεταξύ των αποτελεσμάτων των δύο διαφορετικών συνόλων δεδομένων, φαίνεται πως στην περίπτωση των featured wikipedia articles (dataset A) είναι πολύ υψηλότερο το precision και πολύ χαμηλότερο το recall σε σχέση με τα αντίστοιχα ποσοστά του συνόλου δεδομένων Nguyen2007 (dataset B). Μάλιστα, φαίνεται πως στην πρώτη περίπτωση, οι τιμές precision για ποσοστά 2% και 5% επί των αποτελεσμάτων φτάνουν σε πολύ υψηλά επίπεδα ενώ οι τιμές recall για ποσοστά 90% και 100% επί των αποτελεσμάτων, σε πολύ χαμηλά. Η παρατήρηση αυτή, συνδέεται άμεσα με την διαφορά που παρουσιάζεται στην ποσότητα των προσημειωμένων όρων κάθε συνόλου δεδομένων. Ενώ τα κείμενα των δύο συνόλων δεδομένων είχαν ίδιας τάξης μεγέθους αριθμό λέξεων, οι προσημειωμένοι όροι που χρησιμοποιήθηκαν ως κριτήριο για την αξιολόγηση των αποτελεσμάτων ήταν υποδεκαπλάσιοι στο dataset B σε σχέση με αυτούς του dataset A. Ήταν, επομένως, πολύ πιο εύκολο στο σύνολο των κατά προσέγγιση 251 αποτελεσμάτων να περιέχονται οι 21 προσημειωμένοι όροι. Επίσης, στην περίπτωση του dataset B, ήταν πολύ πιο εύκολο μεγάλο ποσοστό των κατά προσέγγιση 211 όρων να περιέχεται στους 360 όρους που αποτελούσαν απόκριση του συστήματος. Έτσι, εξ ορισμού, τα αποτελέσματα του dataset A δίνουν πληροφορίες για την ικανότητα του συστήματος να αποκλείει τους μη σημαντικούς όρους, ενώ τα αποτελέσματα του dataset B δίνουν πληροφορίες για την ικανότητα του συστήματος να ξεχωρίζει τους πιο σημαντικούς.

Ολοκληρώνοντας το σχολιασμό των αποτελεσμάτων, θα γίνει μία προσπάθεια να ερμηνευθούν τα σφάλματα (ανεπιθύμητοι όροι εντός των αποτελεσμάτων) με βάση παρατηρήσεις που έγιναν κατά την αξιολόγηση. Κατά την ερμηνεία αυτή, θα γίνει για ακόμα μία φορά διάκριση μεταξύ των δύο συνόλων δεδομένων καθώς διαφέρουν και στην ποιότητα των προσημειωμένων φράσεων που παρέχουν.

DATASET A

Στο συγκεκριμένο σύνολο δεδομένων παρουσιάστηκαν αρκετά σφάλματα που σχετίζονται με τον τρόπο που εμφανίζονται οι σημαντικές φράσεις στο κείμενο. Η συγκεκριμένη κατηγορία άρθρων περιέχουν links που έχουν επιλεχθεί από άνθρωπο και είναι υψηλότερης ποιότητας σε σχέση με αυτά των υπόλοιπων άρθρων, για το λόγο αυτό, υπάρχουν φράσεις οι οποίες αν και ο σύνδεσμός τους σχετίζεται με το κείμενο, με τη μορφή που εμφανίζονται δεν αποτελούν ”σημαντικές φράσεις”. Ένα τέτοιο χαρακτηριστικό παράδειγμα, αποτελεί η λέξη ”economy” στο άρθρο με τίτλο ”Germany”. Πρόκειται για μία πολύ γενική λέξη στην οποία δεν περιέχεται όγκος πληροφορίας για το θέμα του κειμένου. Ο σύνδεσμός της, όμως, που είναι το άρθρο ”Economy of Germany” αποτελεί φράση πλούσιας σημασίας. Όμως, η προαναφερθείσα φράση δεν εμφανίζεται πουθενά αυτούσια στο κείμενο, με αποτέλεσμα να είναι αδύνατον να ανιχνευθεί από το σύστημα. Φυσικά, αυτό αποτελεί και μειονέκτημα του συστήματος καθώς δεν έχει τη δυνατότητα να αξιολογήσει τη σημασιολογική υπόσταση μίας φράσης όταν αυτή δεν δίνεται με πλήρη λεκτική εκπροσώπευση αλλά υπονοείται από τα συμφραζόμενα.

Άλλο πρόβλημα αποτελεί η ύπαρξη φράσεων-συνδέσμων στα άρθρα του συνόλου δεδομένων οι οποίες είναι πολύ γενικές και έχουν μικρή σχέση με το θέμα του κειμένου. Πιθανώς

να αποτελούν κύριες λέξεις στη δεδομένη πρόταση στην οποία εμφανίζονται, όμως δεν θα επιλέγονταν ως επιθυμητές έξοδοι του συστήματος. Χαρακτηριστικό παράδειγμα αποτελεί η λέξη ”cause” (αιτία) στο άρθρο με τίτλο ”Asperger syndrome” (ιατρική πάθηση με συγγένεια τον με τον αυτισμό). Η λέξη αυτή αποτελεί σύνδεσμο με το άρθρο με τίτλο ”Aetiology”. Πρόκειται για τον ορισμό της λέξης αιτιολογία και δεν έχει μεγάλη θεματολογική συνοχή με το υπόλοιπο κείμενο. Αντίθετα, η λέξη αυτή θα είχε σημασιολογική αξία για το άρθρο αν αναφερόταν στην έννοια ”cause of Asperger Syndrome”.

Τέλος, υπάρχουν φράσεις-σύνδεσμοι οι οποίες αν και έχουν σημασιολογική συγγένεια με το θέμα του κειμένου, δεν το χαρακτηρίζουν σε μεγάλο βαθμό και είναι διφορούμενο το κατά πόσο θα έπρεπε ή όχι να εντοπίζονται σε υψηλές θέσεις στα αποτελέσματα του συστήματος. Ένα τέτοιο παράδειγμα, αποτελεί η φράση ”A Dictionary of the English Language” στο άρθρο με τίτλο ”Shakespeare”. Η φράση αυτή αποτελεί τίτλο βιβλίου στο οποίο αναφέρονται λόγια του William Shakespear. Υπάρχει, λοιπόν, κάποιος αρκετά σοβαρός σημασιολογικός σύνδεσμος μεταξύ της φράσης αυτής και του θέματος του κειμένου, όμως, ο σύνδεσμος αυτός δεν είναι ιδιαίτερα στενός και αποτελεί υποκειμενικό κριτήριο το κατά πόσο θα έπρεπε ή όχι να εμφανίζεται στα αποτελέσματα.

DATASET B

Σε αυτό το σύνολο δεδομένων, οι προσημειωμένες φράσεις που αποτέλεσαν πρότυπα κατά την αξιολόγηση ήταν πολύ λιγότερες και σίγουρα σχετίζονταν σε μεγάλο βαθμό σημασιολογικά με το κείμενο. Κατά τη σύγκριση, όμως των φράσεων αυτών και των αποτελεσμάτων, εμφανίστηκε πρόβλημα. Το πρόβλημα αυτό σχετίζεται με το γεγονός πως οι φράσεις αυτές, εξήχθησαν κυρίως με στόχο clustering κειμένων. Αυτό, σε συνδυασμό με τον μικρό αριθμό τους, σημαίνει αυτόματα τον αποκλεισμό από την λίστα των φράσεων-στόχω, φράσεων που είχαν μεγάλη σημασιολογική συνοχή με το κείμενο και κατά τον έλεγχο των αποτελεσμάτων σημειώθpassenger deftonesηκαν ως μη αποδεκτές. Επίσης, σημαντικός παράγοντας πρόκλησης σφαλμάτων αποτέλεσε το γεγονός πως πολλές από τις προσημειωμένες φράσεις περιείχαν ρήματα ή ρηματικούς τύπους τα οποία έχουν αποκλειστεί από την αναζήτηση του συστήματος αυτού λόγω σχεδιαστικής απόφασης. Προφανώς ο προαναφερθείς παράγοντας, αποτελεί και αρνητικό χαρακτηριστικό του συστήματος, καθώς αναδεικνύει την αδυναμία του να αντλήσει σημασιολογική πληροφορία από τέτοιες φράσεις.

Ολοκληρώνοντας την ανάλυση παραγόντων σφαλμάτων που παρατηρήθηκαν, πρέπει να σημειωθεί πως και για τα δύο σύνολα δεδομένων σημαντικό ποσοστό σφαλμάτων προκλήθηκε από την μη ύπαρξη κάποιων οντοτήτων στο σημασιολογικό ιστό της DBpedia που να αναπαριστώνται από φράσεις που θεωρούνται φράσεις-στόχοι από την προσημείωση των κειμένων. Όπως έχει αναφερθεί, το σύστημα αναγνωρίζει μόνο φράσεις που αντιπροσωπεύουν τέτοιες οντότητες και η μη ύπαρξή τους αυτόματα προκαλεί την απόρριψη κάποιων φράσεων που ίσως έχουν μεγάλη σημασιολογική εγγύτητα με το θέμα του κειμένου. Ο τύπος αυτός των σφαλμάτων ήταν αναμενόμενος από την αρχή του σχεδιασμού του συστήματος καθώς οφείλεται σε σχεδιαστική απόφαση.

5.1.4 Προτάσεις για βελτίωση και περαιτέρω επέκταση

Οι προτάσεις που θα γίνουν στη συνέχεια σχετικά με τη βελτίωση των ήδη υλοποιημένων, είναι απόρροια παρατηρήσεων κατά την πειραματική εξέταση του συστήματος αλλά και ιδεών που υπήρχαν από την αρχή του σχεδιασμού ή δημιουργήθηκαν κατά τη διάρκεια της υλοποίησης και δεν θεωρήθηκε σκόπιμο να υλοποιηθούν στο πλαίσιο της διπλωματικής αυτής.

1. Καλύτερος χειρισμός της δυνατότητας που παρέχεται από το σημασιολογικό ιστό της DBpedia για αποσαφήνιση (disambiguation). Στην περίπτωση κατά την οποία η ίδια συμβολοσειρά χρησιμοποιείται για την αναπαράσταση πάνω από μίας οντοτήτων της DBpedia, δεν εμφανίζεται ως αποτέλεσμα της αναζήτησης κάποιο αναγνωριστικό URI αλλά η ειδοποίηση ότι πρόκειται για περίπτωση στην οποία απαιτείται αποσαφήνιση. Στο παρόν σύστημα, επιλέγεται η πρώτη οντότητα η οποία εμφανίζεται κατά τη δεύτερη αναζήτηση (αυτό που φαίνεται πως χρησιμοποιείται πιο συχνά). Θα μπορούσε να υπάρχει υψηλότερης ποιότητας επιλογή μεταξύ των δύο, με χρήση μεθόδων αποσαφήνισης.

2. Εναλλακτικές πηγές πληροφορίας για τον εντοπισμό των υποψήφιων οντοτήτων-στόχων ώστε να μειωθεί το ποσοστό των σφαλμάτων που οφείλονται σε μη εντοπισμό των υπό εξέταση φράσεων στα δεδομένα της DBpedia. Τέτοιες πηγές μπορούν να αποτελέσουν είτε οντολογίες μεγαλύτερης εκφραστικότητας είτε διαφορετικά τμήματα του σημασιολογικού ιστού και εξειδικευμένα λεξιλόγια. Επιπλέον, οι πηγές αυτές μόνο ως βοηθητικές για την αποσαφήνιση ή την ανεύρεση νέων αποτελεσμάτων, διαφορετικά δεν θα υπήρχαν τα πλεονεκτήματα της σύνδεσης του κειμένου με το Σημασιολογικό Ιστό. Σημειώνεται, επίσης, πως η χρήση των πηγών αυτών, θα κατέρριπτε την αρχική παραδοχή που έγινε πως ο κόσμος περιγράφεται πλήρως από τις οντότητες της DBpedia και πως όποια έννοια δεν έχει αντίστοιχη περιγραφή δεν αποτελεί αντικείμενο εξέτασης.

3. Μηχανική μάθηση για τον υπολογισμό βαρών. Στο πλαίσιο της διπλωματικής αυτής, ο υπολογισμός των βαρών των κριτηρίων αξιολόγησης ώστε να υπολογιστεί ο γραμμικός τους συνδυασμός έγινε μέσω περιορισμένου αριθμού πειραμάτων και με ακρίβεια πρώτου δεκαδικού ψηφίου (το δεύτερο δεκαδικό ψηφίο υπολογίστηκε κατά προσέγγιση αφού εξετάστηκε περιορισμένος αριθμός τιμών με βήμα 0.05). Θα ήταν ακριβέστερο το αποτέλεσμα, αν αντ' αυτού είχε χρησιμοποιηθεί κάποιο μοντέλο μηχανικής μάθησης (πχ νευρωνικό δίκτυο).

4. Εξέταση και χρήση επιπλέον κριτηρίων. Κατά το σχεδιασμό του συστήματος υπήρχαν επιπλέον ιδέες για μεγέθη τα οποία θα μπορούσαν να χρησιμοποιηθούν ως κριτήρια αξιολόγησης κάθε φράσης αλλά κατά τα πειράματα που διεξήχθησαν παράλληλα με κάποια βήματα υλοποίησης απορρίφθηκαν. Τέτοια κριτήρια είχαν να κάνουν με RDF πληροφορίες που βρίσκονται στα dumps της DBpedia με τίτλο Infobox Properties και Infobox Types. Οι κλάσεις που ανακτούν τα δεδομένα που απαιτούν τα κριτήρια αυτά υπάρχουν υλοποιημένες, απλά δεν χρησιμοποιούνται κατά την εκτέλεση του προγράμματος. Θα μπορούσαν, παρ'όλα αυτά, να χρησιμοποιηθούν ίσως με διαφορετικό τρόπο (ώστε να παραχθούν ικανοποιητικά αποτελέσματα) στο μέλλον.

5. Υψηλότερου επιπέδου ανάλυση του κειμένου. Κατά την ανάλυση των στοιχείων που δίνονται από το κείμενο, ενοποιούνται φυσικά οι εμφανίσεις φράσεων που έχουν εντοπιστεί πάνω από μία φορά, αλλά μόνο στο βαθμό που το επιτρέπει το stemming του POS Tagger. Αυτό σημαίνει πως, η συχνότητα της λέξης mathematics δεν θα αυξηθεί εξαιτίας της ύπαρξης της φράσης ”mathematical term” ή το αντίστροφο. Επιπλέον, η λέξη ”sickness” δεν θα ενοποιηθεί με τη λέξη ”illness”. Φυσικά, η σημασιολογική τους εγγύτητα θα βρεθεί και θα αυξήσει τις πιθανότητες και των δύο να βρίσκονται μέσα στις φράσεις-στόχους. Τα αποτελέσματα θα ήταν υψηλότερης ποιότητας αν ήταν δυνατόν να εντοπιστούν τέτοιου είδους συγγένειες και να αξιοποιηθούν με πιο συγκεκριμένο τρόπο.

6. Χρήση διαφορετικού Part of Speech Tagger. Ένα ενδιαφέρον στοιχείο προς διερεύνηση θα ήταν η μεταβολή της απόδοσης του συστήματος όταν χρησιμοποιείται διαφορετικός POS tagger. Φυσικά, για να γίνει αυτό, πρέπει είτε να βρεθεί κάποιο εργαλείο επισημείωσης το οποίο να συμπεριλαμβάνει στα αποτελέσματά του τη semi-stemmed μορφή κάθε λέξης, είτε να χρησιμοποιηθεί παράλληλα, κάποιο εργαλείο περιστολής λέξεων που να παρέχει την πληροφορία αυτή.

7. Σημασιολογική αξιοποίηση ρημάτων. Κατά την ανάλυση του κάθε κειμένου, το υπάρχον σύστημα δεν αξιοποιεί με κανέναν τρόπο την πληροφορία που δίνεται από ρήματα και ρηματικούς τύπους. Η αξιοποίησή της θα μπορούσε να συμβάλλει θετικά στη βελτίωση των αποτελεσμάτων.

5.2 Σύστημα SWPID

Το σύστημα SWPID προσανατολίστηκε στην αναγνώριση αναφορών σε υπαρκτά πρόσωπα[1] του πραγματικού κόσμου. Για τις ανάγκες των πειραμάτων, και της αξιολόγησης των αποτελεσμάτων, ήταν απαραίτητος ένας αριθμός προσημειωμένων κειμένων, δηλαδή κειμένων, των οποίων οι αναφορές σε πρόσωπα ήταν γνωστές και επομένως υπήρχε η δυνατότητα εύρεσης ποσοστού λάθος αποκρίσεων του συστήματος, και κατ’επέκταση αξιολόγησής του. Για το σκοπό αυτό χρησιμοποιήθηκαν δεδομένα από τον πιλότο Ανοιχτών και Διασυνδεδεμένων Δεδομένων (Linked Open Data Pilot)[35] που υλοποιήθηκε για το έργο EUScreen [2] και τεκμηριώθηκαν ως προς την ορθότητα των εντοπισμένων οντοτήτων από την Κατερίνα Κομνηνού. Συγκεκριμένα χρησιμοποιήθηκαν 800 κείμενα - περιγραφές από τη βάση της EUScreen. Η επιλογή των κειμένων αυτών, οφείλεται σε δύο λόγους. Πρώτον, καθώς η επισημείωση των κειμένων έγινε με χρήση του εργαλείου Spotlight της DBpedia[37], ήταν δυνατή η σύγκριση της

[1] Ως υπαρκτά πρόσωπα νοούνται όλες οι αναφορές σε φυσικά πρόσωπα, με χρήση του επίσημου ονόματος και επωνύμου. Σε περίπτωση που γίνεται αναφορά σε κάποιον άνθρωπο ο οποίος είναι αναγνωρίσιμος μέ το επώνυμο ή το όνομα του μόνο γίνεται επίσης αποδεκτή ως υπαρκτό πρόσωπο. Επίσης αποδεκτές γίνονται αναφορές σε πρόσωπα με συγκεκριμένο αξίωμα, οι οποίες περιέχουν το αξίωμα και τουλάχιστον το όνομα ή το επώνυμο, εάν αυτό αρκεί για την ορθή και αποκλειστική αναγνώριση τους. Σε κάθε περίπτωση δε γίνονται δεκτές αναφορές σε φανταστικούς χαρακτήρες, χαρακτήρες ταινιών, τηλεοπτικών σειρών κτλ. Επίσης κατ εξαίρεση γίνονται αποδεκτά ονόματα αγίων οι οποίοι είναι καταχωρημένοι στις βάσεις δεδομένων ως PERSONS.

[2] eContentplus 518002, EUscreen, Providing online access to Europe‘s television heritage

απόδοσης του SWPID με ένα ένα έτοιμο αντίστοιχο εργαλείο. Συγκεκριμένα, το Spotlight [25] είναι ένα εργαλείο αυτόματης επισημείωσης φράσεων του κειμένου που αποτελούν οντότητες της DBpedia. Ακόμα προσφέρει και την επιλογή αναγνώρισης οντοτήτων που αποτελούν πρόσωπα όπως και το σύστημα της παρούσας διπλωματικής. Καθώς χρησιμοποιεί ως πρωτεύουσα πηγή δεδομένων τα δεδομένα της DBpedia, θεωρήθηκε ως το πιο σχετικό και αξιόπιστο εργαλείο και κατάλληλο για σύγκριση με χρήση ίδιας εισόδου. Δεύτερον, καθώς στο σύνολό τους καθώς αναφέρονται σε πολιτιστικά γεγονότα, τα πρόσωπα που αναφέρονται στα κείμενα είναι σε μεγάλο ποσοστό αναγνωρίσιμες προσωπικότητες του πραγματικού κόσμου, που διαθέτουν συχνές αναφορές σε διαδικτυακά μέσα, και κατα συνέπεια είναι πολύ πιθανό να είναι καταχωρημένα στη wikipedia, στη DBpedia αλλά και στα λεξικά foaf και schema.

Όσον αφορά την αξιολόγηση του συστήματος υιοθετήθηκε η εξής διαδικασία: Για κάθε κείμενο κρατήθηκε ο συνολικός αριθμός των υπαρκτών προσώπων που αναφέρονται ρητά στο κείμενο και από αυτά ο αριθμός των ατόμων που είναι καταχωρημένα σε κάποια από τις βάσεις δεδομένων που χρησιμοποιούν τα εργαλεία (DBpedia, foaf etc). Για κάθε εργαλείο κρατήθηκε ο αριθμός των επιτυχημένων αναγνωρίσεων για κάθε κείμενο, ο αριθμός των λανθασμένων αναγνωρίσεων καθώς και ο αριθμός των πολλαπλών αναγνωρίσεων για ένα πρόσωπο. Το περιεχόμενο των εννοιών αυτών εξηγείται στη συνέχεια:

- ”Επιτυχημένη αναγνώριση”: Επιτυχημένη θεωρείται η αναγνώριση ενός ατόμου όταν το σύστημα δίνει στην έξοδο το αναγνωριστικό URI που αντιστοιχεί σε πρόσωπο το οποίο αναφέρεται στο κείμενο εισόδου, και συνδέει το URI αυτό με τη συμβολοσειρά του κειμένου της εισόδου η οποία αναφέρεται στο συγκεκριμένο πρόσωπο.

- ”Αποτυχημένη αναγνώριση”: Αποτυχημένη θεωρείται μία αναγνώριση όταν το σύστημα δίνει στην έξοδο το αναγνωριστικό URI που αντιστοιχεί σε κάποιο πρόσωπο το οποίο δεν αναφέρεται στο κείμενο εισόδου, και η συμβολοσειρά του κειμένου με την οποία συνδέει το συνδέει το συγκεκριμένο URI είτε δεν περιγράφει ανθρώπινη οντότητα, είτε περιγράφει διαφορετικό πρόσωπο

- ”Πολλαπλή αναγνώριση ενός ατόμου”: Πολλαπλή θεωρείται μία αναγνώριση, όταν το αναγνωριστικό URI που αντιστοιχεί σε ένα συγκεκριμένο πρόσωπο εμφανίζεται στην έξοδο περισσότερες από μία φορές (και κάθε φορά αντιστοιχίζεται στην ίδια ή και σε διαφορετική συμβολοσειρά μέσα στο κείμενο). Η πολλαπλή αναγνώριση καταγράφεται μόνο στην περίπτωση που η αναγνώριση του ατόμου είναι σωστή. Σε περίπτωση λανθασμένης αναγνώρισης εντάσσεται μόνο στην κατηγορία της ”αποτυχημένης αναγνώρισης”.

Στη συνέχεια παρουσιάζουμε τα πειραματικά αποτελέσματα: Στον παραπάνω πίνακα παρουσιάζονται δύο τιμές για precision και recall για κάθε σύστημα. Το ένα σύνολο τιμών έχει υπολογιστεί ως προς τα πρόσωπα που είναι καταχωρημένα στη DBpedia και το άλλο ως προς τα πρόσωπα που αναφέρθηκαν συνολικά στις φράσεις ανεξάρτητα από το αν είναι καταχωρημένα σε κάποια βάση ή όχι.

Πίνακας 5.3: Precision & Recall σε σύγκριση με spotlight

	total		dbpedia	
	spotlight	thesis	spotlight	thesis
recall	0.432	0.519	0.729	0.876
precision	0.358	0.939	0.358	0.939

5.2.1 Σχολιασμός Αποτελεσμάτων και σύγκριση με αποτελέσματα εργαλείου Spotlight

Ως προς το συνολικό αριθμό ατόμων και για τα δύο συστήματα οι τιμές precision και recall είναι αρκετά χαμηλές. Το γεγονός αυτό υποδηλώνει πως και τα δύο συστήματα εξαρτώνται σε μεγάλο βαθμό από το θέμα του υπο-εξέταση κειμένου. Συγκεκριμένα σε θέματα οπου τα αναφερόμενα πρόσωπα είναι σημαίνουσες προσωπικότητες (από τον τομέα της πολιτικής, των τεχνών, της ιστορίας κτλ), η απόδοση των συστημάτων είναι ικανοποιητική. Μη ικανοποιητική απόδοση εντοπίζεται κυρίως όταν πρόκειται για κείμενα της επικαιρότητας που αναφέρονται σε απλούς πολίτες, ή ορισμένους δημοσιογράφους, ανταποκριτές ή παραγωγούς και παρουσιαστές, οι οποίοι, όπως είναι αναμενόμενο δε διαθέτουν καταχώρηση στη wikipedia. Η αδυναμία ικανοποιητικής απόκρισης ανεξαρτήτως θέματος, είναι προφανώς ένα γεγονός μη αμελητέο, το οποίο όμως δε θα πρέπει να μας απασχολήσει στα πλαίσια της διπλωματικής αυτής καθώς πρόκειται για ένα περιορισμό γνωστό εξ αρχής, δεδομένου ότι χρησιμοποιήθηκε μια συγκεκριμένη βάση δεδομένων (η DBpedia) η οποία ουσιαστικά αποτελεί μια περιορισμένη καταγραφή του πραγματικού κόσμου. Η κατασκευή συστήματος με ευρύτερο πεδίο εφαρμογής απαιτεί πηγές δεδομένων που διαθέτουν μια πιο διευρυμένη καταγραφή του κόσμου. Συγκεκριμένα, η κατασκευή ενός συστήματος που θα ανταποκρίνεται ικανοποιητικά ακόμα και σε άρθρα της επικαιρότητας, προϋποθέτει χρήση μη συστηματικοποιημένων και ιεραρχημένων πόρων του δικτύου (πληροφορίες από blogs, ειδησεογραφικούς ιστότοπους etc), ή χρήση διαφορετικής προσέγγισης ως προς το πρόβλημα της εξαγωγής και διαχείρισης της πληροφορίας του κειμένου. Το ζήτημα αυτό αναλύεται περισσότερο στις προτάσεις για περαιτέρω έρευνα που ακολουθούν.

Όσον αφορά στην απόδοση ως προς τα πρόσωπα που έχουν καταχώρηση στη DBpedia, και αποτελούν το κυρίως αντικείμενο και στόχο της παρούσας διπλωματικής, είναι εμφανές πως το εργαλείο spotlight της DBpedia παρουσιάζει σημαντικά χαμηλότερα ποσοστά τόσο σε precision όσο και σε recall. Ειδικότερα όσον αφορά την τιμή του precision (σύνολο σωστών αποτελεσμάτων προς συνολικά αποτελέσματα) παρατηρούμε πως το σύστημα που κατασκευάστηκε έχει ακρίβεια κοντά στο 95% σε αντίθεση με το spotlight του οποίου η ακρίβεια είναι κάτω του 50% (35%). Η σημαντική απόκλιση στην ακρίβεια οφείλεται κυρίως στη μέθοδο επεξεργασίας των φράσεων και συγκεκριμένα το γεγονός οτι το σύστημά μας αξιοποιεί τη γραμματική επισημείωση των λέξεων που έχει γίνει στο πρώτο στάδιο επεξεργασίας και χρησιμοποιεί ειδικούς κανόνες για τις φράσεις οι οποίες αποτελούν αλληλουχία κύριων ονομάτων. (όπως φαίνεται και στον πίνακα του παραρτήματος 3, ο tree-tagger έχει ειδική ετικέτα για τα κύρια

ονόματα). Έτσι αποφεύγεται το γεγονός λέξεις οι οποίες μπορεί να αποτελούν και συνηθισμένα επώνυμα αλλά στο κείμενο χρησιμοποιούνται ως κοινά ουσιαστικά να επισημανθούν λανθασμένα ως υπαρκτά πρόσωπα, φαινόμενο συχνό στα αποτελέσματα του spotlight. Επιπλέον όπως αναφέρεται σε προηγούμενο κεφάλαιο, αν στον πρώτο έλεγχο του μίας φράσης που αποτελεί αλληλουχία κύριων ονομάτων (NP) δε βρεθεί URI η φράση δεν κατακερματίζεται περαιτέρω, προκειμένου να αποφευχθεί ο κίνδυνος να εντοπιστεί πρόσωπο του οποίου το ονοματεπώνυμο είναι partial match και όχι full match. (πχ έχει ίδιο όνομα και διαφορετικό επώνυμο ή το αντίθετο). Στο ίδιο γεονός συμβάλει και το οτι το σύστημα για να αντιστοιχίσει μια συμβολοσειρά σε ένα αναγνωριστικό URI θα πρέπει να ταιριάζουν απόλυτα. Η μικρή απώλεια precision που παρατηρείται στο σύστημα, εκτός από μη συστηματικοποιημένα λάθη οφείλεται και στην χρήση των redirect URI's αντί του αρχικού αναγνωριστικού URI (ενότητα 4.3.2). Υπολογίσθηκε οτι το 47% των σφαλμάτων οφείλονταν στον παράγοντα αυτό. Ωστόσο, η συμβολή της χρήσης των redirect URIs στη βελτίωση του recall και τον εντοπισμό περισσότερων ατόμων είναι πολύ μεγαλύτερη σε σχέση με το σφάλμα που συνεπάγεται (μόνο 16 λάθη σε σύνολο 586 ατόμων που εντοπίστηκαν συνολικά).

Παρατηρούμε πως τα αποτελέσματα ως προς το recall παραμένουν ικανοποιητικά, και καλύτερα από τα αντίστοιχα του spotlight, ωστόσο, δεν είναι τόσο καλά σε σύγκριση με το precision. Το γεγονός αυτό οφείλεται κυρίως στο ότι αποφεύχθηκε η χρήση της πληροφορίας των συνδέσμων αποσαφήνισης (disambiguation links). Συγκεκριμένα, υπάρχουν ορισμένα ονοματεπώνυμα (ή σκέτα επώνυμα) που αντιστοιχούν σε περισσότερα του ενός πρόσωπα (συχνό φαινόμενο σε ονόματα πολιτικών, βασιλέων κτλ). Στην περίπτωση αυτή η wikipedia (και αντίστοιχα η DBpedia) παρέχει τις σελίδες αποσαφήνισης. Δηλαδή σελίδες με συνδέσμους σε όλες τα άρθρα που αντιστοιχούν στο συγκεκριμένο όνομα. Αντίστοιχα η DBpedia παρέχει τα disambiguation URIs. Το εργαλείο spotlight όταν εντοπίζει οντότητα που κάποια από τα disambiguation URIs του αντιστοιχούν σε υπαρκτά πρόσωπα (person) επιλέγει με χρήση συγκεκριμένου αλγόριθμου[25] το κοντινότερο νοηματικά στο θέμα του κειμένου. Παρ' όλα αυτά, αποδείχτηκε ότι ο αλγόριθμος που χρησιμοποιεί το Spotlight δεν είναι ικανοποιητικός, αφού οδηγεί σε μεγάλη απώλεια precision. Στο σύστημα της παρούσας διπλωματικής, τα disambiguation URIs αγνοήθηκαν με κόστος βέβαια στην τιμή του recall. Αποδεικνύεται παρόλα αυτά πως και πάλι το recall ήταν καλύτερο από αυτό του spotlight.

Πίνακας 5.4: Ανάλυση σφαλμάτων

disambiguation errors	tagger errors	phrase construction errors	person identification errors
78.313	6.024	12.048	3.614

Όπως φαίνεται και στον πίνακα 10.4, οι απώλειες που οφείλονται σε disambiguation issues υπολογίστηκαν ότι αποτελούν το 78,31% των απωλειών. Το ποσοστό αυτό, όπως και τα υπόλοιπα που ακολουθούν, υπολογίστηκε στο σύνολο των ατόμων που υπάρχουν ως οντότητες στην DBpedia. Όσον αφορά τα υπόλοιπα σφάλματα, το 12,05% των απωλειών οφείλεται στον μη κατακερματισμό των φράσεων που αναφέρθηκε παραπάνω, το 6,02% σε λάθος επισημεί-

ωση του tagger που χρησιμοποιήθηκε και το 3,61% σε ταυτοποίηση λάθος ατόμου.

5.2.2 Προτάσεις για βελτίωση και περαιτέρω επέκταση

Το σύστημα επιδέχεται περαιτέρω βελτίωσης, κυρίως σε δύο ζητήματα τα οποία αναφέρθηκαν παραπάνω. Το πρώτο είναι το ζήτημα της αποσαφήνισης λέξεων με πολλαπλά νοήματα. Κάτι τέτοιο μπορεί να καταστεί δυνατό με χρήση μεθόδων και αλγόριθμων σημασιολογικής αποσαφήνισης του θέματος του κειμένου ώστε να επιλεγεί η κατάλληλη ερμηνεία της φράσης η οποία θα είναι πιο κοντά σημασιολογικά στο θέμα. Κάτι τέτοιο μπορεί να γίνει με χρήση μεθόδων clustering, ή με προσαρμογή ενός συστήματος με πρόσβαση σε οντολογίες σημασιολογικού ιστού. Ακόμα είναι δυνατή η χρήση των wikipedia categories με τρόπο ανάλογο με [36] και αναζήτηση κοινών κατηγοριών ανάμεσα στο κείμενο και την υπο εξέταση φράση (ονοματεπώνυμο στη συγκεκριμένη περίπτωση). Τέλος η τροποποίηση του κυρίως συστήματος της παρούσας διπλωματικής εργασίας και της εξόδου του ωστε να δίνει σαν έξοδο τα πιο σημαντικά categories και όχι τις πιο σημαντικές φράσεις, και να εντοπίζει τα κοινά με κάθε προτεινόμενη έννοι των αμφιλεγόμενων φράσεων ενδεχομένως να αποτελεί επαρκή λύση η οποία ωστόσο ήταν πέραν του αντικειμένου της παρούσας διπλωματικής. Ένας ικανοποιητικός αλγόριθμος αποσαφήνισης λέξεων αναμένεται να βελτιώσει τα αποτελέσματα ως προς την τιμή του recall χωρίς να επηρεάσει σημαντικά το precision.

Το δεύτερο ζήτημα αφορά τη συνολική βελτίωση της απόδοσης όσον αφορά τον εντοπισμό ατόμων που δεν ανήκουν στη DBpedia. Οι προσπάθειες βελτίωσης στο συγκεκριμένο τομέα μπορούν να κινηθούν σε δύο άξονες. Ο πρώτος αφορά μεθόδους web mining και στοχεύει στη διεύρυνση των πηγών δεδομένων που έχει στη διάθεση του το σύστημα. Πέρα από τη χρήση διαφορετικών συνόλων δεδομένων και οντολογιών, είναι δυνατή και η υιοθέτηση μεθόδων εξαγωγής και επεξεργασίας πληροφοριών από τον παγκόσμιο ιστό και χρήση της πληροφορίας αυτής για εμπλουτισμό των διαθέσιμων δεδομένων. Η χρήση πληροφορίας από τον παγκόσμιο ιστό και η ένταξη της σε ήδη υπάρχουσες οντολογίες είναι μία μέθοδος που εξασφαλίζει την επικαιροποίηση και εγκυρότητα των δεδομένων. Ο δεύτερος άξονας αφορά διαφορετική προσέγγιση στον τομέα του information retrieval από το κείμενο. Συγκεκριμένα, είναι δυνατό να εφαρμοστούν εκτεταμένα γραμματικοί και συντακτικοί αναλυτές, λεξικά ονομάτων και επωνύμων (name thesaurus) για να εντοπιστούν πιθανές αλληλουχίες λέξεων που αποτελούν ονοματεπώνυμα. Σε συνδυασμό με αλγόριθμους μηχανικής μάθησης και χρήση νευρωνικών δικτύων, είναι δυνατό να κατασκευαστεί ένα αντίστοιχο σύστημα το οποίο βέβαια ενώ ενδεχομένως επιτυγχάνει στον εντοπισμό ατόμων μη καταχωρημένων σε βάσεις δεδομένων, θα είναι αδύνατο να διακρίνει τους φανταστικούς (fictional) από τους πραγματικούς χαρακτήρες.

Ολοκληρώνοντας το κεφάλαιο της αξιολόγησης πρέπει να σημειωθεί πως τα αποτελέσματα που παρουσιάστηκαν, και για τα δύο συστήματα, μπορούν να αξιοποιηθούν από άλλες εφαρμογές, οι οποίες να προσανατολίζονται στην ανάρτηση δεδομένων στον Σημασιολογικό Ιστό με στόχο την επέκταση και βελτίωσή του.

Παράρτημα Α΄

Εγκατάσταση

Διαδικασία εγκατάστασης του συστήματος σε λειτουργικό σύστημα linux

Βήμα 1

Δημιουργία directory που προορίζεται για τη χρήση του συστήματος, λήψη και εγκατασταση του απαραίτητου λογισμικού. Συγκεκριμένα απαιτούνται οι εξής ομάδες πακέτων και εργαλείων:

1. **πακετο treetagger**
   Πρόκειται για τον POS-tagger του πανεπιστημίου της Στουτγκαρδης που εχει περιγραφεί αναλυτικά στην ενότητα []. Τα πακέτα και τα αρχεία που απαιτούνται είναι διαθέσιμα εδω:
   http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
   Συγκεκριμένα απαιτείται πέρα από το πακέτο treetagger η λήψη των ”tagging scripts” του αρχείου εγκατάστασης (install-tagger.sh) και των παραμετρικών αρχείων για την αγγλική γλώσσα (english.par) από τον ίδιο ιστότοπο και η προσθήκη όλων των παραπάνω στο directory όπου βρίσκεται ο κώδικας του συστήματος. Τέλος θα πρέπει να εκτελεστεί σε terminal το αρχείο εγκατάστασης (ΕΝΤΟΛΗ: bash install-tagger.sh). Η χρήση του εργαλείου είναι ενσωματωμένη στον κώδικα και δε χρειάζεται κάποια περαιτέρω ενέργεια από το χρήστη. [1]

2. **πακετο jena-api**
   Απαιτείται η λήψη και προσθήκη των ακόλουθων πακέτων στο directory όπου βρίσκεται ο κώδικας του συστήματος.

   - Jena-2.6.4
   - ARQ-2.8.8

   Στο lib του ARQ θα πρέπει να προστεθούν τα πακέτα

[1] Η επιλογή διαφορετικού εργαλείου συνεπάγεται τροποποίηση του κώδικα αφ ενός ωστε να γίνεται κλήση του σωστού εργαλείου και αφ ετέρου ωστε να μην επηρρεάζει ενδεχόμενη διαφορετική παρουσίαση των αποτελεσμάτων του pos-tagging την υπόλοιπη διαδικασία. Γενικώς η χρήση διαφορετικού εργαλείου ενδέχεται να επηρρεάσει τη σωστή λειτουργία και την απόδοση του συστήματος.

- jena-2.6.4
- arq-2.8.8
- slf4j-api-1.5.8
- slf4j-log4j12-1.5.8
- log4j12-1.5.8
- log4j-1.2.14
- xercesImpl-2.7.1
- iri-0.8
- icu4j-3.4.4.jar

Ενώ στο lib του jena τα πακέτα:

- virt_jena
- virtjdbc3
- virtjdbc4

Τα παραπάνω αρχεία βρίσκονται διαθέσιμα εδω:
(http://sourceforge.net/projects/jena/files/)
Σημειώνεται πως συνίσταται η ενημέρωση των πακέτων με την εκάστοτε τελευταία έκδοση. Οι παραπάνω εκδόσεις είναι αυτές που χρησιμοποιήθηκαν κατά την εγκατάσταση και τη χρήση του συστήματος για τις ανάγκες της παρούσας πτυχιακής εργασίας. Επίσης, σε περίπτωση που γίνει εγκατάσταση σε διαφορετικό directory θα πρέπει να τροποποιηθεί το classpath στο αρχείο build του συστήματος.

3. **πακετο virtuoso**
   Απαιτείται η λήψη του πακέτου virtuoso-opensource-2.6.4 και η εγκατάσταση του με τη χρήση των εντολών make και make-install. Στο εξής τόσο η χρήση του συστήματος όσο και η χρήση του server αυτονομα, συνεπάγεται οτι θα είναι ενεργοποιημένος ο server (θα εκτελέιται το ερχείο virtuoso-t)

4. **πακετο κωδικα συστηματων**
   Απαιτείται η ληψη και η προσθήκη στο directory των ακόλουθων αρχείων:
   Συστημα 1
   Συστημα 2

Βήμα 2

Προσαρμογή των παραμέτρων του virtuoso server με βάση το συστημα στο οποίο θα τρέχει. Η προσαρμογή θα γίνεται από το αρχείο virtuoso.ini στις εξής παραμέτρους:

- NumberOfBuffers
  # each buffer caches a 8K page of data and occupies approx. 8700 bytes of memory. It's suggested to set this value to 65% of ram for a db only server

- MaxDirtyBuffers
  #set according to number of buffers
- MaxCheckpointRemap
  #set to 1/4 of the number of buffers
- ServerThreads
  #set according to server needs

Βήμα 3

”Χτίσιμο” της βασης δεδομένων φορτώνοντας τα δεδομένα από καταλληλα διαμορφωμένα αρχεία με RDF triplets.
Τα αρχεία αποθηκεύονται σε κατάλληλα directories τα οποία πρέπει να ανήκουν οπωσδήποτε σε κάποιο επιτρεπτό path το οποίο ορίζεται στο virtuoso.ini.
(παραμετρος DirsAllowed)
Η διαδικασία του ”χτισίματος” υλοποιείται μέσω της isql με τις εξής εντολές :

ld_dir(’path/to/the/directory/with/files/to/load/to/the/graph’,’*.*’,’graph uri’);

(Η εντολή αυτή δέχεται τρία ορίσματα μέσα σε ”. Το πρώτο όρισμα αντιστοιχεί στο directory το οποιο περιέχει τα αρχεία με τις τριπλέτες που θα φορτωθούν στο γράφο. Σε περίπτωση που θέλουμε να φορτώσουμε ένα directory με subdirectories χρησιμοποιούμε την εντολή ld_dir_all(); με αντιστοιχο τρόπο. Το δεύτερο όρισμα αφορά το ονομα και τον τυπο των αρχείων που περιέχουν τα δεδομένα προς φόρτωση. Σε περίπτωση που θέλουμε να φορτώσουμε όλα τα αρχεία που βρίσκονται στο directory επαρκεί η χρήση του ’*.*’. Τέλος η τρίτη παράμετρος προσδιορίζει το όνομα του γράφου με τον οποίο θα συνδεθούν οι νέες τριπλέτες. Με τη χρήση της εντολής οι πληροφορίες σχετικά με τα αρχεία προς φόρτωση, και τους αντίστοιχους γράφους φορτώνονται σε ένα πίνακα του συστήματος με όνομα ”DB.DBA.LOAD_LIST” όπου κρατούνται όλες οι πληροφορίες σχετικά με τα ήδη φορτωμένα αρχεία αλλά και τα αρχεία προς φόρτωση.)

rdf_loader_run();

(Με την εντολή αυτή, διατρέχεται ο πίνακας DB.DBA.LOAD_LIST που αναφέραμε παραπάνω και όσα δεδομένα δεν έχουν φορτωθεί φορτώνονται (συνδεονται) στους αντίστοιχους γράφους.)

Για οποιοδήποτε πρόβλημα παρουσιαστεί κατά τη διάρκεια της εγκατάστασης της βάσης δεδομένων απαιτείται ο χρήστης να έχει υπ’όψιν πως:

- Για διαγραφή οποιουδήποτε γράφου χρησιμοποιείται η εντολή: SPARQL CLEAR GRAPH <graph uri> ; η οποία διαγράφει τελείως το γράφο και όλα τα δεδομένα που ανήκουν σ αυτόν.

- Για το καθάρισμα του πίνακα DB.DBA.LOAD_LIST χρησιμοποιείται η εντολή : Delete from db.dba.load_list;

- Μετά την ολοκλήρωση της φορτωσης των δεδομένων στη βάση, τα δεδομένα είναι έτοιμα για χρήση.

Βήμα 4

”Χτίσιμο” του κώδικα του συστήματος. Για να είναι έτοιμο προς χρήση το σύστημα απαιτείται να τρέξει ο χρήστης το αρχείο Build_System1.txt (με την εντολή bash ...) για το σύστημα CRESTA και Build_System2.txt για το σύστημα SWPID. Τα αρχεία αυτά εκτελούν τις εντολές μεταγλώτισσης του κώδικα, ρύθμισης των classpaths και δημιουργίας των απαραίτητων directories αν αυτά δεν υπάρχουν.

Βήμα 5

Εκτελεση.

Η εκτέλεση του προγράμματος γίνεται μέσω του αρχείου Run_System1.txt για το σύστημα CRESTA και Run_System2.txt για το σύστημα SWPID. Και τα δύο συστήματα στην αρχή της εκτέλεσης τους ζητούν από το χρήστη να προσδιορίσει και να ρυθμίσει τις απαραίτητες παραμέτρους.

Παράρτημα Β΄

Οδηγίες πρόσβασης στα δεδομένα του Virtuoso

Η πρόσβαση στα δεδομένα μπορεί να γίνει με δύο τρόπους. Είτε με απεθείας ερώτηση από την εφαρμογή της isql, είτε μέσω κώδικα και χρήση κάποιου συγκεκριμεου εργαλείου-εφαρμογής διαχείρησης RDF πλαισίων. Στην παρούσα εργασία έγινε χρήση κώδικα σε java με τη χρήση του jena api.
Ο πρώτος τρόπος παρ όλο που μπορεί να εφαρμοστεί τοσο για απλές όσο και για συνθετες αναζητήσεις, αφορά σε μεμονωμένες αναζητήσεις που πρέπει να εισάγει ο χρήστης και συνεπώς ειναι αδύνατο να καλύψει τους σκοπούς της παρούσας διπλωματικής.
Στη συνέχεια παρουσιάζεται η δευτερη μέθοδος και η χρήση του εργαλείου Jena.
Για τη σύνδεση με χρήση java και jena api είναι απαραίτητο να υπάρχουν στο classpath τα πακέτα του δευτερου συνόλου που αναφέρθηκαν στην αρχή της ενότητας. Επιπλέον η κλάση η οποία είναι υπέυθυνη για τη σύνδεση με το server και τη διαχείρηση της ερώτησης προς αυτόν θα πρέπει να κανει import (εκτός των άλλων) τις εξής κλάσεις από τα παραπάνω πακέτα:

- com.hp.hpl.jena.query.*;
- com.hp.hpl.jena.query.Query;
- com.hp.hpl.jena.query.QueryExecution;
- com.hp.hpl.jena.sparql.engine.http.QueryEngineHTTP;
- com.hp.hpl.jena.sparql.resultset.ResultSetFormat;
- com.hp.hpl.jena.sparql.util.Context;
- com.hp.hpl.jena.query.ResultSetFormatter;
- com.hp.hpl.jena.query.QueryExecutionFactory;
- com.hp.hpl.jena.query.QueryFactory;

- com.hp.hpl.jena.query.QuerySolution;
- com.hp.hpl.jena.query.ResultSet;
- com.hp.hpl.jena.rdf.model.Literal;
- com.hp.hpl.jena.rdf.model.RDFNode;
- com.hp.hpl.jena.sparql.core.Prologue;
- com.hp.hpl.jena.reasoner.rulesys.builtins.Regex;
- virtuoso.jena.driver.*;

Όσον αφορά στον κώδικα και στις μεθόδους και αντικείμενα που χρειάζονται:
Η σύνδεση με κάποιο ήδη υπάρχον γράφο με δεδομένα του virtuoso απαιτεί τη δημιουργία ενός instance της κλάσης VirtGraph:

```
VirtGraph graph = new VirtGraph (”graph_name”, ”localhost_address”, ”root_name”, ”root_pass”);
```

Για παράδειγμα:

```
VirtGraph graph = new VirtGraph (”abstracts”, ”jdbc:virtuoso://localhost:1111”, ”dba”, ”dba”);
```

Στη συνέχεια για την υλοποίηση της αναζήτησης σε SPARQL ορίζεται το instance της κλάσης VirtusosoQueryExecution μέσω της μεθόδου create η οποία δέχεται το σώμα της ερώτησης σε SPARQL με τη μορφή String. Ο κώδικας θα έχει τη μορφή:

```
VirtuosoQueryExecution qehttp = VirtuosoQueryExecutionFactory.create (”query_String”, graph);
```

Σημειώνεται πως εναλλακτικά μπορεί να χρησιμοποιηθεί η κλάση Query για την κατασκευή της ερώτησης.
Η επεξεργασία του αποτελέσματος της ερώτησης εξαρτάται από τον τύπο της. Για τις ερωτήσεις τύπου SELECT χρησιμοποιείται η μέθοδος execSelect() η οποία αποθηκεύει τα αποτελέσματα σε μορφή ResultSet ως εξής:

```
ResultSet rshow = qehttp.execSelect();
```

Η διαχείρηση του αποτελέσματος της execSelect() μπορεί να γίνει ποικιλοτρόπως με τις διαθέσιμες μεθόδους του ResultSet είτε με μεθόδους που ορίζει ο προγραμματιστής ανάλογα με τον τύπο του αναμενόμενου/επιθυμητού αποτελέσματος.

Παράρτημα Γ΄

Σύνολο Ετικετών Γραμματικής Επισημείωσης για τον Stuttgard Tree-tagger

Πίνακας Γ΄.1: Part of Speech Tags

POS Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, third
DT	determiner	the
EX	existential there	there is
FW	foreign word	d'hoevre
IN	preposition, subordinating conjunction	in, of, like
IN/that	that as subordinator	that
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
LS	list marker	1)
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
NP	proper noun, singular	John
NPS	proper noun, plural	Vikings
PDT	predeterminer	both the boys

POS Tag	Description	Example
POS	possessive ending	friend's
PP	personal pronoun	I, he, it
PP$	possessive pronoun	my, his
RB	adverb	however, usually, naturally, here, good
RBR	adverb, comparative	better
RBS	adverb, superlative	best
RP	particle	give up
SENT	Sentence-break punctuation	. ! ?
SYM	Symbol	/ [= *
TO	infinitive 'to'	to go
UH	interjection	uhhuhhuhh
VB	verb be, base form	be
VBD	verb be, past tense	was, were
VBG	verb be, gerund/present participle	being
VBN	verb be, past participle	been
VBP	verb be, sing. present, non-3d	am, are
VBZ	verb be, 3rd person sing. present	is
VH	verb have, base form	have
VHD	verb have, past tense	had
VHG	verb have, gerund/present participle	having
VHN	verb have, past participle	had
VHP	verb have, sing. present, non-3d	have
VHZ	verb have, 3rd person sing. present	has
VV	verb, base form	take
VVD	verb, past tense	took
VVG	verb, gerund/present participle	taking
VVN	verb, past participle	taken
VVP	verb, sing. present, non-3d	take
VVZ	verb, 3rd person sing. present	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WP$	possessive wh-pronoun	whose
WRB	wh-abverb	where, when
#	#	#
$	$	$
”	Quotation marks	' ”
“	Opening quotation marks	' ”
(	Opening brackets	({
)	Closing brackets	) }
,	Comma	,
:	Punctuation	- ; : – ...

Παράρτημα Δ΄

SWPID είσοδος-έξοδος

Παρακάτω παρουσιάζεται ένα παράδειγμα εισόδου-εξόδου του συστήματος SWPID.

Είσοδος:

*T*he analytic–continental divide

Contemporary continental philosophy began with the work of Franz Brentano , Edmund Husserl , Adolf Reinach and Martin Heidegger and the development of the philosophical method of phenomenology. This development was roughly contemporaneous with work by Gottlob Frege and Bertrand Russell inaugurating a new philosophical method based on the analysis of language via modern logic (hence the term ”analytic philosophy”).

Analytic and continental philosophers often hold a disparaging view of each others respective approach to philosophy and as a result work largely independent of each other. While analytic philosophy is the dominant approach in most philosophy departments found in English-speaking countries (e.g. United Kingdom, United States, Canada, Australia), as well as Scandinavia, continental philosophy is prevalent throughout the rest of the world (e.g. France, Germany). Some contemporary philosophers argue that this division is harmful to philosophy, and thus attempt a combined approach (e.g. Richard Rorty).

Analytic and continental philosophy share a common Western philosophical tradition up to Immanuel Kant . Afterwards, analytic and continental philosophers differ on the importance and influence of subsequent philosophers on their respective traditions. The German idealism school which developed out of the work of Kant in the 1780s and 1790s and culminated in Georg Wilhelm Friedrich Hegel is considered an important development in philosophy's history by many continental philosophers, but was thought to be repudiated by Russell , Moore , and many analytic philosophers. Four analytic philosophers. From top-left clockwise: Bertrand Russell , Peter Singer , Saul Kripke , Rosalind Hursthouse

Έξοδος:

persons:

name: Adolf Reinach	URI: http://dbpedia.org/resource/Adolf_Reinach
name: Bertrand Russell	URI: http://dbpedia.org/resource/Bertrand_Russell
name: Edmund Husserl	URI: http://dbpedia.org/resource/Edmund_Husserl
name: Franz Brentano	URI: http://dbpedia.org/resource/Franz_Brentano
name: Moore	URI: http://dbpedia.org/resource/G._E._Moore
name: Georg Wilhelm Friedrich Hegel	URI: http://dbpedia.org/resource/Georg_Wilhelm_Friedrich_Hegel
name: Gottlob Frege	URI: http://dbpedia.org/resource/Gottlob_Frege
name: Immanuel Kant	URI: http://dbpedia.org/resource/Immanuel_Kant
name: Martin Heidegger	URI: http://dbpedia.org/resource/Martin_Heidegger
name: Peter Singer	URI: http://dbpedia.org/resource/Peter_Singer
name: Richard Rorty	URI: http://dbpedia.org/resource/Richard_Rorty
name: Rosalind Hursthouse	URI: http://dbpedia.org/resource/Rosalind_Hursthouse
name: Saul Kripke	URI: http://dbpedia.org/resource/Saul_Kripke

Παράρτημα Ε΄

CRESTA είσοδος-έξοδος

Παρακάτω παρουσιάζεται ένα παράδειγμα εισόδου-εξόδου του συστήματος CRESTA.

Είσοδος:
Natural language processing (NLP) is a field of computer science, artificial intelligence (also called machine learning), and linguistics concerned with the interactions between computers and human (natural) languages. Specifically, it is the process of a computer extracting meaningful information from natural language input and/or producing natural language output. In theory, natural language processing is a very attractive method of human–computer interaction. Natural language understanding is sometimes referred to as an AI-complete problem because it seems to require extensive knowledge about the outside world and the ability to manipulate it.

Whether NLP is distinct from, or identical to, the field of computational linguistics is a matter of perspective. The Association for Computational Linguistics defines the latter as focusing on the theoretical aspects of NLP. On the other hand, the open-access journal ”Computational Linguistics”, styles itself as ”the longest running publication devoted exclusively to the design and analysis of natural language processing systems” (Computational Linguistics (Journal))

Modern NLP algorithms are grounded in machine learning, especially statistical machine learning. Research into modern statistical NLP algorithms requires an understanding of a number of disparate fields, including linguistics, computer science, and statistics. For a discussion of the types of algorithms currently used in NLP, see the article on pattern recognition.

Έξοδος για το 50% των αποτελεσμάτων χωρίς τους εξωτερικούς συνδέσμους:

Natural language processing

http://dbpedia.org/resource/Natural_language_processing

Machine learning

http://dbpedia.org/resource/Machine_learning

Computational linguistics

http://dbpedia.org/resource/Computational_linguistics

Association for Computational Linguistics

http://dbpedia.org/resource/Association_for_Computational_Linguistics

Linguistics

http://dbpedia.org/resource/Linguistics

Computer science

http://dbpedia.org/resource/Computer_science

Artificial intelligence

http://dbpedia.org/resource/Artificial_intelligence

Languages

http://dbpedia.org/resource/Language

Natural language understanding

http://dbpedia.org/resource/Natural_language_understanding

Pattern recognition

http://dbpedia.org/resource/Pattern_recognition

Algorithms

http://dbpedia.org/resource/Algorithm

Computers

http://dbpedia.org/resource/Computer

Open-access journal

http://dbpedia.org/resource/Open_access_journal

Έξοδος για τον πρώτο όρο των αποτελεσμάτων συμπεριλαμβανομένων των εξωτερικών συνδέσμων:

Natural Language Processing

URI : http://dbpedia.org/resource/Natural_language_processing

link to wikipedia : http://en.wikipedia.org/wiki/Natural_language_processing

other links :

http://www.cs.technion.ac.il/ gabr/resources/resources.html

http://www.CICLing.org/

http://www.meshlabsinc.com

http://lucid.cpmc.columbia.edu/medlee/

http://nlpapplications.com/

http://clac.cs.concordia.ca

http://nlp.cic.ipn.mx

http://l2r.cs.uiuc.edu/ cogcomp/

http://nlp.stanford.edu/

http://clsp.jhu.edu

http://nlg.isi.edu/

http://www.aclweb.org

http://code.google.com/p/graph-expression/wiki/Examples

http://www.nltk.org/getting-started

http://specgram.com/CLIII.4/08.phlogiston.cartoon.zhe.html

https://kitwiki.csc.fi/twiki/bin/view/FiLT/FiLTWikiEn

http://www.gelbukh.com/clbook/

http://aclweb.org/aclwiki

http://aclweb.org/anthology-new/

http://atoll.inria.fr/passage/

http://www.evalita.org

http://www.mitpressjournals.org/loi/coli

Βιβλιογραφία

[1] Eric Atwell. The brown corpus tag-set. http://www.scs.leeds.ac.uk/ccalas/tagsets/brown.html.

[2] UK Bermingham University. The bank of english. http://www.titania.bham.ac.uk/docs/svenguide.html.

[3] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3), 2011.

[4] L. M. Campbell and S. MacNeill. The semantic web, linked and open data. 2010.

[5] St. W. Charles. Noun phrase extraction. Master's thesis, The University of Tenessee at Cattanooga, 2008.

[6] G. G. Chowdhury. Natural language processing. *Annual Review of Information Science and Technology*, 37:51–89, 2003.

[7] H. Cunningham. Information Extraction, Automatic. *Encyclopedia of Language and Linguistics, 2nd Edition*, 2005.

[8] editor DARPA, editor. *Proceedings of the Seventh Message Understanding Evaluation and Conference (MUC-98)*, 1998. Morgan Kaufmann.

[9] DBpedia. http://www.dbpedia.org, .

[10] DBpedia. Relfinder. http://www.visualdataweb.org/relfinder/relfinder.php, .

[11] W. Fan, L. Wallace, S. Rich, and Z. Zhang. Tapping the power of text mining. *Communications of the ACM*, 49:76–82, 2006.

[12] W. N. FrancisH and H. Kucera. *MANUAL OF INFORMATION to accompany A Standard Corpus of Present-Day*. Brown University, 1979.

[13] Ralph Grishman. Information extraction: Techniques and challenges. In Maria Pazienza, editor, *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology*, volume 1299 of *Lecture Notes in Computer Science*, pages 10–27. Springer Berlin / Heidelberg, 1997.

[14] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*, volume 1 of *Synthesis Lectures on the Semantic Web: Theory and Technology*. Morgan and Claypool, 2011.

[15] A. Hotho, A. Nürnberger, and G. Paaß. A brief survey of text mining. *LDV Forum- GLDV Journal for Computational Linguistics and Language Technology*, 20:19–62, 2005.

[16] Zhaohui Huang, Huajun Chen, Tong Yu, Hao Sheng, Zhaobo Luo, and Yuxin Mao. Semantic text mining with linked data. *Networked Computing and Advanced Information Management, International Conference on*, 0:338–343, 2009.

[17] P. Jackson and I. Moulinier. *Natural Language Processing for Online Applications*. John Benjamins B.V., 2007.

[18] *Tokenization as the Initial Phase in NLP*, Proceedings of Coling, 1992. Jonathan J. Webster; Chunyu Kit.

[19] Tim-Berners Lee. Linked data. `http://www.w3.org/DesignIssues/LinkedData.html`, 2006.

[20] E.D. Liddy. *Natural Language Processing*. Addison Wesley, 2nd edition, 2001.

[21] P. M. Marcus and M. A. Marcinkiewicz. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics*, 19:313–330, 1933.

[22] A. R. Martinez. Part-of-speech tagging. *WIREs Comp Stat*, 4:107–113, 2012.

[23] Olena Medelyan. *Human-competitive automatic topic indexing. oai:cds.cern.ch:1198029*. PhD thesis, Waikato U., Waikato, 2009. Presented on July 2009.

[24] MediaWiki API. `http://www.mediawiki.org/wiki/API:Main_page`.

[25] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, pages 1–8, New York, NY, USA, 2011. ACM.

[26] David Milne and Ian H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 509–518, New York, NY, USA, 2008. ACM.

[27] R. Mitkov. *The Oxford Handbook Of Computational Linguistics*, chapter 11. Oxford University Press, 2005.

[28] Raymond J. Mooney and Un Yong Nahm. Text mining with information extraction. In W. Daelemans, T. du Plessis, C. Snyman, and L. Teck, editors, *Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium*, pages 141–160. Van Schaik: South Africa, 2003.

[29] Marius Pasca. Acquisition of categorized named entities for web search. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, CIKM '04, pages 137–145, New York, NY, USA, 2004. ACM.

[30] J. Perez, M. Arenas, and C. Gutierrez. Semantics and complexity of sparql. *ACM Transactions on Database Systems (TODS)*, 34(16), August 2009.

[31] John Ross Quinlan. Learning efficient classification procedures and their application to chess end games. In *Machine Learning. An Artificial Intelligence Approach*, pages 463–482, 1983.

[32] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, 1994.

[33] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1–47, 2002.

[34] Nigel Shadbolt, Tim Berners-Lee, and Wendy Hall. The semantic web revisited. *IEEE Intelligent Systems*, 21(3):96–101, May 2006.

[35] N. Simou, J. P. Evain, V. Tzouvaras, M. Rendina, N. Drosopoulos, and J. Oomen. Linking europe's television heritage. Museums and the Web, April 2012, San Diego, USA, 2012.

[36] Gerasimos Spanakis, Georgios Siolas, and Andreas Stafylopatis. Exploiting Wikipedia Knowledge for Conceptual Hierarchical Clustering of Documents. *Computer Journal, Section C: Section C Computational Intelligence*, 55(3):299–312, 2011.

[37] Spotlight. http://dbpedia.org/spotlight.

[38] I. Stavrakantonakis, C. Tsinaraki, N. Bikakis, and S. Christodoulakis. Sparql2xquery 2.0: Supporting semantic-based queries over xml data. In *5th International Workshop on Semantic Media Adaptation and Personalization*, 2010.

[39] Suchanek, M. Fabian, Kasneci, Gjergji, Weikum, and Gerhard. YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. 2007.

[40] Szczuka, Marcin, Janusz, Andrzej, Herba, and Kamil. Clustering of rough set related documents with use of knowledge from dbpedia. In *Proceedings of the 6th international conference on Rough sets and knowledge technology*, RSKT'11, pages 394–403, Berlin, Heidelberg, 2011. Springer-Verlag.

[41] Marcin Szczuka, Andrzej Janusz, and Kamil Herba. Clustering of rough set related documents with use of knowledge from dbpedia. In *Proceedings of the 6th international conference on Rough sets and knowledge technology*, RSKT'11, pages 394–403, Berlin, Heidelberg, 2011. Springer-Verlag.

[42] The Stanford Natural Language Processing Group. `http://nlp.stanford.edu/software/tagger.shtml`.

[43] Tree Tagger tagset. `ImprovementsinPart-of-SpeechTaggingwithanApplicationtoGerman`.

[44] Treetagger. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`.

[45] UK UCL. The internet grammar of english. `http://www.ucl.ac.uk/internet-grammar/home.htm`.

[46] Carlos Vicient. Ontology-based information extraction. Master's thesis, URV, June 2011.

[47] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, April 1967.

[48] W3C. `http://www.w3.org/2004/02/skos/`, .

[49] W3C. Web ontology language overview. `http://www.w3.org/TR/owl-features/`, 2004.

[50] W3C. Rdf vocabulary description language 1.0: Rdf schema. `http://www.w3.org/TR/rdf-schema/`, 2004.

[51] W3C Recommendation. Sparql query language for rdf. `http://nlp.stanford.edu/software/tagger.shtml`, January 2008.

[52] Pu Wang and Carlotta Domeniconi. Building semantic kernels for text classification using wikipedia. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 713–721, New York, NY, USA, 2008. ACM.

[53] I. H. Witten. Text mining. In M. P. Singh, editor, *Practical handbook of internet computing*, chapter 14, pages 1–22. Chapman and Hall/CRC Press, Boca Raton, Florida, 2005.

[54] World Wide Web Consortium. `http://www.w3.org`.

Γλωσσάρι

Σημασιολογικός Ιστός Semantic Web. 19, 21, 23, 38, 48, 63

ακρίβεια precision. 107, 113

ανάκληση recall. 107, 113

ανάκτηση πληροφορίας information retrieval. 26

αναγνωριστική περιγραφή resource description. 39

αναγνωριστικό URI resource description URI. 78

αναγνώριση Θέματος topic recognition. 26

αναγνώριση ονοματοδοτημένων οντοτήτων named entity recognition. 26

αναπαράσταση γνώσης knowledge representation. 19

ανθρώπινη κατανάλωση human consumption. 19

αντικείμενο object. 30

αυστηρή εξαγωγή βασισμένης σε οντολογίες strict ontology based extraction. 62

αυτόματη εξαγωγή περίληψης automatic summarization. 26

αυτόματη εξαγωγή συμπερασμάτων reasoning. 36

αυτόματο πεπερασμένων καταστάσεων finite state automaton. 31

αφηρημένος τύπος κλάσης abstract class. 93

γεγονός fact. 34

γραμματική επισημείωση Part of Speech Tagging. 22, 26–29, 33, 75, 89, 99, 103

διασυνδεδεμένα δεδομένα linked data. 36, 58, 63

εισερχόμενες ακμές incoming links. 43

εκδότης δημόσιων δεδομένων public data publisher. 56, 59

εκτίμηση μέγιστης πιθανοφάνειας Maximum Likelihood Estimation. 90

Acronyms

HTTP HyperText Transfer Protocol - Πρωτόκολλο Μεταφοράς Υπερκειμένου. 38

LOD Linked Open Data - Διασυνδεδεμένα Δεδομένα με ανοιχτά δικαιώματα πρόσβασης. 37, 59

OWL Web Ontology Language - Γλώσσα Οντολογίας Διαδικτύου. 36

RDF Resource Description Framework - Πλαίσιο Περιγραφής Πόρων. 36, 55, 56, 60

RDFs Resource Description Framework schema - Σχήμα Πλαισίου Περιγραφής Πόρων. 38, 44

SPARQL SPARQL Protocol and RDF Query Language - Πρωτόκολλο SPARQL και Γλώσσα Επερώτησης RDF. 39

URI Uniform Resource Identifier - Ενιαίο Αναγνωριστικό Πόρων. 36, 56

URL Uniform Resource Locator -Ενιαίος Εντοπιστής Πόρων. 56