Εθνικό Μετσόβιο Πολυτεχνείο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Εξαγωγή σχέσεων μεταξύ οντοτήτων από το αρχείο της εφημερίδας «ΤΑ ΝΕΑ» με χρήση τεχνικών μη-επιβλεπόμενης μάθησης

Διπλωματικη Εργασια

Γεωργιος Θεοφιλου
Νικολαος Παπασαραντοπουλος

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

Εργαστηριο Ευφυων Υπολογιστικων Συστηματων
Αθήνα, Νοέμβριος 2012

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων
Εργαστηριο Ευφυων Υπολογιστικων Συστηματων

Εξαγωγή σχέσεων μεταξύ οντοτήτων από το αρχείο της εφημερίδας «ΤΑ ΝΕΑ» με χρήση τεχνικών μη-επιβλεπόμενης μάθησης

Διπλωματικη Εργασια

Γεωργιος Θεοφιλου
Νικολαος Παπασαραντοπουλος

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

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 9η Νοεμβρίου 2012.

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

Αθήνα, Νοέμβριος 2012

.....
Γεωργιος Θεοφιλου
Διπλωματούχος
Ηλεκτρολόγος Μηχανικός
και Μηχανικός Υπολογιστών Ε.Μ.Π.

.....
Νικολαος Παπασαραντοπουλος
Διπλωματούχος
Ηλεκτρολόγος Μηχανικός
και Μηχανικός Υπολογιστών Ε.Μ.Π.

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα διπλωματική εργασία εκπονήθηκε κατά το ακαδημαϊκό έτος 2011-2012 στο Εθνικό Μετσόβιο Πολυτεχνείο.

Θα θέλαμε να ευχαριστήσουμε τον επιβλέποντα καθηγητή κ. Ανδρέα Σταφυλοπάτη για την εμπιστοσύνη που μας έδειξε με την ανάθεση της εργασίας αυτής, η οποία αποτέλεσε πολύτιμη ευκαιρία να ασχοληθούμε με ένα πολύ ενδιαφέρον θέμα.

Θα θέλαμε επίσης να ευχαριστήσουμε θερμά το δρ. Γεράσιμο Σπανάκη για τις υποδείξεις και την καθοριστική του καθοδήγηση καθώς και για την υποδειγματική συνεργασία που είχαμε, σε όλα τα στάδια εκπόνησης της εργασίας.

ΠΕΡΙΛΗΨΗ

Η παρούσα διπλωματική εργασία έχει ως αντικείμενο τη μελέτη και την ανάπτυξη ενός συστήματος εξαγωγής σχέσεων μεταξύ οντοτήτων από αδόμητο, ποικίλης θεματολογίας και δομής κείμενο, με χρήση τεχνικών μη επιβλεπόμενης μάθησης. Το σύστημα ακολουθεί το πρότυπο του open relation extraction, δηλαδή δεν απαιτεί καμία πληροφορία εισόδου πέρα από το σώμα κειμένου από το οποίο επιχειρεί να εξάγει σχέσεις. Η εξαγωγή σχέσεων μεταξύ οντοτήτων συνίσταται στην συστηματική εξαγωγή τριάδων της μορφής (e_1, r, e_2), όπου e_1, e_2 οντότητες και r η (ρηματική) σχέση με την οποία συνδέονται.

Το σύστημα αντιμετωπίζει κείμενα τα οποία είναι γραμμένα στην ελληνική γλώσσα. Για την υλοποίηση και τον έλεγχο ορθής λειτουργίας του χρησιμοποιήθηκε το αρχείο της εφημερίδας «ΤΑ ΝΕΑ» · μια επιλογή η οποία εξασφάλισε ένα μεγάλου μεγέθους και ποικίλης θεματολογίας και μορφής σώμα κειμένου.

Η εξαγωγή σχέσεων επιτυγχάνεται με τη χρήση τεχνικών συντακτικής ανάλυσης κειμένου και ο διαχωρισμός τους σε θετικές (σημασιολογικά ορθές) ή μη γίνεται με τη χρήση ταξινομητή. Ο ταξινομητής εκπαιδεύεται με ένα σύνολο επισημειωμένων δεδομένων, τα οποία προκύπτουν από την εφαρμογή ενός συνόλου κανόνων.

ΛΕΞΕΙΣ-ΚΛΕΙΔΙΑ

εξαγωγή σχέσεων μεταξύ οντοτήτων, εξαγωγή πληροφορίας, ταξινόμηση, ομαδοποίηση, μηχανική μάθηση, γραμματική επισημείωση όρων

ABSTRACT

The main object of the present thesis is the study and development of a system that attempts to extract relations between entities from large, unstructured and multiple-topic corpora, using non-supervised learning techniques. The system follows the open relation extraction paradigm; hence it does not require additional input data, except the text corpus. Relation extraction is oriented towards the extraction of tuples (e_1, r, e_2), where e_1, e_2 denote entities and r denotes the (verbal) relation that connects the two entities.

The system addresses texts written in greek language. The corpus used as test set was the archive of the greek newspaper "TA NEA", which offered a multiple topic and multiple structure amount of text as input data.

The system first extracts a large number of relations from the input text using parsing techniques and then each relation gets classified as positive (semantically true) or negative by a classifier. The classifier is trained by a training set of data tagged by the system, using a set of rules.

KEY-WORDS

open relation extraction, open information retireval, classification, clustering, machine learning, part of speech tagging

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΤΜΗΜΑΤΑ ΚΩΔΙΚΑ

1 ΕΙΣΑΓΩΓΗ

Στη σύγχρονη μορφή του, ο Παγκόσμιος Ιστός αποτελείται από ένα διαρκώς αυξανόμενο σύνολο πληροφοριών. Η ευκολία με την οποία πλέον αναρτώνται δεδομένα στο διαδίκτυο οδήγησε στη συσσώρευση τεράστιου όγκου δεδομένων, διαθέσιμων προς παγκόσμια πληροφόρηση. Ωστόσο, το μεγαλύτερο μέρος των δεδομένων αυτών βρίσκεται σε μορφή η οποία δεν επιτρέπει καθόλου τη σημασιολογική επεξεργασία από υπολογιστές.

Ένα μεγάλο ποσοστό της πληροφορίας η οποία κυκλοφορεί στον Παγκόσμιο Ιστό αποτελείται από έγγραφα ελεύθερου κειμένου τα οποία στο μεγαλύτερο μέρος τους δεν προορίζονται για αυτοματοποιημένη επεξεργασία, αλλά για ανθρώπινη κατανάλωση. Ως εκ τούτου, είναι αδόμητα και ανοργάνωτα τόσο ως προς το περιεχόμενο, όσο και ως προς τη δομή. Το αποτέλεσμα είναι να περιορίζεται σημαντικά η δυνατότητα εκμετάλλευσης του τεράστιου αυτού όγκου πληροφοριών, καθώς καθίσταται δύσκολη η αναζήτηση και η διαχείρισή της.

Η ανάπτυξη της τεχνητής νοημοσύνης ώθησε την ανάπτυξη επιστημονικών περιοχών όπως αυτές της επεξεργασίας φυσικής γλώσσας, εξόρυξης κειμένου και εξαγωγής πληροφορίας, με σκοπό, εκτός των άλλων, την εξαγωγή δομημένης πληροφορίας από ανθρωπίνως παραγόμενο κείμενο. Η συμβολή των συστημάτων αυτών στη διαχείριση του όγκου πληροφοριών του διαδικτύου είναι καθοριστικής σημασίας. Επιπλέον, ενισχύουν το όραμα με το οποίο εισήχθη το Web 2.0, δηλαδή να μετατραπεί σταδιακά ο Παγκόσμιος Ιστός από ιστός διασυνδεδεμένων εγγράφων σε ιστό διασυνδεδεμένων δεδομένων

Γίνεται λοιπόν αντιληπτό ότι πλέον δίνεται η δυνατότητα δημιουργίας συστημάτων τα οποία δέχονται ως είσοδο κείμενο και εξάγουν δομημένες πληροφορίες, διαχειρίσιμες από υπολογιστικές μηχανές. Στο πλαίσιο της διπλωματικής εργασίας αυτής, αναπτύχθηκε ένα σύστημα το οποίο προσεγγίζει το ζήτημα της εξαγωγής γνώσης από τη σκοπιά εξαγωγής σχέσεων μεταξύ οντοτήτων και ειδικότερα, σε κείμενα τα οποία είναι γραμμένα στην ελληνική γλώσσα.

1.1 ΣΥΣΤΗΜΑ ΕΞΑΓΩΓΗΣ ΣΧΕΣΕΩΝ ΜΕΤΑΞΥ ΟΝΤΟΤΗΤΩΝ

Το σύστημα εξαγωγής σχέσεων μεταξύ οντοτήτων το οποίο αναπτύχθηκε στα πλαίσια της εργασίας αποσκοπεί στην αναγνώριση σημασιολογικών σχέσεων μεταξύ οντοτήτων από αδόμητο κείμενο, το οποίο δεν περιέχει καθόλου μεταδεδομένα. Το σύστημα ακολουθεί την προσέγγιση του open relation extraction, δηλαδή δεν απαιτεί από το χρήστη να προκαθορίσει τη μορφή ή το περιεχόμενο των προς εξαγωγή σχέσεων. Αντιθέτως, προκειμένου να μπορεί να χρησιμοποιηθεί για πολλά και ετερόκλητα σώματα κειμένου, χρησιμοποιεί τεχνικές μηχανικής μάθησης για να επιτύχει την εξαγωγή όσο περισσότερων, σημασιολογικά ορθών σχέσεων.

Τα περισσότερα συστήματα εξαγωγής σχέσεων μεταξύ οντοτήτων τα οποία έχουν δημιουργηθεί στοχεύουν, όπως είναι λογικό, σε κείμενα τα οποία είναι γραμμένα στην αγγλική γλώσσα. Η τεράστια ποσότητα κειμενικού υλικού γραμμένου σε όλες τις γλώσσες η οποία αναρτάται στον Παγκόσμιο Ιστό καθημερινά, δημιουργεί την ανάγκη ανάπτυξης συστημάτων τα οποία αντιμετωπίζουν κείμενα γραμμένα σε λιγότερο διαδεδομένες γλώσσες, όπως η ελληνική. Το σύστημα που περιγράφεται στην εργασία αυτή είναι κατασκευασμένο για να αντιμετωπίζει κείμενα γραμμένα στην ελληνική γλώσσα · γλώσσα η οποία διαφέρει αρκετά ως προς τη σύνταξη και τη δομή από την αγγλική.

Για την υλοποίηση και τον έλεγχο ορθής λειτουργίας του συστήματος ήταν απαραίτητη η παρουσία ενός συνόλου κειμένων τα οποία δεν παρουσιάζουν καμία ομοιογένεια, δομική ή θεματολογίας. Για το σκοπό αυτό χρησιμοποιήθηκε το αρχείο της εφημερίδας «ΤΑ ΝΕΑ», το οποίο δόθηκε σε ηλεκτρονική μορφή.

Το σύστημα εκτελεί μια σειρά βημάτων για την επεξεργασία του αρχείου και τη συστηματική εξαγωγή σχέσεων (τριάδων της μορφής (οντότητα, σχέση, οντότητα)) από το κείμενο που το αποτελεί, καθώς και την απόδοση βαθμολογίας η οποία να αντιπροσωπεύει κατά πόσον θεωρείται έμπιστη (σημασιολογικά ορθή) ή όχι κάθε σχέση.

1.2 ΔΙΑΡΘΡΩΣΗ ΤΟΥ ΚΕΙΜΕΝΟΥ

Στα κεφάλαια που ακολουθούν, αναλύεται το σύστημα εξαγωγής σχέσεων μεταξύ οντοτήτων καθώς και το σχετικό επιστημονικό και τεχνολογικό υπόβαθρο στο οποίο βασίστηκε η ανάπτυξή του. Το κείμενο είναι χωρισμένο σε δύο μέρη. Στο πρώτο μέρος γίνεται παρουσίαση της σχετικής με την εξαγωγή σχέσεων επιστημονικής γνώσης, ενώ το δεύτερο μέρος είναι αφιερωμένο στην περιγραφή και ανάλυση του συστήματος που αναπτύχθηκε στα πλαίσια της παρούσης εργασίας. Πιο συγκεκριμένα:

- Μέρος Ι (Σχετική επιστημονική γνώση)
  - Στο *δεύτερο κεφάλαιο*, με τίτλο *"Εξαγωγή πληροφορίας"*, παρουσιάζεται η θεωρητική θεμελίωση του προβλήματος της εξαγωγής πληροφορίας (information extraction), η συνήθης διαδικασία επίλυσης, καθώς και κάποια μοντέλα τα οποία έχουν αναπτυχθεί για την επίλυσή του. Στη συνέχεια, αναλύεται το ζήτημα της εξαγωγής σχέσεων μεταξύ οντοτήτων (relation extraction) – τόσο η traditional όσο και η open προσέγγιση – και οι τεχνικές που ακολουθούνται για την αντιμετώπισή του. Τέλος, γίνεται μία «ιστορική αναδρομή» στα συστήματα open relation extraction τα οποία έχουν μέχρι τώρα δημιουργηθεί.
  - Στο *τρίτο κεφάλαιο*, με τίτλο *"Αναγνώριση προτύπων"*, παρουσιάζονται θεωρητικά τα προβλήματα της ταξινόμησης (classification) και της ομαδοποίησης (clustering) καθώς και μερικά από τα μοντέλα τα οποία έχουν αναπτυχθεί για την επίλυσή τους. Η λύση των προβλημάτων αυτών έδωσε τη δυνατότητα δημιουργίας εφαρμογών οι οποίες αποτελούν βασικά εργαλεία για την υλοποίηση αποδοτικών συστημάτων εξαγωγής γνώσης.

- Μέρος ΙΙ (Πειραματικό μέρος)
  - Στο *τέταρτο κεφάλαιο*, με τίτλο *Γενική περιγραφή συστήματος* γίνεται μια αναλυτική περιγραφή του συστήματος που αναπτύχθηκε. Αρχικά περιγράφεται η δομή και η μορφή και στην οποία βρίσκεται το σώμα κειμένου το οποίο χρησιμοποιήθηκε (το αρχείο της εφημερίδας «ΤΑ ΝΕΑ») και στη συνέχεια αναλύεται ο σχεδιασμός της δομής του συστήματος, οι αρχές λειτουργίας του, καθώς και η μορφή εξόδου του.
  - Στο *πέμπτο κεφάλαιο*, με τίτλο *Ζητήματα Υλοποίησης* περιγράφονται λεπτομέρειες υλοποίησης του συστήματος. Επίσης, περιγράφονται τα εργαλεία τα οποία χρησιμοποιήθηκαν προκειμένου να γίνει κατανοητός ο τρόπος λειτουργίας τους καθώς και ο ρόλος τους στα πλαίσια υλοποίησης του συστήματος.
  - Στο *έκτο κεφάλαιο*, με τίτλο *Απόδοση συστήματος* παρουσιάζονται και σχολιάζονται τα αποτελέσματα του συστήματος, με τελικό στόχο την αξιολόγησή τους και την ανάδειξη ζητημάτων που επιδέχονται περαιτέρω μελέτης ή/και βελτίωσης.

Μέρος Ι

ΣΧΕΤΙΚΗ ΕΠΙΣΤΗΜΟΝΙΚΗ ΓΝΩΣΗ

2

ΕΞΑΓΩΓΗ ΠΛΗΡΟΦΟΡΙΑΣ

Στο κεφάλαιο αυτό παρουσιάζεται η θεωρητική θεμελίωση του προβλήματος της εξαγωγής πληροφορίας (information extraction), η συνήθης διαδικασία επίλυσης καθώς και κάποια μοντέλα που έχουν αναπτυχθεί για την επίλυση του προβλήματος αυτού. Στη συνέχεια παρουσιάζεται το ζήτημα της εξαγωγής σχέσεων μεταξύ οντοτήτων (relation extraction), οι τεχνικές που ακολουθούνται για την αντιμετώπισή του, καθώς και μια «ιστορική αναδρομή» στα συστήματα open relation extraction που έχουν μέχρι τώρα αναπτυχθεί.

2.1 ΕΞΑΓΩΓΗ ΠΛΗΡΟΦΟΡΙΑΣ

Η εξαγωγή πληροφορίας (information extraction, information retrieval) συνίσταται στην αυτόματη εξαγωγή δομημένης πληροφορίας από αδόμητα και/ή ημιδομημένα έγγραφα τα οποία μπορεί να επεξεργαστεί ένας υπολογιστής.

Τις περισσότερες φορές, το information extraction αφορά κείμενα γραμμένα σε φυσική γλώσσα, οπότε και χρησιμοποιούνται τεχνικές επεξεργασίας φυσικής γλώσσας (natural language processing - NLP). Ωστόσο, πολλές φορές, ο όρος information extraction αναφέρεται και σε άλλου τύπου δεδομένα, όπως εικόνα, μουσική, βίντεο κ.λπ. από τα οποία μπορεί κανείς να εξάγει χρήσιμες πληροφορίες (content extraction) ή να πραγματοποιήσει εργασίες οι οποίες εμπεριέχουν κάποιας μορφής εξαγωγή πληροφορίας (όπως π.χ. το αυτόματο annotation σε multimedia αρχεία).

Επειδή το πρόβλημα του information extraction είναι, εν γένει, δύσκολο, πολλές φορές οι προσεγγίσεις περιορίζονται σε έγγραφα συγκεκριμένης θεματολογίας, όπου η «παραγόμενη πληροφορία» είναι σχετικά αναμενόμενη. Στην περίπτωση αυτή, η διαδικασία του information extraction, πρακτικά, συνίσταται στην αναγνώριση συγκεκριμένου είδους πληροφοριών, όπως κύρια ονόματα (ονόματα ανθρώπων, τοπωνύμια, ονόματα εταιρειών, ημερών, μηνών κ.λπ.), χρονικές πληροφορίες (ημερομηνίες), σχέσεις και γεγονότα.

Τυπικές εργασίες information extraction είναι οι ακόλουθες:

- αναγνώριση ονομάτων/οντοτήτων
  - αναγνώριση κυρίων ονομάτων
  - αναγνώριση αναφορικών σχέσεων μεταξύ ονομάτων
  - αναγνώριση σχέσεων μεταξύ οντοτήτων
- αναγνώριση ημιδομημένων πληροφοριών
  - εξαγωγή πινάκων από έγγραφα
  - εξαγωγή σχολίων από άρθρα

- γλωσσική και λεξικογραφική ανάλυση
  - εξαγωγή ορολογίας

Εν συντομία, στόχος του information extraction είναι η δημιουργία δομημένης πληροφορίας από τον τεράστιο όγκο δεδομένων τα οποία βρίσκονται σε ανθρωπίνως κατανοητή αναπαράσταση. Επιτυγχάνοντας κάτι τέτοιο, μπορεί να καταστεί δυνατή η εξαγωγή νέας πληροφορίας, μέσω συλλογιστικής. Πιθανές εφαρμογές μιας τέτοιας ανάλυσης είναι η αυτόματη εξαγωγή περιλήψεων κειμένων, αυτόματη απάντηση ερωτήσεων ή αυτόματη μετάφραση κ.ά.

2.2 ΤΕΧΝΟΛΟΓΙΚΟ ΥΠΟΒΑΘΡΟ

Στην ενότητα αυτή παρουσιάζεται το τεχνολογικό υπόβαθρο το οποίο χρησιμοποιεί η ερευνητική περιοχή της εξαγωγής πληροφορίας από κείμενο (text information retrieval). Όπως μπορεί κανείς εύκολα να αντιληφθεί, τα εργαλεία για την εργασία της εξαγωγής πληροφορορίων λαμβάνονται από τον ερευνητικό χώρο της επεξεργασίας φυσικής γλώσσας (natural language processing).

2.2.1 *Επεξεργασία Φυσικής Γλώσσας*

Ο όρος *επεξεργασία φυσικής γλώσσας* υποδηλώνει την επεξεργασία και κατανόηση φυσικώς παραγόμενου ανθρωπίνου κειμένου από υπολογιστικές μηχανές. Πρόκειται για κλάδο στον οποίο συνεισφέρουν επιστήμες όπως η πληροφορική και η τεχνητή νοημοσύνη, αλλά και η γλωσσολογία, ακόμη και η ψυχολογία. Μακροπρόθεσμος στόχος της επεξεργασίας φυσικής γλώσσας είναι όχι απλά η ανάλυση κειμένου, αλλά η κατανόησή του από υπολογιστή. Μιας και αυτό, προς το παρόν, μοιάζει εξαιρετικά δύσκολο, έχουν πραγματοποιηθεί εργασίες οι οποίες λύνουν κάποιο μέρος του προβλήματος.

Κάποια προβλήματα τα οποία έχουν λυθεί ως ένα βαθμό και οι λύσεις τους παρουσιάζουν επιστημονικό και χρηστικό ενδιαφέρον, είναι:

- κατάτμηση σε λεκτικές μονάδες (tokenization): εξαγωγή όρων οι οποίοι αποτελούν λεκτικές μονάδες (tokens) από ένα κείμενο.
- συντακτική ανάλυση (parsing): συντακτική ανάλυση ενός κειμένου με βάση κάποιους καθορισμένους συντακτικούς κανόνες.
- χαμηλού επιπέδου συντακτική ανάλυση (chunking): αναγνώριση ουσιαστικών, ονοματικών φράσεων, ρηματικών φράσεων κ.λπ., χωρίς να προσδιορίζεται ο συντακτικός τους ρόλος μέσα στα συμφραζόμενα.
- γραμματική επισημείωση (part-of-speech tagging): προσδιορισμός μέρους του λόγου για κάθε λέξη του κειμένου.
- περιστολή λέξεων (stemming): αναγωγή στη ρίζα κάθε λέξης. Αυτό επιτυγχάνεται με τη χρήση κανόνων, καθώς και με την αφαίρεση προθεμάτων και καταλήξεων. Κύριος στόχος του stemming είναι η αναγνώριση διαφορετικών λέξεων με την ίδια ρίζα.

Ακολουθεί μια πιο αναλυτική παρουσίαση των τρόπων επίλυσης των προαναφερθέντων προβλημάτων τα οποία δεν έχουν προφανή λύση (όπως έχει λ.χ. το tokenization), μιας και οι λύσεις των προβλημάτων αυτών χρησιμοποιούνται ως εργαλεία στη διαδικασία της εξαγωγής σχέσεων μεταξύ οντοτήτων, η οποία είναι και το θέμα της παρούσης εργασίας.

2.2.2 *Part-of-speech Tagging*

Το πρόβλημα της γραμματικής επισημείωσης (part-of-speech tagging) συνίσταται στην κατάταξη κάθε λέξης ενός κειμένου σε μια κατηγορία ανάλογα με το τι μέρος του λόγου είναι. Δεδομένης της φύσης του προβλήματος, οι λύσεις που έχουν προταθεί χρησιμοποιούν ταξινομητές διαφόρων τεχνολογιών (ταξινομητές k πλησιέστερων γειτόνων (k-NN classifiers), ταξινομητές μεγίστης εντροπίας (maximum entropy classifiers) κ.ά.).

Η ανάγκη χρήσης ταξινομητή προκύπτει από το γεγονός ότι για τις λέξεις ενός κειμένου ως μονάδες δεν μπορεί να αναγνωριστεί μονοσήμαντα το μέρος του λόγου στο οποίο ανήκουν, λόγω πολλών αμφισημιών (π.χ. η λέξη “αποταμιεύσεις”, ανάλογα με τα συμφραζόμενα, θα μπορούσε να λειτουργεί ως ρήμα ή ως ουσιαστικό).

Οι part-of-speech taggers, κάποιες φορές, προχωρούν ένα βήμα παραπάνω από την ταξινόμηση λέξης στο σωστό μέρος του λόγου: υπό προϋποθέσεις και ανάλογα με τη γλώσσα στην οποία είναι γραμμένο το κείμενο, μπορούν να μαντέψουν και πρόσθετες ιδιότητες της λέξης, όπως γένος, πτώση, αριθμό κ.λπ.

Αξίζει να αναφερθεί πως, όπως σε πολλά text information retrieval υποπροβλήματα, οι αποδοτικές υλοποιήσεις διαφέρουν από γλώσσα σε γλώσσα. Στην παρούσα εργασία χρησιμοποιήθηκε ένας ελληνικός part-of-speech tagger[37], ο οποίος ουσιαστικά κάνει χρήση ενός maximum entropy classifier.

2.2.3 *Stemming*

Το πρόβλημα του stemming (της αναγωγής στη ρίζα κάθε λέξης) απασχολεί τους επιστήμονες της πληροφορικής εδώ και πολύ καιρό (όχι μόνο στα πλαίσια του information retrieval, αλλά και ως εργαλείο σε μηχανές αναζήτησης και γενικότερα σε εφαρμογές που έχουν να αντιμετωπίσουν κείμενο). Κύριος στόχος του stemming είναι να αποφασίσει αν δυο λέξεις έχουν την ίδια ρίζα (οπότε και σε κάποιες εφαρμογές θεωρούνται “συνώνυμες”) ή όχι.

Παρακάτω παρουσιάζονται κάποιες προσεγγίσεις οι οποίες έχουν μελετηθεί ως λύσεις για το stemming:

- *lookup algorithms*: ψάχνουν τη λέξη σε έναν πίνακα, ο οποίος μπορεί να συμπληρώνεται όσο διαρκεί η διαδικασία του stemming με νέα παράγωγα (π.χ. για τα αγγλικά αν αναγνωριστεί η λέξη “run”, μπορούν να προστεθούν στον πίνακα οι λέξεις “running”, “runs” κ.λπ.). Κύριο μειονέκτημα των αλγορίθμων αυτών είναι πως νέες ή άγνωστες λέξεις δε μπορούν να αναγνωριστούν.

- *prefix/suffix-stripping algorithms*: απομονώνουν τη λέξη από πιθανά προθέματα/επιθέματα τα οποία έχει και ό,τι απομένει παρουσιάζεται ως ρίζα (stem). Μειονέκτημα της λογικής αυτής είναι ότι δε χειρίζεται σωστά λέξεις οι οποίες δεν έχουν κάποιο γνωστό και κωδικοποιημένο πρόθεμα/επίθεμα ή λέξεις οι οποίες παρουσιάζουν ανωμαλίες στην κλίση τους.

- *lemmatisation algorithms*: σε πρώτη φάση πραγματοποιούν part-of-speech tagging και στη συνέχεια, ανάλογα με το μέρος του λόγου στο οποίο ανήκει η λέξη, εφαρμόζουν διαφορετικούς κανόνες για την εύρεση της ρίζας. Είναι προφανές ότι σε περίπτωση λανθασμένου αποτελέσματος του part-of-speech tagger, το αποτέλεσμα του stemming θα είναι και αυτό λανθασμένο.

- *stochastic algorithms*: με τρόπο παρόμοιο με αυτόν που περιγράφηκε για τη γραμματική επισημείωση, οι στοχαστικοί αλγόριθμοι προσπαθούν με στατιστικά μοντέλα να προσδιορίσουν τη ρίζα της λέξης.

- *hybrid algorithms*: συνδυασμοί κάποιων από τις παραπάνω προσεγγίσεις δίνουν καλύτερα και πιο αξιόπιστα αποτελέσματα.

2.2.4 *Μηχανική Μάθηση (Machine Learning)*

Η χρήση τεχνικών μηχανικής μάθησης είναι ευρέως διαδεδομένη σε υλοποιήσεις information retrieval, αλλά και γενικότερα στην επεξεργασία φυσικής γλώσσας. Ήδη έχουμε αναφερθεί σε classifiers (στο πρόβλημα του part-of-speech tagging), οι οποίοι χρησιμοποιούν μηχανική μάθηση. Σε αντίθεση με προγενέστερες θεωρήσεις οι οποίες περιελάμβαναν την κωδικοποίηση ενός τεράστιου συνόλου από κανόνες για την αναγνώριση κειμένου, η χρήση μηχανικής μάθησης εξασφαλίζει ότι οι αλγόριθμοι "μαθαίνουν" από μόνοι τους τους κανόνες, εξετάζοντας ένα σύνολο κειμένων (corpus) το οποίο δεν έχει κάποια προτυποποιημένη μορφή.

Για τους σκοπούς της μηχανικής μάθησης έχουν προταθεί κατά καιρούς διάφορες προσεγγίσεις, με αυτές που βασίζονται σε στατιστικά μοντέλα να επικρατούν. Οι αλγόριθμοι που ακολουθούν στατιστικά μοντέλα παίρνουν ως είσοδο ένα σύνολο από "χαρακτηριστικά" οι οποίες έχουν εξαχθεί από το corpus και στη συνέχεια αναθέτουν υπολογίσιμα "βάρη" σε κάθε ιδιότητα. Το πλεονέκτημα της προσέγγισης αυτής είναι ότι ως έξοδος παράγεται μία σειρά αποκρίσεων (και όχι μία "τελική" απάντηση), καθώς και μία έκφραση της σχετικής βεβαιότητας των απαντήσεων αυτών. Με αυτόν τον τρόπο, αυξάνεται η αξιοπιστία του συστήματος και ως μεμονωμένου συστήματος, αλλά και όταν αυτό αποτελεί μέρος ενός μεγαλύτερου συστήματος.

Επιπλέον, τα συστήματα τα οποία χρησιμοποιούν τεχνικές machine learning παρουσιάζουν πλεονεκτήματα έναντι της θέσπισης κανόνων, τα οποία φαίνονται στον Πίνακα 1.

MACHINE LEARNING	HARD-CODED RULES
εστιάζουν αυτόματα στις πιο "συνηθισμένες" περιπτώσεις	δεν είναι εύκολο να βρεθεί πού πρέπει να δοθεί μεγαλύτερη έμφαση
μπορούν να χειριστούν corpora ασυνήθιστης μορφής ή/και θεματολογίας ή κείμενα με μικρά λάθη	πάρα πολύ δύσκολη η πρόβλεψη και ο χειρισμός ασυνήθιστης ή λανθασμένης εισόδου
η απόδοσή τους αυξάνεται, όσο δίνονται περισσότερα δεδομένα εισόδου	η απόδοσή τους είναι συγκεκριμένη, προκαθορισμένη από τους κανόνες

Πίνακας 1: Σύγκριση προσέγγισης machine learning και προσέγγισης προκαθορισμένων κανόνων

2.3 ΕΞΑΓΩΓΗ ΠΛΗΡΟΦΟΡΙΑΣ ΚΕΙΜΕΝΟΥ (TEXT INFORMATION EXTRACTION)

2.3.1 *Φάσεις text information extraction*

Θεωρούμε πως ένα σύστημα εξαγωγής πληροφορίας κειμένου, αφαιρετικά, εξυπηρετεί την απάντηση ερωτημάτων του χρήστη περί των εγγράφων του corpus. Ένα τέτοιο σύστημα, συνήθως αποτελείται από τρία υποσυστήματα (φάσεις επεξεργασίας): την αναπαράσταση του περιεχομένου των εγγράφων (indexing process), την αναπαράσταση της γνώσης η οποία αναμένεται ως αποτέλεσμα (query formulation process) και τη σύγκριση των δύο αναπαραστάσεων (matching process). Οι φάσεις της διαδικασίας του text information extraction φαίνονται σχηματικά στο Σχήμα 1.

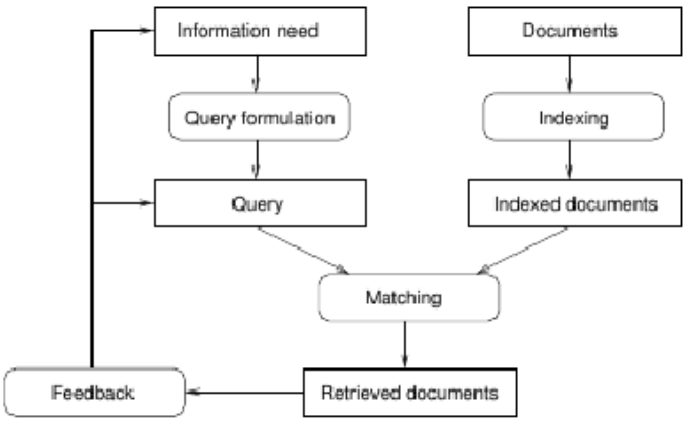
2.3.1.1 *Προεπεξεργασία (indexing process)*

Κατά τη διάρκεια του indexing process, το corpus αποκτά μια πιο εύκολα επεξεργάσιμη μορφή. Τυπικές διαδικασίες προεπεξεργασίας είναι:

- απομάκρυνση ανεπιθύμητων χαρακτήρων ή markup language tags (εάν το κείμενο είναι σε μορφή XML/HTML κ.λπ.),
- tokenization (χωρισμός σε λέξεις-tokens) του κειμένου,
- decapitalization των tokens (εφόσον χρειάζεται),
- stemming (εύρεση ρίζας - αφαίρεση καταλήξεων και προθεμάτων) των tokens του κειμένου,
- απομάκρυνση πολύ συνηθισμένων λέξεων (stopwords), η παρουσία των οποίων δεν επηρεάζει το αποτέλεσμα του information retrieval.

2.3.1.2 *Διαδικασία διατύπωσης ερωτήματος (query formulation process)*

Κατά την διαδικασία διατύπωσης ερωτήματος δημιουργείται η αναπαράσταση του «ερωτήματος» του χρήστη με τρόπο παρόμοιο με αυτόν του indexing process. Σε αυτή τη διαδικασία γίνεται προσπάθεια να κατανοήσει και ο χρήστης καλύτερα τι είναι αυτό που ζητά (μέσω της ανατροφοδότησης (feedback) που δίνεται από το σύστημα και τον επανακαθορισμό του ερωτήματος από το χρήστη, αν χρειαστεί).



Σχήμα 1: Διαδικασία Information Extraction.[18]

2.3.1.3 *Διαδικασία ταιριάσματος (matching process)*

Η σύγκριση μεταξύ των δυο αναπαραστάσεων που δημιουργήθηκαν κατά τις δύο προηγούμενες φάσεις, δίνει μια λίστα αποτελεσμάτων, ταξινομημένη σύμφωνα με κάποια βαθμολογία που έχει δώσει το σύστημα σε κάθε έγγραφο. Ιδανικά, η βαθμολογία αυτή είναι αρκετά ακριβής, ώστε να ελαχιστοποιείται η συμμετοχή του χρήστη στην εύρεση αποτελεσμάτων.

Για την απόδοση βαθμολογίας στα έγγραφα, χρησιμοποιούνται διάφορες τεχνικές, άλλες απλές όπως μέτρηση συχνότητας εμφάνισης όρων στα έγγραφα, και άλλες πιο σύνθετες όπως στατιστικά ή πιθανοτικά μοντέλα. Η σχεδίαση του αλγορίθμου βαθμολόγησης εγγράφων είναι ίσως το πιο σημαντικό κομμάτι της δημιουργίας ενός συστήματος εξαγωγής πληροφορίας, καθώς επηρεάζει σημαντικά την απόδοσή του.

Στη συνέχεια θα ασχοληθούμε με τα στατιστικά και τα πιθανοτικά μοντέλα, μιας και έχει παρατηρηθεί ότι παρουσιάζουν πολύ καλύτερη συμπεριφορά από άλλα.

2.4 ΜΟΝΤΕΛΑ INFORMATION EXTRACTION

2.4.1 *Στατιστικά Μουτέλα*

Το 1957, προτάθηκε για πρώτη φορά ένα μοντέλο εξόρυξης γνώσης [23], το οποίο χρησιμοποιούσε στατιστικές μεθόδους κατά τη matching process .

Σύμφωνα με τη στατιστική προσέγγιση, προκειμένου να γίνει αναζήτηση μέσα σε ένα σύνολο εγγράφων, πρέπει πρώτα να δημιουργηθεί ένα νέο έγγραφο, το οποίο θα μοιάζει αρκετά με τα έγγραφα που αναμένονται ως αποτελέσματα. Ο βαθμός ομοιότητας ανάμεσα στο έγγραφο που έχει δημιουργηθεί και στα έγγραφα του corpus, χρησιμοποιείται για να ταξινομήσει τα αποτελέσματα.

2.4.1.1 *Αναπαράσταση εγγράφων στα στατιστικά μοντέλα*

Ας υποθέσουμε ότι μετά την προεπεξεργασία του corpus, το οποίο αποτελείται από n έγγραφα, απομένουν m διαφορετικοί όροι (terms). Σύμφωνα με τα στατιστικά μοντέλα, οι αναπαραστάσεις του corpus και του query παριστάνονται με διανύσματα ως εξής:

> Σε κάθε term i, κάθε εγγράφου j, ανατίθεται ένα βάρος w_{ij}. Από αυτά τα βάρη συνθέτουμε m διανύσματα $\vec{d_j} = (w_{1j}, w_{2j}, ..., w_{ij}), 1i \in [1, m]$, $j \in [1, n]$ για κάθε έγγραφο. Αντίστοιχα, για την αναπαράσταση του query έχουμε το διάνυσμα $\vec{q} = (w_{q1}, w_{q2}, ..., w_{qi}), i \in [1, m]$.

Στην αναπαράσταση αυτή, θεωρούμε, χάριν απλότητας, πως οι όροι (terms) είναι ανεξάρτητοι μεταξύ τους.

Από τα παραπάνω προκύπτει πως σχέση μεταξύ των διανυσμάτων, υποδηλώνει σχέσεις μεταξύ εγγράφων ή μεταξύ εγγράφων και ερωτήματος. Στη συνέχεια θα αναφερόμαστε σε ένα (τυχαίο) από τα διανύσματα $\vec{d_j}, j \in [1, n]$ απλά ως $\vec{d}$.

2.4.1.2 *Στάθμιση όρων (term weighting)*

Στην προηγούμενη ενότητα δεν αναφερθήκαμε καθόλου στον τρόπο προσδιορισμού των βαρών w_{ij} των διανυσμάτων $\vec{d}$ και $\vec{q}$. Το πρόβλημα ανάθεσης βαρών είναι γνωστό ως στάθμιση όρων (term weighting), και όπως δείχνουν μελέτες ([30],[31]), δεν είναι εύκολα επιλύσιμο πρόβλημα.

ΔΙΑΦΟΡΕΣ ΠΡΟΤΑΣΕΙΣ που έχουν γίνει για τις τιμές των βαρών είναι οι εξής:

- *boolean*: τιμή 1 αν ο όρος υπάρχει στο κείμενο, 0 αν δεν υπάρχει.
- *TF (term frequency)*: συχνότητα εμφάνισης του όρου στο έγγραφο. Η ιδέα είναι ότι όσο περισσότερες φορές εμφανίζεται ένας όρος σε ένα κείμενο, τόσο αυξάνεται η πιθανότητα να σχετίζεται ο όρος με το κείμενο. Μειονεκτήματα της μετρικής αυτής είναι ότι μεροληπτεί υπέρ συνηθισμένων όρων και επίσης δε λαμβάνει υπόψη της το μέγεθος του εγγράφου (έτσι σε μεγάλα έγγραφα, στα οποία είναι λογικό να εμφανίζεται περισσότερες φορές ένα όρος, δίνεται μεγαλύτερη τιμή βάρους).
- *DF (document frequency)*: σε πόσα από τα έγγραφα του corpus εμφανίζεται ένας όρος. Η ιδέα εδώ είναι ότι αν ένας όρος εμφανίζεται σε πολλά έγγραφα, τότε δεν αποτελεί στοιχείο διάκρισης μεταξύ των εγγράφων.
- *TF-IDF (term frequency-inverse document frequency)*: συνδυασμός TF και IDF (αν ο όρος είναι συχνός ή σπάνιος στο corpus).

  Η τιμή του υπολογίζεται ως εξής:

  $$w_{ij} = TF - IDF(i, j, D) = TF(i, j) \times IDF(i, D) \quad (1)$$

  – TF είναι ο λόγος του αριθμού εμφάνισης ενός term σε ένα έγγραφο προς το συνολικό αριθμό terms του εγγράφου.

– IDF είναι ένα μέτρο του αν το term είναι συχό ή σπάνιο στο σύνολο των εγγράφων. Υπολογίζεται από τη σχέση

$$\mathrm{IDF}(i, D) = \log \frac{|D|}{|\{j \in D : i \in j\}|} \tag{2}$$

, όπου $|D|$ ο συνολικός αριθμός των εγγράφων στο corpus και $|\{j \in D : i \in j\}|$ ο αριθμός των εγγράφων στα οποία εμφανίζεται ο όρος t. Αν ο όρος δεν εμφανίζεται στο corpus θα έχουμε διαίρεση με το μηδέν, γι' αυτό συχνά χρησιμοποιείται ως παρανομαστής το $1 + |\{j \in D : i \in j\}|$.

Υψηλή τιμή TF-IDF υποδηλώνει πως ένας όρος εμφανίζεται πιο συχνά σε ένα συγκεκριμένο έγγραφο από το μέσο όρο και γι' αυτό η μετρική αυτή είναι αρκετά πιο αξιόπιστη από τις προαναφερθείσες.

Οι εργασίες των Salton και Yang ([30],[31]), πρότειναν τη χρήση TF-IDF βαρών σε στατιστικά μοντέλα εξαγωγής πληροφορίας. Αξίζει να σημειωθεί πως πολλοί σύγχρονοι weighting αλγόριθμοι είναι παραλλαγές των TF-IDF weighting αλγορίθμων.

2.4.1.3 *Μέτρα σύγκρισης εγγράφων*

Δεδομένης της αναπαράστασης των εγγράφων και του query, απομένει να καθοριστεί ένα κριτήριο με το οποίο θα υπολογίζεται η ομοιότητα των εγγράφων (μεταξύ τους αλλά και με το ερώτημα), προκειμένου ένα μοντέλο να είναι πλήρες. Ο βαθμός ομοιότητας επίσης μπορεί να χρησιμοποιηθεί για να ταξινομηθούν τα αποτελέσματα στην έξοδο ενός συστήματος.

ΚΑΠΟΙΑ ΜΕΤΡΑ ΤΑ ΟΠΟΙΑ ΧΡΗΣΙΜΟΠΟΙΟΥΝΤΑΙ φαίνονται παρακάτω:

- το *εσωτερικό γινόμενο* μεταξύ των δύο διανυσμάτων $\vec{d}, \vec{q}$ (simple matching):

$$\mathrm{score}(\vec{d_j}, \vec{q}) = \sum_{k=1}^{n} d_{kj} \cdot q_k \tag{3}$$

είναι το πρώτο μέτρο σύγκρισης το οποίο προτάθηκε [23].

- *το συνημίτονο της γωνίας* που σχηματίζουν τα δύο διανύσματα $\vec{d}, \vec{q}$ (cosine coefficient):

$$\mathrm{score}(\vec{d_j}, \vec{q}) = \frac{\sum_{k=1}^{n} d_{kj} \cdot q_k}{\sqrt{\sum_{k=1}^{n} (d_{kj})^2} \cdot \sqrt{\sum_{k=1}^{n} (q_k)^2}} \tag{4}$$

- το *Jaccard similarity coefficient*:

$$\mathrm{score}(\vec{d_j}, \vec{q}) = \frac{\sum_{k=1}^{n} d_{kj} \cdot q_k}{\sum_{k=1}^{n} (d_{kj})^2 \cdot \sum_{k=1}^{n} (q_k)^2} \tag{5}$$

- το *Dice similarity coefficient*:

$$score(\vec{d_j}, \vec{q}) = 2 \cdot \frac{\sum_{k=1}^{n} d_{kj} \cdot q_k}{\sum_{k=1}^{n} (d_{kj})^2 + \sum_{k=1}^{n} (q_k)^2} \tag{6}$$

- το *Overlap similarity coefficient*:

$$score(\vec{d_j}, \vec{q}) = \frac{\sum_{k=1}^{n} d_{kj} \cdot q_k}{\min\{\sum_{k=1}^{n} (d_{kj})^2, \sum_{k=1}^{n} (q_k)^2\}} \tag{7}$$

2.4.1.4 *Vector Space Model: ένα παράδειγμα στατιστικού μοντέλου*

Βασισμένοι σε προηγούμενη εργασία του Luhn, οι Salton και McGill,θέλοντας να δώσουν μια πιο θεωρητικά τεκμηριωμένη λύση, παρουσίασαν το 1983 το vector space model [32].

ΣΥΜΦΩΝΑ ΜΕ ΤΟ ΜΟΝΤΕΛΟ ΑΥΤΟ τα διανύσματα $\vec{d}, \vec{q}$ θεωρούνται διανύσματα ενός πολυδιάστατου Ευκλείδειου χώρου, στον οποίο κάθε διάσταση ανατίθεται σε ένα term.

ΩΣ ΜΕΤΡΟ ΣΥΓΚΡΙΣΗΣ για τη φάση του matching ορίζεται το συνημίτονο της γωνίας μεταξύ των δύο διανυσμάτων $\vec{d}$ και $\vec{q}$ (Σχέση 4).

Το συνημίτονο της γωνίας είναι 0 (καμία ομοιότητα), αν τα διανύσματα είναι ορθογώνια στον πολυδιάστατο χώρο και 1 αν η γωνία που σχηματίζουν είναι 0 (έχουν ίδια διεύθυνση).

Επειδή ο πολυδιάστατος χώρος είναι δύσκολο να αναπαρασταθεί, ας θεωρήσουμε ένα τριδιάστατο παράδειγμα. Έστω ότι έχουμε ένα έγγραφο, το οποίο μετά την προεπεξεργασία αποτελείται από τρία terms:

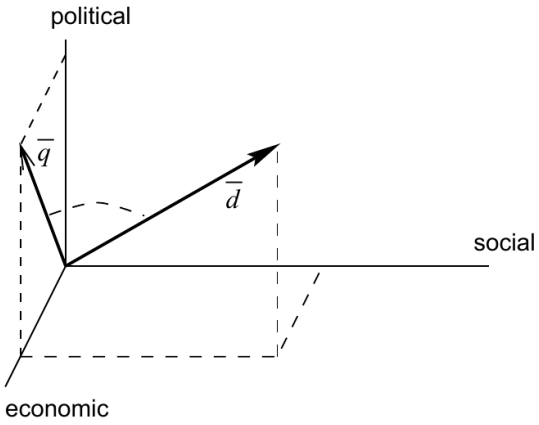
economic, political, social.

Η αναπαράσταση των δύο διανυσμάτων $\vec{d}$ και $\vec{q}$ σε τρισδιάστατο χώρο φαίνεται στο Σχήμα 2.

Παρότι η ιδέα του Ευκλείδειου χώρου βοήθησε αρκετά στην εποπτική διαχείριση του προβλήματος της εξαγωγής γνώσης, ουσιαστικά δεν έχει κάποια ιδιαίτερη θεωρητική βάση. Οι όροι, στην πραγματικότητα, δεν είναι ορθογώνιες διαστάσεις, μιας και εμφανίζουν αλληλεξαρτήσεις.

2.4.2 *Πιθανοτικά μοντέλα*

Στην προσπάθεια να οριστεί και να λυθεί με έναν κομψό και αυστηρό θεωρητικό τρόπο το πρόβλημα του information extraction και, κατά συνέπεια, το πρόβλημα



Σχήμα 2: Vector Space Model - παράδειγμα.

ΠΛΕΟΝΕΚΤΗΜΑΤΑ	ΜΕΙΟΝΕΚΤΗΜΑΤΑ
ισχυρή θεωρητική βάση	χρειάζονται πληροφορίες σχετικότητας (ή υπολογίζονται/μαντεύονται)
παρέχουν τις καλύτερες εκτιμήσεις για τη σχετικότητα εγγράφων	στοιχεία που υποδηλώνουν σχετικότητα μπορεί να μην είναι terms (αν και συνήθως λαμβάνονται υπόψη μόνο terms)
υλοποιούνται παρόμοια με το vector space model	

Πίνακας 2: Πλεονεκτήματα και Μειονεκτήματα probabilistic models

των βαρών (term weighting), πολλές προσεγγίσεις βασίστηκαν στη *θεωρία πιθανοτήτων*. Προϊόντα αυτής της θεώρησης είναι το probabilistic index model [24], το probabilistic retrieval model [29], το 2-Poisson model [10], τα Bayesian network models, τα language models κ.ά.

ΤΑ ΠΙΘΑΝΟΘΕΩΡΗΤΙΚΑ ΜΟΝΤΕΛΑ λειτουργούν με γνώμονα την probabilistic ranking principle του Robertson [29]:

> Αν η έξοδος ενός information extraction συστήματος είναι μία ταξινομημένη κατά φθίνουσα πιθανότητα χρησιμότητας συλλογή εγγράφων, όπου οι πιθανότητες υπολογίζονται όσο ακριβέστερα γίνεται για τα συγκεκριμένα δεδομένα, η συνολική αποδοτικότητα του συστήματος θα είναι η μέγιστη δυνατή για τα συγκεκριμένα δεδομένα.

Πρακτικά, αναλύουν τον υπολογισμό της πιθανότητας κάθε έγγραφο να σχετίζεται με το συγκεκριμένο ερώτημα το οποίο θέτει ο χρήστης. Στη συνέχεια, ταξινομούν τα αποτελέσματα κατά φθίνουσα σειρά πιθανότητας και τα παρουσιάζουν στο χρήστη.

Τα πιθανοθεωρητικά μοντέλα, ουσιαστικά, αντλούν την αξιοπιστία τους από την αξιοπιστία του υπολογισμού των πιθανοτήτων (στο μέτρο που αυτό είναι ρεαλιστικό).

2.4.2.1 *Binary independence retrieval model: ένα απλό πιθανοτικό μοντέλο*

Τα πιθανοτικά μοντέλα προσπαθούν να υπολογίσουν την πιθανότητα ένα συγκεκριμένο έγγραφο d_j να είναι σχετικό με το ερώτημα του χρήστη q, πιθανότητα η οποία συμβολίζεται ως $P(R|q, d_j)$. Για να κατανοηθεί καλύτερα ο τρόπος με τον οποίο χειρίζεται τη θεωρία πιθανοτήτων ένα τέτοιο μοντέλο, παρουσιάζεται ένα σχετικά απλό πιθανοτικό μοντέλο, το binary independence retrieval model (BIR model) [13].

Θεωρούμε ότι οι όροι (terms) είναι ποικιλοτρόπως κατανεμημένοι στα έγγραφα του corpus. Έστω $T = \{t_1, t_2, ..., t_n\}$ το σύνολο των terms σε όλο το corpus και d_j^T το σύνολο των terms, οι οποίοι εμφανίζονται στο έγγραφο d_j. Το σύνολο d_j^T μπορούμε να το συμβολίσουμε ως διάνυσμα $\vec{x} = (x_1, x_2, ..., x_n)$ όπου $x_i = 1$, αν $t_i \in d_j^T$ και $x_i = 0$ διαφορετικά.

Στη συνέχεια, μας ενδιαφέρει να διακρίνουμε τα έγγραφα τα οποία συντίθενται από διαφορετικό σύνολο όρων, γι' αυτό θα υπολογίσουμε την πιθανότητα $P(R|q,\vec{x})$, αντί της $P(R|q, d_j)$. Χρησιμοποιώντας τον κανόνα του Bayes, θα έχουμε:

$$\frac{P(R|q,\vec{x})}{P(\bar{R}|q,\vec{x})} = \frac{P(R|q)}{P(\bar{R}|q)} \cdot \frac{P(\vec{x}|R,q)}{P(\vec{x}|\bar{R},q)} \tag{8}$$

Θεωρώντας πως οι όροι δε σχετίζονται μεταξύ τους, η παραπάνω σχέση μετασχηματίζεται σε:

$$\frac{P(R|q,\vec{x})}{P(\bar{R}|q,\vec{x})} = \frac{P(R|q)}{P(\bar{R}|q)} \cdot \prod_{i=1}^{n} \frac{P(x_i|R,q)}{P(x_i|\bar{R},q)} = \frac{P(R|q)}{P(\bar{R}|q)} \cdot \prod_{x_i=1} \frac{P(x_i=1|R,q)}{P(x_i=1|\bar{R},q)} \cdot \prod_{x_i=0} \frac{P(x_i=0|R,q)}{P(x_i=0|\bar{R},q)} \tag{9}$$

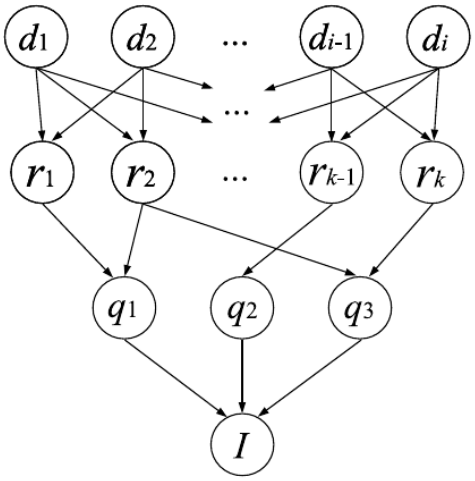
Για απλότητα στο φορμαλισμό, ονομάζουμε $p_i = P(x_i = 1|R,q)$ και $q_i = P(x_i = 1|\bar{R},q)$. Υποθέτουμε πως $p_i = q_i$ για όλους τους όρους οι οποίοι βρίσκονται στο σύνολο q^T, δηλαδή το σύνολο των όρων που απαρτίζουν το ερώτημα q. Έτσι, μετασχηματίζουμε περισσότερο τη σχέση , η οποία γίνεται:

$$\frac{P(R|q,\vec{x})}{P(\bar{R}|q,\vec{x})} = \frac{P(R|q)}{P(\bar{R}|q)} \cdot \prod_{t_i \in d_m^T \cap q^T} \frac{p_i(1-q_i)}{q_i(1-p_i)} \cdot \prod_{t_i \in q^T} \frac{1-p_i}{1-q_i} \tag{10}$$

Το δεύτερο από τα δύο γινόμενα της σχέσης αυτής, όπως και το πρώτο κλάσμα είναι σταθερά για δεδομένο ερώτημα. Συνεπώς, θα μπορούσαμε να τα αγνοήσουμε, μιας και μας ενδιαφέρει η σύγκριση των εγγράφων για συγκεκριμένο ερώτημα. Τελικά, η βαθμολογία η οποία δίδεται στα έγγραφα δίνεται από τον τύπο:

$$score(d_j) = \sum_{t_i \in d_j^T \cap q^T} \log\left(\frac{p_i(1-q_i)}{q_i(1-p_i)}\right) \tag{11}$$

Μετά την εμφάνιση του binary independence retrieval model, ακολούθησαν διάφορα μοντέλα, τα οποία λειτουργούσαν με παρόμοιο τρόπο (είτε "binary", δηλαδή εξετάζοντας την εμφάνιση ή μη ενός όρου σε έγγραφα, είτε χρησιμοποιώντας



Σχήμα 3: Inference Network - παράδειγμα.

"βάρη"). Όπως αποδείχθηκε πειραματικά όμως, δεν παρουσίαζαν σημαντική βελτίωση έναντι του binary independence model.

2.4.2.2 *Bayesian Network Models*

Καθώς ο αριθμός των παραμέτρων και των τιμών πιθανοτήτων που χρειαζόταν να υπολογιστούν στα πιθανοτικά μοντέλα αύξανε όλο και περισσότερο, δημιουργήθηκε η ανάγκη ενός πιο οργανωμένου και αυστηρού μοντέλου. Το 1991, προτάθηκε το inference network model [34], το οποίο βασίζεται στο μηχανισμό του Bayesian Network.

ΤΑ BAYESIAN NETWORKS είναι ακυκλικοί κατευθυνόμενοι γράφοι οι οποίοι κωδικοποιούν σχέσεις πιθανοτικής εξάρτησης μεταξύ τυχαίων μεταβλητών. Η παρουσίαση των πιθανοτικών εξαρτήσεων με το φορμαλισμό του γράφου, επιτρέπει την ανάλυση σύνθετων αλληλοεξαρτήσεων και σχέσεων, αφού όλα τα προβλήματα ανάγονται σε προβλήματα της θεωρίας γραφημάτων.

ΠΡΑΚΤΙΚΑ ένα inference network model οργανώνεται σε τέσσερα επίπεδα:

- επίπεδο κόμβων εγγράφων d
- επίπεδο κόμβων αναπαράστασης r
- επίπεδο κόμβων ερωτημάτων q
- κόμβο information need I

Στο Σχήμα 3 φαίνεται ένα παράδειγμα απλοποιημένου inference network model. Όλοι οι κόμβοι σε αυτό το γράφημα αναπαριστούν τυχαίες δυαδικές μεταβλητές.

ΓΙΑ ΝΑ ΓΙΝΕΙ ΚΑΤΑΝΟΗΤΟΣ ο τρόπος λειτουργίας του μοντέλου, ας εστιάσουμε σε ένα υποσύνολο κόμβων, για παράδειγμα το r_2, q_1, q_3, I, αγνοώντας τους υπόλοιπους. Η από κοινού πιθανότητα των κόμβων που επιλέξαμε είναι

$$P(r_2, q_1, q_3, I) = P(r_2)P(q_1|r_2)P(q_3|r_2, q_1)P(I|r_2, q_1, q_3) \tag{12}$$

Υπάρχουν δύο τρόποι να ικανοποιείται το ενδεχόμενο «η ανάγκη για γνώση ικανοποιήθηκε» ($I = 1$): ο query node q_1 είναι αληθής ή ο query node q_3 είναι αληθής. Όμως, η αλήθεια των ερωτημάτων q_1, q_3 βασίζεται στον κόμβο r_2. Άρα, η πιθανότητα μπορεί να εκφραστεί ως:

$$P(r_2, q_1, q_3, I) = P(r_2)P(q_1|r_2)P(q_3|r_2)P(I|q_1, q_3) \tag{13}$$

Εδώ έχουμε κάνει τις εξής απλοποιήσεις: τα q_1, q_3 είναι ανεξάρτητα δεδομένου του "γονέα" r_2, οπότε $P(q_3|r_2, q_1) = P(q_3|r_2)$ και επίσης το I είναι ανεξάρτητο του r_2, δεδομένων των γονέων του q_1 και q_3, οπότε $P(I|r_2, q_1, q_3) = P(I|q_1, q_3)$.

Γίνεται αντιληπτό ότι ο υπολογισμός των τιμών πιθανοτήτων που χρειάζεται να γίνει είναι χρονοβόρος για μεγάλα δίκτυα και, φυσικά, αυξάνει εκθετικά με τον αριθμό κόμβων. Για το λόγο αυτό, σε όλα τα επίπεδα κόμβων γίνεται κάποιου είδους approximation, προκειμένου το σύστημα να καθίσταται λειτουργικό. Οι Metzler και Croft [25] περιγράφουν τα παρακάτω approximations:

> για το επίπεδο κόμβων εγγράφων, θεωρούμε ότι ασχολούμαστε με ένα μόνο έγγραφο κάθε φορά (για κάθε έγγραφο δημιουργούμε ένα ξεχωριστό inference model, στο οποίο αγνοείται το επίπεδο κόμβων εγγράφων). Για το επίπεδο κόμβων αναπαράστασης, οι πιθανότητες υπολογίζονται, ενώ για το επίπεδο κόμβων ερωτημάτων οι πιθανότητες προσεγγίζονται από κάποιες πιθανοτικές κατανομές με τη χρήση "believe operators". Αυτοί οι τελεστές συνδυάζουν τις τιμές πιθανοτήτων από τους κόμβους αναπαράστασης και από άλλους κόμβους ερωτημάτων με προκαθορισμένο τρόπο.

Αποδεικνύεται ότι με τη χρήση των "believe operators" μπορούμε να καταλήξουμε σε υπολογίσιμη έκφραση για τις πιθανότητες που αναζητούμε.

2.4.2.3 *Probabilistic Models: Logistic Regression*

Σε πολλά πιθανοτικά μοντέλα, ο υπολογισμός των τιμών πιθανοτήτων χρησιμοποιεί logistic regression. Αν θεωρήσουμε πως:

- D όλα τα έγγραφα
- Q όλα τα ερωτήματα χρηστών
- (D_j, Q_k) ζεύγος εγγράφου-ερωτήματος
- x κλάση παρομοίων εγγράφων ($x \subseteq D$)
- y κλάση παρομοίων ερωτημάτων ($y \subseteq Q$)

Η σχετικότητα εγγράφων ορίζεται σα σχέση ως εξής $R = \{(D_i, Q_j)|D_i \in D, Q_j \in Q$, το έγγραφο D_i κρίνεται σε σχέση με το ερώτημα $Q_j\}$.

Η πιθανότητα μπορεί να υπολογιστεί ως:

$$P(R|Q_k, D_j) = c_0 + \sum_{i=1}^{6} c_i \cdot X_i \tag{14}$$

όπου για τον προσδιορισμό των συντελεστών c_i χρησιμοποιείται logistic regression και X_i είναι τα παρακάτω έξι μέτρα:

- X_1: μέση απόλυτη συχνότητα στο ερώτημα (average absolute query frequency)
- X_2: μήκος ερωτήματος (query length)
- X_3: μέση απόλυτη συχνότητα στο έγγραφο (average absolute document frequency)
- X_4: μήκος εγγράφου (document length)
- X_5: μέσο IDF (average inverse document frequency)
- X_6: αριθμός όρων που είναι κοινοί σε έγγραφο και ερώτημα

2.5 ΕΞΑΓΩΓΗ ΣΧΕΣΕΩΝ ΜΕΤΑΞΥ ΟΝΤΟΤΗΤΩΝ (RELATION EXTRACTION)

Η εξαγωγή σχέσεων μεταξύ οντοτήτων (relation extraction) είναι μία από τις εργασίες text information retrieval, η οποία συνίσταται στην αναγνώριση συγκεκριμένων σημασιολογικών σχέσεων μεταξύ δύο ή περισσότερων οντοτήτων σε ένα κείμενο. Για παράδειγμα, αν ως είσοδος έχει δοθεί η πρόταση

«Ο Γιάννης εργάζεται στην Εταιρία1»

η επιθυμητή έξοδος από ένα σύστημα relation extraction είναι

εργάζεταιΣε(Γιάννης, Εταιρία1)

ή πιο γενικά

εργάζεταιΣε(Άνθρωπος, Εταιρία).

2.5.1 *Τεχνικές Relation Extraction*

Για να επιτύχουμε εξαγωγή σχέσεων μεταξύ οντοτήτων, εφαρμόζουμε τεχνικές ανάλογα με το είδος του σώματος κειμένου (corpus) στο οποίο εργαζόμαστε, αλλά και τη γνώση την οποία θέλουμε να εξάγουμε.

Πολλές φορές, ιδιαίτερα όταν το corpus έχει συγκεκριμένη θεματολογία, η εξαγωγή σχέσεων είναι σχετικά απλή. Στην περίπτωση αυτή, χρησιμοποιούμε Traditional Relation Extraction τεχνικές. Ωστόσο, όταν το corpus είναι πολύ μεγάλο και ετερογενές, η τεχνική αυτή δε μπορεί να εφαρμοστεί. Γι' αυτό καταφεύγουμε στην τεχνική του Open Relation Extraction

2.5.1.1 *Traditional Relation Extraction*

Ας υποθέσουμε ότι το το corpus το οποίο θέλουμε να επεξεργαστούμε αποτελείται από φύλλα οικονομικής εφημερίδας. Οι σχέσεις που (πιθανώς) περιγράφονται σε ένα τέτοιο σώμα έχουν να κάνουν με κινήσεις δεικτών χρηματιστηρίων, αγορές/πωλήσεις/συγχωνεύσεις εταιριών κ.λπ.

ΕΝΑ ΣΥΣΤΗΜΑ TRADITIONAL RELATION EXTRACTION δέχεται τις προς αναζήτηση σχέσεις ως είσοδο, όπως και κάποια προκατασκευασμένα patterns ανίχνευσης σχέσεων, τα οποία έχουν ανακαλυφθεί χειροκίνητα ([12], [28], [5]). Τα patterns αυτά, αφορούν τις συγκεκριμένες σχέσεις τις οποίες αναζητούμε.

Στο παράδειγμά μας, ένα traditional relation extraction σύστημα θα έψαχνε να βρει σχέσεις της μορφής

κίνησηΔείκτη(δείκτης, κίνηση)
αγορά(αγοραστής,αγορασθείσαΕταιρεία)
συγχώνευση(εταιρία1, εταιρία2)

,αγνοώντας άλλες πιθανές σχέσεις που δεν τις είχαμε προσδιορίσει κατά την εκτέλεση (π.χ. κάποια πολιτική εξέλιξη η οποία προκάλεσε πτώση του χρηματιστηρίου).

Όπως μπορεί να κατανοήσει κανείς, η χειροκίνητη εργασία που απαιτείται σε ένα τέτοιο σύστημα, αυξάνει γραμμικά συναρτήσει του αριθμού των προς εξαγωγή σχέσεων.

2.5.1.2 *Open Relation Extraction*

Η προαναφερθείσα τεχνική του Traditional Relation Extraction δε μπορεί να λειτουργήσει σε corpora από τα οποία ο αριθμός των διαφορετικών σχέσεων που περιμένουμε είναι μεγάλος, και/ή δε γνωρίζουμε εκ των προτέρων ποιες είναι οι σχέσεις που αναζητούμε. Σε αυτήν την περίπτωση, χρησιμοποιείται η τεχνική του Open Relation Extraction, με την οποία μπορούμε να πάρουμε σχέσεις από ένα corpus, χωρίς να χρειάζεται να προσδιορίσουμε από πριν ποιες είναι αυτές.

Τα open relation extraction συστήματα χρησιμοποιούν τεχνικές μηχανικής μάθησης για να επιτύχουν το σκοπό τους. Αρχικά, αναλύουν ένα (σχετικά) μικρό σύνολο προτάσεων · ανάλυση η οποία στη συνέχεια χρησιμοποιείται για την εκπαίδευση κάποιου classifier. Με τη βοήθεια εργαλείων όπως tokenizers, POS taggers κ.ά. προσδιορίζονται πιθανές σχέσεις από όλο το corpus. Στη συνέχεια ο classifier εφαρμόζεται στο σύνολο των πιθανών σχέσεων και αποφασίζει αν είναι αξιόπιστες (αν πραγματικά είναι σχέσεις) ή όχι.

ΕΝΑ ΣΥΣΤΗΜΑ OPEN RELATION EXTRACTION δέχεται ως μοναδική είσοδο ένα corpus και παράγει ως έξοδο ένα σύνολο σχέσεων.

Μία σύγκριση μεταξύ taditional και open relation extraction φαίνεται στον Πίνακα 3.

	TRADITIONAL	OPEN
είσοδος:	corpus, hand-labelled data	corpus
σχέσεις:	προσδιορισμένες από πριν	ανακαλύπτονται αυτόματα
πολυπλοκότητα:	$\mathcal{O}(D \cdot R)$	$\mathcal{O}(D)$

Πίνακας 3: Σύγκριση open και traditional relation extraction, όπου D ο αριθμός των εγγράφων και R ο αριθμός των αναμενόμενων σχέσεων.

Γίνεται αντιληπτό ότι μια τέτοια προσέγγιση μπορεί να εφαρμοστεί σε αρκετά μεγάλα corpora όπως φύλλα εφημερίδων γενικής θεματολογίας, ή σε ένα σύνολο ετερογενών σελίδων του internet. Η δυνατότητα ενός συστήματος information extraction να απευθύνεται στο σύνολο (ή σε υποσύνολα) του internet, παρουσιάζει εξαιρετικό ενδιαφέρον, μιας και

> το web σταδιακά μετατρέπεται από ιστό παγκόσμιας πληροφόρησης διασυνδεδεμένων εγγράφων σε έναν ιστό στον οποίο τόσο τα έγγραφα, όσο και τα δεδομένα είναι διασυνδεδεμένα [9].

Μέχρι να συμβεί κάτι τέτοιο, αποτελείται, στο μεγαλύτερο κομμάτι του, από αδόμητα έγγραφα τα οποία δεν εμπεριέχουν καθόλου semantic metadata. Η γνώση η οποία περιέχεται στα έγγραφα αυτά, θα μπορούσε να γίνει πιο εύκολα προσβάσιμη αν μετατρέπονταν σε κάποιο «σχεσιακό» format (όπως έχει προταθεί με τη δημιουργία της OWL[4]). Στη μετατροπή αυτή μπορεί να συμβάλει η εξαγωγή σχέσεων.

2.6 ΑΞΙΟΛΟΓΗΣΗ ΣΥΣΤΗΜΑΤΩΝ INFORMATION EXTRACTION

2.6.1 *Μέτρα αξιολόγησης*

Υπάρχουν πολλοί τρόποι με τους οποίους θα μπορούσε κανείς να αξιολογήσει ένα σύστημα information extraction, όπως για παράδειγμα η ευκολία χρήσης, ο τρόπος παρουσίασης των αποτελεσμάτων κ.λπ. Για να μετρήσουμε την αποδοτικότητα του συστήματος, μπορούμε να χρησιμοποιήσουμε πολλά μέτρα. Ωστόσο, τα πιο συνηθισμένα είναι:

- *recall*: ποσοστό σωστής/σχετικής πληροφορίας η οποία ανακτήθηκε

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \tag{15}$$

- *precision*: ποσοστό πληροφορίας που ανακτήθηκε και είναι σωστή/σχετική

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|} \tag{16}$$

- *f-measure* (σταθμισμένος αρμονικός μέσος): μέτρο, το οποίο χρησιμοποιείται συχνά και αποτελεί συνδυασμό precision και recall. Το f-measure ορίζεται ως εξής:

$$F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{17}$$

Η σχέση αυτή δίνει ως αποτέλεσμα αυτό που αποκαλούμε F_1, δηλαδή το f-measure για το οποίο τα precision και recall ζυγίζονται ισομερώς. Στη γενική περίπτωση, για $\beta \geqslant 0, \beta \in \mathbb{R}$, έχουμε τη σχέση

$$F_\beta = \frac{(1+\beta^2) \cdot (precision \cdot recall)}{\beta^2 \cdot precision + recall} \tag{18}$$

Για παράδειγμα, αν $\beta = 2$, τότε παίρνουμε το F_2, δηλαδή το f-measure το οποίο δίνει διπλάσια βαρύτητα στο recall, απ' ότι στο precision.

Ενδιαφέρον παρουσιάζουν και οι ορισμοί του precision και recall, από πιθανοθεωρητική σκοπιά:

- *precision*: η πιθανότητα ένα (τυχαία επιλεγμένο) έγγραφο από αυτά που εξήχθησαν από το σύστημα, να είναι σχετικό με το ερώτημα.
- *recall*: η πιθανότητα ένα (τυχαία επιλεγμένο) σχετικό με το ερώτημα έγγραφο, να έχει εξαχθεί ως αποτέλεσμα από το σύστημα.

2.6.2 *Καμπύλες αξιολόγησης*

Τα precision και recall είναι αλληλοεξαρτώμενα μεγέθη. Για το λόγο αυτό, μετρούμε precision για διαφορετικά επίπεδα recall και δημιουργούμε καμπύλες precision-recall, όπως αυτές που φαίνονται στο Σχήμα 4, προκειμένου να έχουμε μια εποπτική εικόνα της απόδοσης του information extraction συστήματος. Αξίζει να σημειωθεί κάτι που διακρίνεται εύκολα στο σχήμα αυτό: μόνο η παρουσία των καμπυλών δεν είναι αρκετή για να αποφασίσουμε ποιο σύστημα είναι καλύτερο.

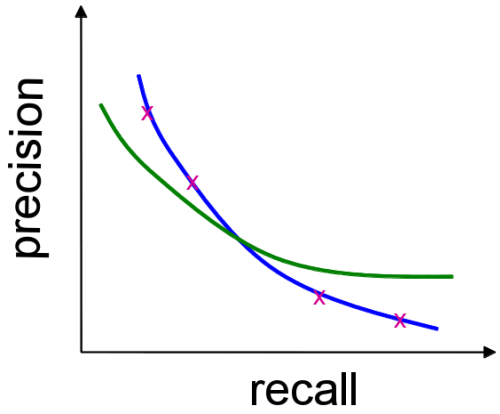
Μιας και τα precision και recall, ως απλοί αριθμοί, δεν φέρουν κάποια ιδιαίτερη πληροφορία για το σύστημα που εξετάζουμε, πολλές φορές, χρησιμοποιούνται παραδοχές προκειμένου να έχουμε πιο "αντικειμενικά" και συγκρίσιμα αριθμητικά αποτελέσματα. Για παράδειγμα, θα μπορούσε κανείς να μετρήσει το precision για συγκεκριμένο αριθμό εξαγόμενων εγγράφων (top 5, top 10, top 20 κ.λπ.), και να εξάγει το σταθμισμένο μέσο των αποτελεσμάτων (εστιάζοντας έτσι σε high precision αποτελέσματα.

2.7 OPEN INFORMATION EXTRACTION: ΣΗΜΑΝΤΙΚΕΣ ΕΡΓΑΣΙΕΣ

Η έννοια του open information extraction εισήχθη πρόσφατα, με τη δημιουργία του TextRunner. Μέχρι το 2007, τα information extraction συστήματα χρησιμοποιούσαν traditional τεχνικές ή τεχνικές οι οποίες, αν μη τι άλλο, χρειάζονταν ως είσοδο προκαθορισμένες σχέσεις προκειμένου να λειτουργήσουν.

2.7.1 TextRunner

Το πρώτο open information extraction σύστημα που δημιουργήθηκε ήταν ο TextRunner [8]. Στο paper στο οποίο παρουσιάστηκε, φαίνονται αποτελέσματα από ένα corpus 9000000 σελίδων του internet. Στο corpus αυτό, ο TextRunner είχε καλύτερη

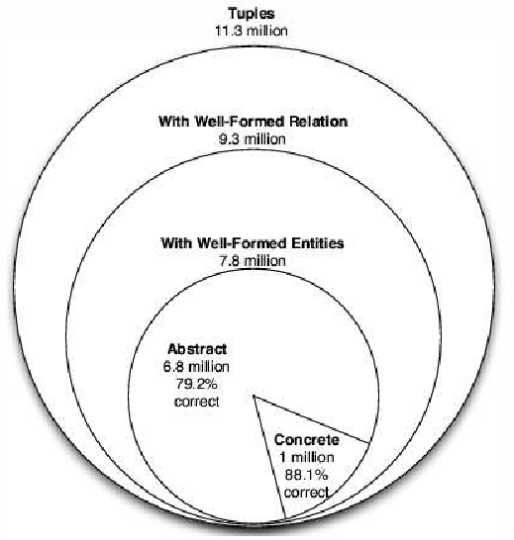


Σχήμα 4: Καμπύλη precision-recall.

απόδοση από τον traditional προκάτοχό του, το σύστημα KnowItAll [16], επιτυγχάνοντας 33% error reduction και εξάγοντας πολύ περισσότερες σχέσεις.

Ο TextRunner, χρησιμοποιεί τρία βασικά modules:

- *self-supervised learner*: με δεδομένο ένα δείγμα από το corpus, ο learner δίνει ως έξοδο έναν classifier, ο οποίος μαρκάρει τις υποψήφιες σχέσεις ως trustworthy ή όχι.
  - μαρκάρει αυτόματα τα training data ως trustworthy ή όχι (χωρίς χειροκίνητη παρέμβαση)
  - χρησιμοποιεί τα training data, τα οποία εξήχθησαν από ένα δείγμα του corpus, προκειμένου να εκπαιδεύσει έναν Naive Bayes Classifier, ο οποίος θα δράσει επί ολοκλήρου του corpus.
- *single-pass extractor*: ο extractor με ένα πέρασμα, εξάγει από το corpus τις υποψήφιες σχέσεις. Για να το επιτύχει αυτό, χρησιμοποιεί έναν part-of-speech tagger μέγιστης εντροπίας και στη συνέχεια έναν noun phrase chunker για να εντοπίσει τις οντότητες και να αποδώσει σε κάθε λέξη μια τιμή πιθανότητας να συμμετέχει στην οντότητα. Κατά τη διαδικασία αυτή, και με τη χρήση ευριστικών, απομακρύνονται προσδιοριστικές προτάσεις ή φράσεις, των οποίων η απουσία δεν αλλοιώνει την εύρεση αληθών σχέσεων (π.χ. η πρόταση “Scientists from many universities are studying...” μπορεί να γίνει “Scientists are studying...”).
- *redundancy-based assessor*: ο assessor αναθέτει μία τιμή πιθανότητας (κατά πόσον είναι trustworthy) σε κάθε σχέση, με βάση το probabilistic model of redundancy [15]. Ο assessor, χρησιμοποιεί εκτός άλλων, και τον αριθμό που υποδηλώνει πόσες φορές εμφανίστηκε μία μεμονωμένη σχέση σε όλο το corpus.



Σχήμα 5: Απόδοση TEXTRUNNER.

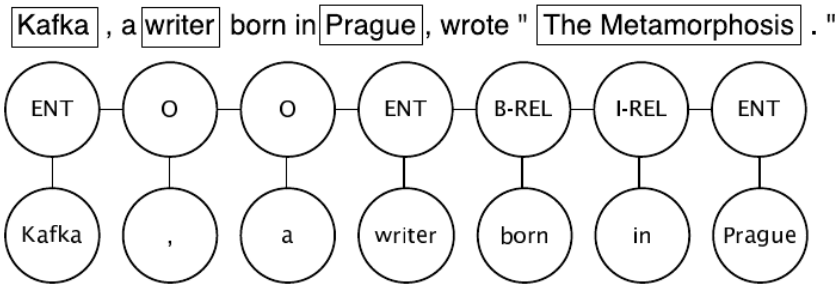
2.7.2 O – CRF

Με την εμφάνιση του TEXTRUNNER, το πρόβλημα του information extraction παρουσιάστηκε ως classification πρόβλημα (μιας και στο επίκεντρο της λύσης ήταν ένας Naive Bayes Classifier).

Οι classifiers μπορούν να κάνουν πρόβλεψη για μία μεταβλητή. Η ιδέα ότι για να μοντελοποιήσουμε περισσότερες αλληλοεξαρτώμενες μεταβλητές, μπορούμε να χρησιμοποιήσουμε γραφικά μοντέλα, οδήγησε στο σύστημα O-CRF [7]. Το σύστημα αυτό χρησιμοποίησε Conditional Random Fields (CRF) [22], δηλαδή μη-κατευθυνόμενα γραφικά μοντέλα εκπαιδευμένα να μεγιστοποιούν τη δεσμευμένη πιθανότητα ενός πεπερασμένου συνόλου labels Y, δεδομένου ενός συνόλου παρατηρήσεων X. Στο σύστημα O-CRF:

- η διαδικασία training είναι self-supervised, όπως και αυτή του TEXTRUNNER. Με τη βοήθεια ευριστικών εξάγονται κάποιες σχέσεις, οι οποίες χρησιμοποιούνται στην εκπαίδευση ενός CRF.
- στη συνέχεια, το O-CRF με ένα μόνο πέρασμα του corpus και με τη χρήση ενός phrase chunker εντοπίζει οντότητες και σχέσεις. Ύστερα ο CRF αποφασίζει αν είναι trustworthy σχέσεις ή όχι.
- τέλος, το O-CRF χρησιμοποιεί τον αλγόριθμο RESOLVER [36] για να εντοπίσει συνώνυμα σχέσεων. Ο RESOLVER αλγόριθμος κάνει χρήση ενός πιθανοτικού μοντέλου για να προβλέψει αν δυο strings αναφέρονται στο ίδιο αντικείμενο.

Το σύστημα O-CRF παρουσιάζει ελαφρώς υψηλότερο precision από τον προκάτοχό του, το TEXTRUNNER, και διπλάσιο recall.



Σχήμα 6: Relation Extraction ως Sequence Labeling: Χρήση CRF για τον εντοπισμό της σχέσης "born in" ανάμεσα στις οντότητες Kafka και Prague.

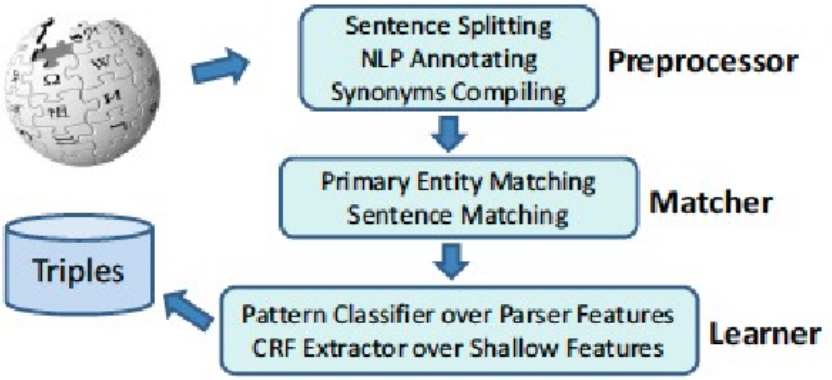
2.7.3 WOE^{pos} και WOE^{parser}

Το σύστημα WOE [35] εκμεταλλεύεται τη δομή της wikipedia για να αυξήσει την απόδοσή του έναντι άλλων συστημάτων information retrieval. Για τη δημιουργία training data ταιριάζει ευριστικά πληροφορίες από σελίδες της wikipedia με πληροφορίες από infoboxes της DBpedia. Στο WOE σύστημα:

- ο *preprocessor* μετατρέπει το κείμενο από τις σελίδες της wikipedia σε προτάσεις και εκτελεί POS-tagging και NP chunking.
- ο *matcher* δημιουργεί τα training data συγκρίνοντας τις πληροφορίες από τη wikipedia page με αυτές από το σχετικό infobox. Για παράδειγμα, από την entry της wikipedia για το Stanford University, θα μπορούσε να ταιριάξει το <established,1891> από το infobox με την πρόταση "The university was founded in 1891 by...", δημιουργώντας έτσι την «υποψήφια» σχέση <arg_1=Staandford University, rel=???, arg_2=1981.
- ο *extractor* τελικά εξάγει όλες τις σχέσεις από το σύνολο του corpus.
  - WOE^{parser}: *extraction με parser features*
    Οι σχέσεις εξάγονται κατά τη διάρκεια του parsing, με χρήση πολλών και σχετικά πολύπλοκων patterns. Το σύστημα WOE^{parser} παρουσιάζει 79% με 90% υψηλότερο f-measure από το TextRunner, αλλά είναι 30 φορές πιο αργό.
  - WOE^{pos}: *extraction με shallow patterns*
    Με παρόμοια λογική με αυτή που δουλεύει ο TextRunner, η εξαγωγή γίνεται με χρήση POS-tagging, κανονικών εκφράσεων κ.λπ. Το σύστημα WOE^{pos} έχει 15% με 34% υψηλότερο f-measure από το TextRunner και περίπου ίδια ταχύτητα με αυτό.

2.7.4 ReVerb

Το πιο σύγχρονο και state-of-the-art σύστημα text infromation extraction, το ReVerb, χρησιμοποιεί μια κάπως διαφορετική, πιο «ολιστική» λογική.



Σχήμα 7: Αρχιτεκτονική του συστήματος WOE.

ΕΝΑ ΑΠΟ ΤΑ ΣΦΑΛΜΑΤΑ ΠΟΥ ΕΜΦΑΝΙΖΑΝ ΤΑ ΠΡΟΗΓΟΥΜΕΝΑ ΣΥΣΤΗΜΑΤΑ (O-CRF, WOE, TEXTRUNNER) ήταν τα incoherent extractions και τα uninformative extractions. Με τον όρο incoherent extractions εννοούμε σχέσεις οι οποίες εξάγονται από το σύστημα, αλλά δεν έχουν κάποιο νόημα (π.χ. από την πρόταση "The Mark 14 was central to the torpedo scandal of the fleet" η σχέση (was central,torpedo)). Με τον όρο uninformative extractions εννοούμε σχέσεις οι οποίες εξάγονται από το σύστημα και παραλείπουν σημαντικές πληροφορίες (π.χ. από την πρόταση "Faust made a deal with the devil" η σχέση (Faust, made, a deal)). Αυτό το είδος σφάλματος προκύπτει από λανθασμένο χειρισμό σχεσιακών φράσεων οι οποίες εκφράζονται με συνδυασμό ρήματος/ουσιαστικού.

Το σύστημα REVERB λαμβάνει υπόψη δύο περιορισμούς σχετικά με τις φράσεις που εκφράζουν σχέσεις:

- *syntactic constraint*: ο περιορισμός αυτός απαιτεί η ρηματική σχέση να ακολουθεί το POS tag pattern που φαίνεται στο Σχήμα 8. Αν σε κάποια πρόταση ικανοποιούνται περισσότερα από ένα τμήματα της κανονικής αυτής έκφρασης, επιλέγεται η μακρύτερη. Επίσης, αν η κανονική έκφραση αυτή ταιριάζει σε περισσότερες από μία κοντινές εκφράσεις, αυτές συγχωνεύονται σε μία ρηματική έκφραση.

- *lexical constraint*: ο περιορισμός αυτός απαιτεί η εξαγόμενη σχέση να μην είναι υπερπροσδιορισμένη σε βαθμό που να καθίσταται άχρηστη. Για παράδειγμα, από την πρόταση

  The Obama administration is offering only modest greenhouse gas reduction targets at the conference.

  η προαναφερθείσα κανονική έκφραση θα απέδιδε ως ρηματική φράση τη φράση

  is offering only modest
  greenhouse gas reduction targets at

  και ως οντότητες το ζεύγος *(Obama administration, conference)*, κάτι που όπως εύκολα διαπιστώνει κανείς δεν είναι ιδιαίτερα χρήσιμο ως γνώση.

$$V \mid VP \mid VW^*P$$
V = verb particle? adv?
W = (noun | adj | adv | pron | det)
P = (prep | particle | inf. marker)

Σχήμα 8: POS tag pattern για το syntactic constraint.

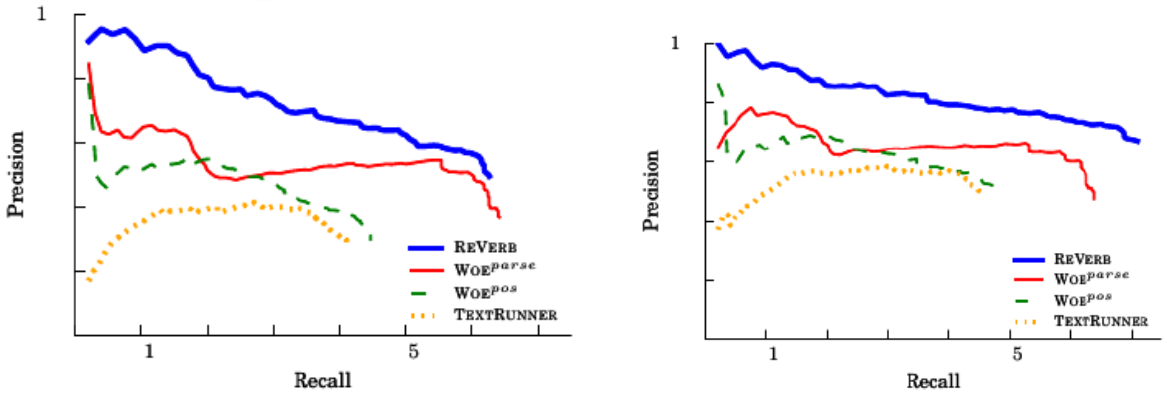
Το σύστημα ReVerb προκειμένου να αποφύγει τα παραπάνω σφάλματα, πραγματοποιεί σε κάθε πρόταση:

- *relation extraction*: ξεκινώντας από τα ρήματα, προσδιορίζει τις σχέσεις ως «ρηματικές εκφράσεις», συμπεριλαμβάνοντας περισσότερες από μία λέξεις σε κάθε φράση.
- *argument extraction*: για κάθε σχέση που προσδιορίστηκε στο προηγούμενο βήμα, προσπαθεί να προσδιορίσει τις ονοματικές εκφράσεις που συμπληρώνουν τη σχέση, δεξιά και αριστερά από τη ρηματική έκφραση.
- *confidence function*: προκειμένου να έχουμε υψηλό precision, εκπαιδεύεται ένας logistic regression classifier ο οποίος αναθέτει σε κάθε σχέση μία τιμή πιθανότητας.

2.7.5 *Σύγκριση προηγούμενων text information retrieval συστημάτων*

Παρακάτω παραθέτουμε μία σύγκριση μεταξύ των συστημάτων που προαναφέρθηκαν, όπως αυτή παρουσιάζεται στο [17].

Όπως φαίνεται στο Σχήμα 9α΄, το σύστημα ReVerb επιτυγχάνει σημαντικά υψηλότερο precision από τα προηγούμενα συστήματα. Στη αναζήτηση ρηματικών εκφράσεων μόνο, επίσης επιτυγχάνει καλύτερο precision και recall από τα προηγούμενα συστήματα (Σχήμα 9β΄).



(α΄) γενική σύγκριση

(β΄) σύγκριση στην αναζήτηση σχέσεων (ρηματικών εκφράσεων).

Σχήμα 9

3

ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ

Ο όρος *αναγνώριση προτύπων (pattern recognition)* χρησιμοποιείται για να δηλώσει την εργασία επισημείωσης μιας παρατήρησης-εισόδου με κάποια ετικέτα (label), η οποία είναι αντιπροσωπευτική της. Σε αντίθεση με το *ταίριασμα προτύπων* (*pattern matching*, τυπική εργασία του οποίου είναι η αναγνώριση κανονικών εκφράσεων) όπου αναζητούνται συγκεκριμένα και προκαθορισμένα patterns στην είσοδο, οι εργασίες pattern recognition προσπαθούν και επιτυγχάνουν (στο μέτρο που αυτό είναι εφικτό) να εξάγουν μία λογική απόκριση για κάθε είδους παρατήρηση-είσοδο.

Παραδείγματα εργασιών pattern recognition είναι το *sequence labeling*, κατά τη διάρκεια του οποίου ανατίθεται μία κατηγορία (class) σε κάθε μέλος μιας ακολουθίας τιμών (π.χ. part of speech tagging), το *parsing*, κατά τη διάρκεια του οποίου εξάγεται ένα συντακτικό δέντρο (parse tree) που περιγράφει τη συντακτική δομή της πρότασης, το *classification* κ.ά..

3.1 ΤΟ ΠΡΟΒΛΗΜΑ ΤΗΣ ΤΑΞΙΝΟΜΗΣΗΣ (CLASSIFICATION)

Το πρόβλημα της *ταξινόμησης (classification)*, όπως ορίζεται στο πλαίσιο της στατιστικής και της μηχανικής μάθησης (machine learning), συνίσταται στον προσδιορισμό της κατηγορίας (από ένα σύνολο κατηγοριών) στην οποία ανήκει μια παρατήρηση. Η κατηγοριοποίηση της παρατήρησης γίνεται με βάση ένα *training* σύνολο αποτελούμενο από παρατηρήσεις για τις οποίες η κατηγορία στην οποία ανήκουν είναι γνωστή.

Κάθε παρατήρηση αναλύεται και επεξεργάζεται ως προς ένα σύνολο μετρήσιμων χαρακτηριστικών (features). Τα χαρακτηριστικά αυτά μπορεί να παίρνουν τιμές ακέραιους αριθμούς, πραγματικούς αριθμούς, κατηγορίες ("κατηγορία Α", "κατηγορία Β", κατηγορία "Γ" κ.λπ.) κ.ά.. Για κάθε παρατήρηση δημιουργείται ένα διάνυσμα από features (feature vector). Αν το σώμα εισόδου είναι εικόνα, το feature vector θα μπορούσε να περιέχει πληροφορίες για τα pixels της εικόνας· αν το σώμα εισόδου είναι αρχείο μουσικής, θα μπορούσε να περιέχει πληροφορίες για κάθε χρονική στιγμή δειγματοληψίας· αν το σώμα εισόδου είναι κείμενο, θα μπορούσε να περιέχει συχνότητα εμφάνισης λέξεων, part of speech tagging πληροφορίες κ.ά..

Για να καταστεί πιο εύκολο το classification, θεωρούμε ότι τα feature vectors που εξάγονται για κάθε παρατήρηση συνιστούν ένα διανυσματικό χώρο, ο οποίος αναφέρεται ως feature space. Γίνεται εύκολα αντιληπτό ότι ο χώρος αυτός είναι πολυδιάστατος (στη γενική περίπτωση αποτελείται από άπειρες διαστάσεις). Για το λόγο αυτό, και για να μπορέσουμε να διατηρήσουμε την υπολογιστική ισχύ που

απαιτείται για την περάτωση του classification σε λογικά πλαίσια, χρησιμοποιούνται τεχνικές dimensionality reduction.

Το classification μπορεί να είναι *binary* ή *multiclass*. Το πρόβλημα του binary classification αναφέρεται στην ταξινόμηση των παρατηρήσεων σε δύο μόνο κλάσεις (A/B, positive/negative κ.λπ.), ενώ το multiclass classification στην ταξινόμηση σε περισσότερες των δύο κατηγοριών.

Η διαδικασία που ακολουθείται κατά το classification θεωρείται supervised, καθότι στηρίζεται στην παρουσία και την εγκυρότητα του training συνόλου. Το σύνολο αυτό μπορεί να δημιουργηθεί χειροκίνητα (οπότε μιλάμε για πλήρως supervised διαδικασία) ή με αυτοματοποιημένες μεθόδους (από υπολογιστή, οπότε και το classification στο σύνολό του θεωρείται semi-supervised διαδικασία).

Τυπικά παραδείγματα εργασίας classification είναι η ένταξη ενός mail σε μία από τις δύο κατηγορίες {junk, non-junk} (binary classification) ή η εξαγωγή διάγνωσης για κάποιον ασθενή με βάση τα παρατηρούμενα συμπτώματα (multiclass classification).

Πολλοί classification αλγόριθμοι μπορούν να εκφραστούν ως γραμμικές συναρτήσεις, οι οποίες σε κάθε νέα παρατήρηση και για κάθε υποψήφια κατηγορία (όπου πιθανώς θα εντάξουν την παρατήρηση), υπολογίζουν μια τιμή βαθμολογίας. Η βαθμολογία αυτή εξάγεται με χρήση του feature vector της παρατήρησης και ενός διανύσματος βαρών. Η κατηγορία στην οποία εντάσσεται τελικά η παρατήρηση είναι αυτή για την οποία υπολογίστηκε η μεγαλύτερη βαθμολογία. Αυτού του τύπου η βαθμολογία είναι γνωστή ως *linear predictor function* και προκύπτει από το εσωτερικό γινόμενο των δύο διανυσμάτων:

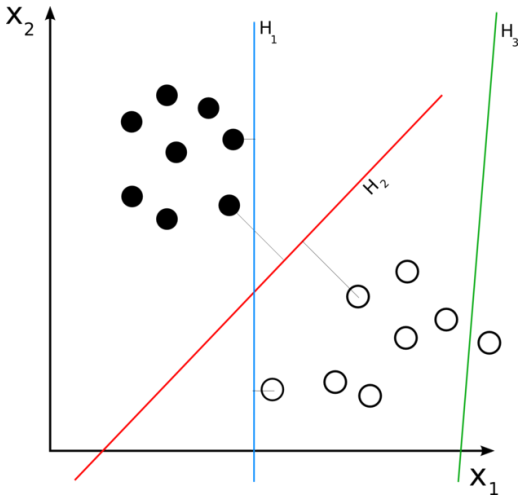
$$score(X_i, k) = \beta_k \cdot X_i \tag{19}$$

όπου X_i το feature vector της προς ταξινόμηση παρατήρησης i και β_k το διάνυσμα βαρών που αντιστοιχεί στην κατηγορία k και το οποίο έχει εξαχθεί κατά το training του classification.

Υπάρχουν και άλλοι αλγόριθμοι οι οποίοι εκφράζονται με τη βοήθεια συναρτήσεων που δεν είναι γραμμικές. Οι πιο συχνά χρησιμοποιούμενοι classifiers σήμερα είναι τα *support vector machines*, ο αλγόριθμος *k-nearest neighbours*, το *Gaussian mixture model*, ο *naive Bayes classifier*, τα *decision trees* και οι *RBF classifiers*. Όπως είναι φυσικό, η απόδοσή τους εξαρτάται από πλήθος παραγόντων και, συνεπακόλουθα, δεν υπάρχει κάποιος από αυτούς ο οποίος να μπορεί να χαρακτηριστεί ως ο «καλύτερος». Στη συνέχεια αναλύεται διεξοδικότερα η λειτουργία των support vector machines (SVMs), μιας και στο πλαίσιο της παρούσης χρησιμοποιήθηκε ένα SVM.

3.2 SUPPORT VECTOR MACHINES (SVM): ΕΝΑ CLASSIFICATION ΜΟΝΤΕΛΟ

Τα support vector machines αποτελούν ένα πολύ κοινό εργαλείο σε εφαρμογές information retrieval και γενικότερα machine learning. Πρόκειται για επιβλεπό-



Σχήμα 10: Παράδειγμα υπερεπιπέδων: Το H_1 διαχωρίζει τις παρατηρήσεις, αλλά με μικρό κενό. Το H_2 διαχωρίζει τις παρατηρήσεις με το μέγιστο δυνατό κενό, ενώ το H_3 δε διαχωρίζει τις παρατηρήσεις.

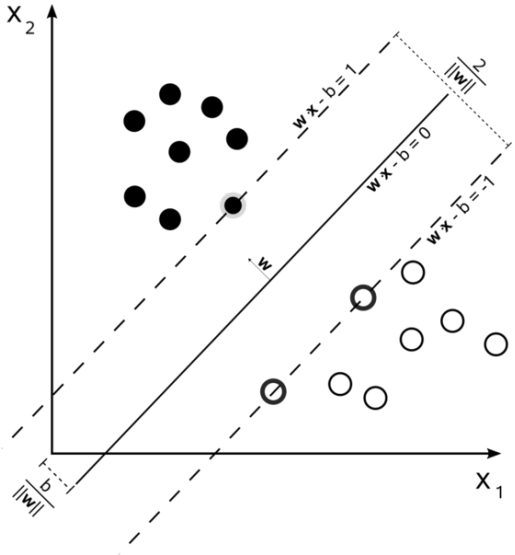
μενα μοντέλα μάθησης, τα οποία σε συνδυασμό με αλγορίθμους μάθησης αναλύουν δεδομένα και αναγνωρίζουν patterns. Κάτι τέτοιο είναι ιδιαίτερα χρήσιμο σε εφαρμογές pattern recognition, όπως classification και regression analysis (ανάλυση παλινδρόμησης).

3.2.1 *Περιγραφή μοντέλου SVM*

Ένα SVM μοντέλο είναι μια αναπαράσταση των παρατηρήσεων του training συνόλου ως σημείων του χώρου, τέτοια ώστε οι παρατηρήσεις που ανήκουν σε διαφορετικές κατηγορίες να διαχωρίζονται από ένα ευδιάκριτο, όσο δυνατόν μεγαλύτερο κενό. Κάθε νέα παρατήρηση αναπαρίσταται στον ίδιο χώρο. Για να γίνει η πρόβλεψη για την κατηγορία στην οποία ανήκει, ελέγχεται σε ποιο από τα δύο μέρη (στα οποία χωρίζει το χώρο το κενό) βρίσκεται το σημείο που την αναπαριστά. Στα επόμενα αναφερόμαστε σε linear binary classification SVM, αλλά το μοντέλο που θα περιγραφεί μπορεί να επεκταθεί τόσο για multiclass classification όσο και για μη-γραμμικό classification.

Πολλές φορές οι κατηγορίες στις οποίες θέλουμε να χωρίσουμε τις παρατηρήσεις δεν είναι εύκολα διαχωρίσιμες στον πολυδιάστατο feature space. Για το λόγο αυτό ο αρχικός χώρος προβάλλεται σε έναν χώρο με περισσότερες διαστάσεις, όπου πιθανώς ο διαχωρισμός γίνεται ευκολότερος. Προκειμένου να κρατηθούν σε λογικά πλαίσια οι απαιτούμενοι υπολογιστικοί πόροι, οι προβολές που χρησιμοποιεί το SVM είναι σχεδιασμένες ώστε να εξασφαλίζεται ο εύκολος υπολογισμός των εσωτερικών γινομένων που απαιτούνται, με χρήση μιας επιλεγμένης kernel function $K(x, y)$ που να ταιριάζει στο πρόβλημα [27].

Το SVM μοντέλο κατασκευάζει ένα υπερπεπίπεδο ή ένα σύνολο υπερεπιπέδων στον πολυδιάστατο χώρο, τα οποία χρησιμοποιούνται για το διαχωρισμό των παρατηρήσεων σε κατηγορίες. Τα υπερεπίπεδα αυτά ορίζονται ως οι γεωμετρικοί τόποι



Σχήμα 11: Υπερεπίπεδο το οποίο εξασφαλίζει το μέγιστο κενό για binary classification.

των σημείων για τα οποία το εσωτερικό γινόμενο με ένα διάνυσμα του χώρου είναι σταθερό. Τα διανύσματα που ορίζουν υπερεπίπεδα μπορούν να επιλεγούν με τέτοιο τρόπο ώστε να είναι γραμμικοί συνδυασμοί των διανυσμάτων εισόδου με παραμέτουρς α_i . Επιλέγοντας έτσι το υπερεπίπεδο, τα σημεία x του feature space τα οποία προβάλλονται στο υπερεπίπεδο προσδιορίζονται από τη σχέση $\sum_i \alpha_i K(x_i, x) = c$, με c σταθερά. Αξίζει να σημειωθεί πως αν η τιμή της $K(x, y)$ μικραίνει όσο το y απομακρύνεται από το x, κάθε στοιχείο του παραπάνω αθροίσματος μετρά την εγγύτητα του σημείου x στο training σημείο x_i. Γίνεται αντιληπτό ότι το παραπάνω άθροισμα μπορεί να χρησιμοποιηθεί ως μέτρο σχετικής εγγύτητας κάθε σημείου του training συνόλου με το κάθε νέο σημείο.

3.2.2 *Μαθηματική περιγραφή μοντέλου SVM*

Δεδομένου ενός training συνόλου παρατηρήσεων $\mathcal{D}$, δημιουργείται ένα σύνολο σημείων της μορφής $\mathcal{D} = \{(x_i, y_i) | x_i \in \mathbb{R}^p, y_i \in \{-1, 1\}\}, 1 \leqslant i \leqslant n$, όπου το y_i δείχνει την κατηγορία (-1 ή 1) στην οποία είναι ενταγμένη η παρατήρηση x_i, η οποία αναπαρίσταται με ένα διάνυσμα p-διαστάσεων. Αυτό που ψάχνουμε είναι το υπερεπίπεδο εκείνο, το οποίο διαχωρίζει τα σημεία για τα οποία $y_i = 1$ από αυτά που έχουν $y_i = -1$, με το μεγαλύτερο δυνατό κενό. Κάθε υπερεπίπεδο μπορεί να παρασταθεί με την εξίσωση $w \cdot x - b = 0$. Στην εξίσωση αυτή, τα w, x, b είναι διανύσματα με το w να είναι το κάθετο διάνυσμα στο υπερεπίπεδο, ενώ το σύμβολο $\cdot$ υπονοεί εσωτερικό γινόμενο.

Αν τα δεδομένα του training συνόλου είναι γραμμικώς διαχωρίσιμα, μπορούμε να επιλέξουμε δύο υπερεπίπεδα τέτοια ώστε να διαχωρίζουν τα σημεία του χώρου και να μην αφήνουν κανένα σημείο στο χώρο ανάμεσά τους. Στη συνέχεια προσπαθούμε να μεγιστοποιήσουμε την απόσταση μεταξύ τους.

Τα εν λόγω υπερεπίπεδα περιγράφονται από τις εξισώσεις

$$w \cdot x - b = 1 \tag{20}$$

και

$$w \cdot x - b = -1 \tag{21}$$

Η απόσταση μεταξύ των δύο υπερεπιπέδων υπολογίζεται με τη χρήση γεωμετρίας ως $\frac{2}{||w||}$. Συνεπώς, προκειμένου να μεγιστοποιήσουμε την απόσταση, επιζητούμε την ελαχιστοποίηση του $||w||$. Επειδή όμως δε θέλουμε κανένα σημείο να βρίσκεται μεταξύ των δύο αυτών υπερεπιπέδων, προσθέτουμε τους ακόλουθους περιορισμούς:

$$w \cdot x_i - b \geqslant 1 \quad \text{και} \quad w \cdot x_i - b \leqslant -1 \tag{22}$$

για τα x_i της μιας και της άλλης κατηγορίας αντίστοιχα. Η σχέση αυτή μπορεί να γραφεί και ως $y_i(w \cdot x_i - b) \geqslant 1, 1 \leqslant i \leqslant n$. Έτσι, καταλήγουμε στο ακόλουθο πρόβλημα βελτιστοποίησης:

> να ευρεθεί η ελάχιστη τιμή του $||w||$ για τις διάφορες τιμές w, b, δεδομένου ότι $y_i(w \cdot x_i - b \geqslant 1$

Το πρόβλημα αυτό δεν είναι εύκολα επιλύσιμο, λόγω της εγγενούς δυσκολίας υπολογισμού της νόρμας $||w||$. Ωστόσο, μπορούμε να αντικαταστήσουμε την $||w||$ με $\frac{1}{2}||w||^2$, η οποία είναι εύκολα υπολογίσιμη, χωρίς να αλλάξουν οι τιμές των w, b που συνιστούν τη λύση του προβλήματος. Χρησιμοποιώντας τη διαπίστωση αυτή, και κάνοντας χρήση πολλαπλασιαστών Lagrange, το πρόβλημα μπορεί να εκφραστεί ως:

$$\min_{w,b} \max_{\alpha \geqslant 0} \{\frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i [y_i(w \cdot x_i - b) - 1]\} \tag{23}$$

Το πρόβλημα πλέον, με τη μορφή που πήρε, μπορεί να λυθεί με κλασικές τεχνικές τετραγωνικού προγραμματισμού. Λόγω της ισχύος της stationary KKT (Karush-Kuhn-Tucker) [21] προϋπόθεσης, μπορούμε να εκφράσουμε τη λύση ως γραμμικό συνδυασμό των x_i:

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{24}$$

Στο άθροισμα αυτό, λίγα α_i θα είναι μεγαλύτερα του μηδενός. Τα αντίστοιχα x_i θα είναι αυτά που ορίζονται ως support vectors, τα οποία βρίσκονται στο κενό και ικανοποιούν τη σχέση $y_i(w \cdot x_i - b) = 1$ και άρα, τις $w \cdot x_i - b = \frac{1}{y_i} \Leftrightarrow w \cdot x_i - b = y_i \Leftrightarrow b = w \cdot x_i - y_i$. Από την τελευταία σχέση μπορούμε να υπολογίσουμε και το b για κάθε i και, τελικά, τη μέση τιμή του b ως

$$b = \frac{1}{N_{SV}} \sum_{i=1}^{N_{SV}} (w \cdot x_i - y_i) \tag{25}$$

3.2.3 *Επεκτάσεις και παραλλαγές του SVM*

3.2.3.1 *Soft Margin*

Το 1995, προτάθηκε μία παραλλαγή του SVM η οποία επέτρεπε την παρουσία σημείων στο κενό ανάμεσα στα υπερεπίπεδα που διαχωρίζουν τα σημεία των δύο κατηγοριών [14]. Η ιδέα είναι ότι αν δε μπορεί να ευρεθεί υπερεπίπεδο (ή υπερεπίπεδα) τα οποίο να διαχωρίζει όλα τα σημεία σαφώς, θα επιλεγεί ένα υπερεπίπεδο το οποίο διαχωρίζει τα σημεία όσο το δυνατό καλύτερα, μεγιστοποιώντας την απόστασή του από τα κοντινότερα πλήρως διαχωρισμένα σημεία. Η μέθοδος αυτή, γνωστή ως *soft margin method* εισάγει τη χρήση μεταβλητών απόκλισης ξ_i, οι οποίες μετρούν το βαθμό αποτυχίας classification του x_i. Η γενική εξίσωση περιγραφής του προβλήματος είναι:

$$y_i(w \cdot x_i - b) \geqslant 1 - \xi_i, 1 \leqslant i \leqslant n \tag{26}$$

Με την τροποποίηση αυτή, ο «στόχος» του μοντέλου από $\min_{w,b}\{\frac{1}{2}||w||^2\}$ διαφοροποιείται σε $\min_{w,\xi,b}\{\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i\}$. Κάνοντας χρήση πολλαπλασιαστών Lagrange όπως και προηγουμένως, το πρόβλημα μετασχηματίζεται σε

$$\min_{w,\xi,b}\max_{\alpha,\beta}\{\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(w \cdot x_i - b) - 1 + \xi_i] - \sum_{i=1}^{n}\beta_i\xi_i\} \tag{27}$$

3.2.3.2 *Μη γραμμικά SVM*

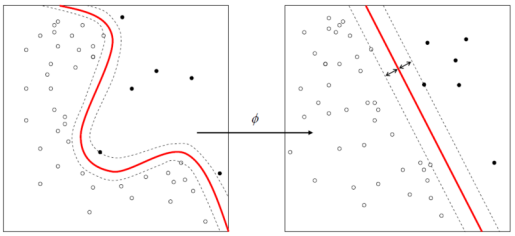
Το 1992 για πρώτη φορά [11] χρησιμοποιήθηκε η ιδέα του *"kernel trick"* (το οποίο πρωτοεμφανίστηκε το 1964 [6]) για τη δημιουργία μη-γραμμικών classifiers. Ουσιαστικά, η νέα ιδέα αυτή αντικαθιστά κάθε εσωτερικό γινόμενο των υπολογισμών με μια μη-γραμμική kernel function. Αυτό επιτρέπει στον αλγόριθμο να «χωρέσει» ένα μεγίστου κενού υπερεπίπεδο σε έναν μετασχηματισμένο feature space. Ο μετασχηματισμός μπορεί να είναι μη γραμμικός και ο μετασχηματισμένος χώρος πολυδιάστατος. Για το λόγο αυτό, το υπερεπίπεδο μπορεί να είναι μη-γραμμικό στον αρχικό feature space.

Μερικές συχνά χρησιμοποιούμενες kernel functions είναι οι εξής:

- *πολυωνυμική (ομογενής)*: $k(x_i, x_j) = (x_i \cdot x_j)^d$
- *πολυωνυμική (μη ομογενής)*: $k(x_i, x_j) = (x_i \cdot x_j + 1)^d$
- *Gaussian radial basis function*: $k(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2)$, για $\gamma > 0$.
- *υπερβολικής εφαπτομένης*: $k(x_i, x_j) = \tanh(\kappa x_i \cdot x_j + c)$, για κάποια $\kappa > 0$ και $c < 0$.

3.2.3.3 *Multiclass SVM*

Μέχρι τώρα έχουν περιγραφεί πρακτικές SVM binary classification. Για την επίλυση multiclass classification προβλημάτων η επικρατούσα προσέγγιση προτείνει



Σχήμα 12: Λειτουργία kernel functions.

την αναγωγή τους σε πολλά binary classification προβλήματα. Συχνές μέθοδοι είναι οι παρακάτω:

- χρήση πολλών binary classifiers οι οποίοι διαχωρίζουν τα στοιχεία μίας κατηγορίας από όλα τα υπόλοιπα (one-versus-all) ή τα στοιχεία μίας κατηγορίας από μια άλλη (one-versus-one) και στη συνέχεια συνδυασμός των αποτελεσμάτων τους με άλλους classifiers.
- *Directed Acyclic Graph SVM*
- *error-correcting output codes*

3.2.4 *Λειτουργία SVM*

Στην πράξη, η λειτουργία ενός SVM θα μπορούσε να χωριστεί σε δύο φάσεις, την training (learning) φάση και την classification φάση.

- *Κατά τη διάρκεια της learning φάσης*, στο SVM δίνεται ως είσοδος ένα σύνολο παραδειγμάτων, για τα οποία έχει επισημειωθεί η κατηγορία στην οποία ανήκουν. Το SVM παράγει ως έξοδο ένα μοντέλο, το οποίο μπορεί να κατατάξει κάθε νέο, μη επισημασμένο παράδειγμα σε μία από τις δύο κατηγορίες. Το μοντέλο αυτό είναι ουσιαστικά μια αναπαράσταση των παραδειγμάτων εισόδου ως σημεία στο χώρο, με τέτοιο τρόπο ώστε τα παραδείγματα τα οποία ανήκουν σε διαφορετική κατηγορία, να διαχωρίζονται από ένα σαφές και όσο δυνατό μεγαλύτερο κενό.
- *Κατά τη διάρκεια της classification φάσης*, το SVM, με δεδομένο το μοντέλο που παρήχθη στην προηγούμενη φάση, μπορεί να ταξινομήσει νέα παραδείγματα σε κατηγορίες. Αυτό το επιτυγχάνει αντιστοιχίζοντας κάθε παράδειγμα σε ένα σημείο του χώρου και ανακαλύπτοντας σε ποια μεριά του κενού του μοντέλου βρίσκεται το κάθε νέο σημείο.

3.3 ΤΟ ΠΡΟΒΛΗΜΑ ΤΗΣ ΟΜΑΔΟΠΟΙΗΣΗΣ (CLUSTERING)

Το unsupervised αντίστοιχο του προβλήματος του classification είναι γνωστό ως *ομαδοποίηση (clustering)*. Κατά το clustering επιδιώκεται η ταξινόμηση των παρατηρήσεων σε σχετικά ευδιάκριτες κατηγορίες (clusters), ώστε οι παρατηρήσεις που ανήκουν στην ίδια κατηγορία να μοιάζουν περισσότερο μεταξύ τους, παρά με αυτές που ανήκουν σε άλλες κατηγορίες. Σε αντίθεση με το classification, το clustering

δεν απαιτεί δεδομένα εισόδου πέραν των παρατηρήσεων (δεν υπάρχουν επισημασμένες “έμπιστες” παρατηρήσεις).

Η έννοια του cluster διαφέρει από υλοποίηση σε υλοποίηση. Τα clusters τα οποία βρίσκουν διαφορετικοί αλγόριθμοι μπορεί να διαφέρουν σημαντικά λόγω αυτής διαφοροποίησης. Τυπικά μοντέλα clusters είναι τα ακόλουθα:

- *Connectivity models*
- *Centroid models*
- *Distribution models*
- *Density models*
- *Subspace models*
- *Group models*
- *Graph-based models*

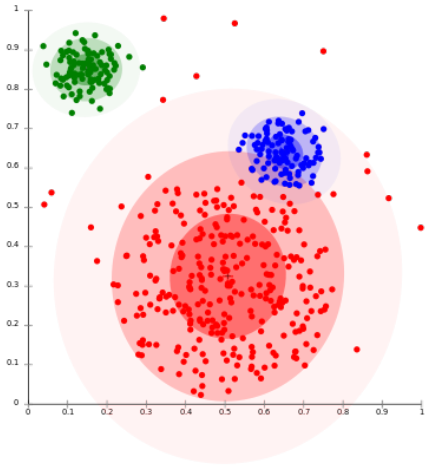
Κατά τη διάρκεια του clustering, οι αλγόριθμοι που χρησιμοποιούνται προσπαθούν να ανακαλύψουν κατά πόσον κάποιες από τις παρατηρήσεις εμφανίζουν ομοιότητες και κατά πόσον μπορούν, ανάλογα με τις ιδιότητες αυτές, να ομαδοποιηθούν. Στη λειτουργία και την αποτελεσματικότητα του εγχειρήματος αυτού παίζει σημαντικό ρόλο η παραμετροποίηση του αλγορίθμου (τι είδους «απόσταση» στο feature space να χρησιμοποιηθεί ως μετρική ομοιότητας ή διαφοράς μεταξύ των παρατηρήσεων, κατώφλια ανεκτικότητας στην ένταξη μιας παρατήρησης σε μία κλάση κ.λπ.). Γίνεται αντιληπτό ότι το clustering δεν είναι μια προκαθορισμένη, αυτοματοποιημένη διαδικασία, αλλά μια διαδικασία ανακάλυψης πληροφορίας, η οποία εμπεριέχει σε μεγάλο βαθμό δοκιμές και βελτιώσεις με βάση τα παραγόμενα αποτελέσματα.

Ανάλογα με το αποτέλεσμα που παράγεται, μπορούμε να διακρίνουμε τις εξής κατηγορίες clustering:

- *hard clustering*: κάθε παρατήρηση ανήκει σε μία κατηγορία ή όχι.
- *soft clustering (fuzzy clustering)*: κάθε παρατήρηση ανήκει σε μία κατηγορία κατά ένα ποσοστό βεβαιότητας.

και κάνοντας έναν πιο αναλυτικό διαχωρισμό:

- *strict partitioning clustering*: κάθε παρατήρηση κατατάσσεται σε μία μόνο κατηγορία.
- *strict partitioning clustering with outliers*: κάποιες παρατηρήσεις οι οποίες δε μπορούν να ενταχθούν σε καμία κατηγορία, δε κατατάσσονται και απλά θεωρούνται outliers.
- *overlapping clustering*: κάθε παρατήρηση μπορεί να ανήκει σε περισσότερες από μία κατηγορίες.



Σχήμα 13: Παράδειγμα clustering: Expectation maximazation αλγόριθμος σε Gaussian-distributed δεδομένα.

- *hierarchical clustering*: δημιουργείται μια ιεραρχία κατηγοριών, έτσι ώστε κάθε παρατήρηση η οποία ανήκει σε μια κατηγορία, να ανήκει και στην κατηγορία-πατέρα αυτής.
- *subspace clustering*: πρόκειται για overlapping clustering, με τη διαφορά ότι εντός ενός μοναδικά καθορισμένου υποχώρου του feature space, οι κατηγορίες δεν επικαλύπτονται.

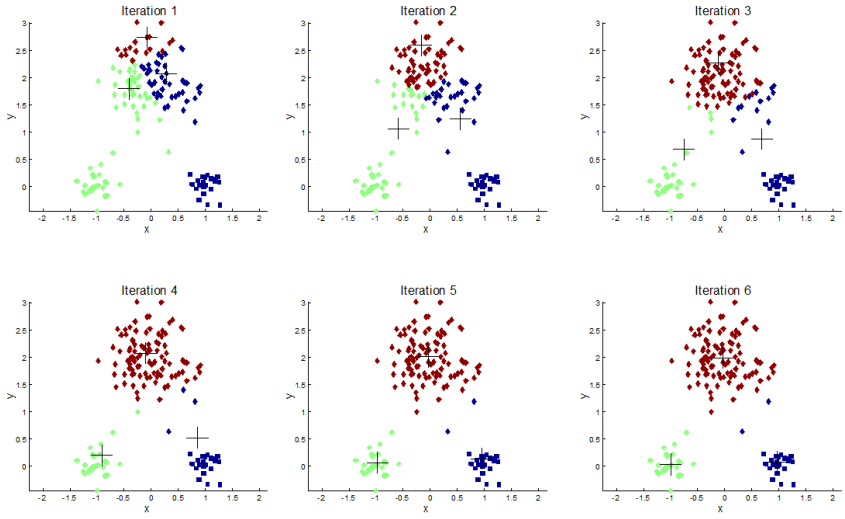
3.4 ΠΑΡΑΔΕΙΓΜΑΤΑ ΑΛΓΟΡΙΘΜΩΝ CLUSTERING

Παρακάτω παρουσιάζονται δυο πολύ συχνά χρησιμοποιούμενοι αλγόριθμοι ομαδοποίησης, ο *k-means* (ο οποίος χρησιμοποιεί centroid model) και ο αλγόριθμος *ιεραρχικής συσσωρευτικής ομαδοποίησης* (ο οποίος χρησιμοποιεί group model).

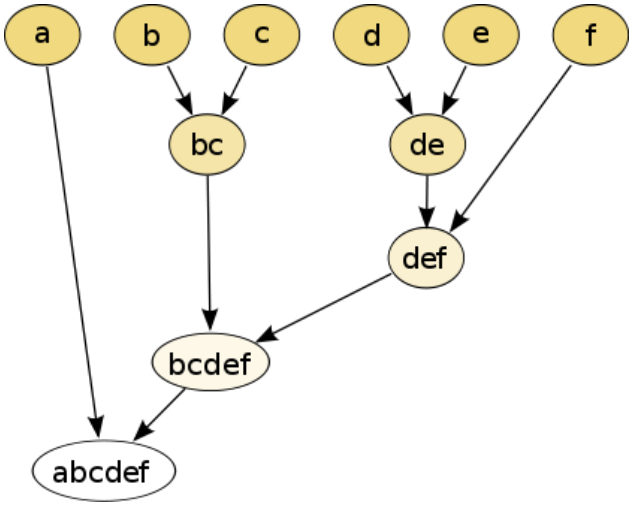
3.4.1 *Αλγόριθμος k-μέσεων (k-means)*

Η ιδέα του αλγορίθμου k-μέσων είναι η εξής: αρχικά επιλέγονται k διανύσματα (όσα και τα clusters), τα οποία καλούνται *κεντροειδή (centroids)*. Στη συνέχεια, κάθε διάνυσμα το οποίο αντιστοιχεί σε παρατήρηση ανατίθεται στο πλησιέστερο κεντροειδές. Για τον υπολογισμό της απόστασης των παρατηρήσεων από τα κεντροειδή χρησιμοποιούνται διάφορα μέτρα, όπως η ευκλείδεια απόσταση, η Manhattan distance ή το μέτρο Jaccard. Σε κάθε επανάληψη του αλγορίθμου, μετά τον υπολογισμό των αποστάσεων, επανυπολογίζονται τα κεντροειδή. Ο αλγόριθμος επαναλαμβάνεται μέχρι να μην υπάρχουν αλλαγές στην επιλογή των κεντροειδών.

Επειδή η αλγόριθμος αυτός είναι ευριστικός, δεν είναι βέβαιο ότι θα συγκλίνει πάντα στη βέλτιστη λύση. Ωστόσο, στην πράξη, συγκλίνει σχεδόν πάντα και μάλιστα από τα πρώτα βήματα.



Σχήμα 14: Παράδειγμα εκτέλεσης αλγορίθμου k-means.



Σχήμα 15: Παράδειγμα αποτελέσματος ιεραρχικής συσσωρευτικής ομαδοποίησης.

Το αποτέλεσμα του αλγορίθμου εξαρτάται σε μεγάλο βαθμό από την αρχική επιλογή των κεντροειδών διανυσμάτων. Αν η αρχική επιλογή των κεντροειδών γίνει με τυχαίο τρόπο, είναι πιθανό κάθε φορά να προκύπτουν διαφορετικά αποτελέσματα. Για να ξεπεραστεί η δυσκολία αυτή, μία λύση είναι να πραγματοποιηθούν πολλές εκτελέσεις του αλγορίθμου. Μία άλλη λύση δόθηκε με παραλλαγή του αλγορίθμου, η οποία είναι γνωστή ως *«διχοτόμος μέθοδος των k-μέσων» (bisecting k-means)* [33]. Η ιδέα της παραλλαγής είναι η εξής: για τη δημιουργία των k αρχικών ομάδων, το σύνολο των παρατηρήσεων χωρίζεται σε δύο σύνολα. Κατόπιν, επιλέγεται ένα σύνολο από αυτά (με κάποιο κριτήριο επιλογής, π.χ. το μεγαλύτερο) και χωρίζεται και αυτό σε δύο σύνολα. Η διαδικασία επαναλαμβάνεται μέχρι να παραχθούν k ομάδες.

3.4.2 *Ιεραρχική συσσωρευτική ομαδοποίηση (Hierarchical Agglomerative Clustering, HAC*

Η *ιεραρχική συσσωρευτική ομαδοποίηση* ξεκινά θεωρώντας κάθε παρατήρηση ως μία ξεχωριστή ομάδα. Σε κάθε εκτέλεση του αλγορίθμου το ζευγάρι των ομάδων που βρίσκονται πιο κοντά συνενώνεται σε μία νέα ομάδα. Είναι προφανές ότι και εδώ απαιτείται ένα μέτρο της εγγύτητας των παρατηρήσεων (ευκλείδεια απόσταση, απόσταση Hamming κ.ά.) αλλά και ένα μέτρο της απόστασης μεταξύ των ομάδων. Για την απόσταση μεταξύ ομάδων έχουν προταθεί διάφορα μέτρα, όπως:

- *min*: η απόσταση μεταξύ των δύο παρατηρήσεων (μία από κάθε ομάδα) που βρίσκονται πιο κοντά
- *max*: η απόσταση μεταξύ των δύο παρατηρήσεων (μία από κάθε ομάδα) που βρίσκονται πιο μακρυά
- *group*: ο μέσος όρος των αποστάσεων των παρατηρήσεων της μιας ομάδας από την άλλη

Η διαδικασία σταματά όταν όλες οι παρατηρήσεις έχουν ενταχθεί σε μία ομάδα. Το αποτέλεσμα είναι ένα δενδροδιάγραμμα το οποίο παρουσιάζει τόσο την ιεραρχική διάταξη των ομάδων και των υποομάδων, όσο και τη σειρά με την οποία οι ομάδες ενώθηκαν.

Μέρος II

ΠΕΙΡΑΜΑΤΙΚΟ ΜΕΡΟΣ

4

ΓΕΝΙΚΗ ΠΕΡΙΓΡΑΦΗ ΣΥΣΤΗΜΑΤΟΣ

Στο κεφάλαιο αυτό θα γίνει παρουσίαση του σχεδιασμού του συστήματος, της γενικής λειτουργίας του, των συγκεκριμένων σκοπών που εξυπηρετεί καθώς και της σημασίας εκπλήρωσής του. Επιπλέον, θα δοθεί η ανάλυση των κρίσιμων αποφάσεων σχετικά με τα μέσα που χρησιμοποιήθηκαν, τις αρχιτεκτονικές επιλογές και τον τρόπο χειρισμού επιμέρους εργασιών.

4.1 ΕΙΣΑΓΩΓΗ

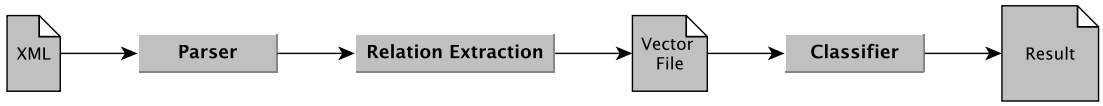
Στόχος του συστήματος είναι η εξαγωγή μη προκαθορισμένων σχέσεων μεταξύ οντοτήτων από κείμενο το οποίο δεν παρουσιάζει αναμενόμενη μορφή ή θεματολογία. Συνεπώς θα μπορούσαμε να το περιγράψουμε ως σύστημα open relation extraction.

Για την υλοποίηση και τη δοκιμή λειτουργίας του συστήματος ήταν απαραίτητη η παρουσία ενός συνόλου κειμένων (corpus) τα οποία δεν παρουσιάζουν καμία ομοιογένεια, δομική ή θεματολογίας. Το σκοπό αυτό επιτέλεσε το αρχείο της εφημερίδας «ΤΑ ΝΕΑ», το οποίο δόθηκε σε ηλεκτρονική μορφή.

Δεδομένης της φύσης του corpus (πολλά ετερόκλητα κείμενα - άρθρα σε κάθε φύλλο, σε αρχείο συγκεκριμένης μορφοποίησης), χρειάστηκε να δημιουργηθούν κάποιες μονάδες (modules) οι οποίες αφορούν αυστηρά σε αρχεία που χρησιμοποιούν μορφοποίηση όμοια με αυτή των αρχείων του corpus. Ωστόσο, ο πυρήνας λειτουργίας του συστήματος μπορεί να εξάγει σχέσεις από κάθε μορφής κείμενο.

Το σύστημα αποτελείται από δύο ανεξάρτητα μέρη. Το πρώτο είναι ένας αναλυτής (parser), ο οποίος δέχεται ως είσοδο ένα αρχείο στη μορφή στην οποία βρίσκονται τα φύλλα του ηλεκτρονικού αρχείου των «Νέων» (XML αρχείο συγκεκριμένης δομής) και παράγει ένα αρχείο (XML) με πολύ πιο απλή δομή, το οποίο είναι εύκολα επεξεργάσιμο.

Το δεύτερο και πιο ουσιαστικό μέρος, δέχεται ως είσοδο ένα αρχείο κειμένου και προσπαθεί να εξάγει σχέσεις μεταξύ οντοτήτων από αυτό. Για να το επιτύχει αυτό, «διαβάζει» το κείμενο και με χρήση γραμματικών και συντακτικών κανόνων εξάγει τριάδες της μορφής $(entity_1, relation, entity_2)$. Στη συνέχεια, για κάθε τριάδα δημιουργεί ένα διάνυσμα χαρακτηριστικών (feature vector). Τα διανύσματα χαρακτηριστικών για όλες τις σχέσεις που εξάγονται από το κείμενο εισόδου, αποθηκεύονται σε ένα αρχείο. Τέλος, καλείται ένας classifier ο οποίος αποδίδει μια βαθμολογία σε κάθε σχέση (κατά πόσον θεωρείται θετική (positive), δηλαδή έμπιστη, σωστή, σημασιολογικά ορθή ή αρνητική (negative). Αξίζει να σημειωθεί πως το σύστημα παρέχει την επιλογή δημιουργίας δεδομένων εκπαίδευσης (training data) · δεδομένων τα οποία μπορούν να χρησιμοποιηθούν για εκπαίδευση του classifier. Για το σκοπό



Σχήμα 16: Γενική περιγραφή του συστήματος.

αυτό, χρησιμοποιεί ένα σύνολο κανόνων και κάποιες εξωτερικές λίστες, προκειμένου να αποφανθεί αν μια σχέση είναι θετική ή αρνητική.

Τελος, μετά την επιτυχή εξαγωγή σχέσεων, το σύστημα δύναται να προβεί σε μία ομαδοποίηση (clustering) των σχέσεων προκειμένου να βρεθούν στην ίδια ομάδα σχέσεις με ίδιο ή παρόμοιο σημασιολογικό περιεχόμενο.

Αξίζει να σημειωθεί ότι το σύστημα είναι σχεδιασμένο για να επεξεργάζεται δεδομένα διατυπωμένα στην ελληνική γλώσσα. Όπως αναφέρθηκε σε προηγούμενη ενότητα, τα συστήματα εξαγωγής σχέσεων τα οποία έχουν δημιουργηθεί μέχρι τώρα, αναφέρονται σε κείμενο γραμμένο στην αγγλική γλώσσα. Η γραμματική και συντακτική δομή της ελληνικής γλώσσας διαφέρει αρκετά από αυτήν τις αγγλικής. Συνεπώς, παρ' ότι οι τεχνικές οι οποίες χρησιμοποιήθηκαν είναι παρόμοιες με αυτές των συστημάτων που έχουν περιγραφεί, η υλοποίηση είναι σημαντικά διαφορετική.

Η αρχιτεκτονική του δεύτερου μέρους του συστήματος ακολουθεί την αρχιτεκτονική τεχνοτροπία του αυλού/φίλτρου. Κατά την εκτέλεσή του δημιουργείται μια ροή δεδομένων από την είσοδο μέχρι την έξοδο, η οποία περνά από ενδιάμεσα στάδια (φίλτρα). Κάθε ένα από αυτά τα στάδια δέχεται δεδομένα από το προηγούμενο στάδιο, εκτελεί μια αυτόνομη εργασία και προωθεί τα δεδομένα στο επόμενο στάδιο.

4.2 ΤΟ ΑΡΧΕΙΟ ΤΗΣ ΕΦΗΜΕΡΙΔΑΣ «ΤΑ ΝΕΑ»

Για την υλοποίηση και τον έλεγχο λειτουργίας του συστήματος, το ρόλο του σώματος κειμένου (corpus) διαδραμάτισε το ηλεκτρονικό αρχείο της εφημερίδας «ΤΑ ΝΕΑ». Πιο συγκεκριμένα, χρησιμοποιήθηκαν αρχεία από φύλλα της καθημερινής και κυριακάτικης έκδοσης των «Νέων», των ετών 2007, 2008 και 2009. Η επιλογή αυτή έγινε προκειμένου να εξασφαλιστεί ένα όσο γίνεται μεγάλο, ανομοιόμορφο, ποικίλης θεματολογίας και δομής σώμα κειμένου.

4.2.1 *Στατιστικά στοιχεία από το αρχείο της εφημερίδας*

Το αρχείο των ετών 2007, 2008, 2009 αποτελείται από:

- 895 φύλλα,
- 238.123 άρθρα,
- 55.423.216 λεκτικές μονάδες (tokens),

- 34.187.520 λεκτικές μονάδες, αν εξαιρεθούν κάποιες πολύ συνηθισμένες λέξεις (stopwords),
- 876.378 διαφορετικές λεκτικές μονάδες (εξαιρώντας τις stopwords).

4.2.2 *Δομή του αρχείου της εφημερίδας*

Το αρχείο της εφημερίδας «ΤΑ ΝΕΑ», όπως δόθηκε, είναι μία συλλογή XML αρχείων. Κάθε αρχείο αντιστοιχεί σε ένα φύλλο της εφημερίδας και είναι οργανωμένο σε δενδρική δομή, παρόμοια με τη δομή που χρησιμοποιείται στο εκτυπωμένο φύλλο.

ΤΑ ΒΑΣΙΚΑ ΔΟΜΙΚΑ ΣΤΟΙΧΕΙΑ ΕΝΟΣ ΦΥΛΛΟΥ εφημερίδας είναι:

- *οι θεματικές ενότητες*: στα «ΝΕΑ» έχουμε «Κύριο Τεύχος», «Ορίζοντες», «Ομάδα» κ.λπ.. Στο εκτυπωμένο φύλλο οι θεματικές ενότητες διαχωρίζονται φυσικά, μιας και για κάθε μια υπάρχει ξεχωριστό τευχίδιο.
- *οι κατηγορίες*: στα «ΝΕΑ» έχουμε: «Πρωτοσέλιδο», «Ελλάδα», «Κόσμος», «Οικονομία» κ.λπ.. Κάθε κατηγορία περιλαμβάνει άρθρα παρόμοιας θεματολογίας.
- *τα άρθρα*, τα οποία με τη σειρά τους δομούνται από υπέρτιτλο, τίτλο, υπότιτλο, μεσότιτλο, περίληψη, εισαγωγή, υπογραφή, παραγράφους, εικόνες και λεζάντες εικόνων. Ένα άρθρο μπορεί να αποτελείται από συνδυασμό των παραπάνω, πιθανώς παραλείποντας κάποια από αυτά. Σίγουρα όμως έχει τίτλο και τουλάχιστο μία παράγραφο κειμένου.

ΣΤΟ ΗΛΕΚΤΡΟΝΙΚΟ ΑΡΧΕΙΟ των «Νέων», για κάθε XML αρχείο, χρησιμοποιούνται λίγα tags, τα οποία ιεραρχούνται δενδρικά:

- `<fr:Issue>` : ένα tag σε κάθε αρχείο, αποτελεί το ριζικό κόμβο του αρχείου.
- `<fr:Nodes>` : περικλείει μία λίστα από `<fr:Node>` tags, κάθε ένα από τα οποία αντιστοιχεί σε μια θεματική ενότητα της εφημερίδας (Κύριο Τεύχος, Ορίζοντες, Ομάδα, nDigital κ.λπ.).
- `<fr:Node>` : κάθε βασικό δομικό στοιχείο της εφημερίδας από τα προαναφερθέντα (θεματική ενότητα, κατηγορία, τίτλος, μεσότιτλος κ.λπ.) περικλείεται από ένα `<fr:Node>` tag. Το attribute `"TypeName"` αποσαφηνίζει τι είδους στοιχείο είναι καθένα.
- `<fr:Segments>` : Τα tags `<fr:Segments>` και `<fr:Segment>` περιέχουν ως attributes IDs τα οποία χρησιμοποιούνται για την οργάνωση του υλικού σε σελίδες και κατηγορίες.
- `<fr:Segment>`
- `<fr:Summary>` : χρησιμοποιείται για την ενθυλάκωση της περίληψης κάθε άρθρου.
- `<fr:ChildNodes>` : περικλείουν τα `<fr:Node>` tags τα οποία αναφέρονται σε δομικά στοιχεία ενός άρθρου (τίτλος, υπέρτιτλος, παράγραφοι κ.λπ.).

- `<fr:Text>` : η καθαρή πληροφορία της εφημερίδας, δηλαδή ό,τι μπορεί κανείς να διαβάσει σε ένα έντυπο αντίγραφο του φύλλου βρίσκεται εντός των `<fr:Text>` tags, με HTML μορφοποίηση.

ΠΑΡΑΚΑΤΩ ΦΑΙΝΕΤΑΙ ΕΝΑ ΑΠΟΣΠΑΣΜΑ από ένα XML αρχείο ενός φύλλου (της 02/01/2007), όπου μπορεί κανείς να διακρίνει τη δομή που περιγράφηκε.

```
<?xml version="1.0" encoding="utf-8"?>
<fr:Issue ID="5531" BrandName="TA NEA" BrandID="2" DateFrom="2007-01-02
    T00:00:00" DateTo="2007-01-02T00:00:00" xmlns:fr="urn:franklin-content-
    issue-s">
  <fr:Nodes>
    <fr:Node ID="3052424" Name="ΚΥΡΙΟ ΤΕΥΧΟΣ" PrototypeName="TA NEA - ΚΥΡΙΟ
    ΤΕΥΧΟΣ" TypeName="Sector" Visibility="30000">
      <fr:Segments>
        <fr:Segment PageID="47669" ContentSegmentID="2754899" />
        <fr:Segment PageID="47670" ContentSegmentID="2754900" />
        <fr:Segment PageID="47671" ContentSegmentID="2754901" />
        <fr:Segment PageID="47672" ContentSegmentID="2754902" />
        <fr:Segment PageID="47673" ContentSegmentID="2754903" />
        <fr:Segment PageID="47674" ContentSegmentID="2754904" />
        <fr:Segment PageID="47675" ContentSegmentID="2754905" />
        <fr:Segment PageID="47676" ContentSegmentID="2754906" />
        <fr:Segment PageID="47677" ContentSegmentID="2754907" />
        <fr:Segment PageID="47678" ContentSegmentID="2754908" />
        <fr:Segment PageID="47679" ContentSegmentID="2754909" />
        <fr:Segment PageID="47680" ContentSegmentID="2754910" />
        <fr:Segment PageID="47681" ContentSegmentID="2754911" />
        <fr:Segment PageID="47682" ContentSegmentID="2754912" />
        <fr:Segment PageID="47684" ContentSegmentID="2754913" />
        <fr:Segment PageID="47685" ContentSegmentID="2754914" />
        <fr:Segment PageID="47686" ContentSegmentID="2754915" />
        <fr:Segment PageID="47694" ContentSegmentID="2754916" />
        <fr:Segment PageID="47697" ContentSegmentID="2754917" />
        <fr:Segment PageID="47723" ContentSegmentID="2754918" />
        <fr:Segment PageID="47724" ContentSegmentID="2754919" />
        <fr:Segment PageID="47725" ContentSegmentID="2754920" />
        <fr:Segment PageID="47726" ContentSegmentID="2754921" />
        <fr:Segment PageID="47727" ContentSegmentID="2754922" />
        <fr:Segment PageID="47728" ContentSegmentID="2754923" />
        <fr:Segment PageID="47729" ContentSegmentID="2754924" />
        <fr:Segment PageID="47730" ContentSegmentID="2754925" />
        <fr:Segment PageID="47731" ContentSegmentID="2754926" />
        <fr:Segment PageID="47732" ContentSegmentID="2754927" />
        <fr:Segment PageID="47733" ContentSegmentID="2754928" />
        <fr:Segment PageID="47735" ContentSegmentID="2754929" />
        <fr:Segment PageID="47736" ContentSegmentID="2754930" />
        <fr:Segment PageID="47737" ContentSegmentID="2754931" />
        <fr:Segment PageID="47738" ContentSegmentID="2754932" />
        <fr:Segment PageID="47739" ContentSegmentID="2754933" />
        <fr:Segment PageID="47740" ContentSegmentID="2754934" />
      </fr:Segments>
      <fr:ChildNodes>
        <fr:Node ID="3052425" Name="ΠΡΩΤΟΣΕΛΙΔΟ" PrototypeName="TA NEA -
    ΚΥΡΙΟ ΤΕΥΧΟΣ - Πρωτοσέλιδο" TypeName="Section" Visibility="30000">
          <fr:Segments>
            <fr:Segment PageID="47669" ContentSegmentID="2754935" />
```

```
47            <fr:Segment PageID="47670" ContentSegmentID="2754936" />
48            <fr:Segment PageID="47676" ContentSegmentID="2754937" />
49            <fr:Segment PageID="47679" ContentSegmentID="2754938" />
50            <fr:Segment PageID="47680" ContentSegmentID="2754939" />
51            <fr:Segment PageID="47682" ContentSegmentID="2754940" />
52            <fr:Segment PageID="47694" ContentSegmentID="2754941" />
53            <fr:Segment PageID="47697" ContentSegmentID="2754942" />
54            <fr:Segment PageID="47730" ContentSegmentID="2754943" />
55          </fr:Segments>
56          <fr:ChildNodes>
57            <fr:Node ID="3068446" Name="Βαρβαρότητα made in U S A " PrototypeName="ΤΑ ΝΕΑ – ΚΥΡΙΟ ΤΕΥΧΟΣ – ΤΟ ΘΕΜΑ" TypeName="Article" Visibility="20000">
58              <fr:Segments>
59                <fr:Segment PageID="47669" ContentSegmentID="2754944" />
60              </fr:Segments>
61              <fr:Summary>Ο κόσμος στη Μέση Ανατολή δεν έγινε ασφαλέστερος μετά την εκτέλεση του Σαντάμ Χουσεΐν. Με αμερικανικό χέρι , οι δήμιοι θανάτωσαν τον δικτάτορα με την ίδια βαρβαρότητα που σκότωνε κι αυτός κάποτε τους εχθρούς του.
  </fr:Summary>
62              <fr:ChildNodes>
63
64                <fr:Node ID="3052584" Name="Title" TypeName="Title" Visibility="0">
65                  <fr:Segments>
66                    <fr:Segment PageID="47669" ContentSegmentID="2638101">
67                      <fr:Text><![CDATA[<b> Βαρβαρότητα </b>]]></fr:Text>
68                    </fr:Segment>
69                  </fr:Segments>
70                </fr:Node>
71                <fr:Node ID="3052585" Name="Τίτλος" TypeName="Title" Visibility="0">
72                  <fr:Segments>
73                    <fr:Segment PageID="47669" ContentSegmentID="2638102">
74                      <fr:Text><![CDATA[<b>made in U S A </b> ]]></fr:Text>
75                    </fr:Segment>
76                  </fr:Segments>
77                </fr:Node>
78                <fr:Node ID="3052586" Name="Subtitle" TypeName="subtitle" Visibility="0">
79                  <fr:Segments>
80                    <fr:Segment PageID="47669" ContentSegmentID="2638103">
81                      <fr:Text><![CDATA[<b> Γενική κατακραυγή </b> <b> στην Ευρώπη </b> ]]></fr:Text>
82                    </fr:Segment>
83                  </fr:Segments>
84                </fr:Node>
85                <fr:Node ID="3052587" Name="Paragraph" TypeName="Paragraph" Visibility="0">
86                  <fr:Segments>
87                    <fr:Segment PageID="47669" ContentSegmentID="2638104">
88                      <fr:Text><![CDATA[ Ο κόσμος στη Μέση Ανατολή δεν έγινε ασφαλέστερος μετά την εκτέλεση του Σαντάμ Χουσεΐν. Με αμερικανικό χέρι , οι δήμιοι θανάτωσαν τον δικτάτορα με την ίδια βαρβαρότητα που σκότωνε κι αυτός κάποτε τους εχθρούς του. ]]></fr:Text>
89                    </fr:Segment>
90                  </fr:Segments>
91                </fr:Node>
```

```
92              <fr:Node ID="3052588" Name="Υπότιτλος" TypeName="subtitle"
    Visibility="0">
93                <fr:Segments>
94                  <fr:Segment PageID="47669" ContentSegmentID="2638105">
95                    <fr:Text><![CDATA[ <b> Βαθαίνει το </b> <b> αδιέξοδο στο
    Ιράκ. <br /></b> <b> Φόβοι για τριχοτόμηση </b> ]]></fr:Text>
96                  </fr:Segment>
97                </fr:Segments>
98              </fr:Node>
99              <fr:Node ID="3052599" Name="Picture" TypeName="Picture"
    Visibility="0">
100               <fr:Segments>
101                 <fr:Segment PageID="47669" ContentSegmentID="2638116" />
102               </fr:Segments>
103             </fr:Node>
104             <fr:Node ID="3052622" Name="Υπότιτλος" TypeName="subtitle"
    Visibility="0">
105               <fr:Segments>
106                 <fr:Segment PageID="47669" ContentSegmentID="2638138">
107                   <fr:Text><![CDATA[ <p><b> Το τέλος ενός αιμοσταγούς
    δικτάτορα </b></p>]]></fr:Text>
108                 </fr:Segment>
109               </fr:Segments>
110             </fr:Node>
111             <fr:Node ID="3052623" Name="Υπογραφή" TypeName="Signature"
    Visibility="0">
112               <fr:Segments>
113                 <fr:Segment PageID="47669" ContentSegmentID="2638139">
114                   <fr:Text><![CDATA[ <b> Του Ρόμπερτ Φισκ </b> ]]></fr:Text
    >
115                 </fr:Segment>
116               </fr:Segments>
117             </fr:Node>
118             <fr:Node ID="3068445" Name="Σοκ από τις εικόνες  του απαγχονισμού
      " TypeName="Article" Visibility="20000">
119               <fr:Segments>
120                 <fr:Segment PageID="47669" ContentSegmentID="2754945" />
121               </fr:Segments>
122               <fr:Summary>Αποτροπιασμό προκαλούν οι εικόνες του  Σαντάμ
    Χουσεΐν με τη θηλιά στον λαιμό, που  έκαναν τον γύρο του κόσμου. Πολιτικοί και
    θρησκευτικοί ηγέτες στην Ευρώπη μιλούν για  εκδήλωση βαρβαρότητας, ενώ
    κυβερνήσεις  σπεύδουν να καταδικάσουν για μια ακόμη  φορά την επιβολή της
    θανατικής ποινής.  </fr:Summary>
123               <fr:ChildNodes>
124                 <fr:Node ID="3052597" Name="Picture" TypeName="Picture"
    Visibility="0">
125                   <fr:Segments>
126                     <fr:Segment PageID="47669" ContentSegmentID="2638114" /
    >
127                   </fr:Segments>
128                 </fr:Node>
129                 <fr:Node ID="3052590" Name="Title" TypeName="Title"
    Visibility="0">
130                   <fr:Segments>
131                     <fr:Segment PageID="47669" ContentSegmentID="2638107">
132                       <fr:Text><![CDATA[ <i> Σοκ από τις εικόνες </i> <i>
    του απαγχονισμού </i> ]]></fr:Text>
133                     </fr:Segment>
```

```
134                     </fr:Segments>
135                   </fr:Node>
136                   <fr:Node ID="3052591" Name="Paragraph" TypeName="Paragraph"
      Visibility="0">
137                     <fr:Segments>
138                       <fr:Segment PageID="47669" ContentSegmentID="2638108">
139                         <fr:Text><![CDATA[ Αποτροπιασμό προκαλούν οι εικόνες
      του Σαντάμ Χουσεΐν με τη θηλιά στον λαιμό, που έκαναν τον γύρο του κόσμου.
      Πολιτικοί και θρησκευτικοί ηγέτες στην Ευρώπη μιλούν για εκδήλωση βαρβαρότητας,
      ενώ κυβερνήσεις σπεύδουν να καταδικάσουν για μια ακόμη φορά την επιβολή της
      θανατικής ποινής. <br /> ]]></fr:Text>
140                       </fr:Segment>
141                     </fr:Segments>
142                   </fr:Node>
143                   <fr:Node ID="3052598" Name="Υπότιτλος" TypeName="subtitle"
      Visibility="0">
144                     <fr:Segments>
145                       <fr:Segment PageID="47669" ContentSegmentID="2638115">
146                         <fr:Text><![CDATA[ <b> Χαοτική εκτέλεση </b> <b> με
      ύβρεις και προσβολές ώς το τέλος </b> ]]></fr:Text>
147                       </fr:Segment>
148                     </fr:Segments>
149                   </fr:Node>
150                 </fr:ChildNodes>
151               </fr:Node>
152             </fr:ChildNodes>
153           </fr:Node>
154           <fr:Node ID="3068436" Name="Να επενδύσουμε στους νέους "
      TypeName="Article" Visibility="20000">
```

Τμ. Κωδ. 1: Τμήμα XML αρχείου του φύλλου της 02/01/2007.

ΓΙΑ ΠΑΡΑΔΕΙΓΜΑ στη γραμμή 57 βλέπουμε ότι ξεκινά ένας κόμβος (node), ο οποίος αντιστοιχεί σε ένα άρθρο. Το ότι ο συγκεκριμένος κόμβος αποτελεί άρθρο, φαίνεται από το attribute "`TypeName="Article"`" . Εντός του tag αυτού, υπάρχει μια αναφορά σε Segments (γραμμές 58–60), μία περίληψη εντός των `<fr:Summary>` tags (γραμμή 61) και στη συνέχεια, μια child node λίστα, η οποία περικλείεται σε `<fr:ChildNode>` tags (γραμμές 62–152). Εντός της λίστας, και για κάθε στοιχείο του άρθρου, υπάρχουν τα αντίστοιχα `<fr:Node>` και `<fr:Text>` tags. Ενδεικτικά, οι γραμμές 63–69 περιγράφουν ένα node, το οποίο έχει "`TypeName="Title"`" και περιέχει ένα `<fr:Text>` tag μέσα στο οποίο βρίσκεται το κείμενο του τίτλου (`<![CDATA[<b>Βαρβαρότητα</b>]]>`). Είναι επίσης εμφανές από τις γραμμές 44 και 4 αντίστοιχα, ότι το εν λόγω άρθρο ανήκει στην κατηγορία “Πρωτοσέλιδο” της θεματικής ενότητας “Κύριο Τεύχος”.

Αξίζει να σημειωθεί ότι, πολλές φορές, εντός ενός άρθρου, υπάρχουν εμφωλευμένα σχετικά άρθρα, τα οποία ακολουθούν την τυπική δομή του άρθρου (π.χ. γραμμές 118–151).

4.3 ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΤΟΥ ΑΡΧΕΙΟΥ ΤΗΣ ΕΦΗΜΕΡΙΔΑΣ

Όπως διαπιστώθηκε, το αρχείο της εφημερίδας βρίσκεται σε μορφή η οποία μπορεί ενδεχομένως να διευκολύνει τη σελιδοποίηση και την εύρεση άρθρων, σίγουρα όμως δε διευκολύνει την επεξεργασία του καθαρού κειμένου. Κάθε αρχείο, περιέχει, εκτός από το κείμενο που συναποτελεί τα άρθρα, πολλές γραμμές με πληροφορίες οι οποίες δεν είναι χρήσιμες για ένα σύστημα εξαγωγής σχέσεων. Για το λόγο αυτό, κρίθηκε σκόπιμη η δημιουργία ενός parser, ο οποίος μετατρέπει το κάθε XML αρχείο (φύλλο της εφημερίδας) σε ένα άλλης μορφής XML αρχείο, που έχει πιο απλή ιεραρχία και πιο εύκολα επεξεργάσιμη μορφή.

Πιο συγκεκριμένα, ο parser για κάθε αρχείο:

- αφαιρεί όλες τις γραμμές και tags του αρχικού αρχείου οι οποίες περιέχουν μη αξιοποιήσιμες για την επεξεργασία κειμένου πληροφορίες (segment IDs, node tags, child node tags κ.λπ.),
- αφαιρεί τα HTML tags τα οποία μορφοποιούν το κείμενο,
- διαχωρίζει τις λέξεις από τα σημεία στίξης (π.χ. μετατροπή του “λαιμό, ” σε “λαιμό , ”) προκειμένου να εντοπίζονται εύκολα τα σημεία στίξης στο κείμενο. Δόθηκε προσοχή ώστε να μη γίνει διαχωρισμός ο οποίος να επηρεάσει τη σημασιολογία της λέξης (π.χ. η λεκτική μονάδα “ό,τι” είναι μία λέξη και δε χρήζει διαχωρισμού, παρόλο που περιέχει κόμμα).
- αναδιατάσσει τα άρθρα: άρθρα τα οποία βρίσκονται εντός του κειμένου άλλων άρθρων εξάγονται ως ξεχωριστά αυτοτελή άρθρα.
- εξάγει ένα νέο αρχείο στο οποίο περιέχονται μόνο τα εξής tags:
  - `<issue>` : ριζικός κόμβος του αρχείου, ένα tag σε κάθε αρχείο.
  - `<category>` : ένα tag για κάθε κατηγορία άρθρων. Ως κατηγορίες θεωρούνται strings της μορφής <θεματική ενότητα - κατηγορία>.
  - `<article>` : ένα tag για κάθε άρθρο. Τα `<article>` tags περικλείονται από `<category>` tags. Εντός κάθε `<article>` tag περιέχονται μόνο δύο tags, τα `<title>` και `<text>` .
  - `<title>` : περιέχει τον τίτλο κάθε άρθρου
  - `<text>` : περιέχει το κείμενο του άρθρου, μαζί με τυχόν μεσότιτλους, λεζάντες εικόνων και άλλο κείμενο.

ΓΙΑ ΠΑΡΑΔΕΙΓΜΑ αν δοθεί ως είσοδος το Τμ. Κωδ. 1, η έξοδος του parser θα έχει την παρακάτω μορφή.

```
1 <?xml version="1.0" encoding="utf-8"?>
2 <Issue>
3 <section name="ΤΑ ΝΕΑ – ΚΥΡΙΟ ΤΕΥΧΟΣ – Πρωτοσέλιδο">
4 <article>
5   <title>Βαρβαρότητα . made in U S A . </title>
6   <text>
```

```
7     Γενική κατακραυγή στην Ευρώπη . Ο κόσμος στη Μέση Ανατολή δεν έγινε
      ασφαλέστερος μετά την εκτέλεση του Σαντάμ Χουσεΐν . Με αμερικανικό χέρι ,
      οι δήμιοι θανάτωσαν τον δικτάτορα με την ίδια βαρβαρότητα που σκότωνε κι αυτός
      κάποτε τους εχθρούς του . . Βαθαίνει το αδιέξοδο στο Ιράκ . Φόβοι για
      τριχοτόμηση .
8     Το τέλος ενός αιμοσταγούς δικτάτορα . Του Ρόμπερτ Φισκ .
9   </text>
10 </article>
11 <article>
12   <title>Σοκ από τις εικόνες του απαγχονισμού . </title>
13   <text>
14     Αποτροπιασμό προκαλούν οι εικόνες του Σαντάμ Χουσεΐν με τη θηλιά στον λαιμό ,
      που έκαναν τον γύρο του κόσμου . Πολιτικοί και θρησκευτικοί ηγέτες στην Ευρώπη
      μιλούν για εκδήλωση βαρβαρότητας , ενώ κυβερνήσεις σπεύδουν να καταδικάσουν
      για μια ακόμη φορά την επιβολή της θανατικής ποινής . . Χαοτική εκτέλεση με
      ύβρεις και προσβολές ώς το τέλος .
15   </text>
16 </article>
17 <article>
```

Τμ. Κωδ. 2: Τμήμα XML αρχείου μετά το parsing του φύλλου της 02/01/2007.

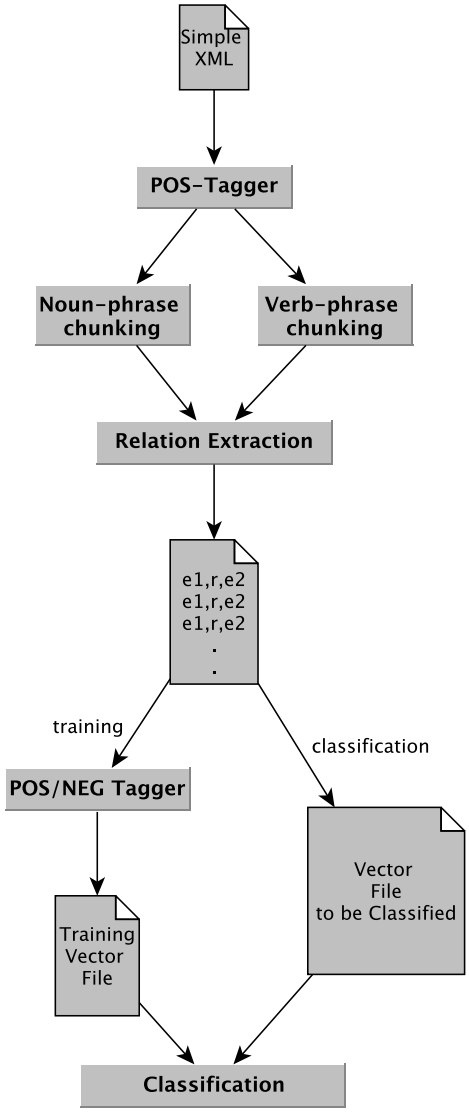
4.4 ΕΞΑΓΩΓΗ ΣΧΕΣΕΩΝ

Το δεύτερο μέρος του συστήματος δέχεται ως είσοδο ένα αρχείο κειμένου στη μορφή εξόδου του parser. Για κάθε δομικό στοιχείο κάθε άρθρου (παράγραφος, τίτλος, μεσότιτλος κ.λπ.) αρχείου εκτελεί μια σειρά βημάτων προκειμένου να εξάγει σχέσεις. Όπως αναφέρθηκε, το μέρος αυτό του συστήματος χρησιμοποιεί το αρχιτεκτονικό μόρφημα του αυλού/φίλτρου. Παρακάτω περιγράφεται κάθε στάδιο επεξεργασίας (φίλτρο) ξεχωριστά:

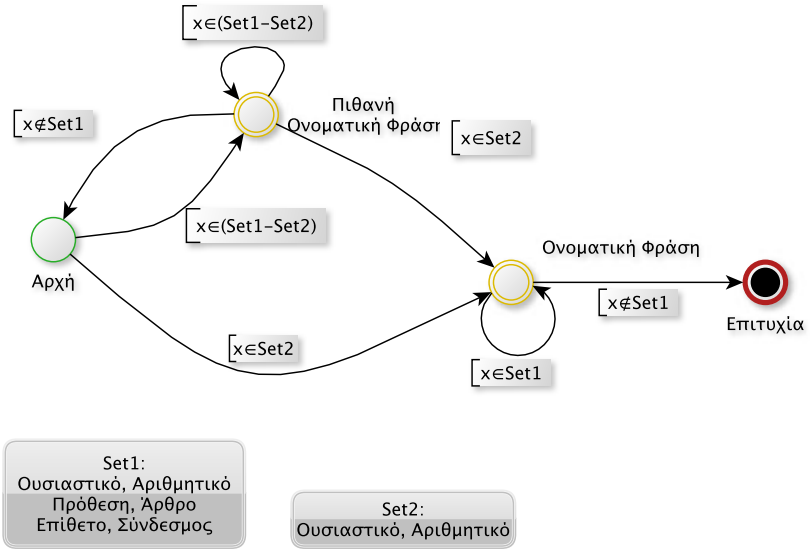
4.4.1 *Γραμματική επισημείωση (part-of-speech tagging)*

Το κείμενο εισόδου είναι απλό κείμενο, το οποίο δεν περιέχει καθόλου μεταδεδομένα ή άλλη πληροφορία η οποία θα βοηθήσει στον εντοπισμό ή στην εξαγωγή σχέσεων. Αρχικά, γίνεται μία κατάτμηση των λεκτικών μονάδων (tokenization) και στη συνέχεια επιχειρείται γραμματική επισημείωση κάθε λέξης (part-of-speech tagging). Για το σκοπό αυτό χρησιμοποιήθηκε ο POS tagger [37]. Μιας και η γραμματική επισημείωση εξαρτάται σε μεγάλο βαθμό από τα συμφραζόμενα, το κείμενο προωθείται στο σύνολό του στον POS tagger, και όχι κάθε λέξη ξεχωριστά. Ο POS tagger χαρακτηρίζει κάθε λέξη του κειμένου ως:

- ρήμα,
- ουσιαστικό,
- επίθετο,
- επίρρημα,
- άρθρο,



Σχήμα 17: Γενική περιγραφή εξαγωγής σχέσεων.



Σχήμα 18: Διάγραμμα καταστάσεων για τον εντοπισμό ονοματικών φράσεων.

- αντωνυμία,
- αριθμητικό,
- πρόθεση,
- μόριο,
- σύνδεσμος,
- σημείο στίξης ή
- άλλο.

Το κείμενο καθώς και η έξοδος-πληροφορία του POS tagger αποθηκεύονται σε ένα διάνυσμα.

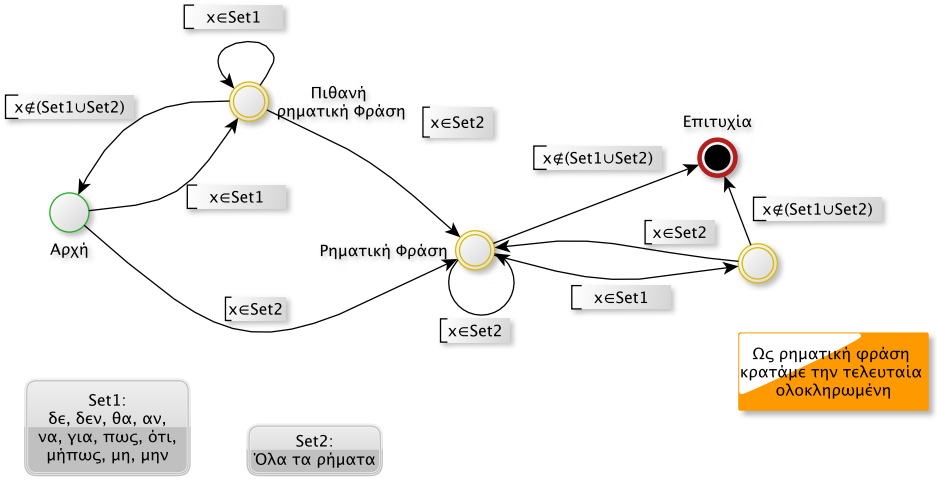
4.4.2 *Αναγνώριση ονοματικών και ρηματικών φράσεων (noun-phrase και verb-phrase chunking)*

Στη συνέχεια, το διάνυσμα προωθείται σε έναν κατατμητή ονοματικών και ρηματικών φράσεων (noun-phrase και verb-phrase chunker), ο οποίος με τη χρήση κανόνων προσπαθεί να χαρακτηρίσει σύνολα λέξεων ως “ονοματικές φράσεις” ή “ρηματικές φράσεις”.

Για την αναγνώριση ονοματικών φράσεων χρησιμοποιείται ο παρακάτω κανόνας:

ονοματική φράση θεωρούμε το σύνολο των συνεχόμενων λέξεων που ικανοποιούν τους παρακάτω περιορισμούς

1. *κάθε ονοματική φράση πρέπει να περιέχει κατ’ ελάχιστο ένα ουσιαστικό ή αριθμητικό*



Σχήμα 19: Διάγραμμα καταστάσεων για τον εντοπισμό ρηματικών φράσεων.

2. *η ονοματική φράση μπορεί να περιέχει επίσης προθέσεις, άρθρα, επίθετα ή συνδέσμους*

Στο Σχήμα 18 φαίνεται το διάγραμμα καταστάσεων για την αναγνώριση ονοματικών εκφράσεων.

Σημειώνεται ότι ο πρώτος περιορισμός αναφέρεται σε «ουσιαστικό ή αριθμητικό», καθώς σε μία πρόταση η παρουσία του αριθμητικού μπορεί να υποδηλώνει ονοματική φράση, ενώ το ουσιαστικό παραλείπεται. Για παράδειγμα, στην πρόταση *«Οι δύο ψηλοί έφυγαν από τη σχολή»*, η λέξη «δύο» είναι αριθμητικό και η λέξη «ψηλοί» επίθετο. Αν ο κανόνας περιοριζόταν μόνο στα ουσιαστικά, δε θα εντοπιζόταν η ονοματική φράση «οι δύο ψηλοί» (εννοείται «άνθρωποι»).

Για την αναγνώριση ρηματικών φράσεων χρησιμοποιείται ο παρακάτω κανόνας:

ρηματική φράση θεωρούμε το σύνολο των συνεχόμενων λέξεων που ικανοποιούν τους παρακάτω περιορισμούς

1. *κάθε ρηματική φράση πρέπει να περιέχει κατ' ανάγκη ένα ρήμα*
2. *η ρηματική φράση μπορεί να περιέχει επίσης τις λέξεις δε, δεν, θα, αν, να, για, πως, ότι, μήπως, μη, μην, έχω, έχεις, έχει, έχουμε, έχετε, έχουν, είχα, είχες, είχε, είχαμε, είχατε, είχαν.*

Το διάγραμμα καταστάσεων για την αναγνώριση ρηματικών φράσεων φαίνεται στο Σχήμα 19.

Αξίζει να σημειωθεί πως οι εργασίες κατάτμησης ονοματικών και ρηματικών φράσεων γίνονται ταυτόχρονα, καθώς μπορεί η παρουσία μιας λέξης η οποία υποδηλώνει την έναρξη ονοματικής φράσης, μπορεί να σημαίνει το τέλος μιας ρηματικής φράσης (παράδειγμα η λέξη "μη" στη φράση *"θέλουν να μη βρέξει"* δηλώνει συνέχεια της ρηματικής φράσης, ενώ στη φράση *"πιέζουν για μη εφαρμογή"* δηλώνει αρχή ονοματικής φράσης). Ένας επιπλέον περιορισμός ο οποίος τίθεται είναι ότι

κάθε λέξη μπορεί να συμμετέχει στο σχηματισμό είτε ονοματικής φράσης είτε ρηματικής φράσης, όχι και των δύο ταυτόχρονα.

Μετά τη λήξη της κατάτμησης ονοματικών και ρηματικών φράσεων, η παραγόμενη πληροφορία αποθηκεύεται χωρίς να χαθεί η πληροφορία της γραμματικής επισημείωσης κάθε λέξης. Έτσι έχουμε πλέον ένα διάνυσμα το οποίο, εκτός από τις λέξεις οι οποίες αποτελούν το κείμενο εισόδου φέρει και την πληροφορία για το μέρος του λόγου κάθε λέξης και την ονοματική ή ρηματική φράση στην οποία ανήκει.

4.4.3 *Εξαγωγή σχέσεων (relation extraction)*

Το διάνυσμα που παρήχθη στην προηγούμενη φάση περιέχει την απαραίτητη πληροφορία για την εξαγωγή σχέσεων. Υπενθυμίζεται ότι, σε πρώτη φάση, το σύστημα εξάγει ένα μεγάλο αριθμό σχέσεων και στη συνέχει αποφασίζει για κάθε μία αν είναι θετική ή αρνητική.

Ως σχέσεις ορίζονται τριάδες της μορφής (e_1, r, e_2)*, όπου* e_1, e_2 *οντότητες και* r *η ρηματική σχέση που συυδέει τις οντότητες.*

Για την επιτυχή εξαγωγή σχέσεων, αρχικά γίνεται η αναγνώριση των ρηματικών σχέσεων (r) και στη συνέχεια η αναγνώριση οντοτήτων.

Η αναγνώριση ρηματικών σχέσεων γίνεται ως εξής:

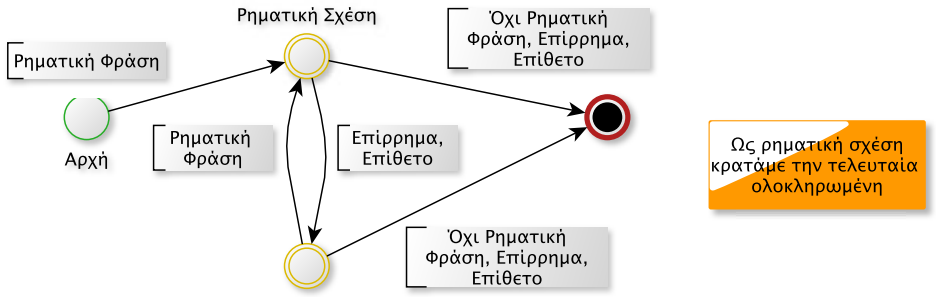
> *ως «ρηματική σχέση» θεωρούμε μία ή περισσότερες ρηματικές φράσεις οι οποίες ενώνονται με λέξη που έχει χαρακτηριστεί ως επίθετο ή επίρρημα.*

Ο ορισμός αυτός είναι αρκετά γενικός για να περιγράψει σχέσεις οι οποίες εκφράζονται με δύο συνενωμένες ρηματικές φράσεις, όπως η σχέση *«...είχαν βγει έξω για να καπνίσουν...»*. Στη σχέση αυτή, ο verb-phrase chunker σημείωσε τη φράση *«είχαν βγει»* ως ρηματική φράση και τη φράση *«για να καπνίσουν»* επίσης ως ρηματική φράση. Με την αναγνώριση του επιρρήματος "έξω" οι δύο ρηματικές φράσεις συνενώθηκαν για να εκφράσουν μία ρηματική σχέση.

ΓΙΑ ΝΑ ΕΠΙΤΕΥΧΘΕΙ Ο ΕΝΤΟΠΙΣΜΟΣ ΡΗΜΑΤΙΚΩΝ ΣΧΕΣΕΩΝ, το σύστημα εντοπίζει μια ρηματική φράση και στη συνέχεια ψάχνει αριστερά της (το διάνυσμα διαβάζεται από δεξιά προς τα αριστερά) για επίθετα και επιρρήματα. Αν εντοπιστούν, ψάχνει για δεύτερη ρηματική φράση. Αν, ενόσω ψάχνει για δεύτερη ρηματική φράση συναντήσει λέξη η οποία δεν είναι επίθετο ή επίρρημα, κρατά ως σχέση μόνο την πρώτη ρηματική φράση. Στο Σχήμα 20 φαίνεται το διάγραμμα καταστάσεων για την αναγνώριση ρηματικών σχέσεων.

ΑΦΟΥ ΕΝΤΟΠΙΣΤΕΙ ΜΙΑ ΡΗΜΑΤΙΚΗ ΣΧΕΣΗ, το σύστημα ψάχνει αριστερά της για την κοντινότερη οντότητα. Η αναγνώριση οντοτήτων γίνεται με τον εξής απλό κανόνα:

Ως οντότητα θεωρείται μία ονοματική φράση.



Σχήμα 20: Διάγραμμα καταστάσεων για τον εντοπισμό ρηματικών σχέσεων.

Αν στην αναζήτηση αριστερά της ρηματικής σχέσης, το σύστημα συναντήσει τελεία ή ερωτηματικό, τότε η ρηματική σχέση εγκαταλείπεται και αναζητείται η επόμενη.

Στο σημείο αυτό αξίζει να παρατηρήσουμε πως κατά τη διάρκεια της υλοποίησης του συστήματος δοκιμάστηκαν διάφορες λύσεις για το πρόβλημα της "πρόβλεψης" της οντότητας στην οποία αναφέρεται η "ορφανή" (χωρίς ονοματική φράση αριστερά της, εντός της ίδιας πρότασης) ρηματική σχέση. Οι προσεγγίσεις αυτές (ενδεικτικά αναφέρονται: να θεωρείται πως η ρηματική σχέση αναφέρεται στην κοντινότερη οντότητα της προηγούμενης πρότασης, να θεωρείται πως η ρηματική σχέση αναφέρεται στην κοντινότερη οντότητα της προηγούμενης πρότασης η οποία συμφωνεί κατά αριθμό με το ρήμα κ.ά.) δεν έδωσαν τα αναμενόμενα αποτελέσματα. Η αποτυχία της πρόβλεψης οφείλεται στο γεγονός ότι στην ελληνική γλώσσα, πολλές φορές το υποκείμενο μιας ρηματικής φράσης εννοείται (συνάγεται από τα συμφραζόμενα ή συνάγεται από τον αριθμό και την κατάληξη του ρήματος).

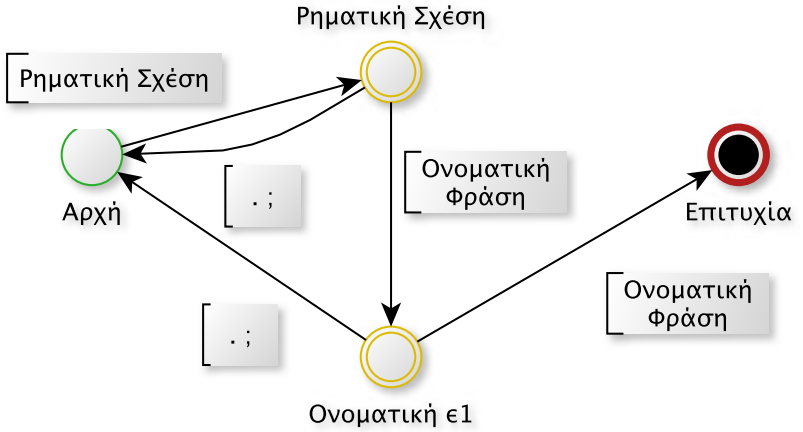
ΑΝ ΕΝΤΟΠΙΣΤΕΙ ΟΝΤΟΤΗΤΑ ΑΡΙΣΤΕΡΑ ΤΗΣ ΡΗΜΑΤΙΚΗΣ ΣΧΕΣΗΣ, το σύστημα ψάχνει δεξιά της για τη δεύτερη οντότητα στην οποία αναφέρεται η ρηματική σχέση. Παρόμοια με παραπάνω, αν συναντήσει τελεία ή ερωτηματικό, εγκαταλείπει τη ρηματική σχέση και συνεχίζει την αναζήτηση της επόμενης. Αν βρεθεί και δεύτερη οντότητα, η συμπληρωμένη πλέον τριάδα (e_1, r, e_2) αποθηκεύεται και συνεχίζεται η αναζήτηση.

Η διαδικασία επαναλαμβάνεται μέχρι να εξαντληθούν όλες οι ρηματικές φράσεις του κειμένου.

4.4.4 *Εξαγωγή διανυσμάτων χαρακτηριστικών (feature vectors)*

Για κάθε μία από τις σχέσεις που εξήχθησαν στο προηγούμενο βήμα, δημιουργείται ένα διάνυσμα χαρακτηριστικών (feature vector). Το διάνυσμα αυτό περιέχει τις εξής πληροφορίες (κάθε χαρακτηριστικό έχει έναν διατακτικό αριθμό):

- αριθμός λεκτικών μονάδων (tokens) της σχέσης (1)
- αριθμός stopwords της σχέσης (2)
- part-of-speech tag της λέξης αριστερά της e_1 (3)



Σχήμα 21: Διάγραμμα καταστάσεων για τον εντοπισμό σχέσης (τριάδας).

- part-of-speech tag της λέξης δεξιά της e_1 (4)
- part-of-speech tag της λέξης αριστερά της e_2 (5)
- part-of-speech tag της λέξης δεξιά της e_2 (6)
- part-of-speech tag της πρώτης λέξης της ρηματικής σχέσης r (7)
- 0/1 (δυαδική τιμή) αν υπάρχει κύριο όνομα στην e_1 (8)
- 0/1 (δυαδική τιμή) αν υπάρχει κύριο όνομα στην e_2 (9)
- το κείμενο της σχέσης (για λόγους αναφοράς)

Τα διανύσματα αυτά αποθηκεύονται σε ένα αρχείο εξόδου, όπου χρησιμοποιείται μία γραμμή για κάθε σχέση. Στο αρχείο αυτό χρησιμοποιείται η κωδικοποίηση <*διατακτικός αριθμός χαρακτηριστικού*>:<*τιμή χαρακτηριστικού*>. Στο Τμ. Κωδ. 3 φαίνεται τμήμα από ένα τέτοιο αρχείο.

```
1  #vector output
2 0 1:5 2:0 3:4 4:6 5:1 6:11 7:1 8:0 9:0 # την( ομάδα, συνέλαβε, τη 17 Ν)
3 0 1:12 2:0 4:1 5:4 6:11 7:1 8:1 9:0 # Το( ναυάγιο της ΔΕΗ, οδήγησε, στην
    παραίτηση του διευθύνοντος συμβούλου της εταιρείας)
4 0 1:6 2:0 4:1 5:1 6:11 7:1 8:0 9:0 # Το( νέφος, αυξάνει, 10% τους θανάτους)
5 0 1:6 2:0 3:11 4:1 5:12 6:6 7:1 8:0 9:0 # υγρασία( και συννεφιά, αυξάνουν,
    τον κίνδυνο)
6 0 1:11 2:0 3:4 4:11 5:1 6:6 7:1 8:0 9:0 # τα( 65 του χρόνια, μεγαλώνει, όμως ο
    φόβος για τη ζωή)
7 0 1:8 2:0 4:1 5:4 6:11 7:1 8:1 9:0 # Ο( Μοχάμεντ Άλι, συμπληρώνει, τα 65 του
    χρόνια)
8 0 1:12 2:0 3:11 4:1 5:1 6:11 7:1 8:0 9:0 # Η( γυμναστική και οι αγώνες, είναι,
    τρόπος ζωής για τις πρωταθλήτριες γιαγιάδες)
```

Τμ. Κωδ. 3: Τμήμα αρχείου με διανύσματα χαρακτηριστικών.

Ας εστιάσουμε στη γραμμή 2 του αρχείου:

- Το 0 στην αρχή δείχνει ότι η σχέση αυτή δεν έχει αποφασιστεί αν είναι θετική (τιμή +1) ή αρνητική (τιμή -1).
- Το "1:5" στη συνέχεια, δείχνει ότι η σχέση αποτελείται από πέντε (5) λεκτικές μονάδες.
- Το "2:0" δηλώνει ότι δεν υπάρχουν καθόλου stopwords στη σχέση.
- Το "3:4" δείχνει ότι η λέξη η οποία βρίσκεται αριστερά της e_1 στο κείμενο είναι επίρρημα[1].
- Το "4:6" δείχνει ότι η λέξη η οποία βρίσκεται δεξιά της e_1 στο κείμενο είναι πρόθεση.
- Το "5:1" δηλώνει ότι η λέξη η οποία βρίσκεται αριστερά της e_2 είναι ρήμα.
- Το "6:11" δηλώνει ότι δεξιά της e_2 βρίσκεται σημείο στίξης.
- Το "7:1" δείχνει ότι η πρώτη λέξη της r είναι ρήμα.
- Το "8:0" δείχνει ότι δεν υπάρχει κύριο όνομα στην e_1.
- Το "9:0" δηλώνει ότι στην e_2 δεν υπάρχει κύριο όνομα.
- Μετά το "#" βρίσκεται το κείμενο της σχέσης.

Σχέσεις για τις οποίες τα χαρακτηριστικά 1, 2, 3, 4, 5, 6 και 7 παίρνουν μηδενικές τιμές, δεν εμφανίζουν καθόλου τα χαρακτηριστικά αυτά στο διάνυσμά τους.

Η ΜΟΡΦΗ ΑΥΤΗ του αρχείου εξόδου επιλέχθηκε για να διευκολύνει το επόμενο στάδιο του classification. Αν ο χρήστης έχει επιλέξει τη δημιουργία δεδομένων εκπαίδευσης για τον classifier, πριν τη δημιουργία του αρχείου διανυσμάτων χαρακτηριστικών, καλείται το στάδιο της βαθμολόγησης σχέσεων, το οποίο περιγράφεται στην επόμενη ενότητα.

4.4.5 *Βαθμολόγηση σχέσεων για τη δημιουργία δεδομένων εκπαίδευσης*

Το στάδιο αυτό εκτελείται μόνον αν ο χρήστης επιθυμεί τη δημιουργία δεδομένων εκπαίδευσης. Όπως αναφέρθηκε παραπάνω, η τελική επιλογή των θετικών σχέσεων γίνεται από έναν classifier. Προκειμένου να μπορεί ο classifier να βαθμολογήσει σωστά τις σχέσεις, πρέπει να εκπαιδευτεί με ένα αρκετά μεγάλο σύνολο σχέσεων οι οποίες είναι ήδη επισημειωμένες ως θετικές ή αρνητικές. Αν επιλεγεί η δημιουργία δεδομένων εκπαίδευσης, το σύστημα επιτελεί μια εργασία παραπάνω: την αναγνώριση των σχέσεων που εξήχθησαν ως θετικές ή αρνητικές με βάση κάποιους κανόνες.

1 έχει γίνει απαρίθμηση των μερών του λόγου τα οποία δίνει ο POS tagger.

οι κανονες με τους οποίους αποφασίζεται αν μια σχέση θα θεωρείται θετική ή αρνητική είναι οι εξής:

Μια σχέση θεωρείται αρνητική, αν:

- *ο συνολικός αριθμός των λεκτικών μονάδων της είναι μικρότερος του 6 και οι οντότητες (e_1, e_2) δεν περιέχουν κανένα κύριο όνομα.*
- *η πρώτη λέξη της r έχει χαρακτηριστεί ως "πρόθεση" από τον POS tagger.*
- *η r αποτελείται από δύο μόνο λέξεις, από τις οποίες η πρώτη έχει χαρακτηριστεί ως σύνδεσμος και η δεύτερη ως ρήμα.*
- *αν η πρώτη λέξη της e_1 έχει χαρακτηριστεί ως επίθετο.*
- *αν η πρώτη λέξη της e_2 έχει χαρακτηριστεί ως σύνδεσμος, χωρίς να είναι η λέξη «και» ή η λέξη «για».*
- *αν η τελευταία λέξη της e_1 είναι η λέξη «και»*

Σε κάθε άλλη περίπτωση, η σχέση θεωρείται θετική.

Για την αναγνώριση κυρίων ονομάτων που απαιτείται από τον πρώτο κανόνα:

- χρησιμοποιήθηκαν εξωτερικές λίστες ελληνικών ονομάτων, επιθέτων και ονομάτων οργανισμών [1]
- για την αναγνώριση κυρίων ονομάτων τα οποία δεν εμπεριέχονται στις λίστες (μεταγραφή ξένων ονομάτων στα ελληνικά, όπως «Σαντάμ Χουσεΐν» ή άλλων), χρησιμοποιήθηκε ο κανόνας:

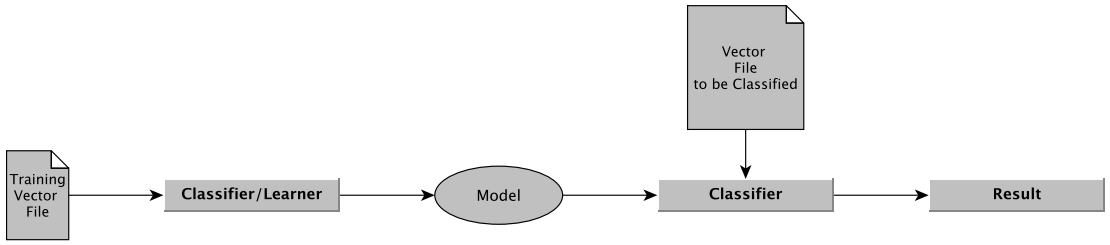
  όποια λέξη ξεκινά από κεφαλαίο γράμμα, χωρίς να είναι πρώτη λέξη πρότασης, θεωρείται κύριο όνομα

  μιας και στην ελληνική γλώσσα τα κύρια ονόματα ξεκινούν με κεφαλαίο γράμμα.

Δεν υπάρχει κάποιο σύνολο δεδομένων στην ελληνική γλώσσα το οποίο να περιέχει σχέσεις που γνωρίζουμε πως είναι αρνητικές ή θετικές και για το λόγο αυτό έπρεπε να κατασκευαστεί κάποιο τέτοιο σύνολο. Για το σκοπό αυτό χρησιμοποιήθηκαν εμπειρικοί κανόνες οι οποίοι παραμένουν όμως λειτουργικοί όσον αφορά στο σκοπό της εξαγωγής σχέσεων. Η επιβεβαίωση από ανθρώπους της ορθότητας ή όχι των εξαγόμενων σχέσεων μπορεί να οδηγήσει σε πιο ακριβές σύνολο δεδομένων εκπαίδευσης.

Όταν δημιουργούνται δεδομένα εκπαίδευσης, στο vector κάθε σχέσης, ο πρώτος αριθμός δείχνει αν η σχέση είναι θετική ή αρνητική. Παράδειγμα αρχείου με διανύσματα χαρακτηριστικών δεδομένων εκπαίδευσης, φαίνεται στο Τμ. Κωδ. 4. Όπως παρατηρούμε, έχει ακριβώς την ίδια δομή με το αρχείο εξόδου που περιγράφηκε στην προηγούμενη ενότητα, με τη διαφορά ότι ο πρώτος αριθμός καθορίζει αν η σχέση θεωρείται θετική (+1) ή αρνητική (-1).

```
1 #vector output
2 +1 1:1 2:0 3:1 4:6 5:4 6:6 7:1 8:0 9:0 # τον( δικτάτορα με την ίδια βαρβαρότητα,
    σκότωνε, τους εχθρούς)
```



Σχήμα 22: Διάγραμμα της διαδικασίας του classification.

```
3 +1 1:1 2:0 3:11 4:11 5:1 6:6 7:1 8:0 9:0 # Με( αμερικανικό χέρι , θανάτωσαν,
      τον δικτάτορα με την ίδια βαρβαρότητα)
4 +1 1:2 2:1 3:11 4:9 5:1 6:11 7:9 8:1 9:1 # (Ο κόσμος στη Μέση Ανατολή, δεν
      έγινε, ασφαλέστερος μετά την εκτέλεση του Σαντάμ Χουσεΐν)
5 +1 1:3 2:1 3:11 4:1 5:1 6:4 7:1 8:0 9:0 # ενώ( κυβερνήσεις, σπεύδουν να
      καταδικάσουν, για μια)
6 +1 1:1 2:0 3:11 4:1 5:1 6:11 7:1 8:1 9:0 # Πολιτικοί( και θρησκευτικοί ηγέτες στην
      Ευρώπη, μιλούν, για εκδήλωση βαρβαρότητας)
7 -1 1:2 2:1 4:11 4:4 5:1 6:11 7:9 8:0 9:0 # Στην( πολιτική, δεν είναι, θέσφατο )
```

Τμ. Κωδ. 4: Τμήμα αρχείου με διανύσματα χαρακτηριστικών και χαρακτηρισμό κάθε σχέσης ως θετικής ή αρνητικής.

4.5 ΕΚΠΑΙΔΕΥΣΗ ΤΟΥ ΤΑΞΙΝΟΜΗΤΗ

Το τελευταίο στάδιο του συστήματος αποτελεί ένας classifier. Όπως ήδη έχει αναφερθεί, η λειτουργία του classifier μπορεί να χωριστεί σε δύο στάδια, αυτό της εκπαίδευσης και αυτό της ταξινόμησης. Για τη σωστή λειτουργία του classifier είναι απραίτητη η εκτέλεση και των δύο σταδίων και μάλιστα με τη σειρά που αναφέρθηκαν. Για το λόγο αυτό, το σύστημα, μετά τη δημιουργία δεδομένων εκπαίδευσης καλεί τον classifier προκειμένου να τον εκπαιδεύσει.

Έξοδος του σταδίου αυτού είναι ένα αρχείο που περιγράφει ένα μοντέλο, με βάση το οποίο ο classifier θα μπορεί στο μέλλον να κατατάσσει νέες σχέσεις στις δύο κατηγορίες (θετικές, αρνητικές).

4.6 ΤΑΞΙΝΟΜΗΣΗ

Ο εκπαιδευμένος, πλέον, classifier μπορεί να βαθμολογήσει κάθε νέα σχέση ως θετική ή αρνητική. Ως είσοδο δέχεται το μοντέλο το οποίο παρήχθη κατά την εκπαίδευσή του, καθώς και ένα αρχείο με διανύσματα χαρακτηριστικών των νέων, αταξινόμητων σχέσεων. Στην έξοδό του παράγεται ένα αρχείο με έναν αριθμό για κάθε σχέση εισόδου, ο οποίος δηλώνει κατά πόσον η σχέση θεωρείται θετική ή αρνητική. Φυσικά, προκειμένου να ληφθεί το τελικό αποτέλεσμα (οι θετικές σχέσεις), πρέπει να οριστούν κατώφλια, με βάση τα οποία τελικά θεωρούμε ότι μία σχέση είναι θετική.

για παραδειγμα, ο classifier μπορεί να αποδώσει σε μια σχέση την τιμή $0,7$. Επιλέγοντας μια ελαστικότητα της τάξης του $\pm 30\%$, η σχέση αυτή θα θεωρηθεί θετική. Αντίθετα, μια σχέση στην οποία έχει αποδοθεί ο αριθμός $0,5$ δε θα θεωρηθεί θετική. Παρόμοια, μια σχέση στην οποία έχει αποδοθεί η τιμή -0.7 θα θεωρηθεί αρνητική.

4.7 ΟΜΑΔΟΠΟΙΗΣΗ

Μετά την επιτυχή εξαγωγή σχέσεων, προκειμένου οι χαρακτηρισμένες ως θετικές σχέσεις να χωριστούν σε ομάδες με ίδιο ή παρεμφερές σημασιολογικό περιεχόμενο, επιχειρείται μια ομαδοποίηση (clustering). Για την επιτυχή ομαδοποίηση των παρατηρήσεων, δημιουργείται ένα νέο σύνολο διανυσμάτων, τα οποία αντλούν πληροφορία αποκλειστικά από το κείμενο που απαρτίζει τις σχέσεις.

Πιο συγκεκριμένα, διατρέχεται το σύνολο των παρατηρήσεων και καταγράφεται πόσοι και ποιοι διαφορετικοί όροι (terms) αποτελούν τις σχέσεις. Προκειμένου να αποφευχθεί σύγχυση μεταξύ των διάφορων τύπων στους οποίους μπορεί να εμφανίζεται μία λέξη (π.χ. υπολογιστής, υπολογιστές, υπολογιστή κ.λπ.) αποθηκεύεται το στέμμα (ρίζα) της λέξης. Για την εύρεση της ρίζας της λέξης χρησιμοποιείται ένας περιστολέας λέξεων [26] για την ελληνική γλώσσα, ο οποίος αποκόπτει την κάθε λέξη από τυχόν καταλήξεις ή προθέματα και φανερώνει τη ρίζα της. Η είσοδος του περιστολέα είναι μία και μόνο λέξη σε κεφαλαία και χωρίς τονισμό. Για το λόγο αυτό από κάθε λέξη κάθε σχέσης αφαιρείται ο τονισμός, γίνεται η μετατροπή από πεζά σε κεφαλαία και στη συνέχεια οδηγείται στον περιστολέα.

Κατά την καταγραφή των όρων, σε ένα hash set, καταγράφεται και το πλήθος των σχέσεων που εμφανίζουν κάθε όρο. Στη συνέχεια, οι σχέσεις διατρέχονται ξανά και για κάθε μία δημιουργείται ένα διάνυσμα το οποίο περιέχει ως χαρακτηριστικά την TF-IDF τιμή για κάθε έναν από τους όρους που βρέθηκε ότι εμφανίζονται στο σύνολο των σχέσεων. Όλα τα διανύσματα όλων των σχέσεων αποθηκεύονται σε ένα αρχείο εξόδου (αρχείο διανυσμάτων χαρακτηριτικών προς ομαδοποίηση).

Στη συνέχεια, το αρχείο αυτό δίνεται ως είσοδος σε ένα σύστημα ομαδοποίησης [2], το οποίο δέχεται ως παράμετρο το επιθυμητό πλήθος ομάδων στις οποίες ζητείται να ομαδοποιήσει τις παρατηρήσεις (σχέσεις). Έξοδος του συστήματος ομαδοποίησης είναι ένα αρχείο το οποίο περιγράφει τα αποτελέσματα της ομαδοποίησης, καθώς και στοιχεία περί ομοιότητας των παρατηρήσεων που κατατάχθηκαν στην ίδια ομάδα.

Το πλήθος των ομάδων στις οποίες γίνεται η κατάταξη των σχέσεων είναι καθοριστικής σημασίας για την ορθή ομαδοποίησή τους. Στα πλαίσια της παρούσης, στις δοκιμές ομαδοποίησης των σχέσεων επιλέχθηκε μεγάλος αριθμός ομάδων (μεγαλύτερος από το μισό του πλήθους των σχέσεων). Η επιλογή αυτή έγινε διότι καθώς το σώμα κειμένου αποτελείται από πολλά και διαφορετικής θεματολογίας άρθρα, η πιθανότητα μεγάλος αριθμός εξαγόμενων σχέσεων να χρίζει κατάταξης στην ίδια ομάδα είναι μικρή. Ένας μικρός αριθμός ομάδων θα οδηγήσει σε ομάδες με αρκετά διαφορετικές σημασιολογικά μεταξύ τους σχέσεις.

Αξίζει να αναφερθεί, κατά τη διάρκεια του ελέγχου της ομαδοποίησης που πραγματοποιήθηκε στα πλαίσια της εργασίας αυτής, επιχειρήθηκε και η ομαδοποίηση με βάση τα διανύσματα τα οποία χρησιμοποιήθηκαν κατά την εξαγωγή και βαθμολογία των σχέσεων. Το αποτέλεσμα αυτής της προσέγγισης ήταν μία ομαδοποίηση για την οποία οι ομάδες αποτελούνταν από (σχεδόν) τελείως διαφορετικές σημασιολογικά σχέσεις · αποτέλεσμα αναμενόμενο, μιας και η "δομική" ομοιότητα (ως προς τα μέρη του λόγου που αποτελούν τις οντότητες και τη ρηματική σχέση, την παρουσία κυρίων ονομάτων κ.λπ.) δε συνεπάγεται σημασιολογική ομοιότητα.

Ένα παράδειγμα σχέσεων οι οποίες ομαδοποιούνται στην ίδια ομάδα (cluster) φαίνεται στο Τμ. Κωδ. 5. Ένα πληρέστερο παράδειγμα της εξόδου ομαδοποίησης με βάση τα διανύσματα TF-IDF βαρών, φαίνεται στο Παράρτημα ii.

```
1 ( στο βάθρο του παγκόσμιου πρωταθλήματος , κάνοντας , την ομάδα των Αμερικανών)
2 ( ο Βλάσης Μάρας , ανέβηκε , στο βάθρο του παγκόσμιου κυπέλλου ενόργανης γυμναστικής)
3 ( Σοφοκλής Σχορτσιανίτης , κέρδισε , τις εντυπώσεις στη διάρκεια του παγκόσμιου
    πρωταθλήματος)
```

Τμ. Κωδ. 5: Παράδειγμα ομάδας η οποία προέκυψε μετά την ομαδοποίηση (clustering).

5

ΖΗΤΗΜΑΤΑ ΥΛΟΠΟΙΗΣΗΣ

Στο κεφάλαιο αυτό αναλύονται κάποια ζητήματα υλοποίησης του συστήματος, η περιγραφή των οποίων κρίνεται απαραίτητη για την κατανόηση της λειτουργίας του.

Για την υλοποίηση του συστήματος χρησιμοποιήθηκε η γλώσσα JAVA 7. Παρότι η εκτέλεση του συστήματος δοκιμάστηκε σε περιβάλλον linux, το γεγονός ότι έχει υλοποιηθεί σε JAVA επιτρέπει την εκτέλεσή του σε όλες τις πλατφόρμες. Επίσης, τα εργαλεία που χρησιμοποιεί το σύστημα έχουν και αυτά υλοποιηθεί σε JAVA ή διαθέτουν εκδόσεις για διαφορετικά λειτουργικά συστήματα.

5.1 ΕΡΓΑΛΕΙΑ

Στην ενότητα αυτή περιγράφονται συνοπτικά τα εργαλεία που χρησιμοποιήθηκαν για την υλοποίηση του συστήματος.

5.1.1 *Java XML Parsers*

Για την εύκολη και αποδοτική διαχείριση των XML αρχείων από τα οποία αποτελείται το αρχείο της εφημερίδας «ΤΑ ΝΕΑ», καθώς και των αρχείων ενδιάμεσης μορφής τα οποία προέκυψαν μετά από την προεπεξεργασία των φύλλων της εφημερίδας, χρησιμοποιήθηκε το πακέτο `javax.xml.parsers` της γλώσσας JAVA, το οποίο περιέχει κλάσεις και μεθόδους διαχείρισης και τεχνολόγησης (parsing) XML αρχείων οποιασδήποτε μορφής.

5.1.2 *Ελληνικός επισημειωτής μερών του λόγου (POS tagger)*

Για την γραμματική επισημείωση χρησιμοποιήθηκε ένας επισημειωτής μερών του λόγου για την ελληνική γλώσσα [37]. Μιας και η εργασία της γραμματικής επισημείωσης είναι μία διαδικασία ταξινόμησης, τον πυρήνα του επισημειωτή αποτελεί ένας classifier. Ο συγκεκριμένος επισημειωτής είναι βασισμένος στον ταξινομητή μεγίστης εντροπίας (maximum entropy classifier) του Πανεπιστημίου Stanford [3].

Ο συγκεκριμένος επισημειωτής έχει δύο τρόπους λειτουργίας:

- *λειτουργία βασικού συνόλου*: κάθε λέξη ταξινομείται σε μία από τις κατηγορίες: ρήμα, ουσιαστικό, επίθετο, επίρρημα, άρθρο, αντωνυμία, αριθμητικό, πρόθεση, μόριο, σύνδεσμος, σημείο στίξης, άλλο.
- *λειτουργία εκτεταμένου συνόλου*: κάθε λέξη κατατάσσεται σε μία από 170 συνολικά κατηγορίες, οι οποίες παρέχουν περισσότερες πληροφορίες από αυτές του βασικού συνόλου, όπως γένος, πτώση, αριθμός κ.λπ.

Στα πλαίσια της παρούσης, χρησιμοποιήθηκε το βασικό σύνολο κατηγοριών, καθώς η επιπλέον πληροφορία δεν ήταν χρήσιμη για την διαδικασία εξαγωγής σχέσεων.

Υπενθυμίζεται ότι, καθώς η αναγνώριση του μέρους του λόγου στο οποίο ανήκει μια λέξη εξαρτάται σε μεγάλο βαθμό από τα συμφραζόμενα, η ορθή λειτουργία του επισημειωτή προϋποθέτει την παρουσία κειμένου στην είσοδο και όχι μίας μόνο λέξης.

Παρ' ότι δεν υπάρχει μεγάλος αριθμός διαθέσιμων γραμματικών επισημειωτών για την ελληνική γλώσσα, ο συγκεκριμένος παρουσιάζει αρκετά καλή απόδοση, Για το λόγο αυτό κατέστη δυνατό να χρησιμοποιηθεί ως βασικό εργαλείο για τη συγκεκριμένη υλοποίηση εξαγωγής σχέσεων μεταξύ οντοτήτων, η οποία στηρίζεται αρκετά στην σωστή επισημείωση μερών του λόγου για κάθε λέξη. Ωστόσο, το γεγονός ότι η λειτουργία του POS tagger βασίζεται σε ταξινομητή σημαίνει, φυσικά, πως το αποτέλεσμά του δεν είναι εγγυημένα σωστό. Η λανθασμένη απόδοση μερών του λόγου στις λέξεις είναι ένας από τους λόγους λανθασμένης εξαγωγής σχέσεων του συστήματος.

5.1.3 *Ελληνικός περιστολέας λέξεων (stemmer)*

Για την εύρεση της ρίζας κάθε λέξης χρησιμοποιήθηκε ο ελληνικός περιστολέας λέξεων [26]. Η υλοποίηση του συγκεκριμένου περιστολέα βασίζεται σε ένα σύμπλεγμα κανόνων, οι οποίοι έχουν προκύψει από βιβλιογραφική έρευνα και παρατήρηση. Οι κανόνες τους οποίους χρησιμοποιεί ο περιστολέας αφορούν στην αφαίρεση χαρακτηριστικών καταλήξεων, προθεμάτων και άλλων στοιχείων από κάθε λέξη, προκειμένου να παραχθεί η ρίζα της ως έξοδος.

Ο stemmer δέχεται ως είσοδο μία και μόνο λέξη κάθε φορά και εξάγει τη ρίζα (στέμμα) της.

5.1.4 SVM^{light} *classifier*

Η εφαρμογή SVM^{light} είναι μία υλοποίηση του μοντέλου SVM για ταξινόμηση παρατηρήσεων. Οι αλγόριθμοι βελτιστοποίησης στους οποίους βασίζεται η λειτουργία του περιγράφονται στα [19], [20]. Το SVM^{light} αποτελείται από δύο ξεχωριστά εκτελέσιμα, το `svm_learn` και το `svm_classfify` .

Το `svm_learn` δέχεται ως όρισμα ένα αρχείο διανυσμάτων κατάστασης συγκεκριμένης μορφοποίησης (η μορφοποίηση αναλύθηκε στην Ενότητα 4.4.4). Το αρχείο αυτό δίνεται στο `svm_learn` με σκοπό να εκπαιδευτεί ο classifier, γι' αυτό περιέχει, εκτός των άλλων χαρακτηριστικών και μία, όσο γίνεται έμπιστη τιμή στη μεταβλητή ως προς την οποία θέλουμε να ταξινομεί ο τελικός ταξινομητής. Στην έξοδό του, παράγεται ένα αρχείο (model file) το οποίο περιγράφει ένα μοντέλο, με βάση το οποίο θα μπορέσει να γίνει η ταξινόμηση των παρατηρήσεων. Το `svm_learn` δέ-

χεται επίσης πολλές παραμέτρους για την επιλογή διαφορετικών kernel functions, επιλογές optimization κ.λπ.

Το `svm_classify` δέχεται ως όρισμα ένα αρχείο διανυσμάτων τα οποία είναι αταξινόμητα και το αρχείο του μοντέλου που παρήχθη από το `svm_learn` . Στην έξοδό του παίρνουμε ένα αρχείο, το οποίο περιέχει τα αποτελέσματα της ταξινόμησης.

5.1.5 CLUTO

Για την ομαδοποίηση σχέσεων μετά το πέρας της εξαγωγής και ταξινόμησής τους χρησιμοποιήθηκε η εφαρμογή CLUTO[2]. Πρόκειται για μία εφαρμογή η οποία ομαδοποιεί παρατηρήσεις οι οποίες περιγράφονται από διανύσματα με πολλά χαρακτηριστικά. Η είσοδος που δέχεται η συγκεκριμένη εφαρμογή είναι ένα αρχείο διανυσμάτων χαρακτηριστικών (για την ομαδοποίηση των σχέσεων χρησιμοποιήθηκαν διανύσματα TF-IDF βαρών) καθώς και έναν αριθμό ο οποίος υποδηλώνει σε πόσες ομάδες επιθυμεί ο χρήστης να γίνει ομαδοποίηση. Στην έξοδό της δίνει αναλυτικά στοιχεία για την ομοιότητα και τα χαρακτηριστικά των ομάδων, όπως επίσης και την ανάλυση της ομαδοποίησης (ποια παρατήρηση εντάσσεται σε ποια ομάδα).

5.2 ΔΟΜΕΣ-ΑΝΤΙΚΕΙΜΕΝΑ

Στην ενότητα αυτή περιγράφονται αναλυτικά οι βασικές δομές δεδομένων (αντικείμενα κλάσεων) οι οποίες χρησιμοποιήθηκαν στο σύστημα.

5.2.1 *Διάνυσμα δεδομένων*

Όπως έχει αναφερθεί σε προηγούμενα κεφάλαια, για κάθε βασικό δομικό στοιχείο άρθρου της εφημερίδας για το οποίο επιχειρείται εξαγωγή σχέσεων, σχηματίζεται ένα διάνυσμα δεδομένων, το οποίο προωθείται από κάθε στάδιο εκτέλεσης του συστήματος στο επόμενο. Το διάνυσμα αυτό, έχει μήκος όσες και οι λεκτικές μονάδες που απαρτίζουν το βασικό στοιχείο. Σε κάθε θέση του διανύσματος αποθηκεύονται οι εξής πληροφορίες:

- η λεκτική μονάδα (token),
- η επισημείωση για το μέρος του λόγου στο οποίο ανήκει η λεκτική μονάδα,
- η επισημείωση για την ονοματική ή ρηματική φράση στην οποία ανήκει η λεκτική μονάδα.

Και τα τρία αυτά στοιχεία αποθηκεύονται ως συμβολοσειρές στο διάνυσμα. Κάθε αντικείμενο της δομής-κλάσης αυτής διάνυσμα δημιουργείται κατά τη φάση της γραμματικής επισημείωσης και αφού περάσει από το στάδιο της κατάτμησης ονοματικών και ρηματικών φράσεων, περιέχει όλη την πληροφορία που χρειάζεται για τη δημιουργία πιθανών σχέσεων.

Η κλάση αυτή περιέχει μεθόδους οι οποίες επιτρέπουν την ανάγνωση στοιχείων του διανύσματος καθώς και την τροποποίησή τους. Αλλαγές στα περιεχόμενα του διανύσματος πραγματοποιούνται μόνο κατά τη διάρκεια της γραμματικής επισημείωσης και της κατάτμησης ονοματικών και ρηματικών φράσεων.

Ο κύκλος ζωής του κάθε αντικειμένου της κλάσης αυτής σταματά όταν πλέον έχουν εξαχθεί όλες οι δυνατές σχέσεις από το υπό εξέταση δομικό στοιχείο του άρθρου. Στη συνέχεια, δημιουργείται νέο αντικείμενο για το επόμενο στοιχείο το οποίο θα διαβαστεί.

5.2.2 *Σχέση*

Κατά τη διάρκεια της αναγνώρισης σχέσεων-τριάδων και για κάθε τριάδα η οποία εξάγεται με βάση τους κανόνες που έχουν αναφερθεί, δημιουργείται ένα αντικείμενο της κλάσης "Σχέση". Το αντικείμενο αυτό περιέχει τα εξής χαρακτηριστικά:

- *Integer tokenNo*: αριθμός λεκτικών μονάδων (tokens) της σχέσης.
- *Integer stopWordsNo*: αριθμός stopwords της σχέσης.
- *Integer posLeftOfE1*: part-of-speech tag της λέξης αριστερά της e_1.
- *Integer posRightOfE1*: part-of-speech tag της λέξης δεξιά της e_1.
- *Integer posLeftOfE2*: part-of-speech tag της λέξης αριστερά της e_2.
- *Integer posRightOfE2*: part-of-speech tag της λέξης δεξιά της e_2.
- *Integer posStartOfR*: part-of-speech tag της πρώτης λέξης της ρηματικής σχέσης r.
- *Integer e1HasPropername*: 0/1 (δυαδική τιμή) αν υπάρχει κύριο όνομα στην e_1.
- *Integer e2HasPropername*: 0/1 (δυαδική τιμή) αν υπάρχει κύριο όνομα στην e_2.
- *Integer e1Start*: η θέση του διανύσματος στην οποία ξεκινά η e1.
- *Integer e1End*: η θέση του διανύσματος στην οποία ξεκινά η e1.
- *Integer rStart*: η θέση του διανύσματος στην οποία ξεκινά η e1.
- *Integer rEnd*: η θέση του διανύσματος στην οποία ξεκινά η e1.
- *Integer e2Start*: η θέση του διανύσματος στην οποία ξεκινά η e1.
- *Integer e2End*: η θέση του διανύσματος στην οποία ξεκινά η e1.

Κάθε αντικείμενο της κλάσης αυτής δημιουργείται μόλις αναγνωριστεί μια νέα σχέση και παύει να είναι χρήσιμο μόλις κατασκευαστεί το διάνυσμα χαρακτηριστικών της σχέσης και γραφεί στο αρχείο εξόδου.

5.3 ΠΕΡΑΙΤΕΡΩ ΑΝΑΛΥΣΗ ΖΗΤΗΜΑΤΩΝ ΥΛΟΠΟΙΗΣΗΣ ΣΥΣΤΗΜΑΤΟΣ

5.3.1 *Θέματα υλοποίησης Parser*

Το στάδιο της προεπεξεργασίας, όπως έχει αναφερθεί, απλώς καλεί μία μέθοδο η οποία με τη χρήση μεθόδων του πακέτου `javax.xml.parsers` διατρέχει το αρχικό XML αρχείο και αποθηκεύει σε ένα νέο αρχείο το απλοποιημένο XML αρχείο το οποίο παράγει. Για την απλοποίηση του XML αγνοούνται πολλά από τα αρχικά tags στην έξοδο, ενώ για τη μορφοποίηση του κειμένου αναγνωρίζονται τα τμήματα που επιδέχονται μορφοποίησης (με τη χρήση κανονικών εκφράσεων) και εξάγονται μορφοποιημένα. Η μέθοδος αυτή καλείται μία φορά για κάθε αρχείο.

Η χρήση του parser στα πλαίσια της παρούσης εργασίας σταμάτησε μετά την επιτυχή κλήση του για κάθε αρχείο-φύλλο της εφημερίδας. Για την επιτυχή τεχνολόγηση (parsing) όλου του αρχείου της εφημερίδας, η μέθοδος κλήθηκε για κάθε φύλλο-αρχείο με τη χρήση ενός βοηθητικού script.

5.3.2 *Θέματα υλοποίησης Εξαγωγής Σχέσεων*

Το δεύτερο μέρος του συστήματος καλείται ανεξάρτητα από τον parser. Η μέθοδος `main` της κεντρικής κλάσης `RelationExtractionSystem` δέχεται ως όρισμα ένα path φακέλου στον οποίο βρίσκονται τα αρχεία για τα οποία ζητείται η εξαγωγή σχέσεων και ένα επιπλέον όρισμα, αν ο χρήστης επιθυμεί τη δημιουργία δεδομένων εκπαίδευσης και την εκπαίδευση του ταξινομητή. Αν ζητούνται δεδομένα εκπαίδευσης, καλείται ο κατασκευαστής της κλάσης `PosNegTagger` , ο οποίος διατρέχει τα αρχεία των εξωτερικών λιστών (λίστες ονομάτων, επωνύμων, ονομάτων οργανισμών) και αποθηκεύει τα δεδομένα τους σε hash sets προκειμένου να είναι γρήγορη η εύρεσή τους, όταν χρειαστούν. Η μέθοδος `main` επίσης, καλεί τον κατασκευαστή της κλάσης `Tagger` .

Στη συνέχεια παρουσιάζονται τα υπόλοιπα βήματα υλοποίησης μαζί με τις κλάσεις και τις μεθόδους που συμμετέχουν σε κάθε βήμα:

- *Κλήση των απαραίτητων μεθόδων για κάθε αρχείο του φακέλου.* Από τη `main` καλείται η μέθοδος `tagMe` της κλάσης `Tagger` για κάθε αρχείο του φακέλου.
- *Κατάτμηση του κειμένου σε βασικά δομικά στοιχεία και επεξεργασία καθενός ξεχωριστά.* Για το σκοπό αυτό καλείται από τη μέθοδο `tagMe` η μέθοδος `parse` της κλάσης `DocumentBuilder` του πακέτου `javax.xml.parsers` . Στη συνέχεια, με τη χρήση των μεθόδων `getDocumentElement` και `getElementsByTagName` , επιλέγονται οι τίτλοι και τα κείμενα των άρθρων, με τη σειρά εμφάνισής τους στο αρχείο εισόδου.
- *Κατάτμηση κειμένου σε λεκτικές μονάδες και γραμματική επισημείωση.* Από κάθε `<title>` και `<text>` tag του αρχείου εισόδου επιλέγεται το κείμενο και προωθείται ως παράμετρος στην κλήση της μεθόδου `smallSetClassify` της κλάσης `SmallSetFunctions` η οποία βρίσκεται στο πακέτο `gr.aueb.cs.nlp.postagger` (υλοποιεί τον γραμματικό επισημειωτή που

χρησιμοποιήθηκε). Η μέθοδος αυτή επιστρέφει ένα διάνυσμα το οποίο σε κάθε θέση του έχει μία λεκτική μονάδα του κειμένου και το αντίστοιχο μέρος του λόγου στο οποίο ανήκει. Στο σημείο αυτό καλείται ο κατασκευαστής της κλάσης `List` ο οποίος δημιουργεί ένα διάνυσμα όπως αυτό που περιγράφηκε παραπάνω (μπορεί να αποθηκεύσει και την πληροφορία από την αναγνώριση ονοματικών και ρηματικών φράσεων).

- *Αναγνώριση ονοματικών και ρηματικών φράσεων.* Η αναγνώριση ονοματικών και ρηματικών φράσεων γίνεται με την κλήση της μεθόδου `findNPVP` της κλάσης `Tagger` . Η μέθοδος αυτή διατρέχει το διάνυσμα και επισημειώνει την έναρξη και λήξη ρηματικών και ονοματικών φράσεων σε αυτό.

- *Εξαγωγή σχέσεων.* Το διάνυσμα προωθείται στη μέθοδο `extractRelations` της κλάσης `Tagger` , η οποία διατρέχει το διάνυσμα από δεξιά προς τα αριστερά, αναζητώντας αρχικά ρηματικές σχέσεις και στη συνέχεια οντότητες, προκειμένου να συμπληρώσει τριάδες της μορφής (e_1, r, e_2). Στην περίπτωση που αναγνωριστεί μία σχέση, καλείται ο κατασκευαστής της κλάσης `Relation`, ο οποίος στη συνέχεια καλεί τις απαραίτητες μεθόδους (`vectorMaker` , `writeVectorToFile` κ.λπ.) προκειμένου να δημιουργηθεί και να εγγραφεί στο αρχείο εξόδου το διάνυσμα της σχέσης που εξήχθη. Η διαδικασία συνεχίζεται μέχρι να διαβαστεί ολόκληρο το διάνυσμα.

- *Βαθμολόγηση δεδομένων εκπαίδευσης.* Σε περίπτωση που κατά την εκτέλεση της `main` έχει δοθεί από το χρήστη το όρισμα που κωδικοποιεί την επιθυμία εξαγωγής δεδομένων εισόδου, η μέθοδος `vectorMaker` καλεί τη μέθοδο `posOrNeg` της κλάσης `PosNegTagger` . Η μέθοδος αυτή υπολογίζει τη βαθμολογία που θα αποδοθεί στη σχέση (+1 ή -1) και στη συνέχεια την επιστρέφει στη μέθοδο `vectorMaker` προκειμένου να εγγραφεί το σωστό διάνυσμα στο αρχείο εξόδου.

- *Εκπαίδευση Ταξινομητή.* Η μέθοδος `main` καλεί την εξωτερική εφαρμογή `svm_learn` με παραμέτρους το όνομα του αρχείου διανυσμάτων χαρακτηριστικών το οποίο δημιουργήθηκε στην προηγούμενη φάση και το επιθυμητό όνομα του αρχείου μοντέλου.

- *Ταξινόμηση Σχέσεων.* Η μέθοδος `main` καλεί την εξωτερική εφαρμογή `svm_classify` με παραμέτρους το όνομα του αρχείου διανυσμάτων χαρακτηριστικών το οποίο δημιουργήθηκε στη φάση εξαγωγής σχέσεων, το όνομα του αρχείου μοντέλου που θα χρησιμοποιεί και το επιθυμητό όνομα του αρχείου εξόδου.

6

ΑΠΟΔΟΣΗ ΣΥΣΤΗΜΑΤΟΣ

Στο κεφάλαιο αυτό δίνονται παραδείγματα λειτουργίας του συστήματος που υλοποιήθηκε και παρουσιάζεται αναλυτικά η απόδοσή του. Για τη μέτρηση της απόδοσης του συστήματος χρησιμοποιήθηκαν τα μέτρα precision, recall και f-measure τα οποία έχουν αναλυθεί στο Κεφάλαιο 2.

6.1 ΜΕΤΡΗΣΗ ΤΗΣ ΑΠΟΔΟΣΗΣ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ

Η έξοδος του τελευταίου βήματος του συστήματος (ταξινομητή) είναι μία ακολουθία αριθμών. Για κάθε σχέση δίνεται ένας αριθμός ο οποίος αντιπροσωπεύει κατά πόσον θεωρείται θετική (σημασιολογικά ορθή) ή αρνητική η προς εξέταση σχέση. Για την ερμηνεία των αποτελεσμάτων του ταξινομητή, χρειάστηκε η θέσπιση κατωφλίων (thresholds) με βάση τα οποία θεωρούμε ότι κάθε σχέση είναι όντως θετική ή αρνητική.

Για παραδειγμα, ο classifier μπορεί να αποδώσει σε μια σχέση την τιμή $0,7$. Επιλέγοντας μια ελαστικότητα (κατώφλι) της τάξης του $\pm 30\%$, η σχέση αυτή θα θεωρηθεί θετική. Αντίθετα, μια σχέση στην οποία έχει αποδοθεί ο αριθμός $0,5$ δε θα θεωρηθεί θετική. Παρόμοια, μια σχέση στην οποία έχει αποδοθεί η τιμή -0.7 θα θεωρηθεί αρνητική.

Για τη μέτρηση της αποδοτικότητας του συστήματος, ακολουθήθηκε η εξής διαδικασία:

- δόθηκε ως είσοδος κάποιο σύνολο ελέγχου,
- οι σχέσεις που εξήχθησαν από το σύνολο αυτό βαθμολογήθηκαν από τον ταξινομητή,
- οι σχέσεις βαθμολογήθηκαν επίσης από το τμήμα του συστήματος το οποίο βαθμολογεί σχέσεις ως θετικές (+1) ή αρνητικές (-1) (το τμήμα αυτό χρησιμοποιείται για την εξαγωγή δεδομένων εκπαίδευσης).
- αντιπαραβλήθηκαν τα παραγόμενα αποτελέσματα. Για κάθε σχέση, ελέγχθηκε κατά πόσον η βαθμολογία του ταξινομητή συμφωνεί με τη βαθμολογία του συστήματος, με τη χρήση κατωφλίων.

Αξίζει να σημειωθεί ότι στην έξοδο του ταξινομητή συνατώνται αριθμοί οι οποίοι είναι μεγαλύτεροι της μονάδας ή μικρότεροι του -1. Οι τιμές αυτές, κατά την αντιπαραβολή των αποτελεσμάτων θεωρήθηκαν ως +1 και -1 αντίστοιχα, χωρίς να εξεταστεί η απόστασή τους από τις τιμές +1,-1.

Στη συνέχεια παρουσιάζονται αναλυτικά τα αποτελέσματα των δοκιμών που διεξήχθησαν για διαφορετικού τύπου διανύσματα και διαφορετικού τύπου ταξινομητές.

Στους πίνακες που περιέχονται στο παρόν κεφάλαιο, η στήλη "thres" αναφέρεται στην τιμή κατωφλίου η οποία χρησιμοποιήθηκε. Η στήλη "classifier" αναφέρεται στο είδος του ταξινομητή που χρησιμοποιήθηκε (γραμμικός, πολυωνυμικός κ.λπ.). Οι τιμές "poly2", "poly3" κ.λπ. της στήλης αυτής υποδηλώνουν πολυωνυμικό ταξινομητή δευτέρου βαθμού, τρίτου βαθμού κ.λπ. Τέλος, η στήλη "c" αναφέρεται στην επιλογή ορισμού τιμής για το tradeoff μεταξύ error και margin κατά τη διάρκεια του classification. Η απόδοση τιμών στην παράμετρο αυτή, όπως γίνεται αντιληπτό, επηρεάζει αρκετά την αποδοτικότητα του εξαγόμενου μοντέλου.

6.2 ΕΛΕΓΧΟΣ ΠΑΡΑΛΛΑΓΩΝ ΤΩΝ ΔΙΑΝΥΣΜΑΤΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Προκειμένου να ελεγχθεί (κατά το δυνατό) η ορθότητα επιλογής των χαρακτηριστικών που χρησιμοποιήθηκαν ως συστατικά για τα διανύσματα χαρακτηριστικών, έγιναν δοκιμές με διαφορετική επιλογή διανυσμάτων.

Σε αυτό το στάδιο, για την εκπαίδευση του συστήματος χρησιμοποιήθηκε ένα σύνολο αποτελούμενο από 20 αρχεία-φύλλα της εφημερίδας (σύνολο εκπαίδευσης). Για τον έλεγχο των παραλλαγών χρησιμοποιήθηκε ένα σύνολο αποτελούμενο από 10 αρχεία-φύλλα της εφημερίδας (σύνολο ελέγχου). Τα αρχεία και για τα δύο σύνολα επιλέχθηκαν τυχαία από το σύνολο του αρχείου της εφημερίδας, έτσι ώστε και τα δύο σύνολα να περιέχουν φύλλα και από τα τρία έτη (2007, 2008, 2009).

Κατά τη διαδικασία αυτή δοκιμάστηκαν τρεις διαφορετικές τιμές ελαστικότητας: 0.2, 0.25, 0.3. Όπως είναι αναμενόμενο, όσο αυξάνεται το μήκος του κατωφλίου, τόσο περισσότερες σχέσεις παρουσιάζουν σύγκλιση αποτελεσμάτων μεταξύ ταξινομητή και συστήματος.

Παρακάτω παρουσιάζεται η απόδοση του συστήματος με διαφορετικού τύπου διανύσματα χαρακτηριστικών. Για κάθε τύπο διανύσματος ελέγχθηκαν διαφορετικού τύπου μοντέλα ταξινόμησης (γραμμικό, πολυωνυμικό, radial).

ΑΡΧΙΚΑ, δοκιμάστηκε η λειτουργία του συστήματος, με διάνυσμα χαρακτηριστικών το οποίο περιέχει τις εξής πληροφορίες (επιλογή I):

- αριθμός λεκτικών μονάδων (tokens) της ρηματικής σχέσης (r)
- αριθμός stopwords της σχέσης
- part-of-speech tag της λέξης αριστερά της e_1
- part-of-speech tag της λέξης δεξιά της e_1
- part-of-speech tag της λέξης αριστερά της e_2
- part-of-speech tag της λέξης δεξιά της e_2
- part-of-speech tag της πρώτης λέξης της ρηματικής σχέσης r
- part-of-speech tag της πρώτης λέξης της οντότητας e_1
- part-of-speech tag της πρώτης λέξης της οντότητας e_2

thr.	classifier	c	accuracy	precision	recall	f-meas
0.2	linear	0.0033000	48.19%	60.52%	63.02%	61.74%
0.2	linear	0.0000200	54.05%	62.48%	76.93%	68.96%
0.2	linear	0.0000025	57.01%	62.89%	85.93%	72.63%
0.2	poly	0.0000000	54.43%	64.88%	68.30%	66.55%
0.2	poly 2	0.0000000	52.24%	63.49%	65.92%	64.68%
0.2	poly 3	0.0000000	54.44%	64.88%	68.31%	66.55%
0.25	linear	0.0033000	50.01%	61.73%	64.88%	63.27%
0.25	linear	0.0000200	57.36%	64.20%	80.79%	71.55%
0.25	linear	0.0000025	60.77%	64.36%	91.59%	75.60%
0.25	poly	0.0000000	56.44%	66.14%	70.39%	68.20%
0.25	poly 2	0.0000000	53.95%	64.60%	67.68%	66.10%
0.25	poly 3	0.0000000	56.42%	66.13%	70.38%	68.19%
0.3	linear	0.0033000	51.72%	62.84%	66.65%	64.69%
0.3	linear	0.0000200	59.72%	65.58%	82.71%	73.16%
0.3	linear	0.0000025	65.50%	66.06%	98.73%	79.16%
0.3	poly	0.0000000	58.52%	67.54%	72.16%	69.77%
0.3	poly 2	0.5000000	55.89%	65.86%	69.61%	67.68%
0.3	poly 3	0.0000000	58.52%	67.54%	72.16%	69.77%

Πίνακας 4: Αποτελέσματα του συστήματος για την επιλογή I του διανύσματος χαρακτηριστικών.

- 0/1 (δυαδική τιμή) αν υπάρχει κύριο όνομα στην e_1
- 0/1 (δυαδική τιμή) αν υπάρχει κύριο όνομα στην e_2
- το κείμενο της σχέσης (για λόγους αναφοράς)

Το καλύτερο (από άποψη f-measure) αποτέλεσμα που έδωσε αυτή η επιλογή διανυσμάτων δόθηκε με γραμμικό SVM μοντέλο και πέτυχε 79.16% f-measure με 98.73% precision και 66.06% recall. Τα αποτελέσματα για τον τύπο αυτό του διανύσματος χαρακτηριστικών φαίνονται αναλυτικά στον Πίνακα 4

ΣΤΗ ΣΥΝΕΧΕΙΑ, δοκιμάστηκε η λειτουργία του συστήματος, με διάνυσμα χαρακτηριστικών αποτελούμενο από (επιλογή II):

- αριθμός λεκτικών μονάδων (tokens) της ρηματικής σχέσης (r)
- αριθμός stopwords της ρηματικής σχέσης (r)
- part-of-speech tag της λέξης αριστερά της e_1
- part-of-speech tag της λέξης δεξιά της e_1

thr.	classifier	c	accuracy	precision	recall	f-meas
0.20	linear	0.004600	44.71	57.99	57.74	57.86%
0.20	poly	0.000000	61.55	65.90	86.08	74.65%
0.20	poly2	0.500000	60.01	65.11	84.43	73.52%
0.20	poly3	0.000000	61.55	65.90	86.08	74.65%
0.20	poly4	0.000000	61.92	66.18	86.08	74.83%
0.20	radial	0.500000	56.65	64.98	73.92	69.16%
0.25	linear	0.004600	45.16	58.39	57.78	58.08%
0.25	poly	0.000000	62.45	66.35	87.04	75.30%
0.25	poly3	0.000000	62.08	66.07	87.02	75.11%
0.25	poly4	0.000000	62.45	66.35	87.04	75.30%
0.25	poly2	0.000000	62.70	66.57	86.93	75.40%
0.25	radial	0.500000	58.82	66.01	77.15	71.15%
0.30	linear	0.010000	45.41	58.55	58.15	58.35%
0.30	linear	4.320000	48.16	60.55	60.76	60.65%
0.30	linear	0.004600	64.41	68.75	84.13	75.67%
0.30	linear	0.000020	64.77	65.42	98.49	78.62%
0.30	linear	0.0000025	65.76	65.76	100.00	79.34%
0.30	poly	0.000000	63.12	66.70	87.68	75.77%
0.30	poly3	0.000000	62.59	66.35	87.48	75.46%
0.30	poly2	0.0000025	63.61	66.51	89.99	76.49%
0.30	poly4	0.000000	63.12	66.70	87.68	75.76%
0.30	poly2	0.000000	63.11	66.79	87.31	75.68%
0.30	radial	0.500000	60.28	66.66	79.21	72.40%

Πίνακας 5: Αποτελέσματα του συστήματος για την επιλογή II του διανύσματος χαρακτηριστικών.

- part-of-speech tag της λέξης αριστερά της e_2
- part-of-speech tag της λέξης δεξιά της e_2
- part-of-speech tag της πρώτης λέξης της ρηματικής σχέσης r
- 0/1 (δυαδική τιμή) αν υπάρχει κύριο όνομα στην e_1
- 0/1 (δυαδική τιμή) αν υπάρχει κύριο όνομα στην e_2
- το κείμενο της σχέσης (για λόγους αναφοράς)

Από το διάνυσμα αυτό λείπουν τα χαρακτηριστικά που αναφέρουν το μέρος του λόγου της πρώτης λέξης των οντοτήτων e_1 και e_2. Το καλύτερο (από άποψη f-measure) αποτέλεσμα που έδωσε αυτή η επιλογή διανυσμάτων δόθηκε με γραμμικό SVM μοντέλο και πέτυχε 78.62% f-measure με 98.49% precision και 65.42% recall. Τα αποτελέσματα για τον τύπο αυτό του διανύσματος χαρακτηριστικών φαίνονται αναλυτικά στον Πίνακα 5.

ΤΕΛΟΣ, ελέγχθηκε η λειτουργία με διάνυσμα χαρακτηριστικών αποτελούμενο από (επιλογή III):

- αριθμός λεκτικών μονάδων (tokens) ολόκληρης της σχέσης
- αριθμός stopwords της σχέσης
- part-of-speech tag της λέξης αριστερά της e_1
- part-of-speech tag της λέξης δεξιά της e_1
- part-of-speech tag της λέξης αριστερά της e_2
- part-of-speech tag της λέξης δεξιά της e_2
- part-of-speech tag της πρώτης λέξης της ρηματικής σχέσης r
- 0/1 (δυαδική τιμή) αν υπάρχει κύριο όνομα στην e_1
- 0/1 (δυαδική τιμή) αν υπάρχει κύριο όνομα στην e_2
- το κείμενο της σχέσης (για λόγους αναφοράς)

Από το διάνυσμα αυτό λείπουν τα χαρακτηριστικά που αναφέρουν το μέρος του λόγου της πρώτης λέξης των οντοτήτων e_1 και e_2. Το καλύτερο αποτέλεσμα που έδωσε η επιλογή αυτή δόθηκε για μοντέλο με radialkernel και πέτυχε 79.57% f-measure με 83.38% precision και 76.10% recall. Τα αποτελέσματα για διαφορετικού τύπου ταξινομητές παρουσιάζονται αναλυτικότερα στον Πίνακα 6.

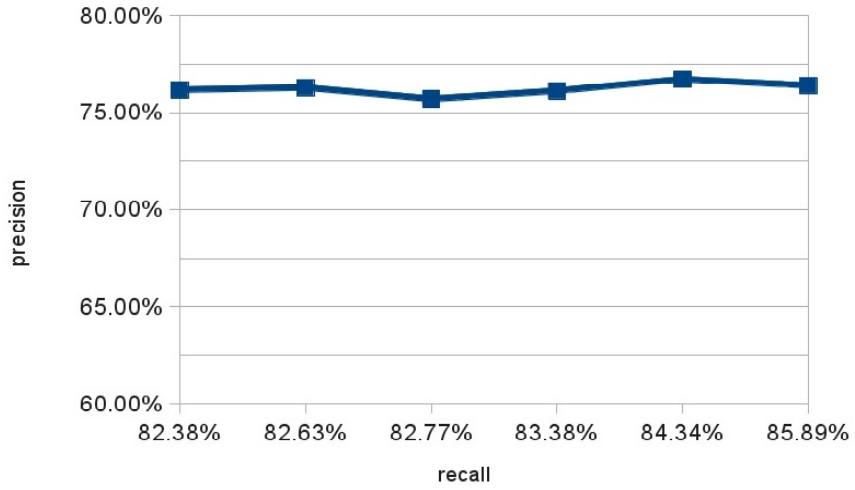
Οι τρεις αυτές επιλογές δεν έδωσαν πολύ διαφορετικά αποτελέσματα. Για το λόγο αυτό, για την διεξαγωγή περαιτέρω ελέγχων επιλέχθηκε η επιλογή III, η οποία περιέχει χαρακτηριστικά τα οποία συμφωνούν περισσότερο με τους κανόνες που θεσπίστηκαν στο σύστημα, στη φάση της εξαγωγής δεδομένων εκπαίδευσης. Όπως γίνεται αντιληπτό από το σχετικό πίνακα, για την επιλογή III, την καλύτερη απόδοση είχε το μοντέλο με το radial kernel, το οποίο και υιοθετήθηκε ως επιλογή για περαιτέρω δοκιμές.

thr.	classifier	c	accuracy	precision	recall	f-meas
0.2	linear	0.00310	47.10%	59.58%	63.03%	61.26%
0.2	poly 2	0.00000	49.15%	60.20%	60.94%	60.57%
0.2	poly 3	0.00000	55.88%	65.16%	72.01%	68.41%
0.2	poly 4	0.00000	58.06%	66.14%	75.39%	70.46%
0.2	radial	0.50000	66.94%	74.17%	77.00%	75.56%
0.25	linear	0.00310	48.81%	60.63%	65.15%	62.81%
0.25	poly 2	0.00000	51.64%	61.86%	64.01%	62.92%
0.25	poly 3	0.00000	58.04%	66.41%	74.37%	70.16%
0.25	poly 4	0.00000	60.48%	67.40%	78.31%	72.45%
0.25	radial	0.50000	69.28%	75.19%	80.13%	77.58%
0.3	linear	0.00310	50.94%	61.90%	67.79%	64.71%
0.3	poly 2	0.00000	54.39%	63.69%	67.08%	65.34%
0.3	poly 3	0.00000	60.29%	67.61%	77.05%	72.02%
0.3	poly 4	0.00000	61.99%	68.17%	80.15%	73.68%
0.3	radial	0.50000	71.60%	76.10%	83.38%	79.57%

Πίνακας 6: Αποτελέσματα του συστήματος για την επιλογή III του διανύσματος χαρακτηριστικών.

6.3 ΑΠΟΔΟΣΗ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ

Σε αυτή την ενότητα παρουσιάζονται τα αποτελέσματα του συστήματος από 6 διαφορετικά σύνολα κειμένου καθένα από τα οποία αποτελείται από 10 τυχαία επιλεγμένα αρχεία-φύλλα της εφημερίδας. Για την ταξινόμηση των σχέσεων χρησιμοποιήθηκε SVM μοντέλο με radial kernel. Τα αποτελέσματα φαίνονται τόσο στον Πίνακα 7 όσο και στο Σχήμα 23.



Σχήμα 23: Απόδοση συστήματος.

set	relations	accuracy	precision	recall	f-meas
1	39775	71.05%	76.15%	82.38%	79.14%
2	40958	71.48%	76.27%	82.63%	79.32%
3	40606	70.91%	75.69%	82.77%	79.07%
4	39654	71.60%	76.10%	83.38%	79.57%
5	37765	72.54%	76.71%	84.34%	80.34%
6	38600	73.34%	76.38%	85.89%	80.86%

Πίνακας 7: Απόδοση του συστήματος.

6.4 ΣΧΟΛΙΑΣΜΟΣ ΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Όπως παρατηρούμε, το σύστημα παρουσιάζει αρκετά καλή απόδοση, συγκρινόμενο με την έξοδο που θα έδιναν οι κανόνες με βάση τους οποίους ταξινομούνται οι σχέσεις ως αρνητικές ή θετικές εντός του συστήματος. Φυσικά, η απόδοση αυτή αντικατοπτρίζει την δημιουργία ενός αρκετά καλού ταξινομητή για νέες σχέσεις.

Στα αποτελέσματα τα οποία παρουσιάστηκαν για τα διαφορετικά σύνολα ελέγχου (τα οποία είναι τυχαία επιλεγμένα και δεν εμφανίζουν καμία σημασιολογική ομοιότητα), παρατηρούνται παραπλήσιες τιμές στα μέτρα απόδοσης. Θα μπορούσε να εξαχθεί λοιπόν το συμπέρασμα ότι η απόδοση του ταξινομητή είναι σταθερή και δεν επηρεάζεται από το σώμα εισόδου.

Η απόδοση του συστήματος το οποίο υλοποιήθηκε, στηρίζεται κυρίως σε δύο στοιχεία: την αποδοτικότητα των εργαλείων, και ειδικότερα του γραμματικού επισημειωτή ο οποίος έπαιξε καθοριστικό ρόλο στην αναγνώριση προτύπων για την εξαγωγή σχέσεων και στην αποδοτικότητα των κανόνων που θεσπίστηκαν για την επισημείωση σχέσεων ως θετικές ή αρνητικές κατά τη διάρκεια της δημιουργίας δεδομένων εκπαίδευσης για τον ταξινομητή.

Στο σημείο αυτό κρίνεται σκόπιμο να υπενθυμίσουμε ότι δεν υπάρχει κάποιο σύνολο δεδομένων στην ελληνική γλώσσα που να περιέχει σχέσεις οι οποίες γνωρίζουμε πως είναι θετικές ή αρνητικές. Οι κανόνες που θεσπίστηκαν κατά την υλοποίηση του συστήματος, παρότι εμπειρικοί εξασφαλίζουν ένα αποδεκτό και σε μεγάλο ποσοστό σημασιολογικά ευσταθές αποτέλεσμα. Η επιβεβαίωση από ανθρώπους της ορθότητας ή όχι των εξαγόμενων σχέσεων μπορεί να οδηγήσει σε ποιο ακριβές σύνολο δεδομένων εκπαίδευσης, το οποίο, συνεπαγωγικά, θα οδηγήσει σε βελτίωση του αποτελέσματος της εξαγωγής σχέσεων.

6.5 ΠΑΡΑΔΕΙΓΜΑ ΕΞΟΔΟΥ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ

Παρακάτω παρουσιάζεται ένα παράδειγμα της εξόδου του συστήματος για ένα τυχαία επιλεγμένο άρθρο από κάποιο φύλλο της εφημερίδας. Για ένα πληρέστερο παράδειγμα, βλ. το Παράρτημα i. Στο παράδειγμα φαίνονται μόνο οι σημειωμένες ως θετικές σχέσεις. Η πρώτη στήλη αναφέρεται στη βαθμολογία που έδωσε το σύστημα στη σχέση, ενώ η δεύτερη στη βαθμολογία που έδωσε ο ταξινομητής. Όπως παρατηρείται, αν θεωρήσουμε μία τιμή κατωφλίου 0.3, οι βαθμολογίες συμφωνούν.

βαθμ. συστ.	βαθμ. ταξιν.	κείμενο
1	0.99958705	(Η ΚΙΝΗΣΗ Παπανδρέου, χαρακτηρίστηκε, από στελέχη όλων των)
1	1.0000961	(Οι πολιτικοί αναλυτές, επισήμαιναν, και τη δήλωση)
1	1.0003492	(και τη δήλωση, έκανε, ο Γιώργος Παπανδρέου εξερχόμενος του Μεγάρου Μαξίμου)
1	1.0002295	(ο Γιώργος Παπανδρέου εξερχόμενος του Μεγάρου Μαξίμου, υπάρχουν, πιέσεις από)
1	0.99981392	(ότι « η Ελλάδα, αδικείται, κινήσεις)
1	0.99981392	(ότι « η Ελλάδα, γίνονται, κινήσεις)
1	1.000597	(με τον Αμερικανό πρέσβη Ντάνιελ Σπέκχαρντ, είχε προχωρήσει, σε διάβημα διαμαρτυρίας)
1	1.0008705	(σε διάβημα διαμαρτυρίας, θέτοντας, εμμέσως θέμα για τη βάση της Σούδας)
1	1.1032129	(ο Γιώργος, αναφέρθηκε, και στη Βουλή)
1	0.80347385	(τον τρόπο με τον, χειρίζεται, το ευαίσθητο για μας θέμα της ονομασίας των Σκοπίων)
1	0.99594887	(ότι η Ελλάδα, έχει ανταποκριθεί, σε κάθε περίπτωση)
1	0.88731263	(ως παράδειγμα τη λειτουργία της βάσης της Σούδας, ξέρουμε, τις επιφυλάξεις)
1	0.88731263	(ως παράδειγμα τη λειτουργία της βάσης της Σούδας, έχει, τις επιφυλάξεις)
1	1.0003989	(Η Ελλάδα, περιμένει, τον σεβασμό των θέσεών)
1	0.75292786	(τις θέσεις, εξέφρασε, ο πρόεδρος του Γιώργος Παπανδρέου)
1	0.7861247	(τη « μεταφορά της διαπραγμάτευσης από τον ΟΗΕ στο ΝΑΤΟ », απηύθυνε, από τη Βουλή ο επικεφαλής του ΣΥΡΙΖΑ Αλέκος Αλαβάνος)
1	0.99992577	(ο Πρωθυπουργός, έκλεισε, και την τελευταία χαραμάδα για την ανεύρεση διεξόδου στην κρίση)
1	0.79627852	(ότι ο αρχηγός της αξιωματικής αντιπολίτευσης, είχε εξασφαλίσει, συναντήσεις με το σύνολο της πολιτειακής και πολιτικής ηγεσίας των Σκοπίων)
1	0.94798079	(Παπανδρέου στα Σκόπια και το πρώτο, είχε προσδιοριστεί, για τις 11 το πρωί τοπική ώρα (12 ώρα Ελλάδος))

Πίνακας 8: Παράδειγμα εξόδου του συστήματος.

Μέρος ΙΙΙ

ΠΑΡΑΡΤΗΜΑΤΑ

ΠΑΡΑΔΕΙΓΜΑ ΕΞΑΓΩΓΗΣ ΣΧΕΣΕΩΝ

Παρακάτω φαίνονται όλες οι σχέσεις που εξήχθησαν από ένα άρθρο, μαζί με τη βαθμολογία που τους απέδωσε ο ταξινομητής του συστήματος. Συνολικά εξήχθησαν 65 σχέσεις, οι οι οποίες φαίνονται στον Πίνακα 9, τον Πίνακα 10 και τον Πίνακα 11.

Το κείμενο του άρθρου (από το φύλλο της 14/05/2007) είναι το εξής:

```
<article>
  <title>Εκλογικά παιχνίδια με την αυτοδυναμία . </title>
  <text>
    Υπουργοί πιέζουν για νέο νόμο , ενώ φουντώνουν τα σενάρια για διπλές κάλπες .
    Να αλλάξει αμέσως ο εκλογικός νόμος ζητούν κορυφαία στελέχη της ΝΔ.. ,
    εκφράζοντας ουσιαστικά την ανησυχία τους για τη φθορά που έχει υποστεί το
    κυβερνών κόμμα . Παρά τις πρόσφατες δεσμεύσεις του Πρωθυπουργού ότι
    θέμα νέου εκλογικού νόμου δεν υφίσταται , υπουργοί και επιτελάρχες της Ρηγίλλης
    δεν έχουν πειστεί και πιέζουν για νέο νόμο , επικαλούμενοι τον
    κίνδυνο της ακυβερνησίας μετά τις εκλογές . .
    ΡΕΠΟΡΤΑΖ : Γιάννης Λ . Πολίτης Διονύσης Νασόπουλος .
    ΟΙ ΦΩΝΕΣ στο κυβερνητικό στρατόπεδο για νέο εκλογικό νόμο πληθαίνουν όσο
    φουντώνει ο προεκλογικός πυρετός στη ΝΔ. .
    Μετά τον Ευάγγ . Μεϊμαράκη , που έχει εισηγηθεί στον Κων . Καραμανλή να
    αλλάξει μονομερώς η ΝΔ. . στην παρούσα Βουλή τον εκλογικό νόμο ,
    υπέρ της επίμαχης αλλαγής τάχθηκαν χθες οι κκ. . Δημ . Σιούφας και Δημ .
    Αβραμόπουλος , καθώς και ο γραμματέας της ΝΔ. . κ . Ελ .
    Ζαγορίτης . Υπέρ ενός νέου εκλογικού νόμου τάχθηκε χθες και ο επίτιμος πρόεδρος
    Ν. Δ . κ . Κων . Μητσοτάκης , διαβλέποντας πως
    υπάρχει στον ορίζοντα το φάσμα της ακυβερνησίας , ενώ έναν νέο νόμο που θα
    εξασφαλίζει ισχυρή αυτοδυναμία στο πρώτο
    κόμμα έχει εισηγηθεί και η πρόεδρος της Βουλής κ . Άννα Ψαρούδα– Μπενάκη .
    Θετικοί με το ενδεχόμενο αλλαγών στον εκλογικό νόμο
    εμφανίζονται και οι κκ. . Γ . Σουφλιάς και Ντόρα Μπακογιάννη , οι οποίοι
    θεωρούν πάντως ότι οι αλλαγές αυτές θα έπρεπε να
    ενταχθούν σε μια ριζική αναμόρφωση του εκλογικού συστήματος , στα πρότυπα του
    γερμανικού συστήματος . .
    Τ α ίδια στελέχη επιχειρούν να ξορκίσουν μια πεντακομματική Βουλή , η
    οποία μπορεί να σταθεί εμπόδιο στην
    αυτοδυναμία του πρώτου κόμματος . Και δεν κρύβουν ότι απώτερος στόχος με τον
    νέο νόμο είναι να στηθούν σε σύντομο
    χρονικό διάστημα εκ νέου κάλπες , εάν η ΝΔ. . είναι πρώτο κόμμα , αλλά δεν έχει
    αυτοδυναμία ή περιοριστεί σε ισχνή κοινοβουλευτική
    πλειοψηφία . Παράλληλα , εκτιμούν ότι με όχημα τον εκλογικό νόμο η ΝΔ. .
    μπορεί να εγκλωβίσει το ΠΑΣΟΚ σε μια σκληρή αντιπαράθεση ,
    ώστε να ξεφύγει και η κυβέρνηση από τη δίνη του σκανδάλου των ομόλογων . Οι
    φωνές στο κυβερνητικό στρατόπεδο για νέο εκλογικό νόμο
    πληθαίνουν όσο φουντώνει ο προεκλογικός πυρετός στη ΝΔ. . Μετά τον Ευάγγ .
    Μεϊμαράκη , που έχει εισηγηθεί στον Κων . Καραμανλή να αλλάξει
    μονομερώς η ΝΔ. . στην παρούσα Βουλή τον εκλογικό νόμο , υπέρ της επίμαχης
    αλλαγής τάχθηκαν χθες οι κκ. . Δημ . Σιούφας και Δημ . Αβραμόπουλος ,
    καθώς και ο γραμματέας της ΝΔ. . κ . Ελ . Ζαγορίτης . Και οι τρεις προκάλεσαν
    το ΠΑΣΟΚ να συμφωνήσει τώρα σε έναν νέο νόμο ,
```

```
25      ώστε οι αλλαγές να ισχύσουν αμέσως και όχι στις μεθεπόμενες εκλογές . Με τη
        χθεσινή παρέμβασή τους δεν έκρυψαν ουσιαστικά την ανησυχία
26      τους για το εκλογικό αποτέλεσμα , αφού το βασικό τους επιχείρημα είναι ότι πρέπει
        να διασφαλιστεί πως από τις εκλογές θα προκύψει
27      μια ισχυρή , αυτοδύναμη κυβέρνηση και να αποφευχθεί το ενδεχόμενο να
        οδηγηθεί η χώρα σε ακυβερνησία . Έσπευσαν μάλιστα να χαρακτηρίσουν
28      την άρνηση του ΠΑΣΟΚ να εξυπηρετήσει τους εκλογικούς σχεδιασμούς της ΝΔ. .
        ένδειξη ηττοπάθειας από την πλευρά της
29      αξιωματικής αντιπολίτευσης . Παράγοντες του Μαξίμου υποστήριζαν χθες ότι το
        μπαράζ δηλώσεων των κορυφαίων υπουργών και του
30      γραμματέα της ΝΔ. . για νέο εκλογικό νόμο είναι συμπτωματικό και δεν
        υποδηλώνει αλλαγή πλεύσης από την πλευρά του Κων . Καραμανλή .
31      Ωστόσο , ενώ μέχρι πρότινος το πρωθυπουργικό επιτελείο έκλεινε με κατηγορηματικό
        τρόπο το θέμα , οι ίδιες πηγές περιορίζονταν χθες
32      να δηλώσουν ότι δεν είναι πιθανό να αναλάβει η κυβέρνηση στην παρούσα
        Βουλή τέτοια νομοθετική πρωτοβουλία . .
33        Κίνηση πανικού . Οι δημόσιες δεσμεύσεις του Πρωθυπουργού , εντός κι
        εκτός Βουλής , ότι δεν θα αλλάξει τον εκλογικό νόμο γιατί
34      σέβεται τους θεσμούς είναι , σύμφωνα με κυβερνητικές πηγές , και ο πιο σοβαρός
        λόγος που ο Κων . Καραμανλής δεν θα αναθεωρήσει
35      τη στάση του , παρά τις ισχυρές πιέσεις που δέχεται . Είναι δεδομένο ότι η
        αξιοπιστία του Πρωθυπουργού θα πληγεί ανεπανόρθωτα
36      εάν κάνει στροφή 180 μοιρών . Και αυτό θα το εκμεταλλευθεί και η αντιπολίτευση
        που θα επιχειρήσει να προβάλει μια κυβερνητική πρωτοβουλία
37      για νέο εκλογικό ως κίνηση πανικού παραδέχεται κυβερνητικό στέλεχος , που
        θεωρεί πάντως ότι η ΝΔ. . θα έχει περισσότερα
38      να κερδίσει από την αλλαγή του εκλογικού νόμου . Σημειώνεται ότι ο
        Πρωθυπουργός είχε με κατηγορηματικό τρόπο κλείσει το θέμα ,
39      κατά τη συζήτηση της πρότασης μομφής που είχε καταθέσει το ΠΑΣΟΚ κατά της
        κυβέρνησης , δηλώνοντας στις 4 Φεβρουαρίου από
40      το βήμα της Βουλής : Εγώ αρνούμαι να ανοίξω τέτοιο θέμα . Γιατί σέβομαι τους
        θεσμούς και τους βάζω πάνω από το κομματικό και
41      προσωπικό μου συμφέρον . . Ωστόσο , ο εφιάλτης της πεντακομματικής
        Βουλής και η σχεδόν βέβαιη , σύμφωνα με τις
42      περισσότερες δημοσκοπήσεις , κοινοβουλευτική εκπροσώπηση του ΛΑΟΣ , έχει
        σημάνει συναγερμό στην κυβέρνηση και τη Ρηγίλλης και ,
43      για πολλούς στη ΝΔ. . αυτή η εξέλιξη ίσως αλλάξει και τα μέχρι σήμερα δεδομένα .
        Όσοι εισηγούνται στο Μαξίμου να αλλάξει τώρα
44      ο εκλογικός νόμος , υποστηρίζουν ότι το θέμα προσφέρει μια μεγάλη ευκαιρία για να
        αλλάξει η πολιτική ατζέντα και να κλιμακωθεί
45      η αντιπαράθεση με το ΠΑΣΟΚ , με την κυβέρνηση να έχει εκ των πραγμάτων το
        πάνω χέρι . Επιτελικά στελέχη της ΝΔ. . λένε ,
46      μάλιστα , ότι η πρόσκληση στο ΠΑΣΟΚ να συμφωνήσει τώρα στην αλλαγή του νόμου
        , από την μια πλευρά κάνει συμμέτοχο την
47      αξιωματική αντιπολίτευση στη σχετική συζήτηση και από την άλλη περιορίζει την
        απήχηση του αιτήματος για άμεση προσφυγή στις
48      κάλπες , αφού στο τέλος θα εμφανιστεί ότι δεν πιστεύει πως θα επικρατήσει στις
        εκλογές . Αυτό επιχείρησε να εκπέμψει χθες
49      και ο γραμματέας της ΝΔ. . κ . Ελ . Ζαγορίτης , ο οποίος μιλώντας στη ΝΕΤ–
        κάλεσε το ΠΑΣΟΚ εάν έχει ψυχή και πιστεύει
50      ότι θα νικήσει στις εκλογές , να προχωρήσει τώρα με τη ΝΔ. . στην αλλαγή του
        νόμου . .
51    </text>
52  </article>
```

Τμ. Κωδ. 6: Άρθρο της εφημερίδας προς επεξεργασία.

βαθμ. ταξιν.	κείμενο
1.0006072	(νέο νόμο, φουντώνουν, τα σενάρια για διπλές κάλπες)
0.99934246	(ο εκλογικός νόμος, ζητούν, κορυφαία στελέχη της Ν.Δ.)
1.0696401	(κορυφαία στελέχη της Ν.Δ., εκφράζοντας, την ανησυχία)
1.0006358	(την ανησυχία, έχει υποστεί, το κυβερνών κόμμα)
0.58812966	(Παρά τις πρόσφατες δεσμεύσεις του Πρωθυπουργού ότι θέμα νέου εκλογικού, δεν υφίσταται, υπουργοί και επιτελάρχες της Ρηγίλλης)
1.00002	(υπουργοί και επιτελάρχες της Ρηγίλλης, δεν έχουν πειστεί, για νέο νόμο)
0.99969455	(υπουργοί και επιτελάρχες της Ρηγίλλης, πιέζουν, για νέο νόμο)
0.99980698	(νέο εκλογικό νόμο, πληθαίνουν όσο φουντώνει, στη Ν.Δ)
1.0514809	(Μεϊμαράκη, έχει εισηγηθεί, στον Κων)
0.82057196	(Καραμανλή, να αλλάξει, η Ν.Δ)
0.9207164	(Υπέρ ενός νέου εκλογικού, τάχθηκε, και ο επίτιμος πρόεδρος της Ν.Δ)
0.99972499	(Μητσοτάκης, διαβλέποντας, στον ορίζοντα το φάσμα της ακυβερνησίας »)
0.99972499	(Μητσοτάκης, υπάρχει, στον ορίζοντα το φάσμα της ακυβερνησίας »)
1.0639808	(ενώ έναν νέο νόμο, θα εξασφαλίζει, ισχυρή αυτοδυναμία στο πρώτο κόμμα)
0.2496636	(ισχυρή αυτοδυναμία στο πρώτο κόμμα, έχει εισηγηθεί, και η πρόεδρος της Βουλής)
1.0009288	(Σουφλιάς και Ντόρα Μπακογιάννη, θεωρούν, ότι οι αλλαγές)
0.88989334	(ότι οι αλλαγές, θα έπρεπε να ενταχθούν, σε μια ριζική αναμόρφωση του εκλογικού συστήματος)
0.82846282	(α ίδια στελέχη, επιχειρούν, » μια πεντακομματική Βουλή)
0.82846282	(α ίδια στελέχη, ξορκίσουν, » μια πεντακομματική Βουλή)
1.0474535	(με τον νέο νόμο, είναι να στηθούν, σε σύντομο χρονικό διάστημα εκ νέου κάλπες)
0.99933803	(πρώτο κόμμα, δεν έχει, σε ισχνή κοινοβουλευτική πλειοψηφία)
0.60323157	(πρώτο κόμμα, περιοριστεί, σε ισχνή κοινοβουλευτική πλειοψηφία)
0.75551403	(το ΠΑΣΟΚ σε μια σκληρή αντιπαράθεση, ώστε να ξεφύγει, και η κυβέρνηση από τη δίνη του σκανδάλου των ομόλογων)
0.99980698	(νέο εκλογικό νόμο, πληθαίνουν όσο φουντώνει, στη Ν.Δ)
1.0514809	(Μεϊμαράκη, έχει εισηγηθεί, στον Κων)
0.82057196	(Καραμανλή, να αλλάξει, η Ν.Δ)
0.99927584	(Και οι τρεις, προκάλεσαν, το ΠΑΣΟΚ)
-0.9999286	(το ΠΑΣΟΚ, να συμφωνήσει, σε έναν νέο νόμο)

Πίνακας 9: Σχέσεις που εξήχθησαν από το Τμ. Κωδ. 6.

βαθμ. ταξιν.	κείμενο
-0.21641369	(ώστε οι αλλαγές, να ισχύσουν, στις μεθεπόμενες εκλογές)
0.70015932	(Με τη χθεσινή παρέμβασή, δεν έκρυψαν, την ανησυχία)
0.63267588	(αφού το βασικό τους επιχείρημα, είναι ότι πρέπει να διασφαλιστεί, πως από τις εκλογές)
0.95944419	(πως από τις εκλογές, θα προκύψει, μια ισχυρή)
-1.2591046	(αυτοδύναμη κυβέρνηση και, να αποφευχθεί, η χώρα σε ακυβερνησία)
-1.2591046	(αυτοδύναμη κυβέρνηση και, να οδηγηθεί, η χώρα σε ακυβερνησία)
-1.0778732	(την άρνηση του ΠΑΣΟΚ, να εξυπηρετήσει, τους εκλογικούς)
1.0006626	(Παράγοντες του Μαξίμου, υποστήριζαν, δηλώσεων των κορυφαίων υπουργών και του γραμματέα της Ν.Δ)
0.97366563	(νέο εκλογικό νόμο, είναι, « συμπτωματικό » και)
0.60366798	(« συμπτωματικό » και, δεν υποδηλώνει, αλλαγή πλεύσης από την πλευρά)
1.000443	(ενώ μέχρι πρότινος το πρωθυπουργικό επιτελείο, έκλεινε, με κατηγορηματικό τρόπο το θέμα)
1.05967	(οι ίδιες πηγές, περιορίζονταν χθες να δηλώσουν, η κυβέρνηση στην παρούσα Βουλή τέτοια νομοθετική πρωτοβουλία)
0.77748307	(οι ίδιες πηγές, δεν είναι, η κυβέρνηση στην παρούσα Βουλή τέτοια νομοθετική πρωτοβουλία)
0.13066952	(οι ίδιες πηγές, να αναλάβει, η κυβέρνηση στην παρούσα Βουλή τέτοια νομοθετική πρωτοβουλία)
0.5394352	(Βουλής, δεν θα αλλάξει, τον εκλογικό νόμο γιατί)
0.42652633	(τον εκλογικό νόμο γιατί, σέβεται, τους θεσμούς)
-1.000209	(τους θεσμούς, είναι, με κυβερνητικές πηγές)
0.99997043	(Καραμανλής, δεν θα αναθεωρήσει, τη στάση)
0.99962921	(ότι η αξιοπιστία του Πρωθυπουργού, θα πληγεί, ανεπανόρθωτα)
-1.0004881	(ανεπανόρθωτα, κάνει, στροφή 180 μοιρών)
0.63314532	(και η αντιπολίτευση, θα επιχειρήσει να προβάλει, μια κυβερνητική πρωτοβουλία για νέο εκλογικό ως κίνηση πανικού »)
0.87273647	(κυβερνητικό στέλεχος, θεωρεί, ότι η Ν.Δ)
0.45926076	(ότι ο Πρωθυπουργός, είχε, με κατηγορηματικό τρόπο)
-0.99995459	(με κατηγορηματικό τρόπο, κλείσει, το θέμα)
1.0004453	(κατά τη συζήτηση της πρότασης μομφής, είχε καταθέσει, το ΠΑΣΟΚ κατά της κυβέρνησης)
-1.1248047	(« Εγώ, αρνούμαι, να ανοίξω τέτοιο θέμα)
0.80456367	(ΛΑΟΣ, έχει σημάνει, στην κυβέρνηση και τη Ρηγίλλης και)
-0.75184481	(στο Μαξίμου, να αλλάξει, ο εκλογικός νόμος)

Πίνακας 10: Σχέσεις που εξήχθησαν από το Τμ. Κωδ. 6. (συνέχεια)

βαθμ. ταξιν.	κείμενο
0.46974512	(ο εκλογικός νόμος, υποστηρίζουν, ότι το θέμα)
0.50846056	(ότι το θέμα, προσφέρει, μια μεγάλη ευκαιρία)
-0.73989635	(μια μεγάλη ευκαιρία, για να αλλάξει, η πολιτική)
-1.0004561	(η πολιτική, να κλιμακωθεί, η αντιπαράθεση με το ΠΑΣΟΚ)
0.27427618	(ότι η πρόσκληση στο ΠΑΣΟΚ, να συμφωνήσει, στην αλλαγή του)
0.63916692	(από την μια πλευρά «, κάνει, συμμέτοχο την αξιωματική αντιπολίτευση » στη σχετική συζήτηση και από την άλλη)
0.74393423	(αφού στο τέλος, θα εμφανιστεί ότι δεν πιστεύει πως θα επικρατήσει, στις εκλογές)
0.43985377	(το ΠΑΣΟΚ «, έχει, ψυχή » και)
0.038648505	(στις εκλογές, να προχωρήσει, με τη Ν.Δ)

Πίνακας 11: Σχέσεις που εξήχθησαν από το Τμ. Κωδ. 6. (συνέχεια)

Παρακάτω φαίνεται ένα τμήμα της εξόδου της ομαδοποίησης 600 σχέσεων με βάση διανύσματα βαρών $TF-IDF$. Η ομαδοποίηση έγινε για 150 ομάδες, αλλά χάριν συντομίας, παρουσιάζονται οι πρώτες 30.

```
1  Cluster
2  Size   ISim   ISDev ESim   ESdev
3  2    +1.000   +0.000   +0.001   +0.000
4  ( τις εξαιρέσεις οι, θα επιβεβαιώσουν, τον κανόνα)
5  ( τις εξαιρέσεις οι, θα επιβεβαιώσουν, τον κανόνα)
6
7
8  Cluster 2
9  Size   ISim   ISDev ESim   ESdev
10 2    +0.952   +0.000   +0.004 +0.003
11 ( ένα από τα συμπεράσματα του γκάλοπ της Κάπα Research, φοβούνται, για τις
       οικονομικές δυσκολίες)
12 ( ένα από τα συμπεράσματα του γκάλοπ της Κάπα Research, είναι, για τις
       οικονομικές δυσκολίες)
13
14
15 Cluster 3
16 Size   ISim   ISDev ESim   ESdev
17 2    +0.934   +0.000   +0.001   +0.001
18 ( στο σπήλαιο του Διρού, ήταν, σκεπασμένος από σταλακτιτικό υλικό)
19 ( στο σπήλαιο του Διρού, βρέθηκε, σκεπασμένος από σταλακτιτικό υλικό)
20
21
22 Cluster 4
23 Size   ISim   ISDev ESim   ESdev
24 2    +0.915   +0.000   +0.004   +0.003
25 ( και των αμυντικών δαπανών κατ', προβεί, και η Τουρκία σε αντίστοιχες κινήσεις)
26 ( και των αμυντικών δαπανών κατ', μπορεί να γίνει, και η Τουρκία σε αντίστοιχες
       κινήσεις)
27
28
29 Cluster 5
30 Size   ISim   ISDev ESim   ESdev
31 2    +0.914   +0.000   +0.004   +0.000
32 ( 32% πως τα πράγματα, αντιμετωπίζει, με αισιοδοξία τα επαγγελματικα θέματα)
33 ( 32% πως τα πράγματα, θα χειροτερέψουν, με αισιοδοξία τα επαγγελματικα θέματα)
34
35
36 Cluster 6
37 Size   ISim   ISDev ESim   ESdev
38 2    +0.910   +0.000   +0.003   +0.000
39 ( από κινητοποίηση αστυνομικών δυνάμεων οι νεαροί, κατευθύνθηκαν, προς την Ιερά
       Οδό)
40 ( από κινητοποίηση αστυνομικών δυνάμεων οι νεαροί, επιβιβάστηκαν, προς την Ιερά
       Οδό)
41
42
```

```
Cluster 7
Size   ISim   ISDev ESim   ESdev
2    +0.905   +0.000   +0.001   +0.000
( πως τα μεμονωμένα εγκληματικά γεγονότα, υπήρξε, προσπάθεια αναμόχλευσης των)
( πως τα μεμονωμένα εγκληματικά γεγονότα, προβλήθηκαν, προσπάθεια αναμόχλευσης
    των)


Cluster 8
Size   ISim   ISDev ESim   ESdev
2    +0.907 +0.000 +0.005   +0.000τους
( υποψηφίους στους κορυφαίους αυτοδιοικητικούς οργανισμούς του νομού στον,
    έκαναυπερασπίστηκα, την απόφαση του)τους
( υποψηφίους στους κορυφαίους αυτοδιοικητικούς οργανισμούς του νομού στον,
    εκλέγομαι, την απόφαση του)


Cluster 9
Size   ISim   ISDev ESim   ESdev
2    +0.908   +0.000   +0.006   +0.001
( και η φυσική ή οικονομική αδυναμία, γίνονται, αφόρητες)
( και η φυσική ή οικονομική αδυναμία, υπάρχουν, αφόρητες)


Cluster 10
Size   ISim   ISDev ESim   ESdev
2    +0.841   +0.000   +0.003   +0.000
( Η βουλευτής από τη μία πλευρά, δημιουργήθηκε, από την κόντρα με τον)
( Η βουλευτής από τη μία πλευρά, επιχειρεί να αποκλιμακώσει, από την κόντρα με
    τον)


Cluster 11
Size   ISim   ISDev ESim   ESdev
2    +0.830   +0.000   +0.007   +0.001
( Το θέμα, εκτιμώ, μεγαλύτερη ευελιξία επιλογών στο μέλλον  )
( Το θέμα, θεωρείται λήξαν, μεγαλύτερη ευελιξία επιλογών στο μέλλον  )


Cluster 12
Size   ISim   ISDev ESim   ESdev
2    +0.807   +0.000   +0.003   +0.001την
( ολοκλήρωση της αναβάθμισης στην ΕΑΒ άλλων 10 παλαιότερων απλών Μιράζ 2000, θα
    ενισχυθούν, οι δυνατότητες του αεροπορικού στόλου)
( Η έναρξη των παραδόσεων των 15 νέων Μιράζ, ανοίγει, τον δρόμο και για την
    ολοκλήρωση της αναβάθμισης στην ΕΑΒ άλλων 10 παλαιότερων απλών Μιράζ 2000)


Cluster 13
Size   ISim   ISDev ESim   ESdev
2    +0.783   +0.000   +0.002   +0.001
(   σενάρια μελλοντικών εξελίξεων  , να μην πιαστούν, στον ύπνο)
( στο υπουργείο Εξωτερικών, επεξεργάζονται,   σενάρια μελλοντικών εξελίξεων  )


Cluster 14
Size   ISim   ISDev ESim   ESdev
4    +0.779   +0.002   +0.007   +0.001
```

95 (την ομάδα των Αμερικανών, αναδεικνύεται, ως πρόσωπο της χρονιάς για το 2006>)
96 (την ομάδα των Αμερικανών, παραμιλάει, ως πρόσωπο της χρονιάς για το 2006>)
97 (την ομάδα των Αμερικανών, αναδεικνύεται, ως πρόσωπο της χρονιάς για το 2006)
98 (την ομάδα των Αμερικανών, παραμιλάει, ως πρόσωπο της χρονιάς για το 2006)
99
100
101 Cluster 15
102 Size ISim ISDev ESim ESdev
103 2 +0.769 +0.000 +0.002 +0.001
104 (στο βρετανικό υπουργείο Εσωτερικών και στις 27 Οκτωβρίου, έστειλε, απαντητική επιστολή στην επιτροπή των Ελληνοκυπρίων)
105 (Η διοίκηση της Άρσεναλ, απευθύνθηκε, στο βρετανικό υπουργείο Εσωτερικών και στις 27 Οκτωβρίου)
106
107
108 Cluster 16
109 Size ISim ISDev ESim ESdev
110 2 +0.756 +0.000 +0.002 +0.002
111 (ο Πρόεδρος της Δημοκρατίας Κάρολος Παπούλιας, έκανε, το ρεβεγιόν της Πρωτοχρονιάς στη Λέσχη Αξιωματικών)
112 (οι φτωχοί Έλληνες, εξέφρασε, ο Πρόεδρος της Δημοκρατίας Κάρολος Παπούλιας)
113
114
115 Cluster 17
116 Size ISim ISDev ESim ESdev
117 2 +0.751 +0.000 +0.004 +0.002
118 (και η βεβαίωση πρόσβασης, επιστρέφεται, η παλιά αίτηση)
119 (επειδή με την υποβολή τής νέας αίτησης, κατατίθεται, και η βεβαίωση πρόσβασης)
120
121
122 Cluster 18
123 Size ISim ISDev ESim ESdev
124 2 +0.710 +0.000 +0.003 +0.003
125 (άριστη γνώση των ελληνικών−, θεώρησα σημαντικό να μάθω, τη γλώσσα της χώρας στην οποία μένω)
126 (ως οικοδόμος− δουλειά, δεν απαιτεί, άριστη γνώση των ελληνικών−)
127
128
129 Cluster 19
130 Size ISim ISDev ESim ESdev
131 2 +0.710 +0.000 +0.005 +0.002
132 (σε οικονομική αδυναμία την τοποθέτηση προστατευτικών συστημάτων, δεν μπορεί, νόμιμο λόγο απαλλαγής ή μη παροχής από το Δημόσιο των οικονομικών μέσων για την ασφάλεια των διερχόμενων οχημάτων από τις αφύλακτες διαβάσεις)
133 (ΟΣΕ, αποδίδοντας, σε οικονομική αδυναμία την τοποθέτηση προστατευτικών συστημάτων)
134
135
136 Cluster 20
137 Size ISim ISDev ESim ESdev
138 2 +0.704 +0.000 +0.003 +0.000
139 (κόσμοςΟ στη Μέση Ανατολή, δεν έγινε, ασφαλέστερος μετά την εκτέλεση του Σαντάμ Χουσεΐν)
140 (η ανησυχία του υπουργείου Εξωτερικών από τις αντιδράσεις, προκαλεί, η εκτέλεση του Σαντάμ Χουσεΐν)
141
142
143 Cluster 21

```
Size   ISim   ISDev ESim   ESdev
3    +0.708   +0.115   +0.007   +0.004
( η κυβέρνηση Ερντογάν, δεν θα πάρει, οποιαδήποτε πρωτοβουλία)
( η κυβέρνηση Ερντογάν, είναι, οποιαδήποτε πρωτοβουλία)
( χέρι, κάνοντας, δύσκολη τη ζωή της κυβέρνησης Ερντογάν )


Cluster 22
Size   ISim   ISDev ESim   ESdev
4    +0.698   +0.038   +0.003   +0.002
( ρυπασμένες στον πλανήτη, κατατάσσεται, η Ελλάδα>)
( με αρκετές χώρες της Ασίας, θεωρούνται, ρυπασμένες στον πλανήτη)
( ρυπασμένες στον πλανήτη, κατατάσσεται, η Ελλάδα)
( με αρκετές χώρες της Ασίας, θεωρούνται, ρυπασμένες στον πλανήτη)


Cluster 23
Size   ISim   ISDev ESim   ESdev
3    +0.704   +0.128   +0.009   +0.005
( Ο Πρωθυπουργός, είναι, πολιτικός)
( ο ανταγωνισμός στην πολιτική, είναι, γένους ουδετέρου)
( ο ανταγωνισμός στην πολιτική, είναι, γένους ουδετέρου)


Cluster 24
Size   ISim   ISDev ESim   ESdev
3    +0.660   +0.147   +0.008   +0.000τις
( τελικές αποφάσεις, θα είναι, ο Μάρτιος)
( Η επιτυχία, δεν είναι, "τέκνο" μιας γενικής απόφασης)
( Η επιτυχία, δεν είναι,    τέκνο    μιας γενικής απόφασης)


Cluster 25
Size   ISim   ISDev ESim   ESdev
3    +0.651   +0.151   +0.001   +0.001
( Η αβρότητα στη συμπεριφορά, δεν αποκλείει, την ανταγωνιστικότητα στην πράξη)
( Η αβρότητα στη συμπεριφορά, δεν αποκλείει, την ανταγωνιστικότητα στην πράξη)
( στη Βουλή, αντιμετωπίζουν, με αβρότητα τις γυναίκες ή)


Cluster 26
Size   ISim   ISDev ESim   ESdev
3    +0.647   +0.153   +0.001   +0.000
( των γυναικών συναδέλφων, εμπεριέχει, και μια χροιά συγκατάβασης)
( των γυναικών συναδέλφων, εμπεριέχει, και μια χροιά συγκατάβασης)
( κανείς γεωστρατηγικές γνώσεις, εμπεριέχει, μια τέτοια εξέλιξη–)


Cluster 27
Size   ISim   ISDev ESim   ESdev
2    +0.626   +0.000   +0.003   +0.004
( Ελβετός, πιλοτάρει, Airbus της Swissair)
( Ο ΙΒ ΡΟΣΙ, είναι, Ελβετός)


Cluster 28
Size   ISim   ISDev ESim   ESdev
3    +0.622   +0.127   +0.001   +0.002
```

```
202 ( Λίγες νεφώσεις οι οποίες βαθμιαία , θα σημειωθούν , τοπικές βροχές και στα ορεινά
      χιονοπτώσεις)
203 ( Λίγες νεφώσεις οι οποίες βαθμιαία , θα πυκνώσουν , τοπικές βροχές και στα ορεινά
      χιονοπτώσεις)
204 ( κατά τη διάρκεια της ημέρας , θα πυκνώσουν , με βροχές το απόγευμα στο Ιόνιο)
205
206
207 Cluster 29
208 Size  ISim  ISDev ESim  ESdev
209 3   +0.625  +0.124  +0.006  +0.003τα
210 ( PM2,5 , είναι , 15 μg/m3 μέση( ετήσια τιμή))τα
211 ( PM2,5 , ισχύει( , 15 μg/m3 μέση( ετήσια τιμή))
212 ( TIM και GT–elecom , είναι , 8%)
213
214
215 Cluster 30
216 Size  ISim  ISDev ESim  ESdev
217 3   +0.611  +0.017  +0.003  +0.001
218 ( με την υπόθεση του Λιβάνου , λέει , χαρακτηριστικά διπλωματική πηγή)
219 ( χαρακτηριστικά διπλωματικές πηγές , θα επηρεάσουν , τους Ευρωπαίους–)
220 ( Ωστόσο τέτοιου είδους ανακατατάξεις στη Μέση Ανατολή , έλεγαν , χαρακτηριστικά
      διπλωματικές πηγές)
```

[1] Greek text mining resources. *Εργαστήριο Ευφυών Υπολογιστικών Συστημάτων, Σχολή Ηλεκτρολόγων Μηχανικών, Εθνικό Μετσόβιο Πολυτεχνείο*. URL http://www.islab.ntua.gr/el/research/text-mining-resources.

[2] Cluto - software for clustering high-dimensional datasets. URL http://glaros.dtc.umn.edu/gkhome/views/cluto.

[3] URL http://nlp.stanford.edu/software/index.shtml.

[4] URL http://www.w3.org/TR/owl2-overview.

[5] E. Agichtein and O. Etzioni. Snowball: Extracting relations from large plain-text collections. *Proceedings of the Fifth ACM International Conference on Digital Libraries*, 2000.

[6] M. Aizerman, E. Braverman, and L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821 – 837, 1964.

[7] M. Banko and O. Etzioni. The tradeoffs between open and traditional relation extraction. *Proceedings of the 46th Annual Meeting of the Association For Computational Linguistics*, pages 28–36, 2008.

[8] M. Banko, M. Cafarella, S. Soderland, M. J. Broadhead, and O. Etzioni. Open information extraction from the web. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 2670–2676, 2007.

[9] T. Berners-Lee, T. Health, and C. Bizer. Linked data - the story so far. *Special Issue on Linked Data, Intenational Journal on Semantic Web and Information Systems*, 2009.

[10] A. Bookstein and D. Swanson. Probabilistic models for automatic indexing. *Journal of the American Society for Information Science*, 25(5):313 318, 1974.

[11] B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. *5th Annual ACM Workshop on COLT*, pages 144 – 152, 1992.

[12] S. Brin. Extracting patterns and relations from the world wide web. *WebDB Workshop at 6th International Conference on Extending Database Technology*, EDBT '98:172–183, 1998.

[13] Yu T. Clement and G. Salton. An effective automatic indexing method. *ACM*, 23(1):76–88, 1976.

[14] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning, 20*, 1995.

[15] D. Downey, O. Etzioni, and S. Soderland. A probabilistic model of redundancy in information extraction. *Proceedings of International Joint Conferences on Artificial Intelligence*, 2005.

[16] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence*, 165(1):91–134, 2005.

[17] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2001.

[18] D. Hiemstra. *Information Retrieval: Searching in the 21st century*. Willey, 2009.

[19] T. Joachims. ch. 11: Making large-scale svm learning practical. advances in kernel methods - support vector learning. *B. Schölkopf and C. Burges and A. Smola*, 1999.

[20] T. Joachims. Learning to classify text using support vector machines. *dissertation*, 2002.

[21] H. W. Kuhn and A. W. Tucker. Nonlinear programming. *Proceedings of 2nd Berkeley Symposium*, pages 481–492, 1951.

[22] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labelling sequence data. *Proceedings of International Conference on Machine Learning*, 2001.

[23] H. Luhn. A statistical approach to mechanised encoding and searching of litary information. *IBM Journal of Research and Development*, 1(4):308–317, 1957.

[24] M. Maron and J. Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery*, 7:216 244, 1960.

[25] D. Metzler and W. Croft. Combining the language model and inference network approaches to retrieval. *Information Processing and Management*, 40(5):735 750.

[26] G.. Ntais. Development of a stemmer for the greek language. *Master Thesis, Stokholm University/Royal Institute of Technology*, 2006.

[27] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes: The Art of Scientific Computing (3rd ed.)*. New York: Cambridge University Press, 2007.

[28] E. Riloff and R. Jones. Learning dictionaries for information extraction by multi-level boot-strapping. *Proceedings of AAAI-99*, pages 1044–1049, 1999.

[29] S. Robertson. The probability ranking principle in information retrieval. *Journal of Documentation*, 33(4):294 304, 1977.

[30] G. Salton. The smart retrieval system: Experiments in automatic document processing. 1971.

[31] G. Salton and C. Yang. On the specification of term values in automatic indexing. *Journal of Documentation*, 29(4):351–372, 1973.

[32] G. Salton and C. Yang. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[33] S. M. Savaresi and D. L. Boley. On the performance of bisecting k-means and pdpp. *Proceedings of the First SIAM International Conference on Data Mining (ICDM-2001)*, pages 1–14, 2001.

[34] H. Turtle and W. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187 222, 1991.

[35] F. Wu and D. Weld. Open information extraction using wikipedia. *The Annual Meeting of the ACL*, 2010.

[36] A. Yates and O. Etzioni. Unsupervised resolution of objects and relations on the web. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2007.

[37] Ε. Κολέλη. Ενας νέος ελληνικός επισημειωτής μερών του λόγου, βασισμένος σε ταξινομητή μεγίστης εντροπίας. *πτυχιακή εργασία, Τμήμα Πληροφορικής Οικονομικού Πανεπιστημίου Αθηνών*, 2011.