ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Ανάπτυξη και αξιολόγηση αλγορίθμων μηχανικής μάθησης και νευρωνικών δικτύων σε δεδομένα κοινωνικών δικτύων με εφαρμογή στις μεταβολές των κρυπτονομισμάτων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Βασίλειος Α. Πασπάλας

Επιβλέπουσα: Θεοδώρα Βαρβαρίγου
Καθηγήτρια Ε.Μ.Π.

Αθήνα, Οκτώβριος 2018

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΕΠΙΚΟΙΝΩΝΙΩΝ, ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ

Ανάπτυξη και αξιολόγηση αλγορίθμων μηχανικής μάθησης και νευρωνικών δικτύων σε δεδομένα κοινωνικών δικτύων με εφαρμογή στις μεταβολές των κρυπτονομισμάτων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Βασίλειος Α. Πασπάλας

Επιβλέπουσα: Θεοδώρα Βαρβαρίγου
Καθηγήτρια Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 30η Οκτωβρίου 2018.

.....	.....	.....
Θεοδώρα Βαρβαρίγου	Δημήτριος Ασκούνης	Συμεών Παπαβασιλείου
Καθηγήτρια Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

.....

Βασίλειος Α. Πασπάλας

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικών Υπολογιστών Ε.Μ.Π.

Περίληψη

Η έλευση των κρυπτονομισμάτων είχε ως αποτέλεσμα την δημιουργία ηλεκτρονικών ανταλλακτηρίων στα οποία διενεργούνται αγοραπωλησίες μεταξύ κρυπτονομισμάτων αλλά και συναλλάγματος, όπως ευρώ, δολαρίων κτλ. Όπως στην αγορά συναλλάγματος και στα Χρηματιστήρια μετοχών, έτσι και σε αυτά τα ανταλλακτήρια υπάρχουν διακυμάνσεις στις τιμές των κρυπτονομισμάτων. Οι συγκεκριμένες τιμές επηρεάζονται από την προσφορά και την ζήτηση των συμμετεχόντων σε αυτά τα ανταλλακτήρια.

Με την εδραίωση των μέσων κοινωνικής δικτύωσης, όπως το Twitter και το Facebook, πολλοί άνθρωποι έχουν επιλέξει ως μέσο ενημέρωσης τις σελίδες που ενημερώνουν μεγάλες δημοσιογραφικές υπηρεσίες, όπως το Bloomberg και το CNN, στις πλατφόρμες αυτές. Αυτό συμβαίνει καθώς με τις ειδοποιήσεις που αποστέλλουν οι πλατφόρμες κοινωνικής δικτύωσης, υπάρχει άμεση και έγκυρη ενημέρωση των χρηστών για τα νέα που διαδραματίζονται ανά τον κόσμο σε κοινωνικό, πολιτικό και οικονομικό επίπεδο.

Στην συγκεκριμένη εργασία, μελετάμε την επιρροή που μπορεί να έχει η δημοσιοποίηση οικονομικών νέων σχετικά με τα κρυπτονομίσματα, συγκεκριμένα με το Bitcoin, στις τιμές αυτών στα ηλεκτρονικά ανταλλακτήρια ανά τον κόσμο. Για την υλοποίησή της χρησιμοποιήσαμε αναρτήσεις που έγιναν στο Twitter από 1/1/2017 έως 31/12/2017 από μεγάλες ειδησεογραφικές επιχειρήσεις που ενημερώνουν τις σελίδες τους στο Twitter. Στη συνέχεια με τη χρήση της επιβλεπόμενης μηχανικής μάθησης, και συγκεκριμένα με τη δημιουργία νευρωνικών δικτύων, δημιουργήσαμε και αξιολογήσαμε αλγορίθμους οι οποίοι προβλέπουν, με μεγάλη επιτυχία, την άνοδο ή την κάθοδο των τιμών των κρυπτονομισμάτων βάσει των νέων που προκύπτουν από τις αναρτήσεις που περιγράψαμε παραπάνω.

Λέξεις κλειδιά

κρυπτονομίσματα, ανταλλακτήριο νομισμάτων, χρηματιστήριο, αλγόριθμοι Επιβλεπόμενης Μηχανικής Μάθησης, εντοπισμός νέων και εξελίξεων, Νευρωνικά Δίκτυα, πρόβλεψη τιμών, Bitcoin, μέσα κοινωνικής δικτύωσης, Twitter, ειδησεογραφικές επιχειρήσεις, algorithmic trading

Abstract

Cryptocurrencies spread had as a result the creation of many electronic digital exchanges, which trade cryptocurrencies or digital currencies for other assets, such as conventional fiat money or other digital currencies. The volatility of the cryptocurrencies' values is influenced by the bid/ask, like it happens to the traditional foreign and stock exchanges.

Many people, due to the proliferation and the spread of social media platforms, like Twitter and Facebook, are getting informed for the latest news from pages that large media organizations, like Bloomberg and CNN, are updating to these platforms. This happens because people are getting instantly and accurately informed for breaking news all over the world via notification systems that social media platforms have developed. The content of these news varies from economic to politics and social news.

This paper focuses on the influence that news might have on the volatility of cryptocurrencies, specifically the Bitcoin's. To implement this approach we used posts that were published on Twitter between 1/1/2017 and 31/12/2017 from large media organizations. Moreover, we used Machine Learning technics, concretely Neural Networks, to implement and evaluate algorithms, which can precisely predict the volatility of the cryptocurrencies' values, based on the above news.

Key Words

Cryptocurrencies, foreign exchange, cryptocurrencies exchange, stock exchange, Machine Learning algorithms, Event Detection, Neural Networks, values prediction, Bitcoin, social media, Twitter, media organizations, algorithmic trading

Στους Γονείς μου

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στο Εργαστήριο Κατανεμημένης Γνώσης και Συστημάτων Πληροφορικής της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών και επισφραγίζει τις σπουδές μου στο Εθνικό Μετσόβιο Πολυτεχνείο.

Θα ήθελα να ευχαριστήσω θερμά την καθηγήτρια μου κα. Θεοδώρα Βαρβαρίγου που μου εμπιστεύθηκε αυτή τη διπλωματική, δίνοντας μου την ευκαιρία να γνωρίσω το εξαιρετικά ενδιαφέρον πεδίο της Μηχανικής Μάθησης.

Ιδιαίτερες ευχαριστίες οφείλω στον υποψήφιο διδάκτορα Γιώργο Παλαιοκρασσά για την πολύτιμη βοήθεια του σε όλα τα στάδια υλοποίησης της εργασίας. Οφείλω να πω ότι ήταν πάντα διαθέσιμος, όποτε και αν τον χρειάστηκα και φρόντιζε να μου αφιερώνει αρκετό από το χρόνο του για να συζητάμε και να επιλύουμε τα προβλήματα που συναντούσα.

Φυσικά ένα ευχαριστώ είναι λίγο για τους φίλους που όλα αυτά τα χρόνια βρίσκονται δίπλα μου και που χωρίς αυτούς δεν θα είχα καταφέρει να φτάσω μέχρι εδώ.

Επίσης, θα ήθελα να ευχαριστήσω ξεχωριστά τον ξάδερφό μου Άγγελο Βατίκαλο, καθώς σε όλα τα στάδια της ζωής μου με εμπνέει και με καθοδηγεί με τις γνώσεις του και τον ξεχωριστό τρόπο σκέψης του.

Τέλος, ένα πολύ μεγάλο ευχαριστώ στην αδελφή μου και στους γονείς μου που με ανέχονται και με στηρίζουν τόσα χρόνια.

Βασίλειος Α. Πασπάλας

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κατάλογος Κώδικα

1 Εισαγωγή

Τα τελευταία χρόνια, η ανάπτυξη του Διαδικτύου έχει οδηγήσει στην δημιουργία νέων αναγκών και συνηθειών οι οποίες, σε μεγάλο βαθμό, έχουν επηρεάσει την καθημερινότητα των ανθρώπων. Συγκεκριμένα, πριν από περίπου μία δεκαετία ξεκίνησε η λειτουργία των μέσων κοινωνικής δικτύωσης και μέσα σε ελάχιστα χρόνια οι πλατφόρμες αυτές είχαν διεισδύσει ολοκληρωτικά στην καθημερινή ζωή των πολιτών αλλά και των επιχειρήσεων. Εν συνεχεία, στο τέλος της πρώτης δεκαετίας του 21^{ου} αιώνα έκαναν την εμφάνιση τους τα πρώτα κρυπτονομίσματα, με το πιο διαδεδομένο από αυτά να είναι το Bitcoin. Με την εξάπλωση τους, περίπου 5 χρόνια αργότερα, άρχισαν να δημιουργούνται ηλεκτρονικές πλατφόρμες οι οποίες εστιάζουν αποκλειστικά και μόνο στην αγοραπωλησία και στην ανταλλαγή κρυπτονομισμάτων με άλλα συναλλάγματα, όπως το ευρώ ή το δολάριο. Αυτό είχε ως αποτέλεσμα την δημιουργία μεγάλου όγκου συναλλαγών, καθώς επίσης και την ραγδαία ανάπτυξη των ανταλλακτηρίων αυτών. Η γιγάντωση των μέσων μαζικής δικτύωσης σε συνδυασμό με την κατακόρυφη ανάπτυξη των κρυπτονομισμάτων και των επιχειρήσεων γύρω από αυτά, γέννησε την ιδέα της αλγοριθμικής ανταλλαγής των νομισμάτων αυτών, βάσει των αναρτήσεων στα μέσα κοινωνικής δικτύωσης.

Η «αλγοριθμική ανταλλαγή» (Algorithmic trading), που βασίζεται στην επιστήμη της μηχανικής μάθησης, αποκτά ολοένα και περισσότερο έδαφος στο χώρο των επιχειρήσεων που ασχολούνται με τις επενδύσεις και τη διαχείριση κεφαλαίων. Πρωτοπόρες εταιρίες του συγκεκριμένου κλάδου επενδύουν ολοένα και μεγαλύτερα κεφάλαια στον τομέα της τεχνητής νοημοσύνης, και συγκεκριμένα στη μηχανική μάθηση, προκειμένου να δημιουργηθούν αλγόριθμοι οι οποίοι μπορούν να προβλέψουν τις μεταβολές στις τιμές μετοχών, δεικτών χρηματιστηρίου, συναλλαγμάτων και άλλων χρηματιστηριακών προϊόντων.

1.1 Αντικείμενο της Διπλωματικής

Το κίνητρο της παρούσας διπλωματικής εργασίας είναι η δημιουργία ενός καινοτόμου συστήματος το οποίο θα μπορεί να προβλέπει με μεγάλη ακρίβεια τις πιθανές μεταβολές των κρυπτονομισμάτων, βασιζόμενο στις αναρτήσεις που πραγματοποιούνται στην πλατφόρμα κοινωνικής δικτύωσης «Twitter» από μεγάλες, δημοσιογραφικές κυρίως, επιχειρήσεις.

Στη συγκεκριμένη εργασία, θα χρησιμοποιήσουμε δεδομένα που έχουμε συλλέξει από την πλατφόρμα του Twitter, με τη χρήση της διεπαφής προγραμματισμού εφαρμογών (API) που έχει αναπτύξει η συγκεκριμένη εταιρία για την άντληση αναρτήσεων από την βάση

δεδομένων της. Τα δεδομένα αυτά, θα τα αποθηκεύσουμε σε μία σχεσιακή βάση δεδομένων και θα προχωρήσουμε στην κατάλληλη επεξεργασία τους προκειμένου να χρησιμοποιηθούν στο επόμενο στάδιο της μελέτης.

Κατόπιν, θα αντλήσουμε συγκεντρωτικά δεδομένα, από τα μεγαλύτερα ανταλλακτήρια παγκοσμίως, για τις ωριαίες και ημερήσιες μεταβολές που παρουσιάζουν οι τιμές των κρυπτονομισμάτων, σε σχέση με την τιμή του Αμερικάνικου Δολαρίου. Η αποθήκευση και αυτών των δεδομένων θα γίνει στους πίνακες της βάσης δεδομένων που αναφέραμε προηγουμένως.

Εν συνεχεία, με την τεχνική της μηχανικής μάθησης θα εκπαιδεύσουμε ένα νευρωνικό δίκτυο το οποίο θα προβλέπει την άνοδο ή την κάθοδο των τιμών των συγκεκριμένων νομισμάτων. Ως είσοδο του συγκεκριμένου νευρωνικού δικτύου θα χρησιμοποιήσουμε πίνακες οι οποίοι αποτυπώνουν το περιεχόμενο αλλά και την εγγύτητα των αναρτήσεων στο χρόνο, σε συνδυασμό με την μεταβολή της τιμής του Bitcoin για την συγκεκριμένη χρονική περίοδο.

Τέλος, θα αξιολογήσουμε τα αποτελέσματα μας βάσει των μετρήσεων που θα εξάγουμε από μετρικές που χρησιμοποιούνται ευρέως για την αξιολόγηση αλγορίθμων μηχανικής μάθησης.

1.2 Οργάνωση κειμένου

Η παρούσα διπλωματική εργασία αποτελείται από τα ακόλουθα κεφάλαια:

Κεφάλαιο 2

Παρουσιάζεται εκτενέστερα η λειτουργία των κρυπτονομισμάτων, των ανταλλακτηρίων καθώς επίσης και ιστορικά δεδομένα που αφορούν την τιμή του Bitcoin και γεγονότα που το επηρέασαν.

Κεφάλαιο 3

Γίνεται παρουσίαση της μηχανικής μάθησης και της θεωρίας πίσω από αυτή την τεχνική.

Κεφάλαιο 4

Παρουσιάζεται η λειτουργία και η ραγδαία επέκταση των κοινωνικών δικτύων, με εκτενέστερη αναφορά στο Twitter.

Κεφάλαιο 5

Γίνεται παρουσίαση των εργαλείων που χρησιμοποιήθηκαν για την υλοποίηση του συγκεκριμένου συστήματος και των επιμέρους πειραμάτων που πραγματοποιήθηκαν.

Κεφάλαιο 6

Παρουσιάζεται η διαδικασία υλοποίησης του συστήματος και τα βήματα που ακολουθήθηκαν για την επιτυχή ολοκλήρωση των πειραμάτων. Επίσης αξιολογείται η απόδοση των αλγορίθμων.

Κεφάλαιο 7

Συνοψίζονται τα συμπεράσματα της μελέτης και αναφέρονται μελλοντικοί προσανατολισμοί έρευνας όσον αφορά στην αλγοριθμική ανταλλαγή βάσει δεδομένων κοινωνικών δικτύων.

2 Κρυπτονομίσματα

Το κρυπτονόμισμα είναι μια ηλεκτρονική μορφή περιουσιακού στοιχείου που χρησιμοποιείται ως μέσω ανταλλαγής (νόμισμα) και το οποίο βασίζεται πάνω στις αρχές της κρυπτογραφίας για τη διασφάλιση των οικονομικών συναλλαγών, τον έλεγχο της δημιουργίας πρόσθετων μονάδων και την επιβεβαίωση της μεταφοράς νομισμάτων. Τα κρυπτονομίσματα κάνουν χρήση μιας Κατανεμημένης Βάσης Δεδομένων [1] ως τον πυλώνα του συστήματος τους, το επονομαζόμενο Blockchain [2], το οποίο χρησιμοποιείται ως μία δημόσια βάση δεδομένων που περιέχει όλες τις συναλλαγές. Το πρώτο επιτυχημένο αποκεντρωμένο κρυπτονόμισμα είναι το Bitcoin το οποίο παρουσιάστηκε το 2009. Λόγω της ανοιχτής φύσης του λογισμικού του, επετράπη σε πολλούς προγραμματιστές να πειραματιστούν με τον κώδικά του και να τον τροποποιήσουν. Έκτοτε δημιουργήθηκε μία πληθώρα νέων κρυπτονομισμάτων στα οποία έχουν γίνει προσπάθειες για να βελτιωθούν ή και να προστεθούν λειτουργίες όπως ταχύτερες συναλλαγές, μεγαλύτερη ανωνυμία κ.α. Το ανώτατο όριο της αγοράς κρυπτογράφησης εκτιμάται ότι θα φτάσει τα $ 1-2 τρισεκατομμύρια το 2018.

2.1 Bitcoin

Το Bitcoin είναι ένα κρυπτονόμισμα το οποίο εφευρέθηκε από ένα άγνωστο πρόσωπο ή μία ομάδα ατόμων που χρησιμοποιεί το όνομα Satoshi Nakamoto [3] και παρουσιάστηκε με τη μορφή ανοιχτού κώδικα λογισμικού το 2009. Τα Bitcoins δημιουργούνται με τη μορφή επιβράβευσης για τη διαδικασία η οποία είναι γνωστή ως εξόρυξη (Mining). Μπορούν να ανταλλαχθούν με άλλα συναλλάγματα, εμπορεύματα και υπηρεσίες. Το 2017, έπειτα από έρευνα που πραγματοποιήθηκε από το πανεπιστήμιο του Cambridge [4], υπολογίζεται ότι 2,9 έως 5,8 εκατομμύρια μοναδικοί χρήστες χρησιμοποίησαν το Bitcoin ως μέσω συναλλαγής.

Το Bitcoin έχει δεχθεί κριτική καθώς χρησιμοποιείται για παράνομες συναλλαγές, απαιτεί μεγάλη κατανάλωση ενέργειας για την εξόρυξη Bitcoins, έχει μεγάλες μεταβολές στην τιμή του, παρατηρούνται κλοπές κατά τις συναλλαγές και επειδή πολλοί θεωρούν πως είναι μία οικονομική φούσκα (economic bubble) [5]. Όμως το Bitcoin χρησιμοποιείται και ως επένδυση, παρόλο που πολλοί ρυθμιστικοί οργανισμοί έχουν προειδοποιήσει τους επενδυτές.

2.1.1 Ιστορία

Δημιουργία

Τον Ιανουάριο του 2009, το δίκτυο του Bitcoin δημιουργήθηκε όταν ο Nakamoto εξόρυξε το πρώτο “block” της αλυσίδας, το οποίο είναι γνωστό ως *genesis block.*

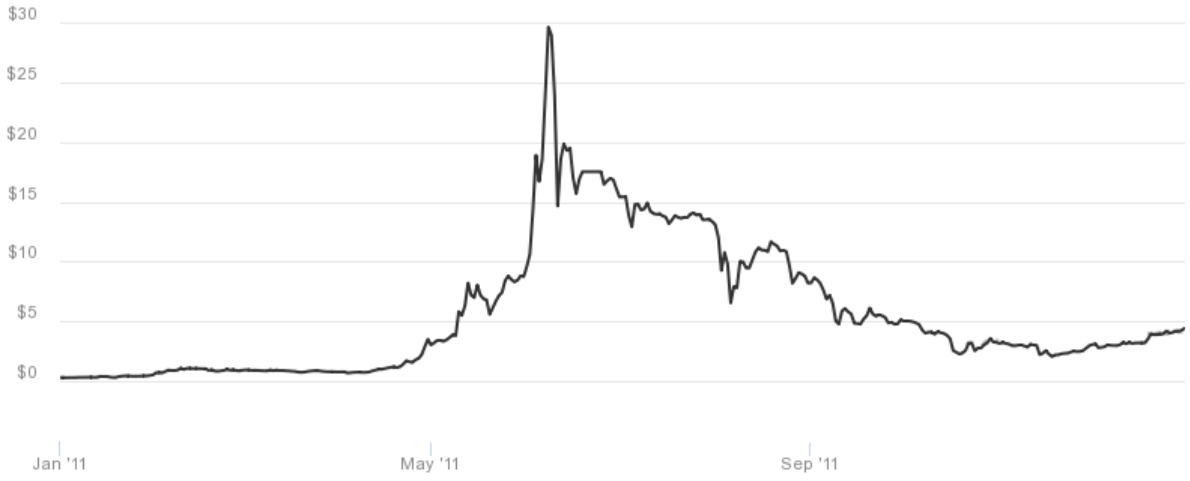
2011-2012

Μέτα τις πρώτες συναλλαγές που πραγματοποιήθηκαν ως απόδειξη λειτουργίας της ιδέας, οι πρώτοι χρήστες του Bitcoin ήταν οι μαύρες αγορές, όπως η Silk Road [6]. Τους πρώτους 30 μήνες λειτουργίας της, ξεκινώντας από το Φεβρουάριο του 2011, η Silk Road δεχόταν αποκλειστικά και μόνο Bitcoin ως μέσω συναλλαγής, συναλλάσσοντας 9.9 εκατομμύρια Bitcoin, αξίας $214 εκατομμυρίων περίπου.

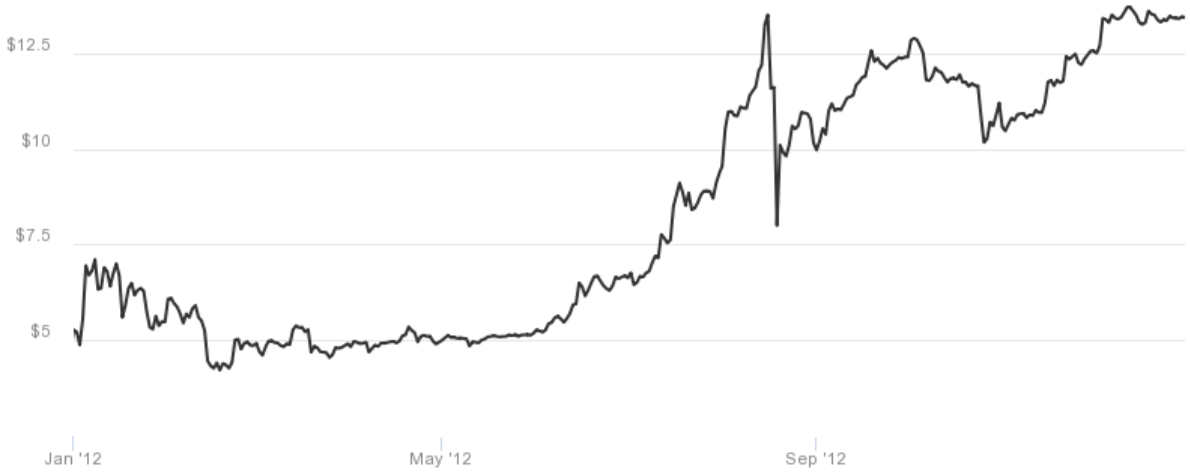
Το 2011, η τιμή του ξεκίνησε στα $0,30 ανά Bitcoin και στις αρχές Ιουνίου κάλπασε στα $31,50. Μέσα σε ένα μήνα έπεσα στα $11,00. Τον επόμενο μήνα, ξανά, υπέστη πτώση και η τιμή του έφτασε στα $7,80 και στη συνέχεια στα $4,77.

Το 2012 η τιμή του Bitcoin ξεκίνησε στα $5,27 και στο τέλος του χρόνου είχε φτάσει τα $13,30. Στις 9 Ιανουαρίου η τιμή του ανέβηκε στα $7,38, αλλά μέσα στις επόμενες 16 μέρες η τιμή του κατέρρευσε κατά 49% και έφτασε τα $3,80. Η τιμή του ανέβηκε ξανά στις 17 Αυγούστου, όμως μέσα στις επόμενες 3 μέρες έπεσε ξανά κατά 57% στα $7,10.

Το ίδρυμα του Bitcoin ιδρύθηκε το Σεπτέμβρη του 2012 προκειμένου να προωθήσει την ανάπτυξη του [7]. Τα δύο παρακάτω σχήματα παρουσιάζουν την τιμή του Bitcoin κατά τα πρώτα χρόνια δημιουργίας του.



Σχήμα 1: Η τιμή του Bitcoin το 2011



Σχήμα 2: Η τιμή του Bitcoin το 2012

2013-2016

Το 2013 η τιμή του Bitcoin ξεκίνησε στα $13,30 και ξεπέρασε τα $770 μέχρι το τέλος του έτους.

Τον Μάρτιο του 2013 το Blockchain χωρίστηκε σε δύο ανεξάρτητες αλυσίδες με διαφορετικούς κανόνες. Τα δύο Blockchain λειτουργούσαν ταυτοχρόνως για 6 ώρες, καθένα από τα οποία είχε το δικό του ιστορικό από τις συναλλαγές. Το σύστημα επανήλθε στην φυσιολογική του λειτουργία όταν η πλειοψηφία του δικτύου έκανε επαναφορά στην έκδοση 0.7 του λογισμικού του Bitcoin. Το ανταλλακτήριο Mt. Gox διέκοψε προσωρινά τις καταθέσεις για το Bitcoin και η τιμή του έπεσε κατά 23% στα $37, όμως μέσα στις επόμενες ώρες η τιμή του επανήλθε στα φυσιολογικά της επίπεδα στα $48.

Το γραφείο Οικονομικών Εγκλημάτων των Ηνωμένων πολιτειών της Αμερικής (FinCEN) θέσπισε κανονισμούς για τα «αποκεντρωμένα εικονικά νομίσματα» όπως το Bitcoin.

Τον Απρίλιο τα ανταλλακτήρια BitInstant και Mt. Gox αντιμετώπισαν καθυστερήσεις στη διεκπεραίωση συναλλαγών λόγω της έλλειψης χωρητικότητας. Αυτό είχε ως αποτέλεσμα την πτώση της τιμής του Bitcoin από τα $266 στα $76 μέχρις ότου να επιστρέψει στα $160 μέσα στις επόμενες έξι ώρες.

Στις 10 Απριλίου η τιμή του Bitcoin εκτινάχτηκε στα $259, αλλά μέσα στις επόμενες τρεις μέρες έχασε το 83% της αξίας του, πέφτοντας στα $45.

Στις 15 Μαΐου του 2013, οι Αμερικάνικες αρχές κατάσχεσαν λογαριασμούς που συσχετίζονταν με το ανταλλακτήριο Mt. Gox καθώς ανακάλυψαν ότι δεν είχε εγγραφεί στα μητρώα του FinCEN.

Η τιμή του Bitcoin έφτασε μέχρι και τα $755 στις 19 Νοεμβρίου, όπου και κατέρρευσε κατά 50% την ίδια μέρα φτάνοντας τα $378. Στις 30 Νοεμβρίου 2013 η τιμή σκαρφάλωσε στα

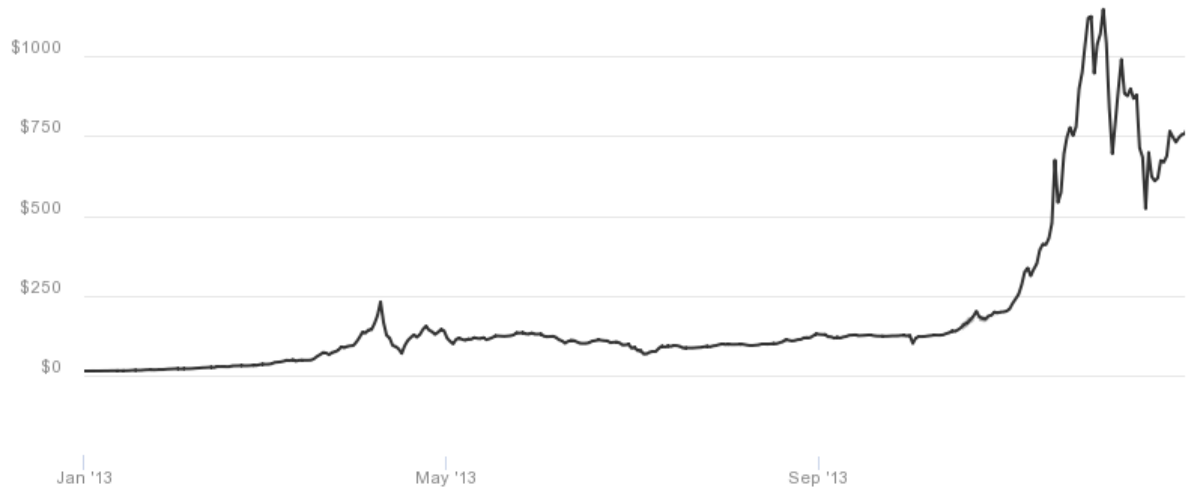
$1.163 όπου και ξεκίνησε η διαρκής πτώση του, χάνοντας μέχρι και το 87% της αξίας του τον Ιανουάριο του 2015 όπου και η τιμή του ήταν $152.

Στις 5 Δεκεμβρίου 2013, η διοίκηση της Τράπεζας της Κίνας απαγόρευσαν στα οικονομικά ιδρύματα της χώρας να χρησιμοποιούν το Bitcoin. Μετά την ανακοίνωση η τιμή του άρχισε να πέφτει και η εταιρία Baidu σταμάτησε να δέχεται Bitcoin για συγκεκριμένες υπηρεσίες της. Η αγορά αγαθών με οποιοδήποτε ηλεκτρονικό νόμισμα ήταν ήδη απαγορευμένη και παράνομη στην Κίνα από το 2009.

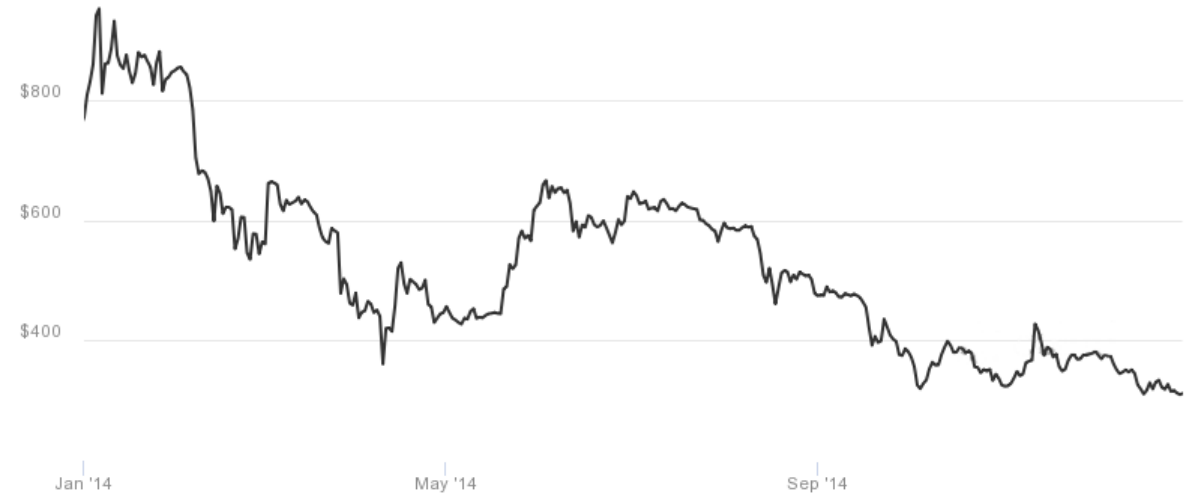
Το 2014 η τιμή του ξεκίνησε στα $770 και έπεσε στα $314 μέχρι το τέλος του έτους.

Το Φεβρουάριο του 2014, το ανταλλακτήριο Mt. Gox, το οποίο ήταν το μεγαλύτερο εκείνη τη στιγμή, ανακοίνωσε ότι 850.000 Bitcoin εκλάπησαν από τους πελάτες του, τα οποία υπολογίζονται στα $500 εκατομμύρια. Η τιμή του Bitcoin έπεσε σχεδόν στο μισό, από $867 στα $439 (πτώση 50%). Η τιμή παρέμεινε χαμηλά μέχρι το τέλος του 2016.

Το 2015 η τιμή του ξεκίνησε στα $314 και έφτασε μέχρι τα $434 στο τέλος του έτους. Στο τέλος του 2016 η τιμή έφτασε στα $998.

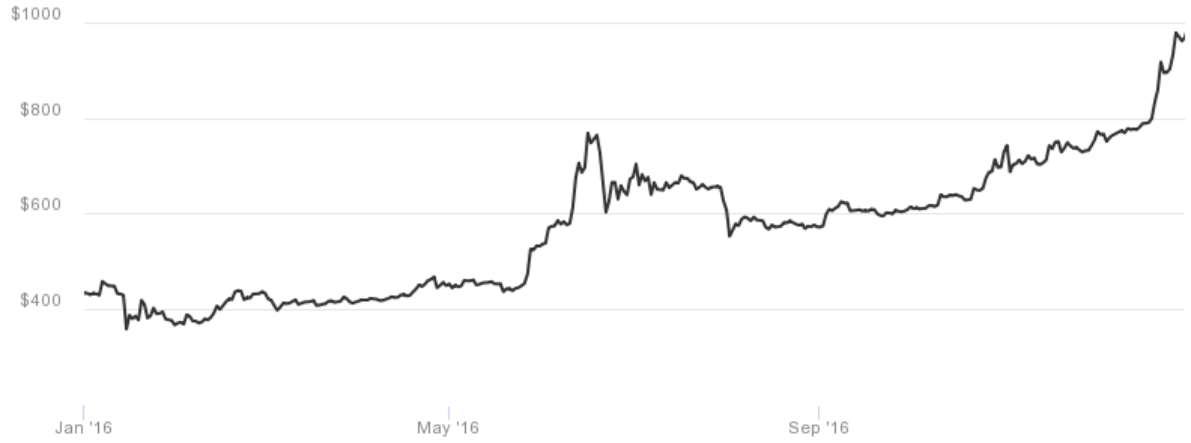


Σχήμα 3: Η τιμή του Bitcoin το 2012



Σχήμα 4: Η τιμή του Bitcoin το 2014

Σχήμα 5: Η τιμή του Bitcoin το 2015

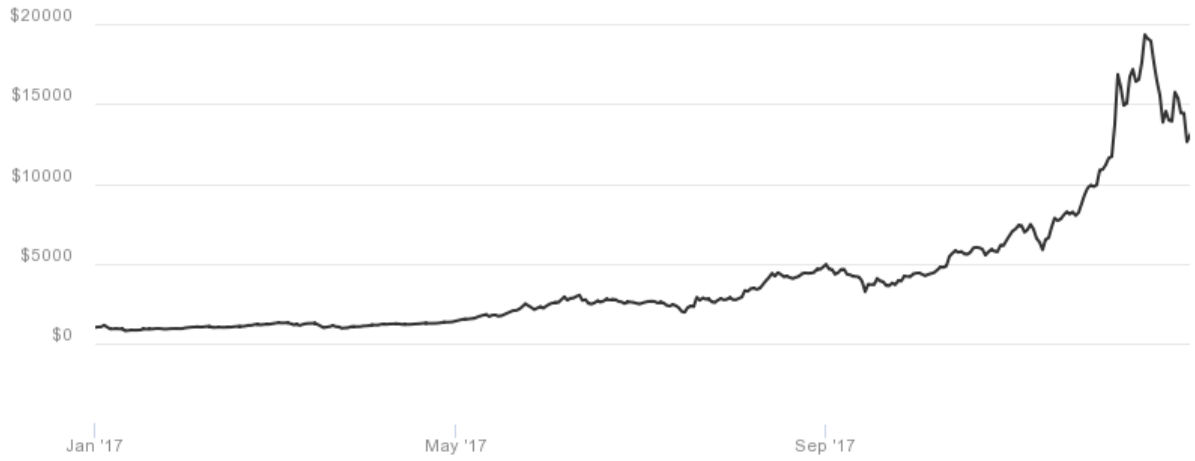


Σχήμα 6: Η τιμή του Bitcoin το 2016

2017

Το 2017 η τιμή του ξεκίνησε στα $998 και στο τέλος του έτους είχε φτάσει τα $13.412,44. Το Δεκέμβριο του 2017 η τιμή του Bitcoin βρισκόταν στα $19.666, η οποία είναι και η υψηλότερη όλων των εποχών.

Η Κίνα απαγόρευσε τις αγορές με Bitcoin, με τις πρώτες ενέργειες να ξεκινούν τον Σεπτέμβρη του 2017.



Σχήμα 7: Η τιμή του Bitcoin το 2017

Γίνεται αντιληπτό πως η τιμή του Bitcoin επηρεάζεται άμεσα από τα γεγονότα και τις αποφάσεις που το αφορούν.

2.1.2 Το λογισμικό

Το Bitcoin αποτελεί στη βάση του ένα λογισμικό ανοιχτού κώδικα (open source protocol). Κατά συνέπεια, ο πηγαίος κώδικας του λογισμικού είναι δημόσιος και διαθέσιμος σε όποιον επιθυμεί να ελέγξει τις λεπτομέρειες της λειτουργίας του. Η ανωτέρω αρχή επιτρέπει σε οποιονδήποτε την ελεύθερη και δωρεάν αντιγραφή και ανάπτυξη δικού του λογισμικού βασισμένου στο υπάρχον.

Το λογισμικό αποτελεί μία μέθοδο για την επίτευξη των παρακάτω κύριων στόχων:

1) Θέσπιση κριτηρίων παραγωγής και συναλλαγής των ανταλλάξιμων μονάδων του λογισμικού (Bitcoin),

2) Διατήρηση των πληροφοριών ιδιοκτησίας των μονάδων των Bitcoin που έχουν ήδη παραχθεί,

3) Δυναμική επιβεβαίωση της εγκυρότητας των παραπάνω, χωρίς την ανάγκη ύπαρξης κεντρικής οντότητας ελέγχου, πιστοποίησης ή διακρίβωσης.

Η χρήση του λογισμικού είναι δωρεάν και διαθέσιμη σε όλες τις χώρες του κόσμου, εφόσον υπάρχει σύνδεση στο διαδίκτυο (Internet). Η βασική λειτουργία του λογισμικού έγκειται στην εκτέλεση συναλλαγών Bitcoin και την αναμετάδοση πληροφοριών ανάμεσα σε κόμβους και την επιβεβαίωση της εγκυρότητάς τους για το υπόλοιπο δίκτυο. Καθώς το λογισμικό είναι ανοιχτού κώδικα, δύνανται να υπάρχουν πάρα πολύ διαφορετικές εκδόσεις και εκδοχές του. Στην ουσία, ο καθένας θα μπορούσε με τις κατάλληλες ικανότητες να δημιουργήσει ένα αντίστοιχο δίκτυο, αντιγράφοντας σε μεγάλο βαθμό το λογισμικό του Bitcoin, προσθέτοντας ή διαφοροποιώντας με ότι κανόνες επιθυμεί. Κατά αυτήν την έννοια, τα συστατικά στοιχεία του λογισμικού, έχουν δημιουργηθεί συναινετικά από προγραμματιστές, ενσωματώνοντας καινοτομίες διαθέσιμες από άλλα λογισμικά ανοιχτού κώδικα, αλλά και νέα στοιχεία που δεν είχαν εμφανιστεί πριν.

Η ισχύς του δικτύου εξασφαλίζεται από την αποδοχή του από τους χρήστες. Το δίκτυο το οποίο αποτελούν οι χρήστες του Bitcoin, αποτελείται από χρήστες της ίδιας εκδοχής του λογισμικού. Αλλαγές στον κώδικα προτείνονται στην κοινότητα, αλλά η συναίνεση της κοινότητας των χρηστών και η αποδοχή τους είναι που δημιουργεί το δίκτυο.

Η μαζική αποδοχή από τους χρήστες οφείλεται:

1) στην διαφάνεια του πηγαίου κώδικα του λογισμικού,

2) στην ακεραιότητα και διαφάνεια των συναλλασσόμενων πληροφοριών,

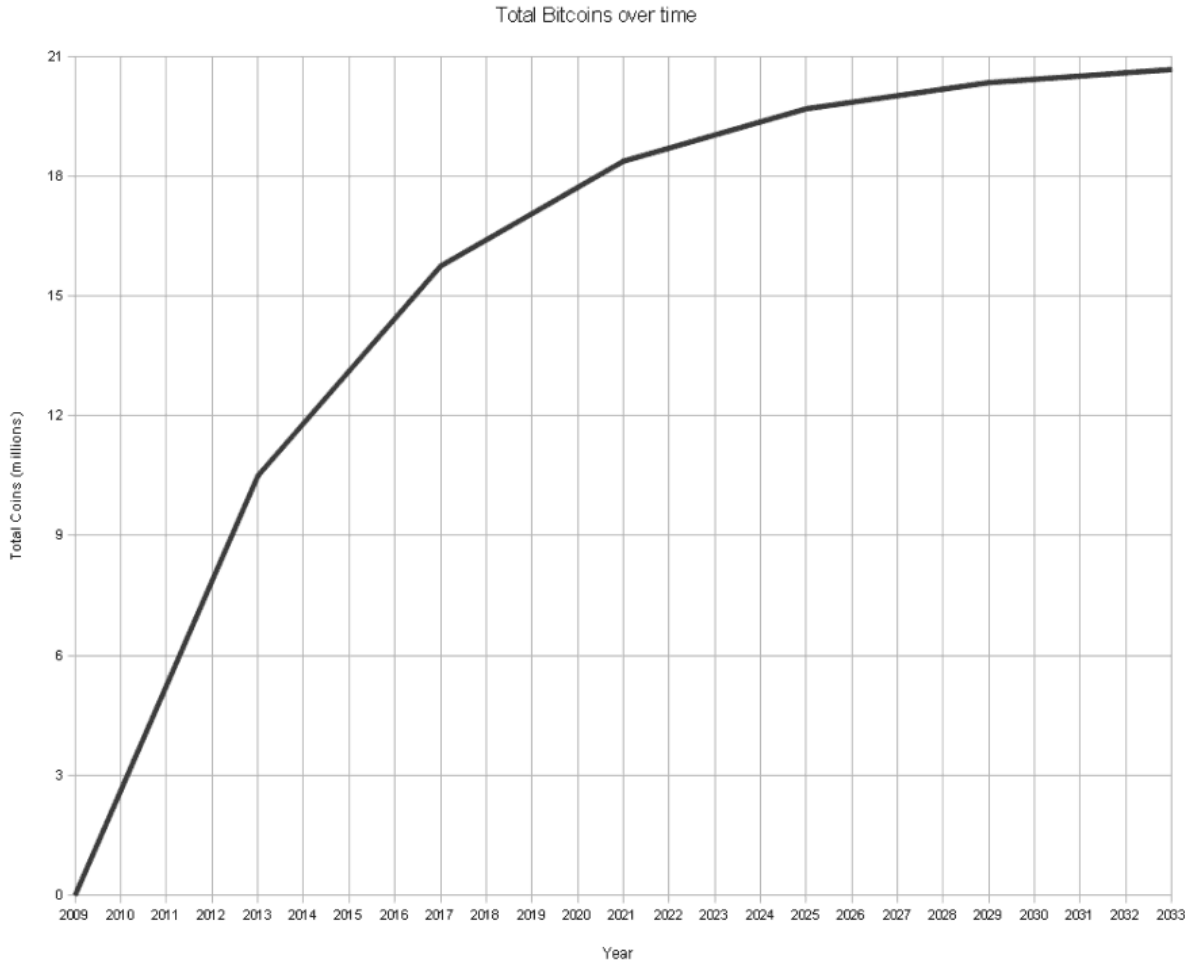
3) στην στιβαρότητα του δικτύου από κακόβουλες επιθέσεις,

4) στην προγραμματισμένα περιορισμένη παραγωγή Bitcoin,

5) στην προστασία που παρέχουν οι κρυπτογραφικοί αλγόριθμοι που χρησιμοποιούνται ενάντια σε κακόβουλη εκμετάλλευση του δικτύου, όπως και άλλοι συμπληρωματικοί λόγοι, είναι συνολικά υπεύθυνοι για την αποδοχή του από τους χρήστες, αλλά και την εξάπλωσή του σε νέους. Αυτό το λογισμικό και οι εξελίξεις του αποτελούν τον πυρήνα του συστήματος συναλλαγής Bitcoin. Η δυνατότητα ανταλλαγής πληροφοριών με ακεραιότητα ανεξαρτήτως αποδέκτη εντός του δικτύου, η περιορισμένη διάθεση και πεπερασμένη ποσότητα των Bitcoin, δημιουργεί τις βασικές προδιαγραφές για ένα δίκτυο ανταλλαγής αξίας. Όποια αξία βρίσκουν οι χρήστες αποτυπώνεται αποκλειστικά στην αξία με την οποία είναι διατεθειμένοι να τα ανταλλάξουν, η οποία με τη σειρά της βασίζεται αποκλειστικά στους νόμους της προσφοράς και της ζήτησης, χωρίς ενδιάμεσα μέρη (χώρες, κεντρικούς εκδότες ή αρχές).

Για να μπορούν να είναι χρήσιμα σαν μέσο συναλλαγής πρέπει να εισάγονται στην κυκλοφορία σταδιακά για την κάλυψη των συναλλακτικών αναγκών, αλλά και να είναι πεπερασμένα σε συνολικό αριθμό. Αυτό επιτυγχάνεται τεχνητά και ο ρυθμός παραγωγής τους όπως και το μέγιστο πλήθος, αποτελούν μέρος των κανόνων του δικτύου. Ο μέγιστος αριθμός που θα παραχθεί ποτέ είναι 21.000.000 και ο ρυθμός παραγωγής τους θα

ελαττώνεται σταδιακά έως περίπου το 2140 οπότε και θα παραχθεί το τελευταίο. Αυτή η μέθοδος σε κάποιο βαθμό, προσομοιάζει την πορεία διάθεσης ενός πολύτιμου μετάλλου (άργυρος, χρυσός) στην παγκόσμια αγορά. Αρχικά, η εξόρυξή του είναι εύκολη και σχετικά μεγάλες ποσότητες είναι πιο εύκολα διαθέσιμες, προοδευτικά όμως γίνεται σπανιότερο έως ότου εξαντληθούν τα αποθέματα του πλανήτη. Ο ρυθμός παραγωγής Bitcoin προσαρμόζεται τεχνητά ώστε να ακολουθεί περίπου την παρακάτω καμπύλη:



Σχήμα 8: Ο ρυθμός παραγωγής του Bitcoin

2.1.3 Βασικά πλεονεκτήματα

Ταχύτητα Συναλλαγών/Διεθνής Φύση: Οι συναλλαγές σε Bitcoin συμβαίνουν άμεσα και ανακοινώνονται ταυτόχρονα σε όλο το δίκτυο ανά τον πλανήτη. Αυτό δεν απαιτεί άλλες υποδομές πέρα από κάποια μορφή του δωρεάν λογισμικού σε υπολογιστή ή σε Smartphone, και σύνδεση στο διαδίκτυο.

Εξαιρετικά Χαμηλό κόστος συναλλαγών: Το παρόν κόστος για κάθε συναλλαγή ανεξαρτήτως μεγέθους ανέρχεται περίπου στα 5€ cents και είναι προαιρετικό, αν δεν υπάρχει βιασύνη επιβεβαίωσης της συναλλαγής. Σε ακόμα πιο σύνθετα δίκτυα υπό την

σκέπη επί μέρους ελεγκτικών δικτύων το κόστος συναλλαγών/αγορών δύναται να προσεγγίσει πολύ χαμηλότερες τιμές. Το ποσό αυτό αποδίδεται αυτόματα στους χρήστες, που εκτελούν τους ελέγχους των συναλλαγών και την επιβεβαίωση της αντικειμενικότητάς του, ως αμοιβή για την επεξεργαστική ισχύ που επενδύουν στην προστασία του δικτύου από κακόβουλες επιθέσεις.

Έλεγχος από το χρήστη/Προστασία από υφαρπαγή: Καθώς ο χρήστης είναι ο μόνος που έχει τη δυνατότητα να εκτελέσει συναλλαγές και εφόσον δεν έχει παραχωρήσει αυτό το δικαίωμα, και έχει προστατεύσει λογικά την πρόσβαση στα Bitcoin του, είναι πρακτικά αδύνατο να κλαπούν ή να υφαρπαχτούν από τρίτους (εφόσον η κρυπτογράφηση δεν παραβιαστεί). Περαιτέρω προβλέψεις επιτρέπουν την δυνατότητα μεταφοράς τους μόνο υπό πολύ ορισμένες συνθήκες, όπως μόνο από ορισμένα προσυμφωνημένα μέρη ταυτόχρονα για την αποφυγή μονομερών εκθέσεων ή μόνο μετά από συγκεκριμένο χρόνο.

Φορητότητα/αντίγραφα ασφαλείας: Ανεξάρτητα από το πλήθος τους, τα Bitcoin και τα «πορτοφόλια» αποθήκευσης ή οι κωδικού πρόσβασης σε αυτά είναι ουσιαστικά πάρα πολύ μικρά σε μέγεθος, και μπορούν να μεταφερθούν εύκολα, να καταγραφούν σε χαρτί, ακόμα και να απομνημονευτούν. Επίσης, κάτι αδύνατο για συμβατικές αξίες, μπορούν να αντιγραφούν ώστε να υπάρχουν αντίγραφα ασφαλείας σε περίπτωση καταστροφής των αρχικών. Βέβαια αν παραβιαστεί οποιοδήποτε από τα αντίγραφα, τα υπόλοιπα είναι επίσης παραβιασμένα.

Διαφάνεια Συναλλαγών/Κανόνων : Όλες οι συναλλαγές που έχουν εκτελεστεί ποτέ στο δίκτυο είναι δημόσια διαθέσιμες και διαφανείς. Έτσι, οποιοσδήποτε μπορεί να εξετάσει οποιαδήποτε διεύθυνση και να δει τις προηγούμενες συναλλαγές που έχουν εκτελεστεί με αυτήν, το πλήθος των BBitcoin που έχουν μετακινηθεί, όπως και το που έχουν σταλεί. Αυτό ισχύει για όλες τις συναλλαγές που έχουν εκτελεστεί ποτέ στο δίκτυο έως την πρώτη. Το ίδιο ακριβώς ισχύει για όλους τους κανόνες σύμφωνα με τους οποίους δουλεύει το λογισμικό και στο οποίο συναινούν οι χρήστες. Δεν υπάρχει κανένας κρυφός κανόνας μέσα στο λογισμικό, και δεν είναι δυνατόν να υπάρξει, καθώς οι χρήστες δεν θα το αποδέχονταν.

Συναινετική Φύση χρήσης/αλλαγών: Η αλλαγή οιουδήποτε χαρακτηριστικού του λογισμικού ή των κανόνων του, έχει ουσιαστικά εφαρμογή μόνο όταν τις δεχτεί η κοινότητα που απαρτίζει το δίκτυο. Με αυτό τον τρόπο αποφεύγονται κακόβουλες αλλαγές που θα μπορούσαν να αλλάξουν θεμελιωδώς το λογισμικό (καθώς η πλειοψηφία των χρηστών θα τις αναγνωρίσει και δεν θα τις δεχτεί), αλλά και μεγάλη ευελιξία και ταχύτητα αντίδρασης σε περίπτωση εντοπισμού σφαλμάτων ή απρόβλεπτων αστοχιών κατά τη λειτουργία. Η ύπαρξη μιας παγκόσμιας, εξειδικευμένης και δραστήριας κοινότητας, που αντιμετωπίζει με επαγγελματισμό την ποιότητα του λογισμικού ενώ είναι απολύτως ανοιχτή σε σχόλια, εισηγήσεις και κριτική από όλα τα μέρη είναι ανεκτίμητη για την βιωσιμότητα του λογισμικού. Αντίστοιχου βεληνεκούς επιτυχημένα εγχειρήματα ανοιχτού λογισμικού αποτελούν το Linux όπως και το Bit torrent.

Αποκεντρωμένη Φύση: Ένα από τα πιο σημαντικά χαρακτηριστικά του δικτύου, είναι η αποκεντρωμένη φύση του, που δεν απαιτεί καμία κεντρική αρχή ελέγχου ή επιβεβαίωσης. Κάθε κόμβος του δικτύου το ενισχύει περαιτέρω, αλλά αν προσβληθεί με κάποιο τρόπο, η

λειτουργία του συνολικού δικτύου δεν επηρεάζεται ανάλογα. Η προσβολή ακόμα και πολύ μεγάλου μέρους των υπολογιστών που απαρτίζουν το δίκτυο δεν θα επηρέαζε σε σημαντικό βαθμό τη λειτουργία του. Ο μόνος τρόπος να σταματήσει να δουλεύει το δίκτυο είναι να αποκοπούν όλοι οι υπολογιστές του δικτύου μεταξύ τους, με δυο λόγια να κοπεί το διαδίκτυο σε όλο τον πλανήτη, κάτι που είναι πέρα από τις δυνάμεις οποιουδήποτε στην παρούσα. Ακόμα και τότε, με την επαναλειτουργία του διαδικτύου, το δίκτυο συνεχίζει ακριβώς εκεί που σταμάτησε. Ακόμα και μόνο ένας υπολογιστής να παραμείνει συνδεδεμένος που περιέχει το αρχείο της αλυσίδας των προηγούμενων συναλλαγών το δίκτυο λειτουργεί κανονικά.

Υποδιαιρέσεις: Κάθε Bitcoin είναι υποδιαιρέσιμο έως 8 δεκαδικά ψηφία (έως 0,00000001) που ονομάζονται Satoshi, επιτρέποντας μικρο-συναλλαγές που δεν είναι δυνατές με άλλα μέσα ή συμβατικά νομίσματα. Η προσθήκη περισσότερων ακόμα δεκαδικών επαφίεται στην συναίνεση του δικτύου αν αυτό χρειαστεί στο μέλλον.

Μη αντιστρέψιμη φύση: Όλες οι συναλλαγές με Bitcoin είναι τελικές και μη αντιστρέψιμες. Αυτό έχει το επιπλέον πλεονέκτημα προς όσους διαθέτουν προϊόντα για Bitcoin ότι δεν είναι δυνατόν να ανακληθούν συναλλαγές όπως π.χ. είθισται στις απάτες με πιστωτικές κάρτες. Αυτό συνήθως δίνει επιπλέον κίνητρα σε επιχειρήσεις να προσφέρουν τα προϊόντα τους σε χαμηλότερες τιμές, εξαιτίας της άμεσης και αμετάκλητης πληρωμής. Από την άλλη, οι χρήστες που εκτελούν αγορές με Bitcoin πρέπει να είναι προσεκτικοί στις επιλογές τους, καθώς ένας πάροχος προϊόντων ή υπηρεσιών που δεν έχει ιστορικό κινήσεων ή έμπιστη παρουσία στην αγορά μπορεί να μην είναι αυτό που δείχνει.

Ιδιωτικότητα συναλλαγών: Κάθε χρήστης μπορεί να δημιουργήσει, μέσω του λογισμικού, σχεδόν απεριόριστο αριθμό διευθύνσεων μέσω τον οποίων να εκτελέσει τις συναλλαγές του. Αυτές οι διευθύνσεις είναι ψευδώνυμες, δεν έχουν δηλαδή κάποια άμεση σχέση με τα πραγματικά στοιχεία ή την τοποθεσία του χρήστη, παρόλο που έχουν αναγνωρίσιμα χαρακτηριστικά ώστε να εντοπίζονται από το δίκτυο. Με αυτό τον τρόπο μπορεί ο χρήστης να διατηρήσει την ιδιωτικότητά του αποπλέκοντας τις συναλλαγές του από τα προσωπικά του στοιχεία. Αυτό δεν συνεπάγεται εξ' ορισμού ανωνυμία συναλλαγών καθώς όλες οι συναλλαγές δημοσιεύονται, και έστω και μία συναλλαγή να έχει γνωστό (δημόσιο) αποδέκτη, ίσως μπορεί να εξαχθεί από συμπληρωματικά στοιχεία η ταυτότητα του χρήστη. Αυτός είναι και ο κύριος λόγος για τον οποίο η χρήση Bitcoin δεν ενδείκνυται για συναλλαγές παράνομων δραστηριοτήτων, ιδιαίτερα μεγάλης κλίμακας, καθώς το ίχνος των συναλλαγών όχι μόνο δεν διαγράφεται με το πέρασμα του χρόνου, αλλά παραμένει διαθέσιμο για εξέταση από όλους, για πάντα.

Ρίσκα και κίνδυνοι:

Απώλεια ιδιωτικών κλειδιών: Το μόνο που ένας κακόβουλος χρήστης χρειάζεται ώστε να αποκτήσει έλεγχο των Bitcoin του χρήστη, είναι η γνώση των ιδιωτικών κλειδιών του. Παρόλο που το λογισμικό ήδη παρέχει ικανή προστασία για το μέσο χρήστη, χρειάζεται

εγρήγορση στην προστασία απέναντι σε ιούς ή κακόβουλο λογισμικό ή άλλου είδους παραβιάσεις (φυσικές ή ψηφιακές). Με μικρή προσπάθεια στην διαφύλαξη των ιδιωτικών κλειδιών, ακόμα και μη εξειδικευμένοι χρήστες, μπορούν να είναι ασφαλείς σε ικανοποιητικό επίπεδο. Διακύμανση ισοτιμίας: Καθώς τα Bitcoin δεν έχουν κάποια κεντρική αρχή να παρεμβαίνει στις διακυμάνσεις στην προσφορά και την ζήτηση όπως συμβαίνει με π.χ. τα κρατικά νομίσματα, είναι επιρρεπές σε μεγαλύτερες διακυμάνσεις της ισοτιμίας του με τα περισσότερα νομίσματα. Επιπλέον παράγοντας που επηρεάζει το παραπάνω φαινόμενο είναι το σχετικά μικρό βάθος της αγοράς, πράγμα που σημαίνει ότι όταν συναλλάσσονται μεγάλοι όγκοι Bitcoin, επηρεάζουν δυσανάλογα τις ισοτιμίες στα ανταλλακτήρια. Αυτό προβλέπεται ότι θα ελαττωθεί με την πάροδο του χρόνου, εφόσον η οικονομία αναπτυχθεί αρκετά ώστε να μπορούν να εμπλακούν και να αναπτυχθούν κατάλληλες υποδομές που ήδη υφίστανται στις κλασσικές κεφαλαιαγορές. Ένας επιπλέον παράγοντας που επηρεάζει την διακύμανση των ισοτιμιών είναι η φύση των Bitcoin, και ειδικότερα το γεγονός ότι μπορούν να μεταφερθούν ταχύτατα οπουδήποτε στον κόσμο. Αυτό προκαλεί πολύ μεγαλύτερη αμεσότητα στις δράσεις και αντιδράσεις μεταξύ προσφοράς και ζήτησης από ότι με συμβατικές αξίες. Ένας τελευταίος παράγοντας είναι οι κερδοσκοπικές πιέσεις που ασκούνται στα ανταλλακτήρια, καθώς έχουν ακόμα σχετικά μικρή ρευστότητα και όγκο συναλλαγών, όπως συμβαίνει αντίστοιχα και στα μικρά ψηφιακά ανταλλακτήρια συναλλάγματος.

Ασαφές νομικό πλαίσιο: Παρόλο που η ευρωπαϊκή νομοθεσία έχει λάβει μέτρα για τη θέσπιση όρων σε ότι αφορά κεντρικά ελεγχόμενα ή εκδιδόμενα ψηφιακά νομίσματα, η αποκεντρωμένη φύση των Bitcoin, όπως και άλλα από τα χαρακτηριστικά τους, εισάγουν νέες παραμέτρους που δεν έχουν εξεταστεί σε όλο τους το εύρος ακόμα, σε καμία χώρα. Εντός της Ευρώπης, η Γερμανία τα έχει καθορίσει ως "ιδιωτικά χρήματα" (Private money), και η Ολλανδία ως κάτι στο οποίο δεν χρειάζεται η παρέμβαση/έλεγχος της κεντρικής τράπεζας της χώρας. Στην ελληνική έννομη τάξη, σύμφωνα με μια πρόσφατη μελέτη, το Bitcoin θα πρέπει να χαρακτηρισθεί ως χρήμα εν ευρεία έννοια[1]. Κατά συνέπεια, όλες οι συναλλαγές σε Bitcoin θα πρέπει να αντιμετωπίζονται ως χρηματικές[1]. Στις ΗΠΑ, η κύρια επίσημη οδηγία (FINCEN) έως τώρα έγκειται στην προσπάθεια αποφυγής εγκληματικών οικονομικών δραστηριοτήτων, με αμφισβητούμενη ως τώρα επιτυχία, πέρα από την επιβράδυνση και δυσχέρανε των επιχειρηματικών δραστηριοτήτων που σχετίζονται με Bitcoin και αφορούν πελάτες από τις ΗΠΑ. Θεωρείται απίθανο κάποια χώρα να απαγορέψει ολοκληρωτικά τις συναλλαγές με Bitcoin, και κάτι τέτοιο είναι εξαιρετικά δύσκολο (έως αδύνατο) να εφαρμοστεί πρακτικά. Δεν αποκλείεται στην προσπάθεια διερεύνησης του περιβάλλοντος όμως, να ισχύσουν οδηγίες με αναδρομική ισχύ (όπως στην περίπτωση των Η.Π.Α.) που να αλλάζουν το τοπίο, κυρίως επηρεάζοντας τις απαιτήσεις από τις επιχειρήσεις αλλά όχι τόσο τους χρήστες και την ιδιοκτησία τους. Το κύριο σημείο στο οποίο προβλέπεται ότι θα ασκηθεί κρατική επίβλεψη είναι το σημείο ανταλλαγής με τα κρατικά νομίσματα και ειδικότερα ότι έχει σχέση με τα KYC και AML νομικά πλαίσια τοπικά και διεθνώς.

Ασφάλεια δικτύου / Νεαρό ηλικίας: Όπως κάθε σύστημα, το δίκτυο του Bitcoin, έχει αδυναμίες και τρωτά σημεία. Τα περισσότερα από αυτά είναι γνωστά από συγγραφής της αρχικής πρότασης του τρόπου λειτουργίας του λογισμικού, μερικώς προβλέψιμα και

αφήνουν περιθώρια αντίδρασης στο δίκτυο και στους χρήστες. Έως τώρα, όσα έχουν προκύψει, έχουν διορθωθεί εντός λίγων ωρών από την εμφάνισή τους, χωρίς ουσιαστικές επιπτώσεις στη λειτουργία του δικτύου, ή της ισοτιμίας με άλλα νομίσματα. Τα τελευταία 8 χρόνια που είναι σε λειτουργία το δίκτυο, έχουν διασαφηνιστεί ακόμα περισσότερο οι πιθανές επιθέσεις που μπορεί να δεχτεί το δίκτυο. Το νεαρό της ηλικίας του δικτύου όμως και η κλιμάκωσή του σε περισσότερους χρήστες, και οι εξελίξεις του λογισμικού, ενδεχομένως να επιφέρουν νέα προβλήματα που δεν έχουν προβλεφθεί έως τώρα. Κάθε χρήστης μπορεί βέβαια να προτείνει λύσεις σε αυτά, και όποιες εύλογες ανησυχίες εμφανίζονται, εξετάζονται με σοβαρότητα από την κοινότητα, και εις βάθος, ώστε να κριθούν ενδεχόμενες διορθωτικές ενέργειες.

Ενδεικτικά, μερικές από τις σοβαρότερες απειλές αποτελούν τα παρακάτω:

1) Έλεγχος μεγάλου μέρος του δικτύου (<51%) από μία κακόβουλη οντότητα. Αυτό θα έχει ως συνέπεια την αυξημένη πιθανότητα για διπλές συναλλαγές (double spending). Σε κάθε περίπτωση δεν επηρεάζονται οι συναλλαγές που έχουν εκτελεστεί πριν από την «επίθεση», θα θιγεί όμως σημαντικά η συνοχή του δικτύου όπως και η εμπιστοσύνη των χρηστών στην στιβαρότητά του.

2) Παραβίαση των αλγόριθμων κρυπτογράφησης του δικτύου. Αυτό όπως και άλλα πρότυπα που χρησιμοποιούνται στην προστασία και λειτουργία του δικτύου, αποτελούν διεθνώς τυποποιημένα και ευρέως χρησιμοποιούμενα πρωτόκολλα. Στο παρελθόν, οι αδυναμίες αυτών έχουν προκύψει σταδιακά με αρκετό χρόνο πρόνοιας ώστε τα ευαίσθητα συστήματα χωρών, τραπεζών και άλλων οργανισμών να μην προσβληθούν. Στην περίπτωση που συμβεί κάτι τέτοιο έκτακτα, πιθανότατα το δίκτυο του Bitcoin έχει σημαντικότερα μικρότερο χρόνο αντίδρασης από ότι τα περισσότερα άλλα σημεία που χρησιμοποιείται.

3) Αντικατάσταση από κάποιο λογισμικό ανώτερης σχεδίασης και μεγαλύτερου δικτύου από το παρόν, χωρίς αλληλουχία με την παρούσα αλυσίδα συναλλαγών. Το ότι το λογισμικό είναι ανοιχτού κώδικα, σημαίνει ότι μπορεί ο καθένας να προτείνει το δικό του, με ότι αλλαγές προτείνει. Από τη δημιουργία του δικτύου έως αυτή τη στιγμή, έχουν προκύψει περισσότερες από 200 προσπάθειες εναλλακτικών δικτύων που εφαρμόζουν διαφοροποιήσεις στο Bitcoin (γνωστά γενικά ως altcoins), εξαιρετικά μικρότερης αποδοχής και χωρίς ουσιαστικές καινοτομίες στο αρχικό λογισμικό. Εφόσον το λογισμικό μπορεί να ενσωματώσει σε οποιαδήποτε φάση του, οποιαδήποτε καινοτομία προκύψει σε άλλο δίκτυο, είναι απίθανο να αντικατασταθεί, δηλαδή να υπερκεραστεί από κάποιο άλλο σε δυναμική και μέγεθος.

4) Άλλες άγνωστες έως τώρα απειλές. Το ενδεχόμενο να προκύψει κάποια νέα απειλή δεν μπορεί να αποκλειστεί, όπως δεν μπορεί να είναι δεδομένη η ικανότητα του δικτύου να αντιδράσει ή να ανακάμψει.

2.1.4 Πως παράγονται τα Bitcoin και τι είναι η εξόρυξη (mining)

Το Bitcoin επειδή είναι αποκεντρωμένο, χρειάζεται τη συνεισφορά τυχαίων υπολογιστών από όλον τον πλανήτη για να επιβεβαιώσει τις συναλλαγές που γίνονται παγκοσμίως. Αυτή η διαδικασία απαιτεί συνολικά τεράστια υπολογιστική δύναμη. Νέα Bitcoin εκδίδονται κάθε δέκα λεπτά τα οποία δίνονται ως ανταμοιβή σε αυτούς που συνεισφέρουν στην επιβεβαίωση των συναλλαγών, ανάλογα με τη συνεισφορά του καθενός. Αυτοί που επιβεβαιώνουν συναλλαγές ώστε να εισπράξουν κάποια ανταμοιβή ονομάζονται miners και η διαδικασία mining, αντίστοιχα.

Κάθε συναλλαγή που γίνεται με Bitcoin περνάει από έλεγχο εγκυρότητας και έπειτα τοποθετείται σε ένα μπλοκ μαζί με άλλες ολοκληρωμένες συναλλαγές. Κάθε μπλοκ που δημιουργείται έχει άμεση σχέση με το αμέσως προηγούμενο άλλα και με όλα τα υπόλοιπα μπλοκ. Με αυτό τον τρόπο δημιουργείται μια αλυσίδα από μπλοκ (Blockchain). Η σχέση κάθε νέου μπλοκ με τα προηγούμενα καθορίζεται από έναν μαθηματικό αλγόριθμο, ο οποίος όμως είναι δύσκολο να δημιουργηθεί.

Κάθε φορά που δημιουργείται ένα νέο μπλοκ δημιουργείται αυτόματα και ένας αριθμός νέων Bitcoin τα οποία μοιράζονται σε αυτούς που θα έχουν λύσει τον αλγόριθμο ανάλογα με τη συνεισφορά του καθενός, αυτή η διαδικασία ονομάζεται mining.

Όσο μεγαλύτερο ποσοστό της συνολικής υπολογιστικής δύναμης διαθέσει κάποιος για τη λύση του αλγορίθμου τόσο μεγαλύτερο ποσοστό από τα καινούργια Bitcoin που δημιουργούνται θα πάρει. Για παράδειγμα, κάποιος που ασχολείται επαγγελματικά με το Bitcoin mining και έχει πολλά miners (κάρτες γραφικών (GPU) που έχουν τροποποιηθεί ώστε να λύνουν τον αλγόριθμο) θα πάρει μεγαλύτερο κομμάτι "της πίτας" των καινούργιων Bitcoin που βγαίνουν στη κυκλοφορία από κάποιον που έχει μόνο ένα miner. Ο αλγόριθμος δημιουργείται πάντα τόσο δύσκολος να επιλυθεί ώστε όλη η υπολογιστική δύναμη (όλα τα miners του πλανήτη) που επιδίδεται στη λύση του να χρειάζεται κατά μέσο όρο 10 λεπτά για να τον λύσει. Κατά συνέπεια με την πάροδο του χρόνου το σύστημα προσαρμόζει τη λύση του αλγορίθμου και την κάνει όλο και πιο δύσκολη, μιας και παράλληλα αυξάνεται η συνολική υπολογιστική δύναμη που διατίθεται στη λύση του (με τη πρόοδο της τεχνολογίας και με τη κατασκευή καινούργιων miners). Αυτό μας φέρνει και σε ένα ακόμη συμπέρασμα: Όλα τα miners με τη πάροδο του χρόνου παράγουν όλο και πιο λίγες υποδιαιρέσεις του Bitcoin, αφού όσο περνάει ο καιρός αποτελούν ολοένα και μικρότερο ποσοστό της συνολικής υπολογιστικής δύναμης που διατίθεται στη λύση του αλγορίθμου, με αποτέλεσμα έτσι να παίρνουν και πιο μικρό κομμάτι "της πίτας". Όταν ένα miner αναπόφευκτα φτάσει σε σημείο να καταναλώνει περισσότερα χρήματα σε ηλεκτρικό απ' όσα παράγει σε Bitcoin τότε δεν έχει νόημα να λειτουργεί πλέον και πρέπει να αντικατασταθεί.

Ο αριθμός των Bitcoin που δημιουργούνται με κάθε νέο block μειώνεται πολύ ελαφρά κάθε φορά. Μέσα σε 4 χρόνια τα Bitcoin που δημιουργούνται με κάθε νέο block πέφτουν στο μισό. Τα τελευταία Bitcoin θα δημιουργηθούν το 2140. Τότε ο συνολικός αριθμός των Bitcoin που υπάρχουν θα είναι 21 εκατομμύρια.

2.1.5 Κρυπτογραφία και συναλλαγές

Το Bitcoin χρησιμοποιεί ευρέως διαδεδομένη ασύμμετρη κρυπτογράφηση. Κατά τις αρχές της, κάθε χρήστης είναι κάτοχος δύο ψηφιακών κλειδιών, ενός ιδιωτικού και ενός δημοσίου. Η ασύμμετρη κρυπτογράφηση έγκειται στο ότι ο ιδιοκτήτης του ιδιωτικού κλειδιού μπορεί να παράγει με συμβατικά μαθηματικά το δημόσιο κλειδί (μαθηματικά αμφίδρομα), αλλά το αντίθετο είναι εξαιρετικά απίθανο (μαθηματικά μονόδρομα με όλα τα γνωστά μέσα). Η ασύμμετρη κρυπτογράφηση χρησιμοποιείται ευρέως για την ασφαλή ψηφιακή μετάδοση και προστασία των περισσοτέρων υψηλά διαβαθμισμένων πληροφοριών κρατικών και μη οργανισμών, αλλά και διεθνών χρηματικών συναλλαγών, πιστωτικών καρτών, κ.α. Η αποστολή και επιβεβαίωση της συναλλαγής Bitcoin, επιτυγχάνεται μέσω της χρήσης των δημόσιων και ιδιωτικών κλειδιών του αποστολέα και του παραλήπτη.

Παράδειγμα συναλλαγής

Ο χρήστης Α επιθυμεί την αποστολή Χ Bitcoin στον χρήστη Β. Για να είναι έγκυρη αυτή η συναλλαγή κατά τους κανόνες του δικτύου, πρέπει να αποδείξει ότι είναι κάτοχός τους και να υποδείξει στο δίκτυο σε ποιόν χρήστη επιθυμεί να μεταφερθούν. Η απόδειξη της κατοχής από τον χρήστη Α γίνεται με την «υπογραφή» με το ιδιωτικό κλειδί του, και η υπόδειξη της «διεύθυνσης» αποστολής είναι το δημόσιο κλειδί του χρήστη Β. Όλες οι συναλλαγές εκτελούνται άμεσα, και υποβάλλονται στο αποκεντρωμένο δίκτυο για επιβεβαίωση της εγκυρότητάς τους. Οι συναλλαγές ομαδοποιούνται (σε blocks) βάσει των κανόνων του δικτύου, και οι επικρατέστερες (με τη μεγαλύτερη εγκυρότητα) ομάδες τοποθετούνται στη συνέχεια μιας αλυσίδας (Blockchain) που ξεκινάει με την πρώτη συναλλαγή που έγινε το 2009 και φτάνουν έως την πιο πρόσφατη. Το δίκτυο είναι σχεδιασμένο κατά τέτοιο τρόπο ώστε να προκύπτει μια τέτοια ομάδα συναλλαγών στην κορυφή της αλυσίδας περίπου κάθε δέκα λεπτά. Κάθε νέα ομάδα συναλλαγών που τοποθετείται στην κορυφή της αλυσίδας, επιβεβαιώνει όχι μόνο της συναλλαγές που περιέχονται σε αυτή, αλλά και την εγκυρότητα των προηγούμενων ομάδων, και άρα την εγκυρότητα όλων των συναλλαγών που έχουν εκτελεστεί έως την πρώτη. Όλη αυτή η αλυσίδα, όπως και η αλληλουχία όλων των συναλλαγών που έχουν εκτελεστεί έως τώρα, είναι δημοσίως διαθέσιμη και προσβάσιμη από οποιονδήποτε, με την μορφή των δημόσιων κλειδιών που έχουν ανταλλάξει Bitcoin αλλά και των ποσών που έχουν διακινηθεί μεταξύ τους. Έχοντας πλέον, επιβεβαιωμένα την νέα ιδιοκτησία του ποσού Χ, ο χρήστης Β με τη χρήση του ιδιωτικού του κλειδιού, μπορεί κατόπιν να αποστείλει

αντίστοιχα το ποσό σε όποιον χρήστη επιθυμεί (γνωρίζει το δημόσιο κλειδί του). Κάθε χρήστης μπορεί να έχει σχεδόν απεριόριστο αριθμό δημόσιων και αντίστοιχων ιδιωτικών κλειδιών, ασφαλισμένα και υπό τον έλεγχό του (στον υπολογιστή ή στο κινητό του ή και σε πολλές άλλες μορφές). Το σύνολο αυτών αποτελεί ένα είδος ψηφιακού πορτοφολιού του οποίου τα ιδιωτικά κλειδιά πρέπει να μείνουν κρυφά για την αποφυγή απώλειας των περιεχόμενων Bitcoin. Εφόσον το δημόσιο κλειδί παράγεται από το ιδιωτικό, και εφόσον το ιδιωτικό κλειδί είναι το μόνο μέσο που επιτρέπει μεταφορά των Bitcoin εκτός πορτοφολιού, αν ο χρήστης απολέσει ή αποκαλύψει το ιδιωτικό κλειδί του, ουσιαστικά χάνει την αποκλειστική κυριότητα των Bitcoin του.

Η ανάπτυξη «νομισμάτων» με καθαρά ψηφιακή ύπαρξη μελετάται από ερευνητές και οργανισμούς (κυρίως στις Η.Π.Α.) από τα πρώτα βήματα του διαδικτύου (Internet). Ένα από τα κύρια προβλήματα που εμφανίστηκαν στην πορεία είναι το πρόβλημα της «διπλής δαπάνης». Η ψηφιακή πληροφορία αναπαράγεται σχετικά εύκολα, και για την αποτελεσματική λειτουργία ενός τέτοιου συστήματος θα πρέπει να υπάρχουν δικλίδες ασφαλείας ώστε να μην είναι δυνατή η αντιγραφή (πλαστογράφηση) κάθε μονάδας συναλλαγής. Ενώ στα κεντρικά ελεγχόμενα συστήματα (χώρες, εταιρίες, κ.λπ.) ο έλεγχος αυτός μπορεί να γίνει από τις ίδιες, δεν υπήρχε ποτέ έως τώρα ένα μέσο να γίνει αυτό σε ένα αποκεντρωμένο σύστημα (χωρίς κεντρικό έλεγχο).

2.2 Ανταλλακτήρια Κρυπτονομισμάτων

Ένα ανταλλακτήριο κρυπτονομισμάτων, ή αλλιώς ανταλλακτήριο ηλεκτρονικών νομισμάτων, είναι μία επιχείρηση που επιτρέπει στους πελάτες της να ανταλλάσσουν κρυπτονομίσματα ή ηλεκτρονικά νομίσματα για άλλα ηλεκτρονικά νομίσματα ή άλλων αξιών, όπως συναλλάγματος.

2.2.1 Γενική Ιδέα

Τα ανταλλακτήρια κρυπτονομισμάτων μπορεί να είναι επιχειρήσεις με φυσική παρουσία ή επιχειρήσεις οι οποίες λειτουργούν μόνο ηλεκτρονικά. Οι επιχειρήσεις με φυσική παρουσία ανταλλάσσουν παραδοσιακούς τρόπους πληρωμής και ηλεκτρονικά νομίσματα. Οι καθαρά ηλεκτρονικές επιχειρήσεις ανταλλάσσουν χρήματα με ηλεκτρονικές πληρωμές και ηλεκτρονικά νομίσματα. Συχνά οι επιχειρήσεις αυτές λειτουργούν σε χώρες που δεν ανήκουν στον Δυτικό Κόσμο, προκειμένου να αποφύγουν κανονισμούς και διώξεις. Ωστόσο, διαχειρίζονται συνάλλαγμα χωρών του Δυτικού Κόσμου και διατηρούν τραπεζικούς λογαριασμούς σε πολλές χώρες προκειμένου να διευκολύνουν τις καταθέσεις σε συναλλάγματα διαφορετικών χωρών. Τα ανταλλακτήρια δέχονται πληρωμές με

πιστωτική κάρτα, μεταφορά χρημάτων και άλλους τρόπους πληρωμής με αντάλλαγμα ηλεκτρονικά νομίσματα ή κρυπτονομίσματα.

2.2.2 Κανονισμοί

Το 2016 αρκετές επιχειρήσεις που λειτουργούσαν στην Ευρωπαϊκή Ένωση απέκτησαν άδειες λειτουργίας από τις αρχές της Ευρωπαϊκής Ένωσης. Η επάρκεια των αδειών αυτών για τη λειτουργία των ανταλλακτηρίων δεν έχει δοκιμαστεί δικαστικά. Το Ευρωπαϊκό Συμβούλιο και το Ευρωπαϊκό Κοινοβούλιο ανακοίνωσαν ότι θα εκδώσουν ακόμα αυστηρότερους κανονισμούς οι οποίοι θα αφορούν τα ανταλλακτήρια.

Το 2018 η Επιτροπή Κεφαλαιαγοράς των Η.Π.Α. υποστήριξε πως «αν μία πλατφόρμα προσφέρει ανταλλαγή ηλεκτρονικών περιουσιακών στοιχείων που είναι τίτλοι και λειτουργεί ως ανταλλακτήριο, τότε θα πρέπει να εγγραφεί στην Επιτροπή Κεφαλαιαγοράς». Η Επιτροπή συναλλαγών Εμπορευμάτων πλέον επιτρέπει την ανταλλαγή παραγώγων των κρυπτονομισμάτων.

Από τις χώρες της Ασίας, στην Ιαπωνία οι κανονισμοί προβλέπουν την χορήγηση άδειας από την Αρχή Χρηματοπιστωτικών Υπηρεσιών για τη λειτουργία ανταλλακτηρίων κρυπτονομισμάτων. Η Κίνα και η Κορέα παραμένουν εχθρικές απέναντι στα κρυπτονομίσματα. Πιο συγκεκριμένα η Κίνα έχει απαγορεύσει την εξόρυξη κρυπτονομισμάτων και έχει παγώσει τους τραπεζικούς λογαριασμούς αυτών που είχαν ασχοληθεί με την διαδικασία αυτή. Η Αυστραλία, ενώ δεν έχει ανακοινώσει ακόμα τους οριστικούς κανονισμούς σχετικά με τα κρυπτονομίσματα, απαιτεί από τους πολίτες της να αποκαλύπτουν τα ψηφιακά περιουσιακά τους στοιχεία προκειμένου να φορολογηθούν ως κεφαλαιουχικά κέρδη.

2.2.3 Τα μεγαλύτερα ανταλλακτήρια

Στα τέλη του 2017, το Bloomberg ανακοίνωσε τα μεγαλύτερα ανταλλακτήρια κρυπτονομισμάτων με βάση τον όγκο συναλλαγών και τον εκτιμώμενο κύκλο εργασιών βασιζόμενο σε δεδομένα του CoinMarketCap. Παρόμοια στατιστικά ανακοινώθηκαν στο Statista με βάση μία έρευνα, η οποία πραγματοποιήθηκε από την Encrybit, που αφορούσε την κατανόηση των προβλημάτων που παρουσιάζονται κατά τη διαδικασία ανταλλαγής των κρυπτονομισμάτων. Σύμφωνα με την έρευνα τα τρία μεγαλύτερα ανταλλακτήρια κρυπτονομισμάτων είναι τα Binance, Huobi και OKEX. Τα υπόλοιπα στοιχεία της έρευνας αφορούσαν τα προβλήματα που αντιμετωπίζουν οι έμποροι κρυπτονομισμάτων καθώς επίσης και τις προσδοκίες αυτών. Η ασφάλεια και οι υψηλές κρατήσεις ανά συναλλαγή ήταν οι μεγαλύτερες ανησυχίες τους. Τα ανταλλακτήρια αυτά είναι κυρίως νέες επιχειρήσεις που ανήκουν σε ιδιώτες και πολλές από αυτές δεν αναφέρουν βασικά

στοιχεία για την επιχείρηση, όπως τον ιδιοκτήτη, οικονομικά στοιχεία ή ακόμα και την τοποθεσία της επιχείρησης. Το News BTC ανακοίνωσε ότι τα ανταλλακτήρια κρυπτονομισμάτων είναι υπέρ της επιβολής κανονισμών προκειμένου να διασφαλιστεί η λειτουργική ασφάλεια και η σταθερότητα των τιμών.

3 Μηχανική μάθηση

Μηχανική μάθηση είναι υποπεδίο της επιστήμης των υπολογιστών που αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη. Το 1959, ο Άρθουρ Σάμουελ ορίζει τη μηχανική μάθηση ως "Πεδίο μελέτης που δίνει στους υπολογιστές την ικανότητα να μαθαίνουν, χωρίς να έχουν ρητά προγραμματιστεί". Η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα και να κάνουν προβλέψεις σχετικά με αυτά. Τέτοιοι αλγόριθμοι λειτουργούν κατασκευάζοντας μοντέλα από πειραματικά δεδομένα, προκειμένου να κάνουν προβλέψεις βασιζόμενες στα δεδομένα ή να εξάγουν αποφάσεις που εκφράζονται ως το αποτέλεσμα.

Η μηχανική μάθηση είναι στενά συνδεδεμένη και συχνά συγχέεται με υπολογιστική στατιστική, ένας κλάδος, που επίσης επικεντρώνεται στην πρόβλεψη μέσω της χρήσης των υπολογιστών. Έχει ισχυρούς δεσμούς με την μαθηματική βελτιστοποίηση, η οποία παρέχει μεθόδους, τη θεωρία και τομείς εφαρμογής. Η Μηχανική μάθηση εφαρμόζεται σε μια σειρά από υπολογιστικές εργασίες, όπου τόσο ο σχεδιασμός όσο και ο ρητός προγραμματισμός των αλγορίθμων είναι ανέφικτος. Παραδείγματα εφαρμογών αποτελούν τα φίλτρα spam (spam filtering), η οπτική αναγνώριση χαρακτήρων (OCR), οι μηχανές αναζήτησης και η υπολογιστική όραση. Η Μηχανική μάθηση μερικές φορές συγχέεται με την εξόρυξη δεδομένων, όπου η τελευταία επικεντρώνεται περισσότερο στην εξερευνητική ανάλυση των δεδομένων, γνωστή και ως μη επιτηρούμενη μάθηση.

Στο πεδίο της ανάλυσης δεδομένων, η μηχανική μάθηση είναι μια μέθοδος που χρησιμοποιείται για την επινόηση πολύπλοκων μοντέλων και αλγορίθμων που οδηγούν στην πρόβλεψη. Τα αναλυτικά μοντέλα επιτρέπουν στους ερευνητές, τους επιστήμονες δεδομένων, τους μηχανικούς και τους αναλυτές να παράγουν αξιόπιστες αποφάσεις και αποτελέσματα και να αναδείξουν αλληλοσυσχετίσεις μέσω της μάθησης από ιστορικές σχέσεις και τάσεις στα δεδομένα.

3.1 Ορισμός

Ο Tom M. Mitchell πρότεινε έναν πιο επίσημο ορισμό που χρησιμοποιείται ευρέως: «Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από εμπειρία E ως προς μια κλάση εργασιών T και ένα μέτρο επίδοσης P, αν η επίδοσή του σε εργασίες της κλάσης T, όπως αποτιμάται από το μέτρο P, βελτιώνεται με την εμπειρία E» [8]. Αυτός ο ορισμός είναι σημαντικός για τον καθορισμό της μηχανικής μάθησης σε βασικό λειτουργικό πλαίσιο παρά με γνωστικούς όρους, ακολουθώντας έτσι την πρόταση του Alan Turing στην εργασία του «Υπολογιστικές

μηχανές και Νοημοσύνη», ότι το ερώτημα αν μπορούν οι μηχανές να σκεφτούν, μπορεί να αντικατασταθεί με το ερώτημα αν μπορούν οι μηχανές να κάνουν αυτό που εμείς (ως σκεπτόμενες οντότητες) μπορούμε να κάνουμε. [9]

Τύποι προβλημάτων και εργασιών

Οι εργασίες μηχανικής μάθησης συνήθως ταξινομούνται σε τρεις μεγάλες κατηγορίες, ανάλογα με τη φύση του εκπαιδευτικού «σήματος» ή την «ανατροφοδότηση» που είναι διαθέσιμα σε ένα σύστημα εκμάθησης. Αυτές είναι:

Επιτηρούμενη μάθηση ή αλλιώς επιβλεπόμενη μάθηση ή μάθηση με επίβλεψη (supervised learning): Το υπολογιστικό πρόγραμμα δέχεται τις παραδειγματικές εισόδους καθώς και τα επιθυμητά αποτελέσματα από έναν «δάσκαλο», και ο στόχος είναι να μάθει έναν γενικό κανόνα προκειμένου να αντιστοιχίσει τις εισόδους με τα αποτελέσματα.

Μη επιτηρούμενη μάθηση ή αλλιώς επίβλεπτη μάθηση ή μάθηση χωρίς επίβλεψη (unsupervised learning): Χωρίς να παρέχεται κάποια εμπειρία στον αλγόριθμο μάθησης, πρέπει να βρει την δομή των δεδομένων εισόδου. Η Μη επιτηρούμενη μάθηση μπορεί να είναι αυτοσκοπός (ανακαλύπτοντας κρυμμένα μοτίβα σε δεδομένα) ή μέσο για ένα τέλος (χαρακτηριστικό της μάθησης).

Ενισχυτική μάθηση: Ένα πρόγραμμα υπολογιστή αλληλεπιδρά με ένα δυναμικό περιβάλλον στο οποίο πρέπει να επιτευχθεί ένας συγκεκριμένος στόχος (όπως η οδήγηση ενός οχήματος), χωρίς κάποιος δάσκαλος να του λέει ρητά αν έχει φτάσει κοντά στο στόχο του. Ένα άλλο παράδειγμα είναι να μάθει να παίζει ένα παιχνίδι εναντίον κάποιου αντιπάλου.

Μεταξύ της επιτηρούμενης και της μη επιτηρούμενης μάθησης είναι η ημι-επιτηρούμενη μάθηση, όπου ο δάσκαλος δίνει ένα ελλιπές εκπαιδευτικό σήμα: ένα σύνολο εκπαίδευσης με κάποια (συχνά πολλά) από τα αποτελέσματα στόχους να λείπουν. Η Μεταγωγή είναι μια ειδική περίπτωση της αρχής αυτής, όπου το σύνολο των καταστάσεων του προβλήματος είναι γνωστό κατά το χρόνο εκμάθησης, όμως ένα μέρος των στόχων λείπουν.

Μεταξύ άλλων κατηγοριών μηχανικής μάθησης, υπάρχει ακόμα μία διαδικασία εκμάθησης (meta learning) που μαθαίνει στην μηχανή (να αναπτύσσει) τις δικές της επαγωγικές μεθόδους, βασιζόμενο στην προηγούμενη εμπειρία. Η Αναπτυξιακή μάθηση (Developmental robotics), η οποία έχει αναπτυχθεί για την εκμάθηση από ρομπότ, δημιουργεί τη δική της ακολουθία μαθησιακών καταστάσεων, ώστε το ρομπότ συσσωρευτικά αποκτά ποικιλία δεξιοτήτων μέσω της αυτόνομης αυτοεξερεύνησης και της κοινωνικής αλληλεπίδρασης με ανθρώπους εκπαιδευτές και χρησιμοποιώντας μηχανισμούς καθοδήγησης, όπως η ενεργητική μάθηση, η ωρίμανση και η μίμηση.

Μια άλλη κατηγοριοποίηση των προβλημάτων μηχανικής μάθησης προκύπτει όταν κάποιος θεωρήσει το επιθυμητό αποτέλεσμα του συστήματος μηχανικής μάθησης.:

Στην ταξινόμηση, τα δεδομένα εισόδου χωρίζονται σε δύο ή περισσότερες κλάσεις, και η μηχανή πρέπει να κατασκευάσει ένα μοντέλο, το οποίο θα αντιστοιχίζει τα δεδομένα σε μία ή περισσότερες (multi-label ταξινόμηση) κλάσεις. Αυτό συνήθως εμπίπτει στην επιτηρούμενη μάθηση. Τα φίλτρα «Ανεπιθύμητης Αλληλογραφίας» είναι ένα παράδειγμα ταξινόμησης, όπου οι είσοδοι είναι μηνύματα ηλεκτρονικού ταχυδρομείου ή άλλα μηνύματα και οι κλάσεις είναι "Ανεπιθύμητο" και "όχι Ανεπιθύμητο ".

Στην παλινδρόμηση, επίσης πρόβλημα επιτηρούμενης μάθησης, τα αποτελέσματα είναι συνεχή και όχι διακριτά.

Στην συσταδοποίηση, ένα σύνολο εισόδων πρόκειται να χωριστεί σε ομάδες. Σε αντίθεση με την ταξινόμηση, οι ομάδες δεν είναι γνωστές εκ των προτέρων, καθιστώντας αυτόν τον διαχωρισμό τυπική εργασία μη επιτηρούμενης μάθησης.

Στην εκτίμηση πυκνότητας βρίσκει την κατανομή των δεδομένων εισόδου σε κάποιο χώρο.

Σε προβλήματα μείωσης διαστασιμότητας (dimensionality reduction), τα δεδομένα απλοποιούνται και αντιστοιχίζονται σε ένα χώρο λιγότερων διαστάσεων. Το στατιστικό μοντέλο θεμάτων (Topic modeling) είναι ένα σχετικό πρόβλημα, όπου η μηχανή καλείται να βρει έγγραφα που καλύπτουν παρόμοια θέματα από ένα σύνολο εγγράφων γραμμένων σε φυσική γλώσσα.

3.2 Ιστορία και σχέσεις με άλλους τομείς

Ως επιστημονικό εγχείρημα, η μηχανική μάθηση αναπτύχθηκε από την αναζήτηση για την τεχνητή νοημοσύνη. Ήδη από την πρώιμη περίοδο της έρευνας στον τομέα της τεχνητής νοημοσύνης σε ακαδημαϊκό επίπεδο, το ζήτημα της κατασκευής μηχανών που θα μάθαιναν από δεδομένα απασχόλησε τους ερευνητές. Προσπάθησαν να προσεγγίσουν το πρόβλημα με διάφορες συμβολικές μεθόδους, καθώς και με τα λεγόμενα νευρωνικά δίκτυα. Αυτά ήταν ως επί το πλείστον perceptrons και μοντέλα, που όπως διαπιστώθηκε αργότερα ήταν επανεφευρέσεις των γενικευμένων γραμμικών μοντέλων της στατιστικής. Επίσης χρησιμοποιήθηκε η πιθανοθεωρητική λογική, ιδιαίτερα στην αυτοματοποιημένη ιατρική διάγνωση.

Ωστόσο, μια αυξανόμενη έμφαση σε προσεγγίσεις που βασίζονται στην λογική γνώση προκάλεσε ένα ρήγμα μεταξύ Τεχνητής Νοημοσύνης και Μηχανικής μάθησης. Τα πιθανοθεωρητικά συστήματα μαστίζονταν από θεωρητικά και πρακτικά προβλήματα απόκτησης δεδομένων και αναπαράστασής τους. Από το 1980, έμπειρα συστήματα επικράτησαν στο πεδίο της Τεχνητής Νοημοσύνης, και ο ρόλος της στατιστικής υποχώρησε. Η εργασία σε συμβολική/βασισμένη σε γνώση εκμάθηση συνεχίστηκε εντός της Τεχνητής Νοημοσύνης, οδηγώντας στον επαγωγικό λογικό προγραμματισμό, αλλά οι κατευθυντήριες γραμμές της στατιστικής ήταν τώρα έξω από το χώρο της τεχνητής νοημοσύνης, στην αναγνώριση προτύπων και στην ανάκτηση πληροφοριών. Η έρευνα για νευρωνικά δίκτυα εγκαταλείφτηκε από την Τεχνητής Νοημοσύνης και την Επιστήμη Υπολογιστών τον ίδιο

περίπου καιρό. Η ίδια επίσης κατεύθυνση ακολουθήθηκε πέρα από την Τεχνητής Νοημοσύνης και την πληροφορική, από ερευνητές άλλων ειδικοτήτων, συμπεριλαμβανομένων των Hopfield, Rumelhart και Hinton. Η επιτυχία ήρθε στα μέσα της δεκαετίας του 1980 με την επανεφεύρεση της μεθόδου ανάστροφης μετάδοσης (backpropagation).

Η Μηχανική μάθηση, αναδιοργανώθηκε ως ένα ξεχωριστό πεδίο, που άρχισε να ακμάζει κατά τη δεκαετία του 1990. Η προσοχή μετατοπίστηκε από τις συμβολικές προσεγγίσεις που κληρονόμησε από την Τεχνητή Νοημοσύνη, που στόχο είχαν την αντιμετώπιση επιλύσιμων προβλημάτων πρακτικής φύσης, και δόθηκε έμφαση σε μεθόδους και μοντέλα της στατιστικής και της θεωρίας πιθανοτήτων. Επίσης επωφελήθηκε από την διαθεσιμότητα ψηφιοποιημένων πληροφοριών και της δυνατότητας να διανεμηθούν μέσω του Διαδικτύου.

Η Μηχανική μάθηση και η εξόρυξη δεδομένων συχνά χρησιμοποιούν τις ίδιες μεθόδους και επικαλύπτονται σημαντικά. Μπορούν να διακριθούν ως εξής:

Η μηχανική μάθηση εστιάζει στην πρόβλεψη, που βασίζεται σε γνωστές ιδιότητες που απορρέουν από το σύνολο εκπαίδευσης.

Η εξόρυξη δεδομένων εστιάζει στην ανακάλυψη ιδιοτήτων μη γνωστών εκ των προτέρων. Αυτό είναι το βήμα ανάλυσης στην Ανακάλυψη Γνώσης από βάσεις δεδομένων.

Οι δύο τομείς επικαλύπτονται με πολλούς τρόπους. Η εξόρυξη δεδομένων χρησιμοποιεί πολλές μεθόδους μηχανικής μάθησης, αλλά συχνά με διαφορετικούς στόχους. Από την άλλη πλευρά και η μηχανική μάθηση χρησιμοποιεί μεθόδους εξόρυξης δεδομένων, όπως η μη επιτηρούμενη μάθηση, ή στο στάδιο προεπεξεργασίας για να βελτιώνει την ακρίβεια της μάθησης. Ένα μεγάλο μέρος της σύγχυσης μεταξύ των δύο ερευνητικών τομέων (που συχνά έχουν ξεχωριστά συνέδρια και περιοδικά, με το ECML PKDD να αποτελεί σημαντική εξαίρεση) προκύπτει από τις βασικές υπόθεσεις πάνω στις οποίες και οι δύο δουλεύουν. Όμως, στην μηχανική μάθηση η απόδοση συνήθως αξιολογείται ως προς την ικανότητα αναπαραγωγής γνώσης, την οποία ήδη κατέχουμε, ενώ στην ανακάλυψη γνώσης και την εξόρυξη δεδομένων το κλειδί είναι η ανακάλυψη γνώσης που δεν προκατέχουμε. Στην πρώτη περίπτωση μια μέθοδος επιτηρούμενης μάθησης μπορεί να έχει καλύτερα αποτελέσματα, ενώ σε μία τυπική διεργασία Ανακάλυψης Γνώσης και Εξόρυξης δεδομένων οι επιτηρούμενες μέθοδοι μάθησης δεν λειτουργούν εξαιτίας της μη διαθεσιμότητας συνόλου εκπαίδευσης.

Η μηχανική μάθηση συνδέεται επίσης με την βελτιστοποίηση: πολλά προβλήματα μάθησης διατυπώνονται ως η ελαχιστοποίηση της συνάρτησης απώλειας από ένα σύνολο δεδομένων εκπαίδευσης. Η συνάρτηση απώλειας εκφράζει τη διαφορά μεταξύ των προβλέψεων του εκπαιδευμένου μοντέλου και των πραγματικών καταστάσεων του προβλήματος. Η διαφορά των δύο τομέων απορρέει από τον στόχο της γενίκευσης: ενώ οι αλγόριθμοι βελτιστοποίησης μπορούν να ελαχιστοποιήσουν την απώλεια ενός συνόλου εκπαίδευσης, η μηχανική μάθηση εστιάζει στην ελαχιστοποίηση της απώλειας σε άγνωστες καταστάσεις.

Σχέση με την στατιστική

Η μηχανική μάθηση και η στατιστική είναι δύο στενά συνδεδεμένοι επιστημονικοί τομείς. Σύμφωνα με τον Michael Jordan, οι ιδέες της μηχανικής μάθησης, από τις μεθοδολογικές αρχές μέχρι τα θεωρητικά εργαλεία, προϋπάρχουν στην στατιστική. Ο ίδιος επίσης πρότεινε τον όρο Επιστήμη Δεδομένων για το συνολικό πεδίο. [10]

Ο Leo Breiman διέκρινε δύο υποδείγματα στατιστικής μοντελοποίησης: το μοντέλο δεδομένων και το αλγοριθμικό μοντέλο. Το αλγοριθμικό μοντέλο ταυτίζεται σχεδόν με αλγορίθμους μηχανικής μάθησης όπως τα Τυχαία Δάση.

Τέλος, ορισμένοι στατιστικολόγοι υιοθετούν μεθόδους μηχανικής μάθησης, με αποτέλεσμα την δημιουργία ενός ανασυνδυασμένου τομέα που ονομάζεται στατιστική μάθηση.

3.3 Θεωρία

Ο βασικός στόχος ενός μαθητευόμενου είναι να γενικεύει την εμπειρία του. Σε αυτό το πλαίσιο, γενίκευση είναι η ικανότητα μιας μηχανής μάθησης να αποδίδει με ακρίβεια σε καινούριες, πρωτόγνωρες εργασίες, αφού πρώτα έχει εκπαιδευτεί σε ένα σύνολο δεδομένων εκπαίδευσης. Γενικά τα προς εκπαίδευση παραδείγματα προέρχονται από κάποια άγνωστη κατανομή πιθανότητας, η οποία θεωρείται αντιπροσωπευτική του χώρου των καταστάσεων, και η μηχανή πρέπει να κατασκευάσει ένα γενικό μοντέλο που θα επιτρέπει την παραγωγή προβλέψεων σε καινούριες καταστάσεις με επαρκή ακρίβεια.

Η υπολογιστική ανάλυση των αλγορίθμων των μηχανών μάθησης και η απόδοσή τους είναι ένας κλάδος της θεωρητικής πληροφορικής, γνωστός ως Υπολογιστική θεωρία μάθησης. Επειδή τα εκπαιδευτικά σύνολα είναι πεπερασμένα και το μέλλον αβέβαιο, η θεωρία μάθησης δεν εγγυάται πάντα την απόδοση των αλγορίθμων. Αντ΄αυτού είναι συχνή η χρήση των πιθανοθεωρητικών ορίων της απόδοσης.

Το πόσο καλά ένα μοντέλο, που έχει εκπαιδευτεί σε υπαρκτά παραδείγματα, μπορεί να προβλέψει άγνωστες καταστάσεις ονομάζεται γενίκευση. Για την καλύτερη δυνατή γενίκευση, η πολυπλοκότητα της υπόθεσης θα πρέπει να είναι αντίστοιχη της πολυπλοκότητας της συνάρτησης των δεδομένων.

Πέρα όμως από την απόδοση, οι θεωρητικοί της υπολογιστικής μάθησης μελετούν την χρονική πολυπλοκότητα καθώς και το κατά πόσο είναι εφικτή η μάθηση. Στην υπολογιστική θεωρία μάθησης ένας υπολογισμός θεωρείται εφικτός αν μπορεί να επιτελεστεί σε πολυωνυμικό χρόνο. Υπάρχουν δύο είδη αποτελεσμάτων αναφορικά με την χρονική πολυπλοκότητα. Τα θετικά αποτελέσματα που σημαίνουν ότι μια συγκεκριμένη κλάση αντιστοιχίσεων μπορούν να επιτευχθούν σε πολυωνυμικό χρόνο και τα αρνητικά αποτελέσματα που δείχνουν το αντίθετο.

Υπάρχουν πολλές ομοιότητες μεταξύ της μηχανικής μάθησης και της στατιστικής συμπερασματολογίας, αν και χρησιμοποιούν διαφορετικούς όρους.

3.4 Προσεγγίσεις

Εκμάθηση με δέντρο απόφασης

Η εκμάθηση με δέντρο απόφασης χρησιμοποιεί ένα δέντρο απόφασης ως προγνωστικό μοντέλο, το οποίο αντιστοιχίζει παρατηρήσεις σχετικά με ένα στοιχείο σε συμπεράσματα σχετικά με την τιμή στόχο του αντικειμένου.

Εκμάθηση με Κανόνες συσχέτισης

Η εκμάθηση με κανόνες συσχέτισης είναι μια μέθοδος ανακάλυψης ενδιαφερουσών σχέσεων μεταξύ των μεταβλητών σε μεγάλες βάσεις δεδομένων.

Τεχνητά νευρωνικά δίκτυα

Ένας αλγόριθμος εκμάθησης Τεχνητού νευρωνικού δικτύου, που συνήθως ονομάζεται "νευρωνικό δίκτυο" (NN), είναι ένας αλγόριθμος μάθησης, που εμπνέεται από τη δομή και τις λειτουργικές πτυχές των βιολογικών νευρωνικών δικτύων. Η δομή των υπολογισμών βασίζεται σε μια ομάδα εσωτερικά διασυνδεδεμένων τεχνητών νευρώνων, οι οποίοι επεξεργάζονται την πληροφορία και εκτελούν υπολογισμούς επικοινωνώντας μεταξύ τους. Τα σύγχρονα νευρωνικά δίκτυα είναι εργαλεία μη γραμμικής στατιστικής μοντελοποίησης δεδομένων. Συνήθως χρησιμοποιούνται για τη μοντελοποίηση σύνθετων σχέσεων μεταξύ δεδομένων εισόδου και εξόδου, για την ανακάλυψη προτύπων στα δεδομένα, ή για τον εντοπισμό στατιστικής δομής σε μία άγνωστη κοινή κατανομή πιθανότητας μεταξύ των παρατηρούμενων μεταβλητών.

Βαθιά Μάθηση

Η πτώση των τιμών του υλικού των τελευταίων ετών καθώς και η ανάπτυξη των καρτών γραφικών (GPU) για προσωπική χρήση, οδήγησε στην ανάπτυξη της ιδέας της Βαθιάς Μάθησης. Αυτή η προσέγγιση προσπαθεί να μοντελοποιήσει τον τρόπο που ο ανθρώπινος εγκέφαλος επεξεργάζεται το φως και τον ήχο και τα μετατρέπει σε όραση και ακοή. Ορισμένες επιτυχείς εφαρμογές της Βαθιάς μάθησης είναι η μηχανική όραση και η αναγνώριση ομιλίας.

Επαγωγικός λογικός προγραμματισμός

Ο Επαγωγικός λογικός προγραμματισμός (ILP) είναι μια προσέγγιση που διέπει την μάθηση και χρησιμοποιεί λογικό προγραμματισμό ως τρόπο παρουσίασης των παραδειγμάτων εισόδου, του γνωστικού υποβάθρου και των υποθέσεων. Δεδομένης μιας κωδικοποίησης του γνωστικού υποβάθρου και ενός συνόλου παραδειγμάτων που παρουσιάζονται σαν λογική βάση γεγονότων, το σύστημα ΕΛΠ παράγει το υποτιθέμενο λογικό πρόγραμμα που περιέχει όλα τα θετικά και κανένα αρνητικό παράδειγμα. Ο επαγωγικός προγραμματισμός είναι ένας σχετικός τομέας που λαμβάνει υπόψιν κάθε είδος προγραμματιστικής γλώσσας για την αναπαράσταση υποθέσεων (και όχι μόνο λογικό προγραμματισμό), όπως τα συναρτησιακά προγράμματα.

Μηχανές διανυσμάτων υποστήριξης

Οι μηχανές διανυσμάτων υποστήριξης είναι ένα σύνολο μεθόδων επιτηρούμενης μάθησης που χρησιμοποιούνται για την ταξινόμηση και την παλινδρόμηση. Σ' αυτήν την περίπτωση δίνεται ένα σύνολο παραδειγμάτων εκπαίδευσης και κάθε φορά δηλώνεται σε ποια από τις δύο κατηγορίες ανήκει το παράδειγμα. Μία μηχανή διανυσμάτων υποστήριξης κατασκευάζει ένα μοντέλο που προβλέπει αν το νέο παράδειγμα εμπίπτει στην μία κατηγορία ή την άλλη.

Ομαδοποίηση

Η ομαδοποίηση είναι η διαδικασία κατά την οποία ένα σύνολο παρατηρήσεων χωρίζεται σε υποσύνολα έτσι ώστε οι παρατηρήσεις που ανήκουν στην ίδια ομάδα (cluster) είναι όμοιες, σύμφωνα με κάποιο ή κάποια προκαθορισμένα κριτήρια, ενώ οι παρατηρήσεις που προέρχονται από διαφορετικά υποσύνολα είναι ανόμοιες. Διαφορετικές τεχνικές κατηγοριοποίησης οδηγούν σε διαφορετικές υποθέσεις σχετικά με τη δομή των δεδομένων, οι οποίες συχνά καθορίζονται από κάποιο μέτρο ομοιότητας και αξιολογούνται για παράδειγμα ως προς την εσωτερική συνοχή (ομοιότητα μεταξύ των μελών του ίδιου cluster) και το διαχωρισμό ανάμεσα σε διαφορετικές ομάδες. Άλλες μέθοδοι βασίζονται στην εκτιμώμενη πυκνότητα και την συνεκτικότητα των γραφημάτων. Η ομαδοποίηση είναι μία μέθοδος μη επιτηρούμενης μάθησης και μία τεχνική η οποία χρησιμοποιείται επίσης στην στατιστική ανάλυση δεδομένων.

Δίκτυα Bayes

Ένα δίκτυο Bayes, ένα δίκτυο εμπιστοσύνης ή ένα άκυκλο γραφικό μοντέλο είναι ένα πιθανοθεωρητικό γραφικό μοντέλο που απεικονίζει ένα σύνολο τυχαίων μεταβλητών και

την μεταξύ τους υποθετική ανεξαρτησία διαμέσου ενός κατευθυνόμενου άκυκλου γράφου. Για παράδειγμα, ένα δίκτυο Bayes μπορεί να αναπαραστήσει την πιθανοθεωρητική σχέση μεταξύ ασθενειών και συμπτωμάτων. Δεδομένων των συμπτωμάτων, το δίκτυο μπορεί να χρησιμοποιηθεί για να υπολογίσει τις πιθανότητες παρουσίας διαφόρων ασθενειών.

Ενισχυτική μάθηση

Η Ενισχυτική μάθηση ασχολείται με το πώς ένα υποκείμενο (πράκτορας) θα πρέπει να δράσει σε ένα περιβάλλον, έτσι ώστε να μεγιστοποιηθεί κάποια έννοια μακροπρόθεσμης ανταμοιβής. Οι αλγόριθμοι ενισχυτικής μάθησης προσπαθούν να βρουν μια πολιτική που αντιστοιχίζει τις καταστάσεις του περιβάλλοντος με τις ενέργειες που ο πράκτορας θα πρέπει να επιτελέσει σε αυτές τις καταστάσεις. Η ενισχυτική μάθηση διαφέρει από τα προβλήματα επιτηρούμενης μάθησης αφού τα σωστά ζεύγη δεδομένων εισόδου/εξόδου δεν παρουσιάστηκαν ποτέ, ούτε οι βέλτιστες δυνατές ενέργειες έχουν ρητά διορθωθεί.

Εκμάθηση με μέτρο ομοιότητας

Σε αυτή την κατηγορία προβλημάτων δίνονται στην μηχανή μάθησης ζεύγη παραδειγμάτων που θεωρούνται όμοια και ζεύγη που θεωρούνται ανόμοια. Τότε η μηχανή μάθησης πρέπει να μάθει μια συνάρτηση ομοιότητας (ή μια συνάρτηση μετρικής απόστασης), που μπορεί να προβλέψει αν δύο καινούρια αντικείμενα είναι όμοια. Πρόκειται για μια τεχνική που χρησιμοποιείται σε συστήματα σύστασης.

Γενετικοί αλγόριθμοι

Ένας γενετικός αλγόριθμος (GA) είναι μια ευρετική αναζήτηση που μιμείται τη διαδικασία της φυσικής επιλογής, και χρησιμοποιεί μεθόδους όπως αυτή της μετάλλαξης και της διασταύρωσης προκειμένου να δημιουργήσει καινούρια γονότυπα με την ελπίδα εύρεσης αποτελεσματικών λύσεων σε ένα συγκεκριμένο πρόβλημα. Στη μηχανική μάθηση, γενετικοί αλγόριθμοι χρησιμοποιήθηκαν τη δεκαετία του 1980 και του 1990. Αντίστροφα, τεχνικές μηχανικής μάθησης έχουν χρησιμοποιηθεί για την βελτίωση της απόδοσης γενετικών και εξελικτικών αλγορίθμων.

3.5 Ηθική

Η Μηχανική Μάθηση, θέτει μια σειρά από ηθικά ζητήματα. Τα συστήματα τα οποία έχουν εκπαιδευτεί σε σύνολα δεδομένων που συλλέγονται με προκαταλήψεις μπορεί να εμφανίζουν αυτές τις προκαταλήψεις κατά τη χρήση, ψηφιοποιώντας πολιτιστικές προκαταλήψεις όπως ο θεσμικός ρατσισμός και ο ταξικός διαχωρισμός. Έτσι η υπεύθυνη συλλογή δεδομένων είναι ένα κρίσιμο κομμάτι της μηχανικής μάθησης.

4 Κοινωνικά Δίκτυα

Τα κοινωνικά δίκτυα είναι ένα σύνολο αλληλεπιδράσεων και διαπροσωπικών σχέσεων. Ο όρος σήμερα χρησιμοποιείται επίσης για να περιγράψει ιστοσελίδες οι οποίες επιτρέπουν την διεπαφή ανάμεσα στους χρήστες, πχ. με σχόλια, φωτογραφίες, άλλες πληροφορίες από σχετική βιβλιογραφία. Οι πιο γνωστές από αυτές τις ιστοσελίδες είναι το Facebook, Twitter, Instagram και Linkedin.

Οι ιστότοποι αυτοί αποτελούν εικονικές κοινότητες όπου οι χρήστες μπορούν να επικοινωνούν και να αναπτύσσουν επαφές μέσα από αυτές.

Ένα κοινωνικό δίκτυο είναι μια κοινωνική δομή που αποτελείται από ένα σύνολο παραγόντων, όπως άτομα ή οργανισμούς. Στο διαδίκτυο, τα κοινωνικά δίκτυα είναι μία πλατφόρμα που συντηρείται για την δημιουργία κοινωνικών σχέσεων μεταξύ των ανθρώπων, που συνήθως αποτελούν ενεργά μέλη του κοινωνικού δικτύου, με κοινά ενδιαφέροντα ή δραστηριότητες.

Οι ιστότοποι κοινωνικής δικτύωσης είναι οργανωμένες ιστοσελίδες στο διαδίκτυο με περισσότερο ομαδοκεντρικό χαρακτήρα που παρέχουν, στην συντριπτική τους πλειοψηφία, μία σειρά από βασικές και δωρεάν υπηρεσίες όπως τη δημιουργία προφίλ, το ανέβασμα εικόνων και βίντεο, τον σχολιασμό σε ενέργειες που γίνονται από άλλα μέλη του δικτύου ή μίας ομάδας, την άμεση ανταλλαγή μηνυμάτων και πολλά άλλα.

4.1 Ιστορία

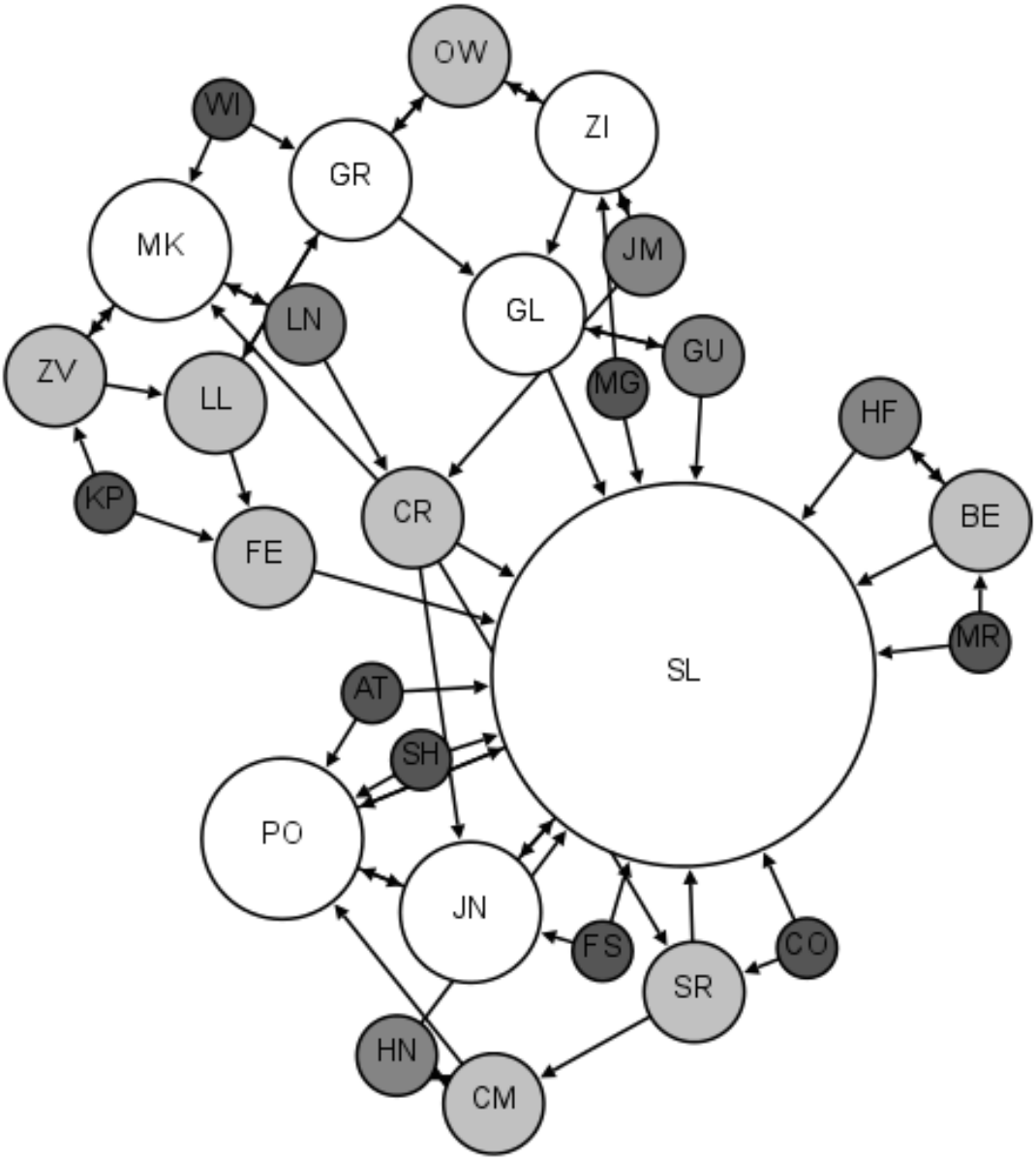
Στα τέλη της δεκαετίας του 1890, τόσο ο Émile Durkheim όσο και ο Ferdinand Tönnies προωθούσαν την ιδέα των κοινωνικών δικτύων στις θεωρίες τους και την έρευνα των κοινωνικών ομάδων. Ο Tönnies ισχυρίστηκε ότι οι κοινωνικές ομάδες μπορούν να υπάρχουν ως προσωπικοί και άμεσοι κοινωνικοί δεσμοί που είτε συνδέουν άτομα που μοιράζονται ίδιες αξίες και πεποιθήσεις (Gemeinschaft, Γερμανικά, συνήθως μεταφρασμένα ως "κοινότητα") [11] είτε απρόσωπες, επίσημες και οργανικές κοινωνικές σχέσεις (Gesellschaft, ως "κοινωνία"). Ο Durkheim έδωσε μια μη-εξατομικευμένη εξήγηση των κοινωνικών γεγονότων υποστηρίζοντας ότι τα κοινωνικά φαινόμενα προκύπτουν όταν οι αλληλεπιδράσεις μεταξύ ατόμων αποτελούν μια πραγματικότητα που δεν μπορεί πλέον να ληφθεί υπόψη από την άποψη των ιδιοτήτων των μεμονωμένων παραγόντων . Ο Georg Simmel, που γράφει στη στροφή του εικοστού αιώνα, επεσήμανε τη φύση των δικτύων και

την επίδραση του μεγέθους του δικτύου στην αλληλεπίδραση και εξέτασε την πιθανότητα αλληλεπίδρασης σε χαλαρά πλεγμένα δίκτυα και όχι σε ομάδες.

Σημαντικές εξελίξεις στον τομέα παρατηρήθηκαν κατά την δεκαετία του 1930 από ερευνητές ψυχολογίας, ανθρωπολογίας και μαθηματικών, οι οποίοι εργάζονταν ανεξάρτητα μεταξύ τους [12]. Στην ψυχολογία, στη δεκαετία του 1930, ο Jacob L. Moreno ξεκίνησε συστηματική καταγραφή και ανάλυση της κοινωνικής αλληλεπίδρασης σε μικρές ομάδες, ιδιαίτερα σε τάξεις διδασκαλίας και σε ομάδες εργασίας (βλέπε κοινωνιομετρία). Στην ανθρωπολογία, το θεμέλιο για τη θεωρία των κοινωνικών δικτύων είναι το θεωρητικό και εθνογραφικό έργο του Bronislaw Malinowski, Alfred Radcliffe-Brown και Claude Lévi-Strauss. Μια ομάδα κοινωνικών ανθρωπολόγων που συνδέονται με τον Max Gluckman και τη σχολή του Μάντσεστερ, συμπεριλαμβανομένου του John A. Barnes, J. Clyde Mitchell και Elizabeth Bott Spillius, θεωρούνται οι πρώτοι οι οποίοι εκπόνησαν εργασίες από τις οποίες πραγματοποιήθηκαν αναλύσεις δικτύων, διερευνώντας δίκτυα κοινοτήτων στη Νότιο Αφρική, την Ινδία και το Ηνωμένο Βασίλειο. Συγχρόνως, ο βρετανικός ανθρωπολόγος S.F. Nadel κωδικοποίησε μια θεωρία της κοινωνικής δομής, κάτι που επηρέασε την ανάλυση δικτύων. Στην κοινωνιολογία, το έργο του Talcott Parsons που έγινε στις αρχές της δεκαετίας του '30 έβαλε τις βάσεις για τη λήψη μιας σχεσιακής προσέγγισης στην κατανόηση της κοινωνικής δομής. Αργότερα, με βάση τη θεωρία του Parsons, ο κοινωνιολόγος Peter Blau παρείχε μια ισχυρή ώθηση για την ανάλυση των σχεσιακών δεσμών των κοινωνικών μονάδων με το έργο του για τη θεωρία των κοινωνικών ανταλλαγών.

Μέχρι την δεκαετία του 1970, ένας μεγάλος αριθμός μελετητών εργάστηκε για να συνδυάσει διαφορετικές διαδρομές και παραδόσεις. Μια ομάδα αποτελούνταν από κοινωνιολόγους και μαθητές του τμήματος Κοινωνικών σχέσεων του Πανεπιστημίου του Χάρβαρντ. Επίσης, ανεξάρτητα ενεργός στο Τμήμα Κοινωνικών Σχέσεων του Χάρβαρντ την εποχή εκείνη ήταν ο Charles Tilly, ο οποίος επικεντρώθηκε στα δίκτυα της πολιτικής και κοινοτικής κοινωνιολογίας και στα κοινωνιολογικά κινήματα, και του Stanley Milgram, οποίος ανέπτυξε την εργασία “έξι βαθμών χωρισμού”. Ο Mark Granovetter και Barry Wellman είναι μεταξύ των πρώην φοιτητών του White που επεξεργάστηκαν και προώθησαν την ανάλυση των κοινωνικών δικτύων [13].

Στα τέλη της δεκαετίας του 1990, η ανάλυση κοινωνικών δικτύων γνώρισε μεγάλη αναγνώριση από ομάδες κοινωνιολόγων, πολιτικών επιστημόνων και φυσικών όπως ο Duncan J. Watts, Albert-László Barabási, Peter Bearman, Nicholas A. Christakis, James H. Fowler, και άλλους, αναπτύσσοντας και χρησιμοποιώντας νέα μοντέλα και μεθόδους για αναδείξουν τα διαθέσιμα δεδομένα σχετικά με τα κοινωνικά δίκτυα, καθώς και "ψηφιακά ίχνη" σχετικά με τα δίκτυα τύπου "πρόσωπο με πρόσωπο" (face-to-face).



Σχήμα 9: Κοινωνιόγραμμα 2ης τάξης του Moreno

4.2 Twitter

Το Twitter είναι μία Αμερικάνικη διαδικτυακή πλατφόρμα που παρέχει υπηρεσίες ενημέρωσης και κοινωνικής δικτύωσης στην οποία οι χρήστες της μπορούν να αναρτούν και να αλληλεπιδρούν με μηνύματα τα οποία είναι γνωστά ως “tweets”. Μόνο οι εγγεγραμμένοι χρήστες επιτρέπεται να αναρτούν tweets, ενώ οι μη εγγεγραμμένοι μπορούν μόνο να διαβάζουν “tweets”. Η πρόσβαση των χρηστών στην πλατφόρμα μπορεί να γίνει μέσω της ιστοσελίδας της εταιρίας, μέσω μηνυμάτων SMS ή μέσω της εφαρμογής για φορητές συσκευές. Το Twitter έχει έδρα στην Καλιφόρνια του Σαν Φρανσίσκο και έχει περισσότερα από 25 γραφεία ανά τον κόσμο. [14]

Η ίδρυση της πλατφόρμας πραγματοποιήθηκε από τους Jack Dorsey, Noah Glass, Biz Stone, και Evan Williams το 2016 και η λειτουργία της ξεκίνησε τον Ιούλιο του ίδιου έτους. Η υπηρεσία άρχισε να διαδίδεται και να χρησιμοποιείται με πολύ γρήγορο ρυθμό σε παγκόσμιο επίπεδο. Το 2012 περισσότεροι από 100 εκατομμύρια χρήστες αναρτούσαν 340 εκατομμύρια “tweets” την ημέρα και η πλατφόρμα διαχειριζόταν πάνω από 1,6 δισεκατομμύρια αναζητήσεις ημερησίως. Το 2013 ήταν μία από τις 10 πιο δημοφιλής ιστοσελίδες και το περιέγραφαν ως το «SMS του διαδικτύου». Μέχρι το 2016 το Twitter είχε περισσότερους από 319 εκατομμύρια ενεργούς χρήστες. Την ημέρα των Αμερικάνικων προεδρικών εκλογών το 2016, αποδείχθηκε ότι το Twitter είναι το μεγαλύτερο μέσω μετάδοσης έκτακτων ειδήσεων, με περισσότερα από 40 εκατομμύρια “tweets”, που αφορούσαν τις εκλογές, να έχουν αναρτηθεί μέχρι τις 10 μμ.

4.2.1 Χρήση

Το Μάρτιο του 2018 το Twitter βρισκόταν στην δωδέκατη θέση ανάμεσα στις ιστοσελίδες με την υψηλότερη επισκεψιμότητα, όπως αξιολογήθηκε από την πλατφόρμα της Alexa. Οι καθημερινοί χρήστες του εκτιμάται ότι ποικίλουν, καθώς η εταιρία δεν ανακοινώνει πληροφορίες και στατιστικά σχετικά με τους ενεργούς χρήστες της. Το Φεβρουάριο του 2009 το blog Compete.com κατέταξε το Twitter στην $3^{η}$ θέση ως το μέσω κοινωνικής δικτύωσης που χρησιμοποιούταν περισσότερο, βασιζόμενο στην μέτρηση 6 εκατομμυρίων μοναδικών χρηστών και 55 εκατομμυρίων επισκέψεων. Η πλατφόρμα γνώρισε αύξηση της τάξης του 1.382%, από 475000 μοναδικούς επισκέπτες το Φεβρουάριο του 2008 σε 7 εκατομμύρια το Φεβρουάριο του 2009. Ο ετήσιος ρυθμός ανάπτυξης του Twitter μειώθηκε από 7,8% το 2015 σε 3,4% το 2017. Τον Απρίλιο του 2017 το blog statista.com αξιολόγησε το Twitter ως το $10^{ο}$ πιο πολυχρησιμοποιημένο μέσο κοινωνικής δικτύωσης, βασιζόμενο σε στοιχεία από 319 εκατομμυρίων μηνιαίων χρηστών. Η παγκόσμια βάση χρήσης του Twitter το 2017 ήταν 328 εκατομμύρια χρήστες.

Δημογραφικά Στοιχεία

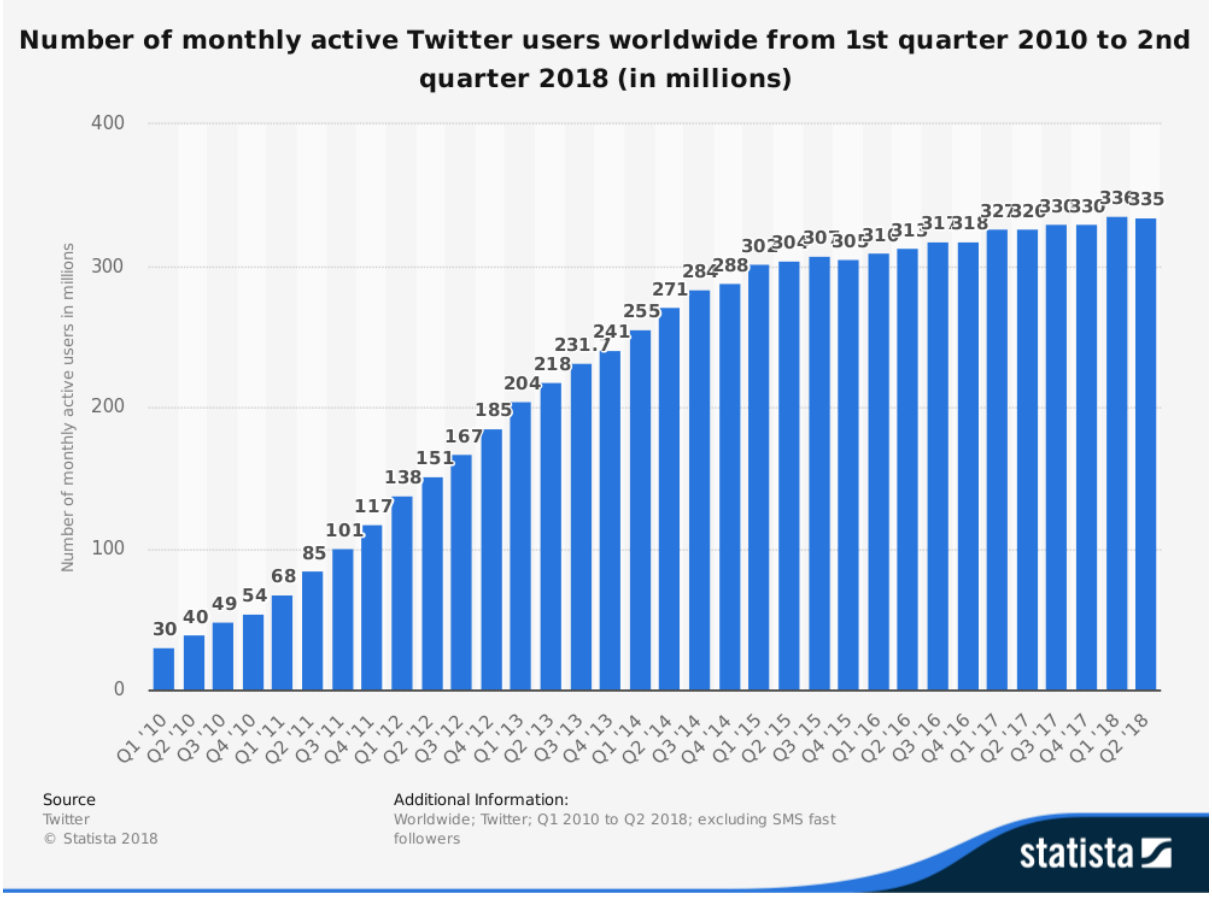
Το 2009 το Twitter χρησιμοποιούταν κυρίως από άτομα μεγαλύτερης ηλικίας, που ίσως να μην είχαν κάποια προηγούμενη εμπειρία από άλλα μέσα κοινωνικής δικτύωσης. Με βάση τα στοιχεία του comScore, μόνο το 11% των χρηστών της πλατφόρμας είναι μεταξύ 12 και 17 ετών. Το comScore απέδωσε αυτό το φαινόμενο στην «πρώιμη περίοδο προσαρμογής» του Twitter, όταν η πλατφόρμα έγινε γνωστή κυρίως για τα επιχειρηματικά νέα, όπου προσέλκυσε άτομα μεγαλύτερης ηλικίας.

Βάσει των στοιχείων του Sysomos τον Ιούνιο του 2009 το ποσοστό των γυναικών που είχαν εγγραφεί στην πλατφόρμα ήταν ελάχιστα μεγαλύτερο από αυτό των αντρών. Πιο συγκεκριμένα, το ποσοστό των γυναικών ήταν 53% και αυτό των αντρών 47%. Επίσης το Sysomos δήλωσε ότι η πόλη της Νέας Υόρκης ήταν εκείνη με τους περισσότερους χρήστες, συγκριτικά με όλες τις υπόλοιπες.

Στις 7 Σεπτεμβρίου του 2011, το Twitter ανακοίνωσε ότι είχε 100 εκατομμύρια ενεργούς χρήστες, οι οποίοι συνδέονταν τουλάχιστον μία φορά το μήνα στην πλατφόρμα, και 50 εκατομμύρια χρήστες οι οποίοι χρησιμοποιούσαν την πλατφόρμα καθημερινά.

Σε άρθρο που δημοσιεύτηκε στις 6 Ιανουαρίου του 2012, επιβεβαιώθηκε ότι το Twitter ήταν το μεγαλύτερο μέσο κοινωνικής δικτύωσης στην Ιαπωνία και το Facebook βρισκόταν στην $2^{η}$ θέση.

Το Μάρτιο του 2014 το Twitter ανακοίνωσε ότι είχε 255 εκατομμύρια ενεργούς χρήστες και 198 εκατομμύρια ενεργούς χρήστες οι οποίοι χρησιμοποιούσαν την πλατφόρμα μέσω του κινητού τους. Το 2013 υπήρχαν παραπάνω από 100 εκατομμύρια καθημερινά ενεργοί χρήστες οι οποίοι αναρτούσαν περίπου 500 εκατομμύρια Tweets κάθε μέρα.



Σχήμα 10: Ενεργοί χρήστες ανά μήνα στο Twitter

4.2.2 Σελίδες Οικονομικών Νέων

Το Twitter, όπως αναφέραμε και προηγουμένως, χρησιμοποιείται και ως μέσο ενημέρωσης από πολλούς χρήστες του, καθώς οι μεγαλύτερες εταιρίες ειδήσεων φροντίζουν να ενημερώνουν τους ακόλουθους τους μέσω της συγκεκριμένης πλατφόρμας. Για το λόγο αυτό το Twitter, όπως και κάθε άλλο μέσω μαζικής ενημέρωσης, μπορεί να επηρεάσει τις αγορές, είτε αυτές είναι τα χρηματιστήρια μετοχών, οι αγορές ομολόγων και εμπορευμάτων ή οι τιμές των κρυπτονομισμάτων, όπως και μελετάμε σε αυτήν την εργασία. Για την υλοποίηση της διπλωματικής εργασίας, χρησιμοποιήσαμε Tweets που ανάρτησαν μεγάλες δημοσιογραφικές επιχειρήσεις στις σελίδες τους στο Twitter. Πιο συγκεκριμένα ανακτήσαμε δεδομένα από τις παρακάτω επιχειρήσεις / λογαριασμούς:

Bloomberg https://twitter.com/business, https://twitter.com/BW

Το Bloomberg είναι μία ιδιωτική επιχείρηση η οποία εξειδικεύεται στην παροχή δεδομένων και ενημερώσεων σε χρηματοπιστωτικές εταιρείες και οργανισμούς. Ιδρύθηκε τον Οκτώβριο του 1981 από τους Michael Bloomberg, Thomas Secunda, Duncan MacMillan και Charles Zegar. Η επιχείρηση απασχολεί 19.000 εργαζόμενους σε 192 γραφεία. Τα κεντρικά γραφεία της βρίσκονται στην πόλη της Νέας Υόρκης. Ο τζίρος της επιχείρησης το 2014 ανήλθε στα 9 δισεκατομμύρια δολάρια.

The Guardian https://twitter.com/guardian, https://twitter.com/guardiannews

Η The Guardian είναι μία Βρετανική εφημερίδα που ασχολείται με νέα γενικού περιεχομένου. Ιδρύθηκε το 1821, από τον John Edward Taylor, υπό την ονομασία The Manchester Guardian και μετονομάστηκε The Guardian το 1959. Τα κεντρικά γραφεία της επιχείρησης βρίσκονται στο Λονδίνο. Η ηλεκτρονική της εφημερίδα είχε περισσότερους από 42,6 εκατομμύρια αναγνώστες τον Οκτώβριο του 2014.

The New York Times https://twitter.com/nytimes, https://twitter.com/nytimesbusiness

Η The New York Times είναι μία Αμερικάνικη εφημερίδα, με έδρα την Νέα Υόρκη, η οποία έχει παγκόσμια επιρροή και αναγνώστες σε όλο τον κόσμο. Ιδρύθηκε το 1851 από τους Henry Jarvis Raymond και George Jones και έχει κερδίσει 125 βραβεία Pulitzer. Η συγκεκριμένη εφημερίδα βρίσκεται στην $2^{η}$ θέση στις Ηνωμένες Πολιτείες της Αμερικής, βάσει των πωλήσεων της, και στην $17^{η}$ παγκοσμίως.

Washington Post https://twitter.com/washingtonpost

Η Washington Post είναι ημερήσια Αμερικάνικη εφημερίδα, η οποία ιδρύθηκε το Δεκέμβριο του 1877 από τον Stilson Hutchins. Είναι η μεγαλύτερη εφημερίδα της πολιτείας Ουάσιγκτον και δίνει έμφαση σε πολιτικά νέα και αποφάσεις.

HuffPost https://twitter.com/HuffPost

Η HuffPost είναι Αμερικάνικη εφημερίδα και blog και έχει πολλές τοπικές και διεθνείς εκδόσεις. Ιδρύθηκε το 2015 από τους Arianna Huffington, Kenneth Lerer, Jonah Peretti και Andrew Breitbart. Η θεματολογία της εφημερίδας ποικίλει και οι βασικότερες κατηγορίες είναι: Νέα, κάλυψη πολιτικών γεγονότων, επιχειρηματικά νέα, τεχνολογία, τοπικά νέα κτλ.

The Associated Press https://twitter.com/AP

Η The Associated Press είναι ένα Αμερικάνικο πρακτορείο ειδήσεων με έδρα τη Νέα Υόρκη. Ιδρύθηκε το 1846 και παρέχει υπηρεσίες μετάδοσης ειδήσεων. Έχει κερδίσει 52 βραβεία Poulitzer. Απασχολεί 3200 εργαζόμενους και ο τζίρος της το 2015 ανήλθε στα 568,13 εκατομμύρια δολάρια.

The Telegraph https://twitter.com/Telegraph

Η The Telegraph είναι μία Βρετανική εφημερίδα που εκδίδεται στο Λονδίνο. Ιδρύθηκε το 1855 από τον Arthur B. Sleigh. Είναι γνωστή για την δημοσιοποίηση και την αποκάλυψη μεγάλων πολιτικών και οικονομικών σκανδάλων.

TIME Inc. https://twitter.com/TIME

Η Time είναι μία διεθνής επιχείρηση μαζικής ενημέρωσης που ιδρύθηκε το 1922 από τους Henry Luce και Briton Hadden και έχει βάση την πόλη της Νέας Υόρκης. Κατέχει και διαχειρίζεται περισσότερα από 100 περιοδικά και ο τζίρος της το 2015 ανήλθε στα 3,1 δισεκατομμύρια δολάρια. Απασχολεί 7.200 υπαλλήλους.

CNN https://twitter.com/CNN, https://twitter.com/cnni, https://twitter.com/CNNMoney, https://twitter.com/cnnbrk

Το CNN είναι ένα Αμερικάνικο κανάλι ειδήσεων το οποίο ιδρύθηκε το 1980 από τον Ted Turner. Τα κεντρικά γραφεία της επιχείρησης βρίσκονται στην Ατλάντα.

Business Insider https://twitter.com/businessinsider

Η Business Insider είναι μία Αμερικάνικη οικονομική ηλεκτρονική εφημερίδα η οποία έχει διαφορετικές εκδόσεις για πολλές χώρες και εκδίδεται σε πολλές γλώσσες. Ιδρύθηκε το 2009 από τον Kevin P. Ryan και τα νέα της αφορούν οικονομικά νέα αλλά και νέα επιχειρήσεων.

Reuters Business https://twitter.com/ReutersBiz

Το Reuters είναι ένα διεθνές πρακτορείο ειδήσεων το οποίο έχει έδρα το Λονδίνο. Ιδρύθηκε το 1851 από τον Paul Reuter. Η σελίδα Reuters Business επικεντρώνεται κυρίως σε επιχειρηματικά και οικονομικά νέα.

BBC News https://twitter.com/BBCBusiness, https://twitter.com/BBCBreaking, https://twitter.com/BBCNews

Το BBC News είναι ένα τμήμα του Βρετανικού καναλιού BBC, το οποίο ασχολείται με τη συλλογή και αναμετάδοση σημαντικών ειδήσεων. Η έδρα του βρίσκεται στο Λονδίνο και απασχολεί 3500 εργαζόμενους.

The Economist https://twitter.com/TheEconomist

Το The Economist είναι ένα εβδομαδιαίο περιοδικό με έδρα το Λονδίνο. Ιδρύθηκε το 1843 από τον James Wilson. Εκδίδει περισσότερα από 1.5 εκατομμύριο αντίτυπα.

Financial Times https://twitter.com/FT, https://twitter.com/FinancialTimes

Οι Financial Times είναι μια εφημερίδα που δίνει έμφαση στα νέα και τις αποφάσεις των επιχειρήσεων καθώς επίσης και σε γενικότερα οικονομικά νέα. Ιδρύθηκε το 1888.

Fortune Magazine https://twitter.com/FortuneMagazine

Το Fortune Magazine είναι ένα Αμερικάνικο περιοδικό το οποίο έχει έδρα την πόλη της Νέας Υόρκης και ανταγωνίζεται το Bloomberg Businessweek και το Forbes. Ιδρύθηκε το 1929 και εκδίδει νέα επιχειρήσεων και κατατάξεις σχετικά με την οικονομική δύναμη και επιρροή τους.

The Independent https://twitter.com/Independent

Η The Independent είναι μία Βρετανική ηλεκτρονική εφημερίδα η οποία ιδρύθηκε το 1986 στο Λονδίνο.

Άλλες σελίδες από τις οποίες ανακτήσαμε Tweets είναι οι: Wall Street Journal, MarketWatch, Bitcoin, blockchain, coindesk, cointelegraph, Bitcoin News, Bitcoin Magazine, Entrepreneur και Crypto Coins News

Η παρουσία των επιχειρήσεων και σελίδων αυτών στο Twitter αποτυπώνεται αναλυτικά στον παρακάτω πίνακα:

Όνομα	Tweets	Ακόλουθοι	Σελίδα
The Associated Press	225.000	12.900.000	https://twitter.com/AP
BBC Breaking News	34.900	38.300.000	https://twitter.com/BBCBreaking
BBC Business	107.000	1.900.000	https://twitter.com/BBCBusiness
BBC News	364.000	9.520.000	https://twitter.com/BBCNews
Bitcoin	20.900	903.000	https://twitter.com/Bitcoin
Bitcoin Magazine	4.093	531.000	https://twitter.com/BitcoinMagazine
Blockchain	7.877	702.000	https://twitter.com/blockchain
Bitcoin News	10.100	363.000	https://twitter.com/btctn
Bloomberg	381.000	4.900.000	https://twitter.com/business
Business Insider	525.000	2.470.000	https://twitter.com/businessinsider
Businessweek	66.300	1.580.000	https://twitter.com/BW
CNN	204.000	40.400.000	https://twitter.com/CNN
CNN Breaking News	63.000	54.300.000	https://twitter.com/cnnbrk
CNN International	168.000	7.910.000	https://twitter.com/cnni
CNNMoney	144.000	1.730.000	https://twitter.com/CNNMoney
CoinDesk	52.300	762.000	https://twitter.com/coindesk
Cointelegraph	17.900	424.000	https://twitter.com/cointelegraph
CCN	24.600	193.000	https://twitter.com/CryptoCoinsNews
Entrepreneur	142.000	3.390.000	https://twitter.com/Entrepreneur
Financial Times	238.000	6.070.000	https://twitter.com/FinancialTimes
FORTUNE	178.000	2.250.000	https://twitter.com/FortuneMagazine
Financial Times	215.000	3.300.000	https://twitter.com/FT
The Guardian	470.000	7.300.000	https://twitter.com/guardian
Guardian news	181.000	2.830.000	https://twitter.com/guardiannews
HuffPost	513.000	11.400.000	https://twitter.com/HuffPost
The Independent	757.000	2.560.000	https://twitter.com/Independent
MarketWatch	241.000	3.590.000	https://twitter.com/MarketWatch
The New York Times	334.000	42.000.000	https://twitter.com/nytimes
NYT Business	186.000	778.000	https://twitter.com/nytimesbusiness
Reuters Business	169.000	1.980.000	https://twitter.com/ReutersBiz
The Telegraph	329.000	2.520.000	https://twitter.com/Telegraph
The Economist	153.000	23.400.000	https://twitter.com/TheEconomist
TIME	299.000	15.400.000	https://twitter.com/TIME
Washington Post	285.000	12.800.000	https://twitter.com/washingtonpost
WSJ Business News	67.300	1.470.000	https://twitter.com/WSJbusiness
WSJ Markets	57.600	489.000	https://twitter.com/WSJmarkets

Πίνακας 1: Σελίδες επιχειρήσεων από τις οποίες ανακτήσαμε δεδομένα

5 Εργαλεία και τεχνολογίες

Στο κεφάλαιο αυτό παρουσιάζονται οι κυριότερες τεχνολογίες που χρησιμοποιήθηκαν για την εκπόνηση των αλγορίθμων, καθώς επίσης και για την συλλογή των απαραίτητων δεδομένων από το Twitter και τα ανταλλακτήρια. Συγκεκριμένα παρουσιάζεται η σύγχρονη γλώσσα προγραμματισμού Python καθώς και διάφορες βιβλιοθήκες, όπως το Tweepy και το scikit-learn, τα οποία χρησιμοποιήθηκαν για να ολοκληρωθούν με επιτυχία όλα τα στάδια του πειράματος της πρόβλεψης των γεγονότων αλλά και των τιμών των κρυπτονομισμάτων. Κατόπιν, παρουσιάζονται τα API που χρησιμοποιήθηκαν από το Twitter για την συλλογή των Tweets καθώς επίσης και το API του Cryptocompare.com, μιας διαδικτυακής πλατφόρμας η οποία περιέχει ενημερώσεις για τις τιμές των κρυπτονομισμάτων και ιστορικά δεδομένα για αυτές. Τέλος, για την αποθήκευση των δεδομένων και των αποτελεσμάτων χρησιμοποιήσαμε μία SQL βάση δεδομένων της Microsoft.

5.1 Η γλώσσα προγραμματισμού Python

Η Python είναι μια αντικειμενοστραφής διερμηνευόμενη γλώσσα γενικού σκοπού με δυναμική σημασιολογία (semantics) και δημιουργήθηκε από τον Guido van Rossum και κυκλοφόρησε δημοσίως το 1991. Αποτελεί γλώσσα υψηλού επιπέδου και η δημιουργία της βασίστηκε στην εύκολη αναγνωσιμότητα του κώδικα ενώ ταυτόχρονα διακρίνεται για την πλούσια εκφραστικότητα της.

Έχει μια μεγάλη κύρια βιβλιοθήκη η οποία καλύπτει ένα τεράστιο εύρος πεδίων, κάνοντας την ιδανική για χρήση σε οποιαδήποτε σχεδόν εφαρμογή. Συμβάλλει στην εξαιρετικά γρήγορη ανάπτυξη εφαρμογών αλλά σε καμία περίπτωση δεν υστερεί σε λειτουργίες που προσφέρουν οι υπόλοιπες εξίσου διαδεδομένες γλώσσες προγραμματισμού όπως η C ή η Java. Επιπροσθέτως, υποστηρίζει πολύ υψηλού επιπέδου δομές δεδομένων, παρέχει αυτόματη διαχείριση μνήμης και είναι φιλική σε οποιοδήποτε λειτουργικό σύστημα. [15]

Σχήμα 11: Λογότυπο της Python

5.1.1 Κύρια Χαρακτηριστικά της Python

Όπως αναφέραμε και προηγουμένως η Python αποτελεί αντικειμενοστραφή γλώσσα προγραμματισμού. Επιπροσθέτως ένα άλλο πολύ σημαντικό χαρακτηριστικό της είναι ότι συνδυάζει πολλαπλά πρότυπα προγραμματισμού, συμπεριλαμβανομένων του προστακτικού και του συναρτησιακού προγραμματισμού. Εν συνεχεία οι υψηλού επιπέδου δομές δεδομένων που προσφέρει, οι δυναμικοί τύποι κωδικοποίησης, η αυτόματη διαχείριση μνήμης καθώς επίσης και μια μεγάλη πρότυπη βιβλιοθήκη την καθιστούν μια από τις πιο διαδεδομένες γλώσσες προγραμματισμού. Επίσης, ένα άλλο αξιοσημείωτο χαρακτηριστικό της Python είναι ότι προάγει την εύκολη συντήρηση και επαναχρησιμοποίηση του πηγαίου κώδικα, αφού δίνει την δυνατότητα της ομαδοποίησης του σε μονάδες και πακέτα. Τέλος, σημαντικά γνωρίσματα της είναι και τα παρακάτω:

- Υποστηρίζει τις εξαιρέσεις.
- Δυνατότητα ενσωμάτωσης σε μια εφαρμογή ώστε να λειτουργεί σαν RESTful υπηρεσία διαδικτύου.
- Συμβατότητα με όλες τις κύριες πλατφόρμες υλικού και λογισμικού. - Χρησιμοποιεί διερμηνέα και είναι scripting language.
- Δημιουργία μικρότερων σε μέγεθος προγραμμάτων σε σχέση με άλλες γλώσσες προγραμματισμού. - Πληθώρα IDEs : IDLE, Ipython, PythonAnywhere (online) κτλ

5.1.2 Η βιβλιοθήκη scikit-learn για την Python

Το scikit-learn είναι ένα δωρεάν βιβλιοθήκη μηχανικής μάθησης για την Python. Υποστηρίζει αλγορίθμους classification, regression και clustering συμπεριλαμβανομένων των αλγορίθμων support vector machines, random forests, gradient boosting, k-means και DBSCAN. Έχει σχεδιαστεί να λειτουργεί με τις βιβλιοθήκες της Python NumPy και SciPy. [16]

Σχήμα 12: Λογότυπο του scikit-learn

5.1.3 Η βιβλιοθήκη Tweepy για την Python

Η βιβλιοθήκη Tweepy, είναι μία υλοποίηση ανοιχτού κώδικα η οποία δίνει την δυνατότητα στην Python να επικοινωνεί με το Twitter μέσω της χρήσης του API του. Επιτρέπει την πρόσβαση στο Twitter μέσω της OAuth πιστοποίησης που υποστηρίζει το Twitter. [17]

5.2 Microsoft SQL Server

Ο Microsoft SQL Server είναι ένα σύστημα διαχείρισης σχεσιακών βάσεων δεδομένων το οποίο έχει αναπτύξει η Microsoft. Όπως και τα υπόλοιπα αντίστοιχα συστήματα έτσι και σε αυτό κύρια λειτουργία του είναι η αποθήκευση και ανάκτηση δεδομένων από άλλες εφαρμογές, οι οποίες μπορεί να εκτελούνται στον ίδιο υπολογιστή ή σε άλλον κάποιον άλλον στο ίδιο δίκτυο.

Σχήμα 13: Λογότυπο του Microsoft SQL Server

5.3 Twitter API

Το Twitter προσφέρει, σε προγραμματιστές, εταιρείες και οργανισμούς, πρόσβαση στα δεδομένα της πλατφόρμας της μέσω του API της. Το συγκεκριμένο API προσφέρει ένα μεγάλο εύρος από ενότητες από τις οποίες μπορεί κάποιος να αντλήσει δεδομένα αλλά και να καλέσει συναρτήσεις και διαδικασίες προκειμένου να εκτελέσει κάποιες λειτουργίες. Οι πέντε βασικές ενότητες είναι οι εξής: [18]

- Λογαριασμούς και Χρήστες
- Tweets και απαντήσεις
- Απευθείας μηνύματα
- Διαφημίσεις
- Εργαλεία για εκδότες και προγραμματιστές

5.4 Cryptocompare API

Το API που προσφέρει ο ιστότοπος Cryptocompare.com είναι από τα καλύτερα δωρεάν για την συλλογή των τιμών των κρυπτονομισμάτων, τόσο σε πραγματικό χρόνο όσο και σε ιστορικά δεδομένα. Η συγκεκριμένη πλατφόρμα συλλέγει δεδομένα από 90 διαφορετικά ανταλλακτήρια ανά τον κόσμο για περισσότερα από 1800 κρυπτονομίσματα.

Σχήμα 14: Λογότυπο του Cryptocompare

6 Σχεδιασμός και υλοποίηση Πειράματος

Στο κεφάλαιο αυτό θα παρουσιάσουμε τις λεπτομέρειες που αφορούν τον σχεδιασμό και την υλοποίηση του συστήματος και των αλγορίθμων που αναπτύχθηκαν προκειμένου να επιτύχουμε τον στόχο μας, ο οποίος είναι αρχικώς να εντοπίσουμε σημαντικά γεγονότα που αφορούν το Bitcoin και αφετέρου να προβλέψουμε την μεταβλητότητα της τιμής του κρυπτονομίσματος αυτού. Για την αποτελεσματικότερη διαχείριση των δεδομένων μας δημιουργήσαμε μία βάση δεδομένων, στον Microsoft SQL Server 2016, στην οποία αποθηκεύαμε και ανακτούσαμε δεδομένα που χρειαζόμασταν σε κάθε στάδιο.

6.1 Μακροσκοπική Αρχιτεκτονική Συστήματος

Η υλοποίηση του συστήματος μας χωρίστηκε σε 5 βασικά στάδια, σε καθένα από τα οποία αντιμετωπίσαμε διαφορετικές δυσκολίες και προκλήσεις. Παρακάτω θα περιγράψουμε την γενικότερη αρχιτεκτονική και την διασύνδεση των σταδίων αυτών προκειμένου να οδηγηθούμε στο επιθυμητό αποτέλεσμα.

Ο σχεδιασμός περιλαμβάνει τα παρακάτω στάδια:

1. **Συλλογή απαραίτητων δεδομένων**
   1.1. Σύνολο δεδομένων (Dataset) με τα IDs των Tweets που αφορούν το Bitcoin
   1.2. Υλοποίηση εφαρμογής για την συλλογή των Tweets με χρήση του Twitter API
   1.3. Υλοποίηση εφαρμογής για την συλλογή των τιμών του Bitcoin από την πλατφόρμα Cryptocompare.com με χρήση του API της

2. **Ομαδοποίηση των Tweets βάσει των σημαντικότερων γεγονότων που αφορούσαν το Bitcoin (Annotate Dataset)**

3. **Προ-επεξεργασία των δεδομένων**
   3.1. Απλοποίηση των Tweets, που ανακτήσαμε στο στάδιο 1.1 και 1.2, σε μορφή που να είναι φιλική για τους αλγορίθμους της μηχανικής μάθησης, προκειμένου να εξάγουν ασφαλέστερα, εγκυρότερα και ταχύτερα συμπεράσματα
   3.2. Υπολογισμός ομοιότητας μεταξύ των Tweets. Για να μπορέσουμε να ομαδοποιήσουμε τα Tweets με τη βοήθεια των αλγορίθμων της μηχανικής μάθησης, πρέπει να υπολογίσουμε την ομοιότητα μεταξύ των Tweets βάσει της χρονικής και νοηματικής απόστασής τους.

3.3. Υπολογισμός της μεταβλητότητας των τιμών του Bitcoin με βάση τις τιμές που ανακτήσαμε στο στάδιο 1.3

4. Υλοποίηση αλγορίθμων μηχανικής μάθησης

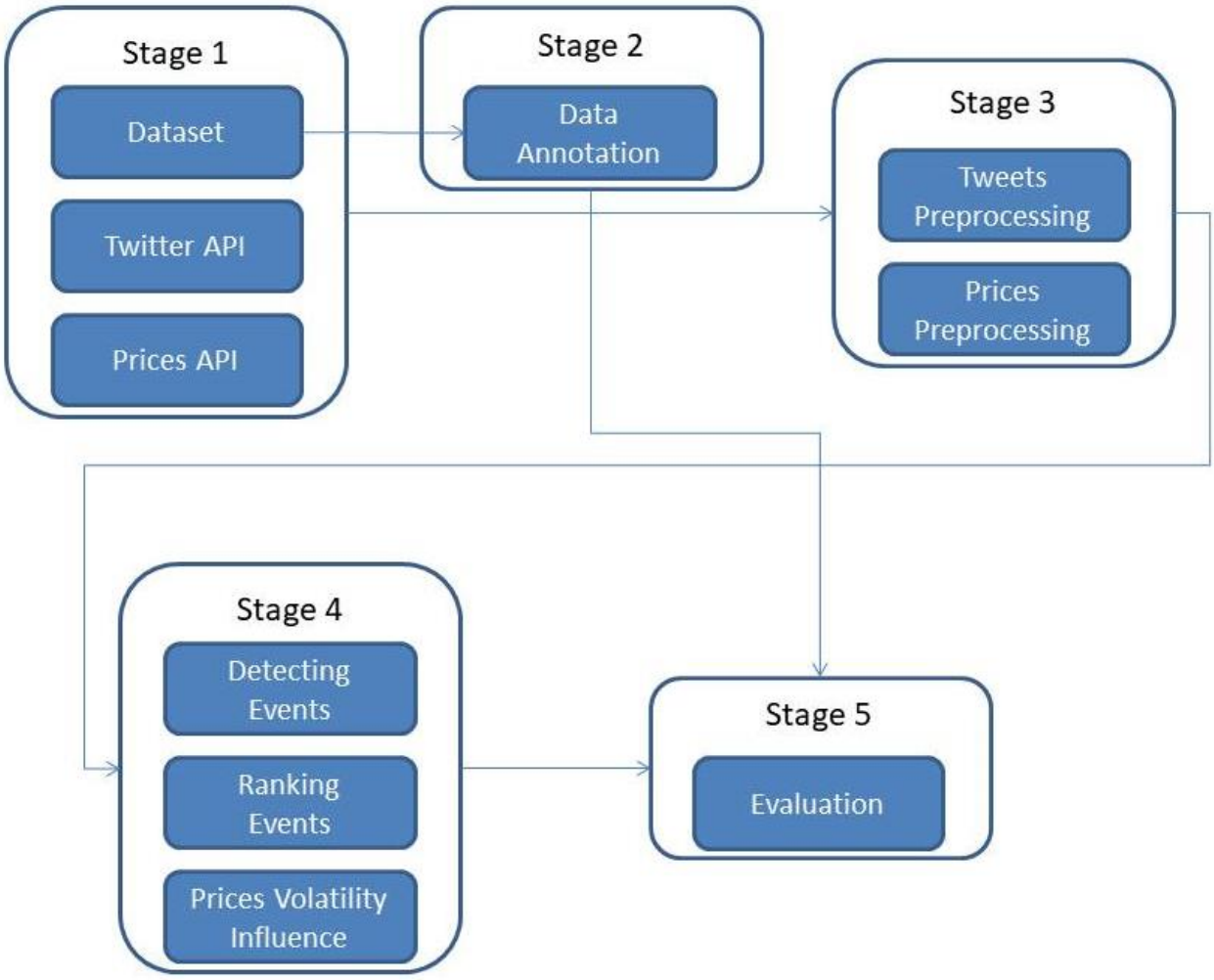
4.1. Εντοπισμός σημαντικότερων γεγονότων

4.2. Κατάταξη γεγονότων βάσει της σημαντικότητας τους

4.3. Μελέτη της επιρροής των γεγονότων αυτών στην τιμή του Bitcoin

5. Αξιολόγηση Αλγορίθμων

Στο παρακάτω σχήμα απεικονίζεται η Μακροσκοπική αρχιτεκτονική του συστήματος



Σχήμα 15: Μακροσκοπική Αρχιτεκτονική Συστήματος

6.2 Υλοποίηση Συστήματος

Στο κεφάλαιο αυτό θα αναλύσουμε εκτενέστερα τις λειτουργίες που υλοποιήσαμε για την αποπεράτωση και αξιολόγηση του συστήματος μας. Αναλυτικότερα, θα εστιάσουμε στα

στάδια και στις ενέργειες που πραγματοποιήσαμε σε αυτά προκειμένου να επιτύχουμε το επιθυμητό αποτέλεσμα.

6.2.1 Συλλογή Δεδομένων

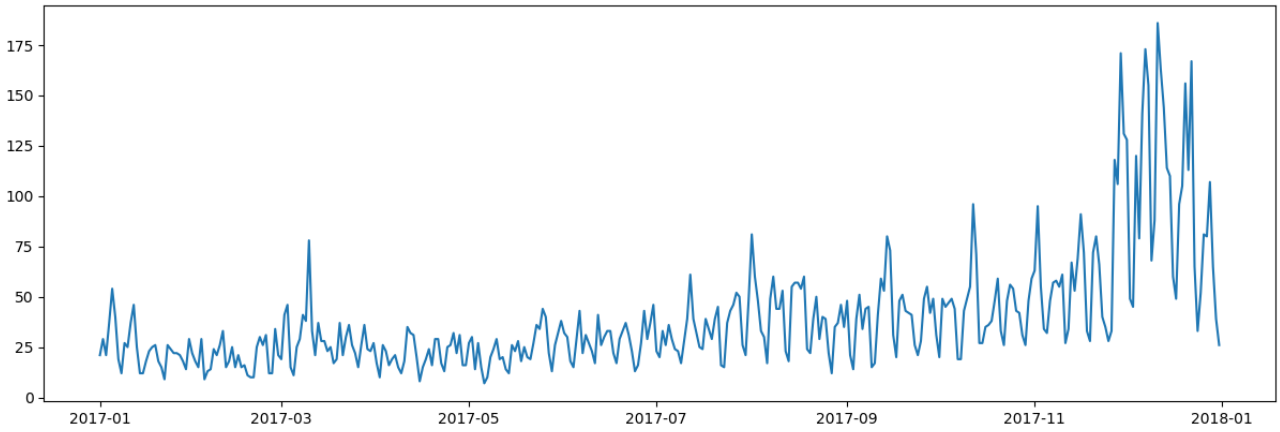
Στο πρώτο στάδιο, όπως φαίνεται και στο σχήμα 6.1, ασχοληθήκαμε με τα δεδομένα που θα χρειαζόμασταν για να πραγματοποιήσουμε το πείραμα του εντοπισμού των γεγονότων και της πρόβλεψης της τιμής του Bitcoin. Για την ολοκλήρωση του σταδίου αυτού χρειάστηκαν τρία διαφορετικά βήματα, όπως αναφέραμε και στην μακροσκοπική αρχιτεκτονική του συστήματος, τα οποία είναι:

1. Επιλογή Συλλογής Δεδομένων (Dataset)
2. Σύνδεση με το Twitter API και συλλογή των Tweets
3. Σύνδεση με το Cryptocompare API και συλλογή ιστορικών δεδομένων της τιμής του Bitcoin

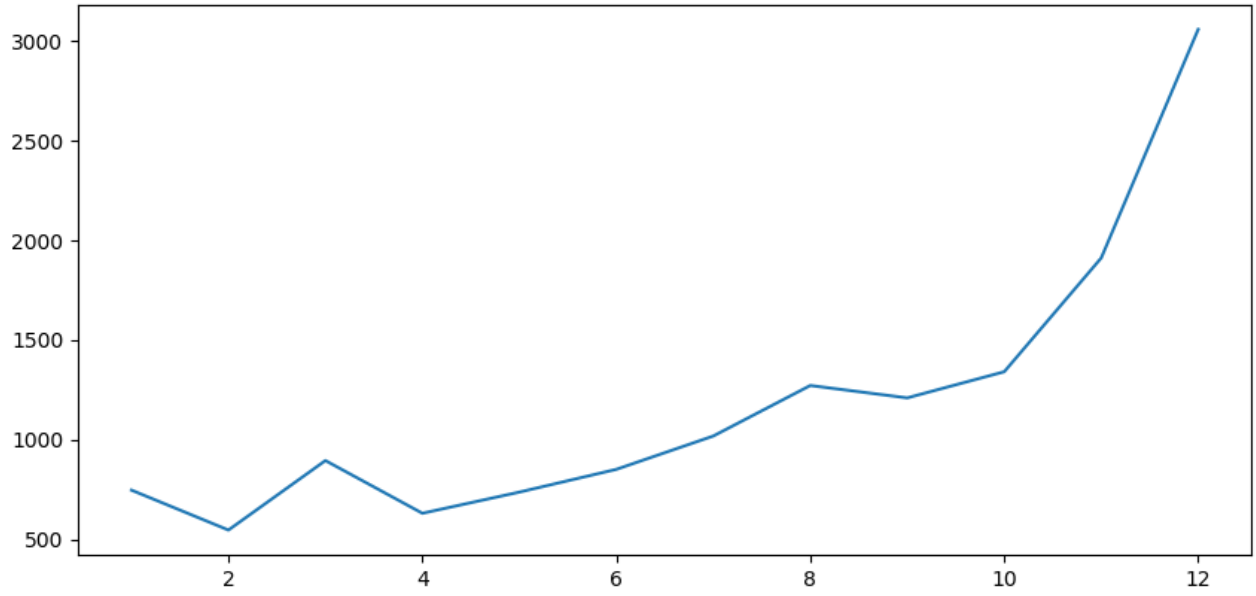
6.2.1.1 Επιλογή Συλλογής Δεδομένων (Dataset)

Ως συλλογή δεδομένων χρησιμοποιήσαμε τα IDs από τα Tweets που αναρτήθηκαν από τις σελίδες που αναφέραμε στο Κεφάλαιο 4 και τον πίνακα 4.1. Η επιλογή των συγκεκριμένων σελίδων έγινε προκειμένου να αποφευχθούν Tweets τα οποία είναι παραπλανητικά, όπως είναι τα Tweets που περιέχουν διαφημίσεις, αλλά και Tweets τα οποία περιέχουν ψευδείς ειδήσεις (Fake News) και τα οποία μπορεί να προκαλούσαν αποσταθεροποίηση του συστήματος, κάτι που θα οδηγούσε σε εξαγωγή λάθος συμπερασμάτων. Οι συγκεκριμένες σελίδες, όπως γίνεται αντιληπτό και στο κεφάλαιο 4.2.2, φημίζονται για την αξιοπιστία τους, την εγκυρότητά τους και την άμεση ανταπόκρισή τους σε έκτακτα γεγονότα. Τα Tweets αυτά αναρτήθηκαν στο διάστημα μεταξύ 1/1/2017 και 31/12/2017. Η επιλογή της χρονιάς 2017 έγινε λόγω των μεγάλων μεταβολών που παρουσιάστηκαν στην τιμή του Bitcoin, καθώς επίσης και λόγω των πολλών νέων και γεγονότων που συνέβησαν και αφορούσαν το κρυπτονόμισμα αυτό. Το συγκεκριμένο Dataset περιείχε 18.877 Tweets και

ο καταμερισμός τους στο παραπάνω διάστημα φαίνεται στα παρακάτω διαγράμματα:



Σχήμα 16: Διάγραμμα καταμερισμού Tweets ανά ημέρα



Σχήμα 17: Διάγραμμα καταμερισμού Tweets ανά μήνα

6.2.1.2 Σύνδεση με το Twitter API και συλλογή των Tweets

Στη συνέχεια υλοποιήσαμε την διασύνδεση με το Twitter API προκειμένου να ανακτήσουμε τις απαραίτητες πληροφορίες των Tweets που είχαμε στη συλλογή δεδομένων, καθώς όπως αναφέραμε το μόνο που διαθέταμε ήταν το ID του κάθε Tweet.

Στην συγκεκριμένη εργασία έπρεπε να λάβουμε υπόψιν μας τα όρια που θέτει το ίδιο το Twitter για την ανάκτηση δεδομένων από το API του. Για το λόγο αυτό δημιουργήσαμε έναν πίνακα στη βάση μας, στον οποίο αποθηκεύαμε τις διαθέσιμες κλήσεις που είχαμε περιθώριο να εκτελέσουμε στο API του Twitter. Η πλατφόρμα του API του Twitter στην απάντηση κάθε κλήσης της μας επέστρεφε στα Headers (metadata) της τον διαθέσιμο

αριθμό κλήσεων που είχαμε, καθώς επίσης και το χρονικό όριο στο οποίο μπορούσαμε να εκτελέσουμε αυτές τις κλήσεις. Η χρήση της βάσης δεδομένων στο συγκεκριμένο σημείο μας βοήθησε στην περίπτωση που για οποιονδήποτε λόγο, όπως απώλεια σύνδεσης στο διαδίκτυο, σφάλματα στον κώδικα κτλ, αποτύγχανε η εφαρμογή και χάναμε όλα τα δεδομένα στη μνήμη της.

Εν συνεχεία αφού είχαμε επιτυχώς συλλέξει το εκάστοτε Tweet, το αποθηκεύαμε στην βάση δεδομένων σε πίνακα όπου κρατούσαμε τα στοιχεία που χρειαζόμασταν για κάθε Tweet. Στο τέλος της συγκεκριμένης διαδικασίας, είχαμε και τα 18.877 Tweets στην παρακάτω μορφή στην βάση δεδομένων μας:

PostID	Username	TweetDate	user_id	text
815392114691313665	CryptoCoinsNews	2017-01-01 03:00:04.000	1856523530	Opinion: Nigeria Needs Bitcoin #Bitcoin Regulation @billpaspalas ...
815407211446935552	coindesk	2017-01-01 04:00:03.000	1333467482	The latest Bitcoin Price Index is 964.26 USD https://t.co/IzUu2wyP...
815437445542055941	CryptoCoinsNews	2017-01-01 06:00:11.000	1856523530	2016 Review: There's New Momentum For Bitcoin and the Blockc...
815467609344929792	coindesk	2017-01-01 08:00:03.000	1333467482	The latest Bitcoin Price Index is 966.29 USD https://t.co/IzUu2wyP...
815512874688933888	coindesk	2017-01-01 10:59:55.000	1333467482	Bitcoin may be booming, but another crypto-coin had a bigger 20...
815521728340193280	BTCTN	2017-01-01 11:35:06.000	3367334171	Coming Up: Drop Zone – a Hyper-Local Bitcoin-Based Market Ser...
815543106175176708	coindesk	2017-01-01 13:00:03.000	1333467482	Did this man create bitcoin? The verdict's still out, yet it was one o...
815545806606503936	CryptoCoinsNews	2017-01-01 13:10:47.000	1856523530	The European Union Wants to Identify Bitcoin Users https://t.co/1...
815557262957834240	coindesk	2017-01-01 13:56:18.000	1333467482	What Will the Bitcoin Price Be in 2017? https://t.co/q49ttv0WE0 htt...
815584165718589440	Cointelegraph	2017-01-01 15:43:12.000	2207129125	Vitalik Buterin: #Bitcoin More Likely Than #Ethereum to Split in 20...
815588936663834625	Bitcoin	2017-01-01 16:02:10.000	357312062	What Will the Bitcoin Price Be in 2017? https://t.co/MbAkmok0uP
815592146375966720	Cointelegraph	2017-01-01 16:14:55.000	2207129125	Rassah of @MyceliumCom : What ETF Will Bring to #Bitcoin Table...
815605057852993537	Cointelegraph	2017-01-01 17:06:13.000	2207129125	#Microsoft to Add Extensive Support For #Bitcoin, Describes it as ...
815605528688635904	BTCTN	2017-01-01 17:08:05.000	3367334171	American Black Cross Helps Political Prisoners With Bitcoin https:/...
815622322681999360	CryptoCoinsNews	2017-01-01 18:14:49.000	1856523530	The New Year Could Bode Well for Bitcoin and Blockchains https:/...
815664004563697668	coindesk	2017-01-01 21:00:27.000	1333467482	When we asked analysts to predict bitcoin's price movements for ...
815664912433172480	BTCTN	2017-01-01 21:04:04.000	3367334171	Bitcoin Breaks $1,000 as Exchanges Break Volume Records Worldwi...
815665831539515397	coindesk	2017-01-01 21:07:43.000	1333467482	Bitcoin Price Tops $1,000 in First Day of 2017 Trading https://t.co/F...
815671045881495556	business	2017-01-01 21:28:26.000	34713362	Forget bitcoin and mobile payments. Cash still rules the world htt...
815679392001228800	Bitcoin	2017-01-01 22:01:36.000	357312062	Bitcoin Price Tops $1,000 in First Day of 2017 Trading https://t.co/...
815708947831263234	coindesk	2017-01-01 23:59:02.000	1333467482	The latest Bitcoin Price Index is 997.75 USD https://t.co/IzUu2wyP...
815716378132946944	BTCTN	2017-01-02 00:28:34.000	3367334171	A Look At Bitcoin Bubbles, When Will the Next One Be? https://t.c...
815739873374175233	FortuneMagazine	2017-01-02 02:01:56.000	25053299	Legal sparring continues in Bitcoin user's battle with IRS tax sweep...
815769597206990848	coindesk	2017-01-02 04:00:02.000	1333467482	The latest Bitcoin Price Index is 1,009.93 USD https://t.co/IzUu2wy...
815814914984046592	coindesk	2017-01-02 07:00:07.000	1333467482	CoinDesk asked experts to forecast bitcoin's price in the year ahea...
815843713712799744	MarketWatch	2017-01-02 08:54:33.000	624413	Bitcoin hits milestone of $1,000 as 2017 begins https://t.co/rZ5D7E...
815875303998455808	CryptoCoinsNews	2017-01-02 11:00:05.000	1856523530	Bitcoin Starts 2017 at the $1000 https://t.co/Y2AhASh60M https://t...
815876452511944704	BTCTN	2017-01-02 11:04:39.000	3367334171	Why Volume Is Exploding at Mexican Bitcoin Exchange Bitso https:...
815880330162868226	Cointelegraph	2017-01-02 11:20:03.000	2207129125	CNBC: Forget About Dow, #Bitcoin to Reach $1,000 https://t.co/tn...

Σχήμα 18: Δείγμα της συλλογής των Tweets

6.2.1.3 Σύνδεση με το Cryptocompare API και συλλογή ιστορικών δεδομένων της τιμής του Bitcoin

Το τελευταίο βήμα του πρώτου σταδίου, αφορούσε την ανάκτηση ιστορικών δεδομένων σχετικά με την τιμή του Bitcoin από την πλατφόρμα Cryptocompare με τη χρήση του API της. Για την υλοποίηση της διαδικασίας αυτής αναπτύξαμε μία εφαρμογή η οποία συνδεόταν με το συγκεκριμένο API και αντλούσε τις ωριαίες τιμές κάθε ημέρας.

Ομοίως και σε αυτό το στάδιο έπρεπε να συμμορφωθούμε με τα όρια του API της συγκεκριμένης πλατφόρμας, καθώς επίσης και να διασφαλίσουμε πως σε περίπτωση αποτυχίας δεν θα χρειαζόταν να ξεκινήσουμε την άντληση των δεδομένων από την αρχή. Για να ξεπεράσουμε το $1^{ο}$ εμπόδιο εργαστήκαμε με πανομοιότυπο τρόπο όπως και με το API του Twitter. Δημιουργήσαμε, δηλαδή, έναν πίνακα όπου αποθηκεύαμε τα όρια που μας έθετε μετά από κάθε κλήση το API του Cryptocompare. Για την αντιμετώπιση του $2^{ου}$ εμποδίου, δημιουργήσαμε έναν πίνακα στον οποίο αποθηκεύαμε τις ημέρες από τις οποίες επιτυχώς αντλούσαμε δεδομένα ανά ώρα για την τιμή του Bitcoin.

Όπως γίνεται αντιληπτό και σε αυτή την περίπτωση αφού αντλούσαμε τα ωριαία δεδομένα ανά ημέρα, στη συνέχεια τα αποθηκεύαμε σε έναν πίνακα στη βάση δεδομένων μας. Το σύνολο των δεδομένων μας ανέρχεται σε 8.760, όσες δηλαδή και οι ώρες του κάθε έτους. Η τιμή του Bitcoin, σε αντίθεση με άλλους χρηματιστηριακούς δείκτες και εμπορεύματα, μεταβάλλεται όλο το 24ωρο. Στο τέλος της συγκεκριμένης διαδικασίας είχαμε τα δεδομένα στη βάση δεδομένων στην παρακάτω μορφή:

time	closeValue	high	low	openValue	volumefrom	volumeto	DateTimeValue
1483221600	965,73	966,36	962,97	964,93	443,25	427355,17	2017-01-01 00:00:00.000
1483225200	965,31	966,02	963,12	965,73	933,69	904240,10	2017-01-01 01:00:00.000
1483228800	967,72	970,51	966,09	965,31	773,31	749952,17	2017-01-01 02:00:00.000
1483232400	964,41	967,36	963,91	967,72	785,40	762670,91	2017-01-01 03:00:00.000
1483236000	961,55	962,92	959,34	964,41	450,41	435835,74	2017-01-01 04:00:00.000
1483239600	962,82	965,60	961,39	961,55	700,45	676237,07	2017-01-01 05:00:00.000
1483243200	963,20	965,55	962,12	962,82	546,21	527906,26	2017-01-01 06:00:00.000
1483246800	964,30	966,00	962,79	963,20	416,66	406083,29	2017-01-01 07:00:00.000
1483250400	957,35	958,60	955,63	964,30	407,54	392742,65	2017-01-01 08:00:00.000
1483254000	959,27	960,89	955,70	957,35	564,88	544161,71	2017-01-01 09:00:00.000
1483257600	961,75	962,32	960,33	959,27	390,30	376360,65	2017-01-01 10:00:00.000
1483261200	961,20	961,56	959,30	961,75	309,14	297731,33	2017-01-01 11:00:00.000
1483264800	960,46	961,41	952,47	961,20	1772,90	1709340,51	2017-01-01 12:00:00.000
1483268400	970,00	971,89	965,51	960,46	2231,94	2171186,38	2017-01-01 13:00:00.000
1483272000	972,58	973,96	971,13	970,00	1451,88	1415767,02	2017-01-01 14:00:00.000
1483275600	966,80	968,68	965,11	972,58	629,70	612025,92	2017-01-01 15:00:00.000
1483279200	969,68	970,52	967,88	966,80	456,86	444281,13	2017-01-01 16:00:00.000
1483282800	967,86	969,35	964,27	969,68	1195,91	1161638,08	2017-01-01 17:00:00.000
1483286400	979,24	980,36	974,62	967,86	3027,10	2968947,68	2017-01-01 18:00:00.000
1483290000	984,52	988,66	969,58	979,24	9134,01	9012076,10	2017-01-01 19:00:00.000
1483293600	994,27	994,49	988,03	984,52	2702,94	2685507,04	2017-01-01 20:00:00.000
1483297200	993,95	998,04	992,97	994,27	3006,08	3000898,61	2017-01-01 21:00:00.000
1483300800	999,64	999,90	993,89	993,95	1862,96	1863959,83	2017-01-01 22:00:00.000
1483304400	1002,50	1003	997,29	999,64	3952,28	3961113,89	2017-01-01 23:00:00.000
1483308000	1002,45	1004	999,15	1002,50	2175,57	2186563,14	2017-01-02 00:00:00.000
1483311600	998,20	1002	994,70	1002,45	2190,08	2193466,12	2017-01-02 01:00:00.000
1483315200	1000,13	1001	995,30	998,20	984,52	986069,54	2017-01-02 02:00:00.000
1483318800	994,56	999,14	992,96	1000,13	919,09	919166,61	2017-01-02 03:00:00.000
1483322400	997,98	998,88	993,49	994,56	786,48	786765,36	2017-01-02 04:00:00.000
1483326000	1008,62	1009	997,57	997,98	3194,95	3215574,57	2017-01-02 05:00:00.000
1483329600	1006,84	1010	1001	1008,62	1921,92	1942506,97	2017-01-02 06:00:00.000
1483333200	1005,92	1008	1004	1006,84	796,56	806142,88	2017-01-02 07:00:00.000
1483336800	1006,08	1006	1001	1005,92	757,44	763673,36	2017-01-02 08:00:00.000
1483340400	1005,84	1007	1003	1006,08	965,39	975288,58	2017-01-02 09:00:00.000
1483344000	1004,72	1009	1003	1005,84	2085,78	2109657,06	2017-01-02 10:00:00.000
1483347600	1009,72	1010	1003	1004,72	1499,80	1513385,08	2017-01-02 11:00:00.000
1483351200	1014,67	1018	1010	1009,72	3121,42	3174829,81	2017-01-02 12:00:00.000
1483354800	1018,51	1019	1011	1014,67	2796,59	2852912,84	2017-01-02 13:00:00.000

Σχήμα 19: Δείγμα της συλλογής της τιμής του Bitcoin ανά ώρα

6.2.2 Ομαδοποίηση των Tweets βάσει των σημαντικότερων γεγονότων που αφορούσαν το Bitcoin (Annotate Dataset)

Στο δεύτερο στάδιο της εργασίας έπρεπε να ομαδοποιήσουμε, χειροκίνητα, τα Tweets που είχαμε συλλέξει στο προηγούμενο στάδιο. Για να επιτύχουμε ένα αξιόλογο αποτέλεσμα ομαδοποιήσαμε 2.035 Tweets σε 41 διαφορετικά γεγονότα. Προκειμένου να φτάσουμε σε αυτό το αποτέλεσμα, αναζητήσαμε στο διαδίκτυο τα πιο σημαντικά γεγονότα του 2017 που αφορούσαν το Bitcoin. Εν συνεχεία, «τρέχοντας» κατάλληλα ερωτήματα (Queries) στον SQL Server καταφέραμε να φιλτράρουμε τα αποτελέσματα για κάθε διαφορετικό γεγονός και να κατατάξουμε τα Tweets στο γεγονός που αντιστοιχούν. Παρακάτω παρατίθεται πίνακας με τα γεγονότα που εντοπίσαμε και τον αριθμό των Tweets που σχετίζονται με το κάθε γεγονός:

EventID	**Γεγονός**	**Tweets**
9	11/01. Anxieties over China crackdown [19]	100
34	09/02. Chinese Central Bank Warns Bitcoin Exchanges [20]	49
51	03/03. Bitcoin is extending its lead over gold [21]	32
10	10/03. SEC denies permission for a bitcoin ETF [22]	110
50	16/03. Bloomberg - Someone wants to stick a fork in bitcoin [23]	60
11	01/04. Japan declares bitcoin legal currency [24]	29
29	23/05. Bitcoin Scaling Consensus, debate and Segwit2x [25]	105
31	23/05. Japanese Budget Airlines Will Accept Bitcoin In 2017, Install BTMs At Airports [26]	10
28	31/05. Chinese Exchanges Resume Withdrawals, Bitcoin Likely to Surge [27]	12
49	13/07. Morgan Stanley says investors shouldn't buy Bitcoin. [28]	9
43	26/07. Greek police arrest Russian man who US says laundered $4 billion through bitcoin [29]	6
12	01/08. The big bitcoin split [30]	102
27	11/08. Institutional Investors Can No Longer Ignore Bitcoin: Goldman Sachs [31]	13
13	04/09. China bans ICOs [32]	76
15	12/09. Jamie Dimon Bitcoin Statement [33]	127
17	14/09. Bitcoin Exchange BTCChina will Close by October [34]	11
40	15/09 BREAKING: China's Bitcoin Exchanges Receive Shutdown Orders and Closure [35]	87
41	28/09. Japan to regulate #Bitcoin [36]	53
33	02/10. Goldman Sachs is weighing a new trading operation dedicated to bitcoin, other digital currencies [37]	54
44	04/10. Greek court rules to extradite Russian bitcoin fraud suspect to the United States. [38]	7

24	10/10. Russia Rejects Cryptocurrency as Authorities Block Access to Exchanges [39]	13
26	01/11. Bitcoin is skyrocketing because 2 of the biggest exchange groups in the world are launching bitcoin futures — here's what that means [40]	66
42	01/11. South Korea to regulate #Bitcoin as commodity [41]	15
20	08/11. Segwit2x Fork Cancelled [42]	32
36	20/11. CME Group Plans to Launch Bitcoin Futures [43]	33
23	21/11. Tether Allegedly Hacked For $30 Mln [44]	10
46	29/11. Bitcoin rises above $10,000 for the first time [45]	72
21	29/11. BREAKING NEWS – NASDAQ TO LAUNCH BITCOIN FUTURES IN 2018 [46]	12
37	01/12. BREAKING: Bitcoin futures will be allowed to start trading [47]	44
14	01/12. CFTC approves bitcoin futures [48]	16
38	06/12. Big banks push back on launch of bitcoin futures [49]	30
48	06/12. Bitcoin mining service NiceHash says hackers emptied its wallet [50]	11
47	07/12. Bitcoin climbs past $US14,000 as demand continues to skyrocket [51]	38
39	11/12. Bitcoin futures are about to go live, and they could change the game for cryptocurrencies [52]	161
32	11/12. Bitcoin Markets Really Like CBOE Futures, Prices Spike Sharply [53]	76
16	15/12. Bitcoin futures are about to get another big boost [54]	44
30	20/12. North Korea likely behind a massive cyber attack on a South Korean bitcoin exchange that caused it to collapse [55]	29
25	22/12. The Sharks Are Beginning to Circle Bitcoin and It's Down 30% [56]	122
22	28/12. Bitcoin falls as South Korea announces crackdown - as it happened [57]	28
45	Nigeria Regulations (Topic) [58]	26
35	Terrorism, Money Laundering, Drugs (Topic) [59]	105

Πίνακας 2: Γεγονότα που ομαδοποιήθηκαν με σειρά εμφάνισης του γεγονότος

Με το τέλος της διαδικασίας αυτής είχαμε πλέον στη βάση δεδομένων το γεγονός (Event) στο οποίο ανήκει κάθε Tweet. Αν κάποιο Tweet δεν το είχαμε κατατάξει σε κάποιο γεγονός η τιμή του ID που αφορούσε το Event ήταν μηδέν. Παρακάτω παρουσιάζεται δείγμα του

πίνακα με τα Tweets:

PostID	Username	TweetDate	user_id	text	EventID
815392114691313665	CryptoCoinsNews	2017-01-01 03:00:04.000	1856523530	Opinion: Nigeria Needs Bitcoin #Bi...	45
816587000853757956	CryptoCoinsNews	2017-01-04 10:08:07.000	1856523530	Bitcoin Usage Gains Traction in Ind...	35
817045998690181120	CryptoCoinsNews	2017-01-05 16:32:00.000	1856523530	China Intervenes, Yuan Soars, Bitco...	9
817052459550052352	ReutersBiz	2017-01-05 16:57:41.000	15110357	Dramatic bitcoin rally nosedives as ...	9
817194773715030018	FortuneMagazine	2017-01-06 02:23:11.000	25053299	How a China crackdown caused bit...	9
817304055663968256	Cointelegraph	2017-01-06 09:37:26.000	2207129125	Indonesian #Bitcoin Market Rises, ...	35
817342688961986560	CryptoCoinsNews	2017-01-06 12:10:57.000	1856523530	Breaking: China's Central Bank Wei...	9
817352588404420609	coindesk	2017-01-06 12:50:17.000	1333467482	China's Central Bank Issues Warnin...	9
817370926769541120	BTCTN	2017-01-06 14:03:09.000	3367334171	Bitcoin Growing Fast In Unbanked I...	0
817377319102935044	ReutersBiz	2017-01-06 14:28:33.000	15110357	Bitcoin extends losses, slides anoth...	9
817396821563756544	ReutersBiz	2017-01-06 15:46:03.000	15110357	Bitcoin plunges another 12 percent...	9
817415401416101888	coindesk	2017-01-06 16:59:53.000	1333467482	Bloomberg - Here's why bitcoin bu...	9
817416610801733633	Cointelegraph	2017-01-06 17:04:41.000	2207129125	#China Warns #Bitcoin Users, Pani...	9
817445703446315008	coindesk	2017-01-06 19:00:17.000	1333467482	Barron's - China's Big Bet on Bitcoi...	9
817533252000944128	business	2017-01-07 00:48:10.000	34713362	Here's why bitcoin buyers are nerv...	9
817576128592441344	coindesk	2017-01-07 03:38:33.000	1333467482	China to Restrict Bitcoin Marketing...	9
817626859919839232	coindesk	2017-01-07 07:00:08.000	1333467482	China-based bitcoin exchange BTC...	9
817717558224756736	coindesk	2017-01-07 13:00:32.000	1333467482	Following a People's Bank of China...	9
817787947072286722	CryptoCoinsNews	2017-01-07 17:40:15.000	1856523530	Do Not Mention Devaluation, Chin...	9
818070540359450629	Cointelegraph	2017-01-08 12:23:10.000	2207129125	#Media Spread Wrong "#China Ba...	9
818079484616310790	CryptoCoinsNews	2017-01-08 12:58:42.000	1856523530	Dutch Public Prosecutor to Tackle ...	0
818334905809666049	business	2017-01-09 05:53:40.000	34713362	Bitcoin extends loss after warning f...	9
818367008706273280	business	2017-01-09 08:01:14.000	34713362	Bitcoin, the people's liberation curr...	9
818435311260889090	ReutersBiz	2017-01-09 12:32:38.000	15110357	Big China bitcoin exchange says no...	9
818455933655322625	coindesk	2017-01-09 13:54:35.000	1333467482	Indonesia's AML Watchdog Links B...	0
818464528694411266	CryptoCoinsNews	2017-01-09 14:28:44.000	1856523530	2-3 Years Before Bitcoin Regulation...	9
818521186602799104	business	2017-01-09 18:13:52.000	34713362	Bitcoin, the people's liberation curr...	9
818592402382475266	coindesk	2017-01-09 22:56:52.000	1333467482	Bitcoin's Price Volatile at $900 as C...	9
818685623468965890	coindesk	2017-01-10 05:07:17.000	1333467482	China's BTCC Welcomes Greater Bit...	9
818714155981144064	coindesk	2017-01-10 07:00:40.000	1333467482	Islamic State (IS) militants are now ...	35
818736656471695362	CryptoCoinsNews	2017-01-10 08:30:04.000	1856523530	Terrorists Use Bitcoin And PayPal In...	35
818744214502645760	coindesk	2017-01-10 09:00:06.000	1333467482	Bloomberg - Bitcoin, the people's li...	9
818804678226018304	coindesk	2017-01-10 13:00:22.000	1333467482	Bloomberg – No Free Ride for Bitco...	9
818811913660284929	BTCTN	2017-01-10 13:29:07.000	3367334171	China's Smart Money is Staying in ...	9
818812950534295553	coindesk	2017-01-10 13:33:14.000	1333467482	OKCoin Joins Calls for Bitcoin Regu...	9
818822887872282624	FinancialTimes	2017-01-10 14:12:44.000	4898091	China probes bitcoin amid capital f...	9
818857479635345409	FT	2017-01-10 16:30:11.000	18949452	China investigates bitcoin exchang...	9
818868789475414016	businessinsider	2017-01-10 17:15:07.000	20562637	China isn't cracking down on bitco...	9

Σχήμα 20: Δείγμα της συλλογής των Tweets μετά την Ομαδοποίηση

6.2.3 Προ-επεξεργασία των δεδομένων

Στο συγκεκριμένο στάδιο της υλοποίησης μετασχηματίσαμε τα δεδομένα που είχαμε συλλέξει, συγκεκριμένα τα Tweets και τις τιμές των κρυπτονομισμάτων, με τέτοιο τρόπο ώστε να είναι συμβατά με τους αλγόριθμους μηχανικής μάθησης που επιλέξαμε παρακάτω να «τρέξουμε», προκειμένου να εξάγουμε κάποια σημαντικά συμπεράσματα από τα δεδομένα αυτά.

6.2.3.1 Προ-επεξεργασία των Tweets (Tweets Preprocessing)

Για την προ-επεξεργασία των Tweets ακολουθήσαμε τα παρακάτω βήματα:

1. Αφαιρέσαμε τις επισημάνσεις άλλων χρηστών από κάθε Tweet (για παράδειγμα το Tweet: “Indonesian #Bitcoin Market Rises, Rapid Increase in User Base Reported https://t.co/YdQ5pRd3AH - By @iamjosephyoung” μετά την αφαίρεση των επισημάνσεων έγινε: “Indonesian #Bitcoin Market Rises, Rapid Increase in User Base Reported https://t.co/YdQ5pRd3AH - By”
2. Απομονώσαμε τα Hashtags από κάθε Tweet και τα κρατήσαμε σε ξεχωριστή οντότητα. Επομένως στο προηγούμενο Tweet κρατήσαμε το Hashtag #Bitcoin
3. Αφαιρέσαμε ειδικούς χαρακτήρες (όπως το «Enter», !, &, κτλ), επομένως το παραπάνω Tweet έγινε: «Indonesian Bitcoin Market Rises, Rapid Increase in User Base Reported https://t.co/YdQ5pRd3AH»
4. Αφαιρέσαμε τα link από το κείμενο των Tweets. Επομένως από προηγούμενο παράδειγμα το Tweet μας πλέον είναι έτσι: «Indonesian Bitcoin Market Rises, Rapid Increase in User Base Reported»
5. Μετατρέψαμε όλους τους χαρακτήρες σε πεζούς. Επομένως το Tweet το οποίο έχουμε σαν αναφορά έγινε: «indonesian bitcoin market rises, rapid increase in user base reported»
6. Με τη βοήθεια των βιβλιοθηκών της Python nltk (Natural Language Toolkit) και gensim αφαιρέσαμε stopwords όπως το “and”, “to”, “from” κτλ καθώς επίσης επεξεργαστήκαμε τις λέξεις με το stemming εργαλείο του gensim προκειμένου να αφαιρέσουμε τον πληθυντικό αριθμό από τις λέξεις και το χρόνο από τα ρήματα. Στο τέλος αυτού του βήματος το παράδειγμά μας είχε αυτή τη μορφή: «indonesian market rise rapid increas user base report»
7. Τέλος, αφαιρέσαμε τις λέξεις που εμφανίζονται μία φορά στο σύνολο των λέξεων όλων των Tweets. Στο δικό μας παράδειγμα επειδή η λέξη indonesian δεν εμφανίστηκε σε άλλο Tweet η τελική μας πρόταση είχε αυτή τη μορφή «market rise rapid increas user base report»

Και σε αυτό το βήμα, αποθηκεύσαμε τα δεδομένα μας σε έναν ξεχωριστό πίνακα στην βάση δεδομένων μας. Δείγμα από τα δεδομένα του πίνακα αυτού μετά το τέλος της

επεξεργασίας όλων των Tweets παρουσιάζεται στο παρακάτω σχήμα:

PostID	text	hashtags
823869424994308096	trade fee see volum dive china	
823893167753887744	ceo china largest big exchang say regul inevit	
823895910031781889	price unfaz china exchang add fee	
823900685511127040	chines exchang enforc trade fee	
823929899190747136	trade plung china exchang transact fee	
823950349262356485	hold fear chines crackdown fade	
824005382553108482	hold fear chines crackdown fade	
824244230126272512	china central_bank continu exchang inspect	
824379022100545540	fall volum trader stick china exchang	
824486770289086464	china communist keep free	
824851932577615872	china base exchang huobi announc updat trade fee polici today	
824995703818768385	busi insid chines love	
825749447959261184	china launch research studi test blockchain	china blockchain
826022887517720577	chang china mean	
826023208998440961	resili chines exchang start charg transact fee	
826046998075817984	price find stabil market factor china crackdown fee financ financi	finance financial china fees2017
826052707433738241	china recent move attempt find place world stage	
826112961630584832	link china recent central_bank move industri home	
826854729015779328	price china south_korea	china
827457199706365957	china trade war stop price pass year	uschina
828591435997929476	dutch author look deem money_laund	
829152818032283648	countdown legal method payment japan month	
829216714743230464	trade chines yuan drop third day	
829249384659447809	slump china central_bank hold close door meet domest exchang	
829291463473233920	break exchang hold door meet china central_bank	
829303962079809536	drop sharpli suddenli news china	
829335719046844416	slump china central_bank call meet domest exchang	
829350238804725761	chines citizen pboc	
829373124877643778	price fall chines author meet exchang	
829382241633632256	nigerian run take nigeria	nigerians nigeria nigeriansprotest
829502859947417600	fall news china central_bank hold close door meet local cryptocurr exchang	
829553090051338240	chines central_bank warn exchang	
829560522676776961	china central_bank go	
829618955585277952	price roller coaster panic china show	china us
829687888891744256	newsflash price crash chines exchang okcoin amp huobi paus withdraw	
829695652154195968	break news forc stop btc ltc withdraw price plung	
829695844836315136	tank chines exchang block withdraw	
829732758234812416	chines regul pressur exchang	

Σχήμα 21: Δείγμα των Tweets μετά την προ-επεξεργασία τους

6.2.3.2 Υπολογισμός ομοιότητας μεταξύ των Tweets

Το επόμενο βήμα που έπρεπε να υλοποιήσουμε για να μπορέσουμε να εκτελέσουμε τους αλγόριθμους της μηχανικής μάθησης, ήταν να υπολογίσουμε την ομοιότητα μεταξύ των Tweets. Η ομοιότητα αυτή εξαρτάται από την χρονική απόσταση μεταξύ των Tweets, την ομοιότητα του κειμένου και του νοήματος τους, τα κοινά Hashtags που έχουν και από τον συγγραφέα του Tweet. Παρακάτω θα αναλύσουμε πως υπολογίσαμε την κάθε ομοιότητα ξεχωριστά.

Χρονική ομοιότητα:

Στην περίπτωση της χρονικής ομοιότητας χρησιμοποιήσαμε μία γνησίως φθίνουσα συνάρτηση, η οποία παίρνει ως είσοδο τις ημερομηνίες δημιουργίας των Tweets και επιστρέφει την ομοιότητα βασιζόμενη στην χρονική διασπορά των 2 αυτών Tweets που συγκρίνουμε κάθε φορά. Επιλέξαμε αυτή η συνάρτηση να έχει εκθετική μείωση. Ο μαθηματικός της τύπος είναι ο εξής:

$$f_t(t_u, t_v) = e^{-|t_u - t_v|/q}$$

Η τιμή της συνάρτησης λαμβάνει την μέγιστη τιμή της όταν $t_u \approx t_v$ και φθίνει στο μηδέν, με ρυθμό που καθορίζει η παράμετρος q και όσο η διαφορά $|t_u - t_v|$ αυξάνεται.

Ομοιότητα Hashtags:

Όταν δύο Tweets έχουν κοινά Hashtags αυξάνουν κατά ένα βαθμό την ομοιότητά τους, για κάθε ένα κοινό Hashtag.

Ομοιότητα Συγγραφέα:

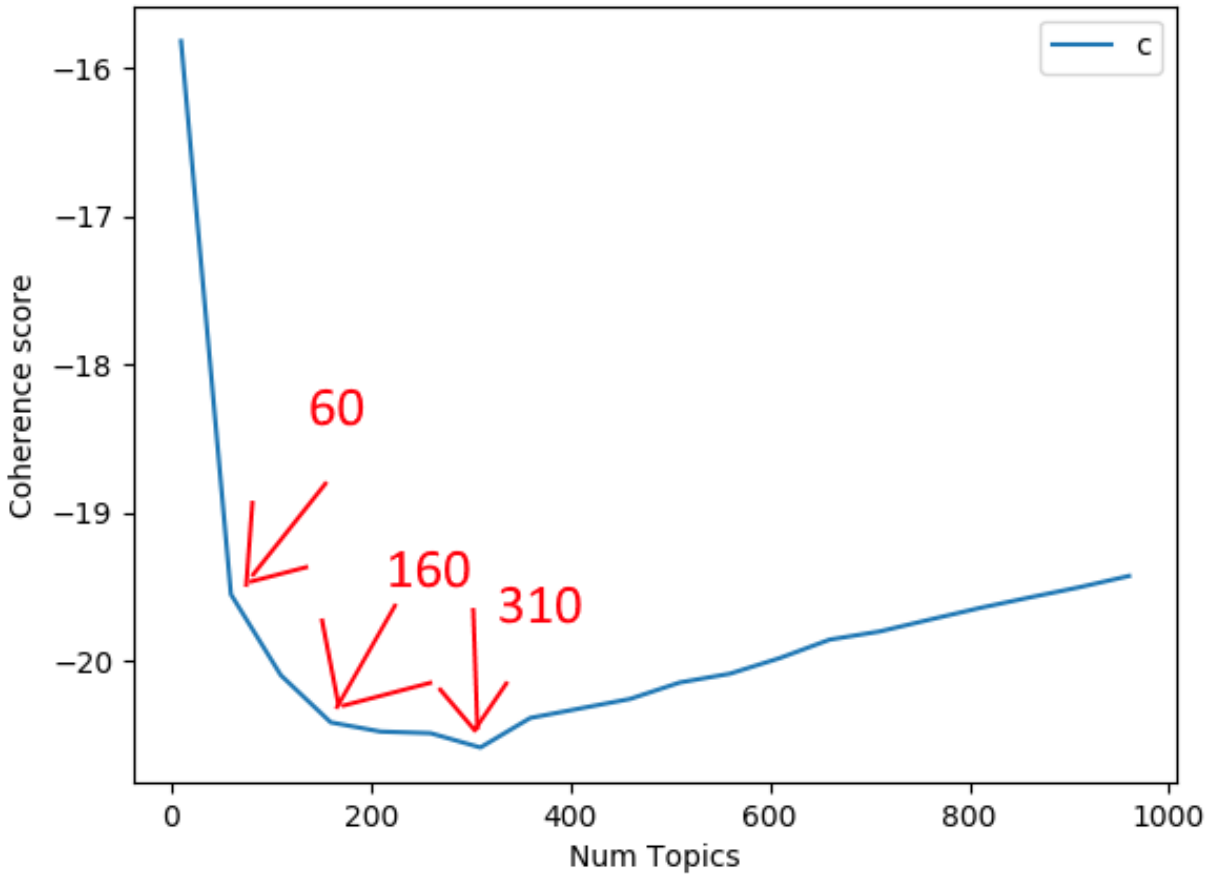
Όταν δύο Tweets μεταξύ τους έχουν κοινό συγγραφέα αυξάνουν την ομοιότητά τους κατά ένα βαθμό.

Ομοιότητα Κειμένου:

Για να υπολογίσουμε την ομοιότητα μεταξύ δύο Tweets χρησιμοποιήσαμε την βιβλιοθήκη gensim, η οποία περιλαμβάνει υλοποιήσεις που αφορούν την επεξεργασία κειμένου φυσικής γλώσσας (Natural Language Processing) και τα WordEmbedings. Η χρήση της συγκεκριμένης βιβλιοθήκης προϋποθέτει την δημιουργία ενός σώματος κειμένων το οποίο εν συνεχεία επεξεργάζεται από ένα νευρωνικό δίκτυο δύο επιπέδων. Το νευρωνικό αυτό βασιζόμενο στην εμφάνιση της κάθε λέξης μέσα στα κείμενα (στην συγκεκριμένη περίπτωση τα Tweets) ομαδοποιεί τα κείμενα βάσει της συνάφειάς τους. [60]

Το πρόβλημα που αντιμετωπίσαμε στην συγκεκριμένη διαδικασία, ήταν ότι δεν γνωρίζαμε σε πόσα λογικά τμήματα χωρίζονται τα Tweets που διαθέταμε. Για να βρούμε τον αριθμό αυτόν τρέξαμε τον αλγόριθμο για διαφορετικό πιθανό αριθμό τμημάτων και με τη βοήθεια της τιμής της συνοχής (coherence value) βρήκαμε πως το dataset μας χωρίζεται περίπου σε 160 νοηματικά τμήματα. Η επιλογή αυτή γίνεται με την μέθοδο του «αγκώνα» για τον ιδανικό αριθμό τμημάτων (clusters). Στη συγκεκριμένη μέθοδο δημιουργούμε την γραφική παράσταση για τις διαφορετικές τιμές των clusters συναρτήσει του coherence score και επιλέγουμε την πρώτη τιμή που ελαχιστοποιεί την συνάρτηση. [61]

Παρακάτω παρουσιάζουμε το διάγραμμα για την περίπτωση που εξετάζουμε:



Σχήμα 22: Διάγραμμα ελαχιστοποίησης Coherence score / αριθμός νοηματικών ενοτήτων

Στη συνέχεια, πάλι με τη χρήση του gensim και αφού ξέραμε τον αριθμό των τμημάτων, υπολογίσαμε την ομοιότητα μεταξύ όλων των Tweets.

Και σε αυτό το στάδιο αποθηκεύσαμε στη βάση δεδομένων όλους τους υπολογισμούς μας για κάθε ζεύγος από Tweets. Ο πίνακας αυτός μετά το τέλος όλων των παραπάνω βημάτων υπολογισμού ομοιότητας είχε την παρακάτω μορφή:

TweetID1	TweetID2	SentenceSimilarity	TimeSimilarity	HashtagsSimilarity	AuthorSimilarity
875822451070033920	886380149645746177	0,77283239364624023	2517152	4	1
875822451070033920	881949238971781122	0,92185544967651367	1460741	3	1
875822451070033920	885249287961153538	0,69966864585876465	2247534	3	1
875822451070033920	885492852930772992	0,63777589797973633	2305604	3	1
877219014572998656	887040889100673026	0,61742377281188965	2341718	3	1
877219014572998656	885249287961153538	0,51734840869903564	1914567	3	1
892825679900835841	892894993924460544	0,46280378103256226	16526	3	1
881949238971781122	885492852930772992	0,68763303756713867	844863	3	1
881949238971781122	885249287961153538	0,76158803701400757	786793	3	1
881949238971781122	886380149645746177	0,69573312997817993	1056411	3	1
885249287961153538	885492852930772992	0,96297407150268555	58070	3	1
885249287961153538	887040889100673026	0,629341721534729	427151	3	1
885249287961153538	886380149645746177	0,71203207969665527	269618	3	1
885492852930772992	886380149645746177	0,65758275985717773	211548	3	1
840367819925651456	840570763031789569	0,99665969610214233	48385	3	1
839962511592026118	840367819925651456	0,44870084524154663	96633	2	1
839962511592026118	840570763031789569	0,45625588297843933	145018	2	1
821063367091748864	825749447959261184	0,37964224815368652	1117249	2	0
821063367091748864	910509993136394240	0,35618934035301208	21322138	2	1
821063367091748864	944304779177742336	0,31544092297554016	29383043	2	1
821063367091748864	910243446606254080	0,39260220527648926	21258588	2	1
825749447959261184	910509993136394240	0,57907384634017944	20204889	2	0
825749447959261184	910243446606254080	0,67041182518005371	20141339	2	0

Σχήμα 23: Δεδομένα ομοιότητας στη βάση δεδομένων

6.2.3.3 Υπολογισμός της μεταβλητότητας των τιμών του Bitcoin

Στη συνέχεια δημιουργήσαμε ένα View στη βάση δεδομένων στο οποίο υπολογίζεται και απεικονίζεται η μεταβλητότητα του Bitcoin ανά ημέρα. Η μορφή των δεδομένων στο

συγκεκριμένο View απεικονίζεται στο παρακάτω σχήμα:

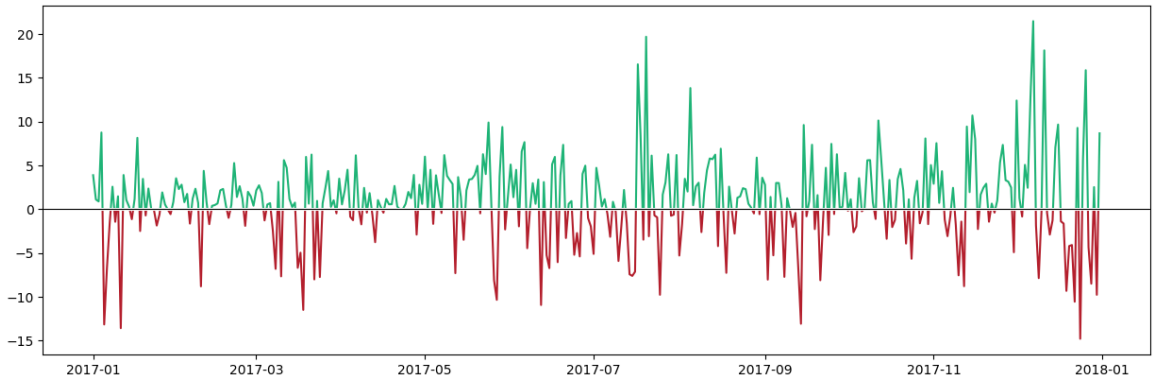
openValue	DateTimeValue	DateInt	closeValue	maxValue	minValue	Volatility
964,93	2017-01-01 00:00:00.000	42734	1002,50	1003,67	952,47	3,893546682142746
1002,50	2017-01-02 00:00:00.000	42735	1013,67	1034,28	992,96	1,114214463840399
1013,67	2017-01-03 00:00:00.000	42736	1022,74	1029,76	1008,70	0,894768514408042
1022,74	2017-01-04 00:00:00.000	42737	1112,41	1143,65	1022,18	8,767624225120754
1112,41	2017-01-05 00:00:00.000	42738	966,03	1157,70	875,51	-13,158817342526586
966,03	2017-01-06 00:00:00.000	42739	901,30	1028,49	855,09	-6,700620063559103
901,30	2017-01-07 00:00:00.000	42740	887,55	904,76	805,10	-1,525574170642405
887,55	2017-01-08 00:00:00.000	42741	910,41	939,59	877,17	2,575629541997634
910,41	2017-01-09 00:00:00.000	42742	897,39	910,95	871,69	-1,430124888786371
897,39	2017-01-10 00:00:00.000	42743	910,68	916,38	884,71	1,480961454885836
910,68	2017-01-11 00:00:00.000	42744	786,99	917,03	752,45	-13,582158387139281
786,99	2017-01-12 00:00:00.000	42745	817,79	822,27	742,95	3,913645662587835
817,79	2017-01-13 00:00:00.000	42746	826,47	832,80	771,90	1,061397180205187
826,47	2017-01-14 00:00:00.000	42747	828,99	839,91	814,42	0,304911249047152
828,99	2017-01-15 00:00:00.000	42748	819,58	828,57	808,95	-1,135116225768706
819,58	2017-01-16 00:00:00.000	42749	829,16	836,17	814,18	1,168891383391493

Σχήμα 24: Δείγμα της συλλογής της τιμής του Bitcoin ανά ημέρα

Οι παραπάνω τιμές μπορούν να απεικονιστούν με τη βοήθεια των «κεριών» (candlesticks), τα οποία χρησιμοποιούνται για την απεικόνιση των τιμών των μετοχών και άλλων εμπορευμάτων και βοηθούν στην ευκολότερη κατανόηση της μεταβλητότητας αλλά ταυτόχρονα απεικονίζονται και οι μέγιστες και ελάχιστες τιμές στην διάρκεια της ημέρας. Οι ημέρες που είναι χρωματισμένες με κόκκινο χρώμα υποδεικνύουν πτώση στην τιμή, ενώ αντιστοίχως οι ημέρες με πράσινο χρώμα υποδεικνύουν άνοδο. Παρακάτω παραθέτουμε το αντίστοιχο διάγραμμα του Bitcoin για το έτος 2017:

Σχήμα 25: Απεικόνιση της ημερήσιας μεταβλητότητας με "κεριά"

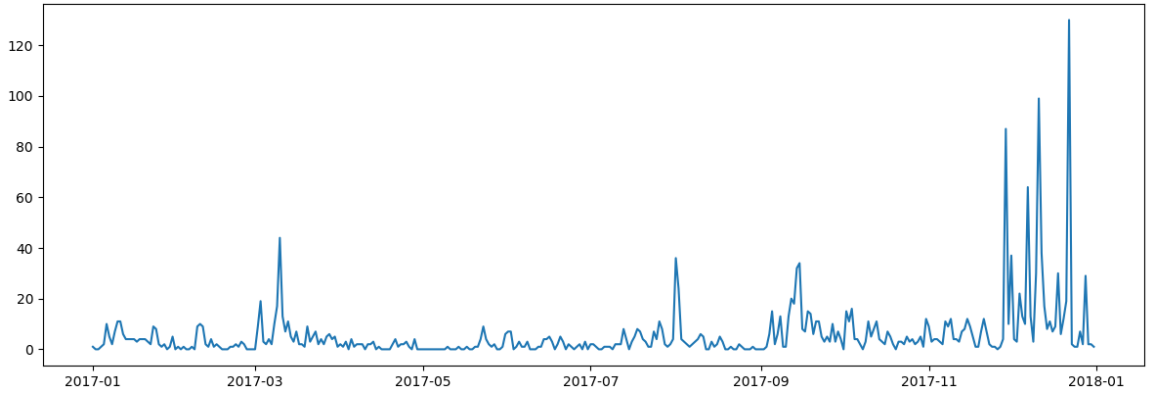
Το παρακάτω διάγραμμα αντιστοιχεί στην μεταβλητότητα του Bitcoin το 2017:



Σχήμα 26: Μεταβλητότητα Bitcoin για το 2017

Παρατηρούμε ότι κατά τη διάρκεια του 2017 η τιμή του Bitcoin είχε μεγάλες μεταβολές που έφταναν σε άνοδο το 20% και σε πτώση το -14%.

Εν συνεχεία, αναλύσαμε τις μεταβολές της τιμής του Bitcoin με τα γεγονότα που έχουμε ήδη ομαδοποιήσει, για να διαπιστώσουμε αν κάποια από αυτά είχαν πράγματι επίδραση, είτε θετική είτε αρνητική, στην τιμή του κρυπτονομίσματος. Αρχικά απεικονίσαμε τα γεγονότα σε ένα γράφημα που δείχνει το πλήθος των ομαδοποιημένων Tweets ανά ημέρα.



Σχήμα 27: Πλήθος ομαδοποιημένων Tweets ανά ημέρα

Κατόπιν παρατηρώντας τις ημέρες που έχουμε έξαρση πολλών Tweets διαπιστώσαμε πως η τιμή του κρυπτονομίσματος είναι ιδιαίτερα ευαίσθητη στην ανακοίνωση και δημοσίευση νέων που αφορούν την λειτουργία του, τις αποφάσεις κρατών και τραπεζών σχετικά με τους κανονισμούς που θα επιβληθούν για την ρύθμισή του, τις αποφάσεις επενδυτικών οίκων και οίκων αξιολόγησης σχετικά με τις δυνατότητες ανταλλαγής και επένδυσης στο Bitcoin, καθώς επίσης και στην αναδημοσίευση απόψεων ανθρώπων της αγοράς και της τεχνολογίας αναφορικά με το κρυπτονόμισμα.

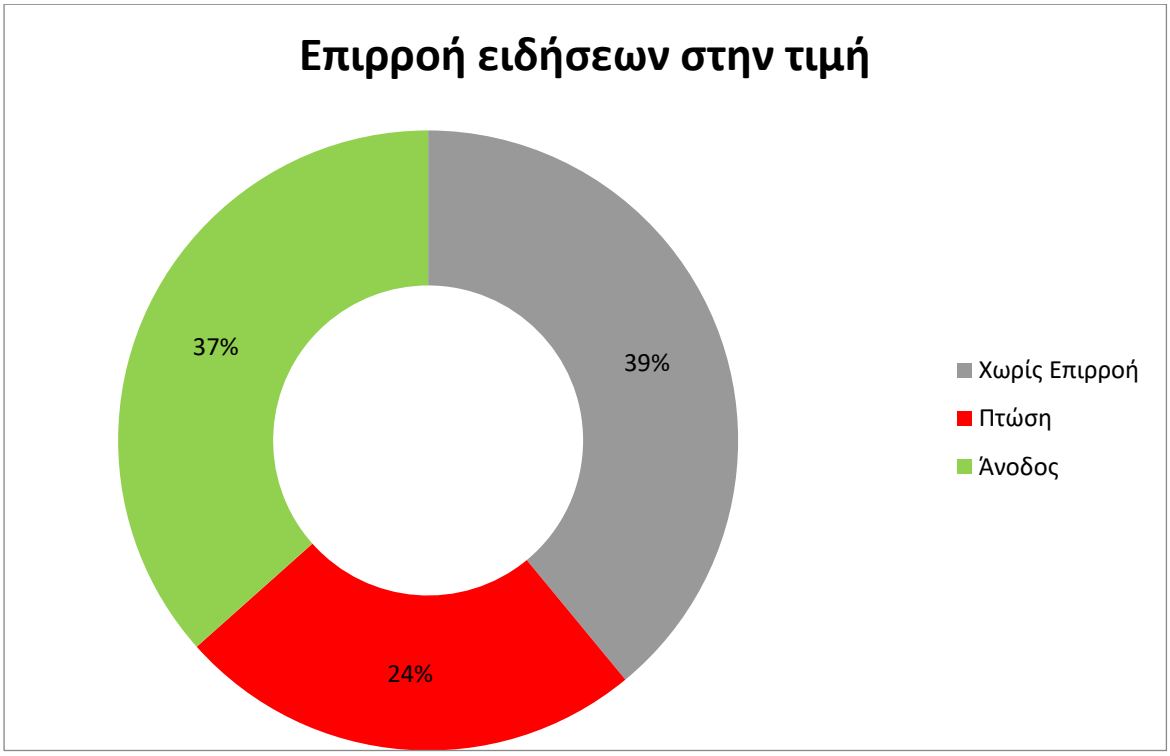
Παρακάτω θα παρουσιάσουμε σε ποιά από τα γεγονότα που εντοπίσαμε υπήρξε σημαντική μεταβολή στην τιμή του Bitcoin:

EventID	Topic	Volatility
9	11/01. Anxieties over China crackdown [19]	**-13,58%**
34	09/02. Chinese Central Bank Warns Bitcoin Exchanges [20]	**-8,80%**
51	03/03. Bitcoin is extending its lead over gold [21]	1,85%
10	10/03. SEC denies permission for a bitcoin ETF [22]	**-7,66%**
50	16/03. Bloomberg - Someone wants to stick a fork in bitcoin [23]	-6,69%
11	01/04. Japan declares bitcoin legal currency [24]	2,15%
29	23/05. Bitcoin Scaling Consensus, debate and Segwit2x [25]	**9,91%**
31	23/05. Japanese Budget Airlines Will Accept Bitcoin In 2017, Install BTMs At Airports [26]	-
28	31/05. Chinese Exchanges Resume Withdrawals, Bitcoin Likely to Surge [27]	-
49	13/07. Morgan Stanley says investors shouldn't buy Bitcoin. [28]	-
43	26/07. Greek police arrest Russian man who US says laundered $4 billion through bitcoin [29]	-
12	01/08. The big bitcoin split [30]	**-5,29%**
27	11/08. Institutional Investors Can No Longer Ignore Bitcoin: Goldman Sachs [31]	-
13	04/09. China bans ICOs [32]	**-5,26%**
15	12/09. Jamie Dimon Bitcoin Statement [33]	-
17	14/09. Bitcoin Exchange BTCChina will Close by October [34]	**-13,08%**
40	15/09. BREAKING: China's Bitcoin Exchanges Receive Shutdown Orders and Closure [35]	**9,62%**
41	28/09. Japan to regulate #Bitcoin [36]	-
33	02/10. Goldman Sachs is weighing a new trading operation dedicated to bitcoin, other digital currencies [37]	1,14%
44	04/10. Greek court rules to extradite Russian bitcoin fraud suspect to the United States. [38]	-
24	10/10. Russia Rejects Cryptocurrency as Authorities Block Access to Exchanges [39]	-
26	01/11. Bitcoin is skyrocketing because 2 of the biggest exchange groups in the world are launching bitcoin futures — here's what that means [40]	**7,55%**
42	01/11. South Korea to regulate #Bitcoin as commodity [41]	**7,55%**
20	08/11. Segwit2x Fork Cancelled [42]	2,45%
36	20/11. CME Group Plans to Launch Bitcoin Futures [43]	2,92%
23	21/11. Tether Allegedly Hacked For $30 Mln [44]	-
46	29/11. Bitcoin rises above $10,000 for the first time [45]	-
21	29/11. BREAKING NEWS – NASDAQ TO LAUNCH BITCOIN FUTURES IN 2018 [46]	-
37	01/12 BREAKING: Bitcoin futures will be allowed to start	**12,42%**

	trading [47]	
14	01/12. CFTC approves bitcoin futures [48]	**12,42%**
38	06/12. Big banks push back on launch of bitcoin futures [49]	**12,68%**
48	06/1. Bitcoin mining service NiceHash says hackers emptied its wallet [50]	-
47	07/12. Bitcoin climbs past $US14,000 as demand continues to skyrocket [51]	-
39	11/12. Bitcoin futures are about to go live, and they could change the game for cryptocurrencies [52]	**18,15%**
32	11/12. Bitcoin Markets Really Like CBOE Futures, Prices Spike Sharply [53]	**18,15%**
16	15/12. Bitcoin futures are about to get another big boost [54]	**7,06%**
30	20/12. North Korea likely behind a massive cyber attack on a South Korean bitcoin exchange that caused it to collapse [55]	**-9,32%**
25	22/12. The Sharks Are Beginning to Circle Bitcoin and It's Down 30% [56]	**-10,56%**
22	28/12. Bitcoin falls as South Korea announces crackdown - as it happened [57]	**-8,50%**
45	Nigeria Regulations (topic) [58]	-
35	Terrorism, Money Laundering, Drugs (Topic) [59]	-

Πίνακας 3: Παρουσίαση μεταβλητότητας σε συνδυασμό με τα γεγονότα

Όπως εύκολα μπορεί να διαπιστώσει κάποιος, τα περισσότερα από τα γεγονότα που εντοπίσαμε, συγκεκριμένα το 61% αυτών, έπαιξαν σημαντικό ρόλο στην μεταβολή της τιμής του κρυπτονομίσματος. Το παρακάτω διάγραμμα μας επιβεβαιώνει με ακρίβεια τον συλλογισμό αυτόν.



Σχήμα 28: Επιρροή ειδήσεων στην τιμή του Bitcoin

6.2.4 Ομαδοποίηση Tweets με χρήση του αλγορίθμου K-Means

Με την ολοκλήρωση της ομαδοποίησης του Dataset, στο προηγούμενο στάδιο, αυτό που χρειαζόταν να υλοποιήσουμε ήταν ένας αλγόριθμος μηχανικής μάθησης ο οποίος θα μπορούσε μόνος του να ομαδοποιήσει τα Tweets που είχαμε συλλέξει. Για την υλοποίηση αυτό του βήματος χρησιμοποιήσαμε τον αλγόριθμο μη επιτηρούμενης μάθησης **K-Means**.

6.2.4.1 Περιγραφή Αλγορίθμου

Έχοντας ένα σύνολο παρατηρήσεων, όπου η κάθε παρατήρηση είναι ένα διάνυσμα d-διαστάσεων, ο αλγόριθμος K-Means στοχεύει στο διαχωρισμό του συνόλου n παρατηρήσεων σε k(< n) συλλογών **S** = {S_1, S_2, ..., S_k} έτσι ώστε να ελαχιστοποιηθεί το άθροισμα των τετραγώνων των αποστάσεων μέσα στη συλλογή. [62]

Τυπικά, ο στόχος είναι να βρεθεί:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}\in S_i} \|\mathbf{x}-\boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i|\,\mathrm{Var}\,S_i$$

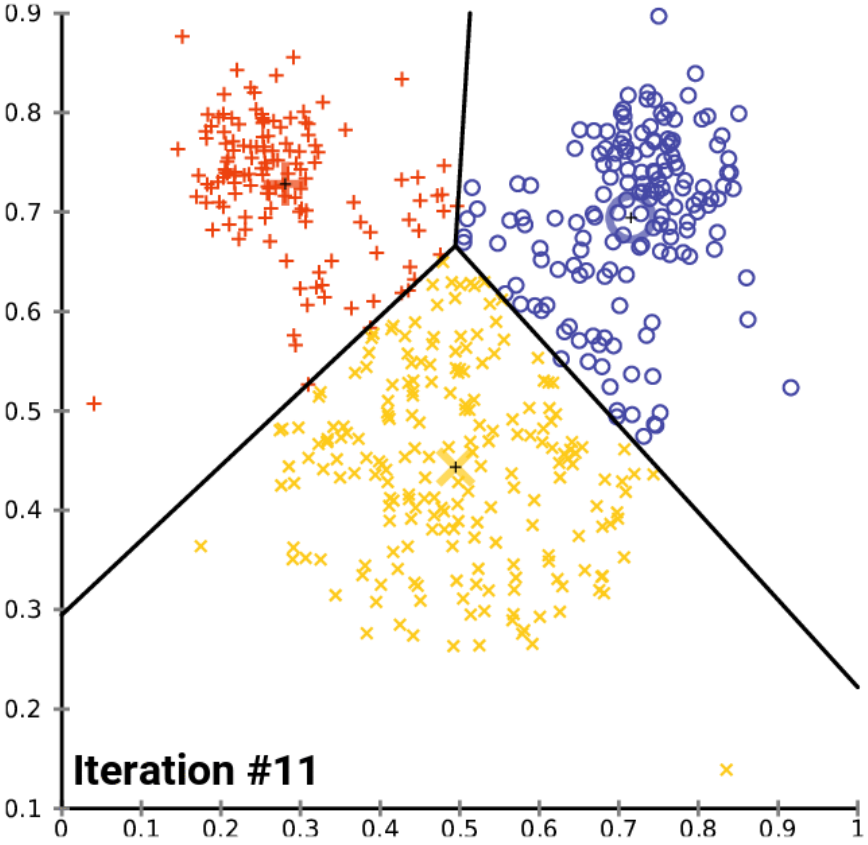
όπου μ_i είναι ο μέσος των σημείων στη συλλογή S_i. Αυτό ισοδυναμεί με την ελαχιστοποίηση των ζευγών τετραγωνικών αποκλίσεων σημείων στην ίδια συλλογή:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \frac{1}{2|S_i|} \sum_{\mathbf{x},\mathbf{y}\in S_i} \|\mathbf{x}-\mathbf{y}\|^2$$

Το ισοδύναμο μπορεί να συναχθεί από την ταυτότητα:

$$\sum_{\mathbf{x}\in S_i} \|\mathbf{x}-\boldsymbol{\mu}_i\|^2 = \sum_{\mathbf{x}\neq\mathbf{y}\in S_i} (\mathbf{x}-\boldsymbol{\mu}_i)(\boldsymbol{\mu}_i-\mathbf{y})$$

Επειδή η συνολική διασπορά είναι σταθερή, αυτό επίσης ισοδυναμεί με την μεγιστοποίηση το αθροίσματος των τετραγωνικών αποκλίσεων μεταξύ των σημείων διαφορετικών συλλογών.



Σχήμα 29: Διαδικασία διαχωρισμού παρατηρήσεων

6.2.4.2 Υλοποίηση Αλγορίθμου

Για την υλοποίηση του αλγορίθμου και την αξιολόγηση του, χρησιμοποιήσαμε την βιβλιοθήκη scikit της Python, η οποία περιλαμβάνει τον αλγόριθμο clustering K-Means. Οι παρατηρήσεις μας ήταν τα ζεύγη ομοιότητας που είχαμε υπολογίσει στο προηγούμενο βήμα, επομένως κάθε παρατήρηση ήταν ένα διάνυσμα 2035 διαστάσεων και οι παρατηρήσεις 2035 και αυτές. Η ομοιότητα κάθε ζεύγους αποτελούσε το άθροισμα της ομοιότητας κειμένου, της χρονικής ομοιότητας, της ομοιότητας των hashtags και του αρθογράφου. Η συνάρτηση που υπολόγιζε την τιμή αυτή για κάθε ζεύγος δίνεται παρακάτω:

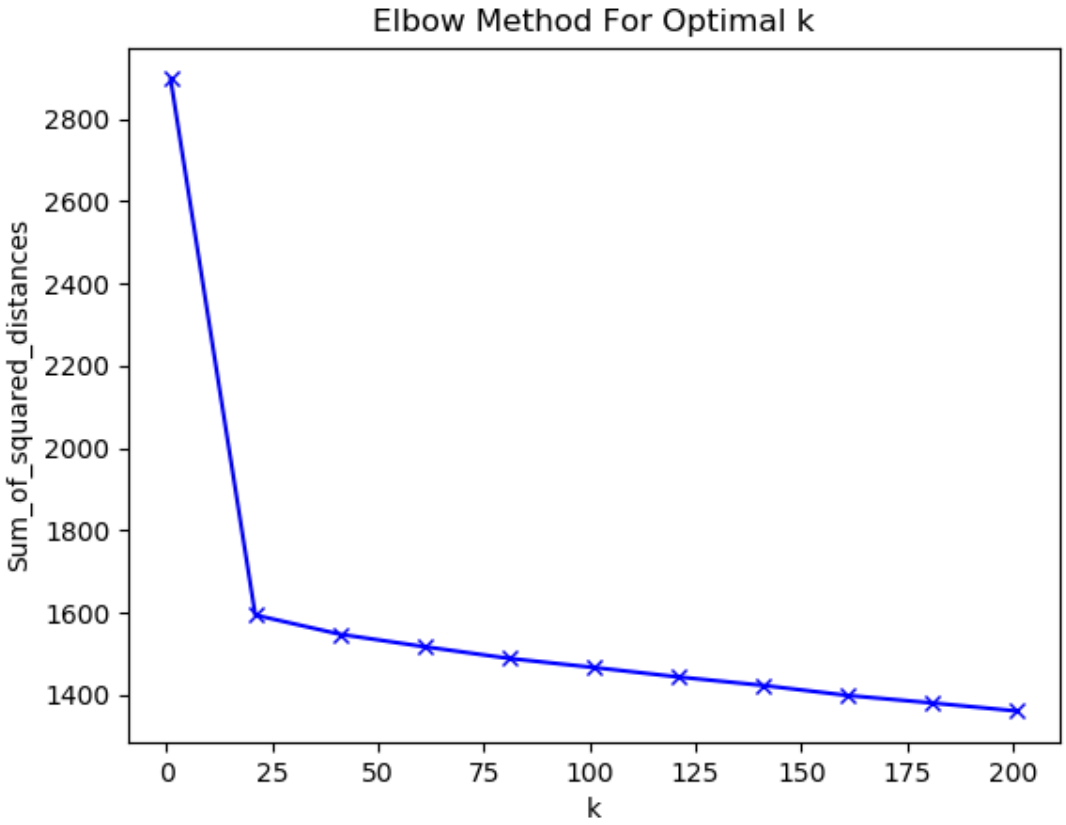
$$sim(m_u, m_v) = a_1 \cdot sim_{author}(m_u, m_v) + a_2 \cdot sim_{hashtag}(m_u, m_v) + a_3 \cdot sim_{text} + a_4 \cdot f_t(t_u, t_v)$$

όπου τα βάρη α_i θα πρέπει να βελτιστοποιηθούν κατά τη διαδικασία εκπαίδευσης του συστήματος.

Όπως και στον αλγόριθμο για τον καθορισμό των νοηματικών ενοτήτων, έτσι και στην περίπτωση αυτού του αλγορίθμου, το πρόβλημα που αντιμετωπίσαμε είναι ότι απαιτεί να

γνωρίζεις από πριν τον αριθμό των clusters. Για να αντιμετωπίσουμε το πρόβλημα αυτό εργαστήκαμε με παρόμοιο τρόπο όπως και στον αλγόριθμο ομαδοποίησης κειμένων. Πιο αναλυτικά διατρέξαμε τον αλγόριθμο για διαφορετικό αριθμό clusters και υπολογίσαμε το άθροισμα των τετραγώνων των αποστάσεων. Εν συνεχεία, δημιουργήσαμε την γραφική παράσταση των τιμών αυτών και καταλήξαμε στο συμπέρασμα πως ο ιδανικός αριθμός clusters, βάσει των δεδομένων του αλγορίθμου, είναι 35. Ο αριθμός αυτός είναι πολύ κοντά στον αριθμό που είχε προκύψει με την χειροκίνητη ομαδοποίηση, ο οποίος ήταν 41. [61]

Παρακάτω παραθέτουμε το διάγραμμα που προέκυψε μετά το τέλος των υπολογισμών:



Σχήμα 30: Διάγραμμα υπολογισμού βέλτιστου αριθμού Clusters για τον KMeans

Η απόκλιση αυτή στον αριθμό των clusters οφείλεται στο γεγονός ότι πολλά γεγονότα είναι συναφή μεταξύ τους όσον αφορά τη χρονική και τη νοηματική απόκλιση.

6.2.4.3 Αξιολόγηση Αλγορίθμου

Για την αξιολόγηση του αλγορίθμου συγκρίναμε τα αποτελέσματα της ομαδοποίησης του αλγορίθμου με την χειροκίνητη ομαδοποίηση που είχαμε υλοποιήσει, χρησιμοποιώντας τις παρακάτω μετρήσεις:

- Normalized Mutual Information: [63]

$$NMI(\Omega, C) = \frac{I(\Omega,C)}{[H(\Omega)+H(C)]/2}$$

όπου Ω = {ω1, ω2,…, ωκ} είναι το σύνολο των clusters, C = {c1,c2,…,ck} είναι το σύνολο των κλάσεων, $I(\Omega, C)$ είναι αμοιβαία πληροφορία μεταξύ του Ω και του C, H(Ω) και H(C) είναι οι εντροπίες του Ω και του C αντιστοίχως.

- Η μέτρηση F1-Measure υπολογίζεται από το Precision και το Recall από τον τύπο: [64]

$$\text{F1-Measure} = 2 * \frac{Precision*Recall}{Precision+Recall},$$

όπου το Precision και Recall επεξηγούνται παρακάτω.

Μία αληθώς θετική (TP) απόφαση αναθέτει δύο όμοια αντικείμενα στο ίδιο cluster, μία αληθώς αρνητική (TN) απόφαση αναθέτει δύο ανόμοια αντικείμενα σε διαφορετικά cluster. Υπάρχουν δύο ειδών λάθη που μπορούν να συμβούν. Μία ψευδώς θετική (FP) απόφαση αναθέτει δύο ανόμοια αντικείμενα στο ίδιο cluster, ενώ μία ψευδώς αρνητική (FN) απόφαση αναθέτει δύο όμοια αντικείμενα σε διαφορετικά clusters.

$$Precision = \frac{TP}{TP+FP} \qquad Recall = \frac{TP}{TP+FN} \text{ [65]}$$

Παρακάτω παρουσιάζουμε την απόδοση του αλγορίθμου για τις διαφορετικές τιμές του χρονικού παραθύρου q:

Παράθυρο q (σε ημέρες)	**NMI**	**F1-Measure**
15	0,7283	0,5465
10	0,7625	0,5753
5	**0,8227**	**0,6226**
1	0,6941	0,5372

Πίνακας 4: Πίνακας αξιολόγησης του αλγόριθμου KMeans

Παρατηρούμε ότι τα βέλτιστα αποτελέσματα ομαδοποίησης του dataset μας τα έχουμε όταν q=5 μέρες. Αυτό συμβαίνει διότι στο συγκεκριμένο dataset η εμφάνιση των νέων γίνεται περίπου με συχνότητα μίας εβδομάδος.

6.2.5 Πρόβλεψη ημερήσιας μεταβλητότητας με χρήση Νευρωνικού Δικτύου

Στο τελευταίο στάδιο της συγκεκριμένης διπλωματικής εργασίας είχαμε ως στόχο να προβλέψουμε τη μεταβολή της τιμής του Bitcoin σε επίπεδο ημέρας. Για να επιτύχουμε το στόχο μας υλοποιήσαμε ένα νευρωνικό δίκτυο δύο επιπέδων, το οποίο δέχεται ως είσοδο τα Tweets και στην έξοδο του επιστρέφει αν θα ανέβει ή θα πέσει η τιμή του Bitcoin.

6.2.5.1 Περιγραφή Αλγορίθμου

Η ιδέα, όπως περιγράψαμε και προηγουμένως, ήταν να δίνουμε τα Tweets ως είσοδο στο νευρωνικό δίκτυο και εκείνο να μας επιστρέφει αν θα ανέβει ή θα πέσει η τιμή του. Για να το επιτύχουμε αυτό δημιουργήσαμε έναν πίνακα 365 γραμμών, όσες δηλαδή και οι ημέρες του έτους 2017, και σε κάθε στήλη αναθέσαμε μία λέξη από το σύνολο των λέξεων που υπήρχαν στο Dataset μας. Επομένως, οι διαστάσεις του πίνακα αυτού ήταν 365 X #Αριθμός_λέξεων_Dataset. Η αξία της κάθε τιμής του πίνακα υπολογιζόταν από την παρακάτω σχέση:

$$f(d,w) = \sum_{t=1}^{tweets} \big(\mathrm{g(w,t,d)}\big)$$

όπου d είναι η μέρα που εξετάζουμε, επομένως d ∈ [1, 365], και w είναι ο αριθμός της στήλης που αντιστοιχεί η λέξη που εξετάζουμε, με g(w,t,d) να ισούται με:

$$g(w,t,d) \propto \begin{cases} e^{-(d-t.date)/q}, & if\ t.date \le d \\ 0, & if\ t.date > d \end{cases}$$

Η τιμή της συνάρτησης λαμβάνει την μέγιστη τιμή της όταν $d \approx t.date$ και φθίνει στο μηδέν, με ρυθμό που καθορίζει η παράμετρος q και όσο η διαφορά $(d - t.date)$ αυξάνεται.

Με τον τρόπο αυτό γνωρίζαμε ποιες λέξεις επικρατούσαν ανά ημέρα και σε ποιες υπήρχε μεγάλη έξαρση. Ο παρακάτω πίνακας αποτελεί παράδειγμα, τα νούμερα είναι ενδεικτικά, για την καλύτερη κατανόηση της συγκεκριμένης τεχνικής:

	china	warn	bank	futures	cboe	cme	contract
9-Ιαν	0	0	0	0	0	0	0
10-Ιαν	5	3	8	0	0	0	0
11-Ιαν	90	80	110	0	0	0	0
12-Ιαν	80	75	108	0	0	0	0
13-Ιαν	30	15	22	0	0	0	0
…							
12-Δεκ	0,005	0,005	0,005	0	0	0	0
13-Δεκ	0,005	0,005	3	10	4	3	8
14-Δεκ	0,005	0,005	150	350	142	187	250
15-Δεκ	0,005	0,005	120	180	60	70	132
16-Δεκ	0,005	0,005	40	35	20	22	50

Πίνακας 5: Παράδειγμα πίνακα εισόδου Νευρωνικού Δικτύου

Κάθε νευρωνικό δίκτυο εκτός από τον πίνακα με τα παραδείγματα, χρειάζεται και έναν πίνακα με τα αποτελέσματα για να εκπαιδευτεί και στη συνέχεια να αξιολογηθεί βάσει αυτών. Επειδή, όμως, είναι αλγόριθμος classification επιτηρούμενης μάθησης, ο πίνακας αποτελεσμάτων περιείχε την τιμή 1 για κάθε ημέρα που η τιμή του Bitcoin παρουσίασε άνοδο και μηδέν για τις ημέρες που η τιμή του παρουσίασε πτώση. Η παρακάτω συνάρτηση υπολογίζει την τιμή αυτή συναρτήσει της ημέρας:

$$f(d) \propto \begin{cases} 1, & if\ p[d] > 0 \\ 0, & if\ p[d] \leq 0 \end{cases}$$

όπου p είναι ο πίνακας που περιέχει τις ημερήσιες μεταβολές του Bitcoin ανά ημέρα.

6.2.5.2 Υλοποίηση Αλγορίθμου

Για την υλοποίηση του αλγορίθμου και την αξιολόγηση του χρησιμοποιήσαμε και σε αυτήν την περίπτωση την βιβλιοθήκη scikit της Python, η οποία περιέχει αλγορίθμους νευρωνικών δικτύων. Για την συμπλήρωση του πίνακα εισόδου χρησιμοποιήσαμε τις λέξεις όπως τις είχαμε επεξεργαστεί στο στάδιο της προ-επεξεργασίας δεδομένων για να έχουμε καλύτερα και ακριβέστερα αποτελέσματα.

Πιο αναλυτικά, οι μοναδικές λέξεις που είχαμε πριν από το στάδιο της προ-επεξεργασίας των Tweets ήταν 13.852 ενώ μετά το τέλος της προ-επεξεργασίας μειώθηκαν σε 3.153 λέξεις. Αυτό οδήγησε στη βελτίωση της απόδοσης καθώς επίσης και στην ακρίβειας του αλγορίθμου. Η κάθε μία από αυτές τις λέξεις, όπως αναφέραμε και προηγουμένως, αποτελεί ένα διαφορετικό «χαρακτηριστικό» (attribute) του συστήματός μας και το τελικό πλήθος τους μας οδήγησε στην επιλογή των κρυφών επιπέδων του νευρωνικού. Καταλήξαμε ότι το πλήθος τους θα πρέπει να κυμαίνεται μεταξύ 500 και 1000 για κάθε διαφορετικό επίπεδο. Το πλήθος των επιπέδων επιλέξαμε να είναι δύο.

Αρχικά εφαρμόσαμε τον αλγόριθμο σε όλο το πλήθος των Tweets, ενώ στη συνέχεια επιλέξαμε μόνο τα Tweets που είχαμε κατατάξει σε κάποιο γεγονός. Με αυτόν τον τρόπο ελέγξαμε την απόδοση του συστήματος στο θόρυβο, δηλαδή την απόδοσή του όταν υπάρχουν Tweets τα οποία δεν επηρεάζουν με κάποιον τρόπο την μεταβολή της τιμής, όπως για παράδειγμα Tweets τα οποία μπορεί να περιλαμβάνουν διαφημίσεις.

Για την υλοποίηση του νευρωνικού δικτύου, όπως και για την ανάπτυξη όλων των υπόλοιπων αλγορίθμων, κάναμε χρήση της βιβλιοθήκης της Python, scikit.

6.2.5.3 Αξιολόγηση Αλγορίθμου

Για την εκπαίδευση του συγκεκριμένου συστήματος χρησιμοποιήσαμε τα Tweets των πρώτων 328 ημερών του έτους και για την αξιολόγηση του τις υπόλοιπες 37, στις οποίες όμως παρουσιάστηκε το μεγαλύτερο πλήθος Tweets. Αυτό σημαίνει πως για το συνολικό Dataset χρησιμοποιήσαμε το 72% των Tweets για εκπαίδευση του συστήματος και το υπόλοιπο 28% για την αξιολόγηση του, ενώ για το Annotated Dataset χρησιμοποιήσαμε το 62% για εκπαίδευση του νευρωνικού δικτύου και το υπόλοιπο 38% για την αξιολόγηση του.

Παρακάτω γίνεται παρουσίαση των αποτελεσμάτων για κάθε διαφορετική σχεδίαση των κρυφών επιπέδων του νευρωνικού δικτύου:

Επίπεδο 1	**Επίπεδο 2**	**Precision**	**Recall**	**F1-Score**
500	-	0,71	0,71	0,71
1000	500	0,78	0,78	0,77
1000	**1000**	**0,81**	**0,79**	**0,78**

Πίνακας 6: Αξιολόγηση αλγορίθμου για το σύνολο του Dataset

Επίπεδο 1	**Επίπεδο 2**	**Precision**	**Recall**	**F1-Score**
250	-	0,75	0,56	0,68
250	250	0,78	0,78	0,72
500	**500**	**0,83**	**0,82**	**0,82**

Πίνακας 7: Αξιολόγηση αλγορίθμου για το Annotated Dataset

Όπως μπορούμε να δούμε και από τα αποτελέσματα της αξιολόγησης, στην περίπτωση που χρησιμοποιήσαμε μόνο τα ομαδοποιημένα Tweets η διαφορά στην βελτίωση των αποτελεσμάτων είναι ελάχιστη. Αυτό σημαίνει πως το νευρωνικό δίκτυο έχει πολύ καλή συμπεριφορά στον θόρυβο που δημιουργούν Tweets που δεν επηρεάζουν άμεσα την μεταβολή της τιμής του Bitcoin.

7 Επίλογος

7.1 Σύνοψη και Συμπεράσματα

Στην εργασία αυτή μελετήσαμε διεξοδικά τις δυνατότητες των αλγορίθμων μηχανικής μάθησης, τόσο επιβλεπόμενης όσο και μη επιβλεπόμενης, και τις διάφορες εφαρμογές που μπορεί οι αλγόριθμοι αυτοί να έχουν. Πιο αναλυτικά, μελετήσαμε αρχικά αν μπορούμε να ομαδοποιήσουμε νέα εφημερίδων, σχετικά με το κρυπτονόμισμα Bitcoin, που δημοσιεύονται στο μέσο κοινωνικής δικτύωσης Twitter. Στη συνέχεια με τα ίδια δεδομένα προσπαθήσαμε να προβλέψουμε την μεταβολή της τιμής του συγκεκριμένου κρυπτονομίσματος.

Για την εκπόνηση της πρώτης εργασίας, κάναμε χρήση αλγορίθμων μη επιβλεπόμενης μηχανικής μάθησης και για να μπορέσουμε να αξιολογήσουμε την απόδοση των αλγορίθμων αυτών ομαδοποιήσαμε πρώτα χειροκίνητα μεγάλο μέρος του Dataset. Με το τέλος της αξιολόγησης του συστήματος αυτού οδηγηθήκαμε στο συμπέρασμα ότι μπορούμε να κατηγοριοποιήσουμε και να εντοπίσουμε νέα και ειδήσεις που δημοσιεύονται στο Twitter με πολύ μεγάλη επιτυχία.

Το δεύτερο σκέλος της εργασίας αυτής, περιελάμβανε την πρόβλεψη της μεταβλητότητας της τιμής του Bitcoin, συναρτήσει των νέων, των ειδήσεων και των εξελίξεων που δημοσιεύουν οι μεγαλύτερες ηλεκτρονικές εφημερίδες. Για την επίτευξη του στόχου αυτού χρησιμοποιήσαμε αλγορίθμους επιβλεπόμενης μηχανικής μάθησης και συγκεκριμένα Νευρωνικά Δίκτυα. Με την κατάλληλη επεξεργασία των δεδομένων καταφέραμε να δημιουργήσουμε ένα σύστημα το οποίο μπορεί και προβλέπει με μεγάλη ακρίβεια την μεταβλητότητα του κρυπτονομίσματος.

7.2 Μελλοντικές προεκτάσεις

Πραγματοποιήσαμε μία διεξοδική έρευνα γύρω από τη μεταβλητότητα της τιμής του Bitcoin συναρτήσει των γεγονότων που το αφορούν. Οδηγηθήκαμε στο συμπέρασμα πως, όπως είναι αναμενόμενο, η τιμή του επηρεάζεται άμεσα από έκτατες αποφάσεις που λαμβάνονται και από γεγονότα που συμβαίνουν ανά τον κόσμο και αφορούν το κρυπτονόμισμα. Με βάση το συμπέρασμα αυτό μπορούμε να χρησιμοποιήσουμε και άλλους οδηγούς που μπορεί να επηρεάζουν την τιμή του, όπως είναι για παράδειγμα η μεταβλητότητα των χρηματιστηρίων των μεγαλύτερων οικονομιών του πλανήτη.

Επιπρόσθετα, μπορούμε να επεκτείνουμε την έρευνα αυτή και στα χρηματιστηριακά προϊόντα, όπως μετοχές εταιριών, συναλλάγματα και εμπορεύματα. Μια τέτοια μελέτη θα μπορούσε να οδηγήσει στην δημιουργία ενός συστήματος έγκαιρης και έγκυρης πρόβλεψης της τιμής των προϊόντων αυτών.Η έρευνα αυτή θα περιλαμβάνει και την μελέτη γεγονότων και Tweets που δεν αφορούν μόνο τα χρηματιστηριακά προϊόντα αλλά έχουν και κοινωνικοπολιτικές επεκτάσεις, όπως για παράδειγμα ένα Tweet ενός παγκόσμιου ηγέτη.

Τέλος, η μελέτη που αφορά την ομαδοποίηση νέων και τον εντοπισμό έκτακτων γεγονότων θα μπορούσε να έχει πολλαπλές εφαρμογές εκτός των χρηματιστηριακών προϊόντων. Πιο συγκεκριμένα, ένα τέτοιο σύστημα θα μπορούσε να συνεισφέρει τα μέγιστα σε υπηρεσίες που αφορούν την υγεία και την ασφάλεια των πολιτών. Για παράδειγμα, στην αντιμετώπιση μιας πιθανής επιδημίας στην έναρξή της, ή στην αντιμετώπιση μιας πυρκαγιάς πριν επεκταθεί και απειλήσει κατοικίες και ανθρώπινες ζωές.

Βιβλιογραφία

[1] «Distributed Databases,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Distributed_database.

[2] «Blockchain,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Blockchain.

[3] «Satoshi Nakamoto,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Satoshi_Nakamoto.

[4] «Global Cryptocurrency Benchmarking Study,» [Ηλεκτρονικό]. Available: https://www.jbs.cam.ac.uk/fileadmin/user_upload/research/centres/alternative-finance/downloads/2017-global-cryptocurrency-benchmarking-study.pdf.

[5] «'Only good for drug dealers': More Nobel prize winners snub bitcoin,» [Ηλεκτρονικό]. Available: https://finance.yahoo.com/news/good-drug-dealers-nobel-prize-winners-snub-bitcoin-184903784.html?guccounter=1.

[6] «Silk Road (marketplace),» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Silk_Road_(marketplace).

[7] «The Bitcoin Boom,» [Ηλεκτρονικό]. Available: https://www.newyorker.com/tech/elements/the-bitcoin-boom.

[8] T. Mitchell, «Machine Learning,» 1997.

[9] S. Harnad, «The Annotation Game: On Turing (1950) on Computing, Machinery, and Intelligence,» 2008. [Ηλεκτρονικό]. Available: https://eprints.soton.ac.uk/262954/.

[10] M. Jordan, «statistics and machine learning,» 2014.

[11] F. Tönnies, «Community and Society,» 1887.

[12] J. P. Scott, «Social Network Analysis: A Handbook,» Sage Publications, 2000.

[13] N. Mullins, «heories and Theory Groups in Contemporary American Sociology,» New York: Harper and Row, 1973.

[14] «About Twitter, Inc,» [Ηλεκτρονικό]. Available: https://about.twitter.com/en_us/company.html.

[15] «About Python,» [Ηλεκτρονικό]. Available: https://www.python.org/about/.

[16] «About scikit-learn,» [Ηλεκτρονικό]. Available: http://scikit-learn.org/stable/about.html.

[17] «Tweepy Documentation,» [Ηλεκτρονικό]. Available: https://tweepy.readthedocs.io/en/v3.5.0/.

[18] «Twitter API Docs,» [Ηλεκτρονικό]. Available: https://developer.twitter.com/en/docs.html.

[19] J. Garber, «Bitcoin is getting demolished,» Business Insider, 11 Ιανουάριος 2017. [Ηλεκτρονικό]. Available: https://www.businessinsider.com/bitcoin-price-january-11-2017-2017-1.

[20] «Chinese Central Bank Warns Bitcoin Exchanges,» ccn, 9 Φεβρουάριος 20117. [Ηλεκτρονικό]. Available: https://www.ccn.com/chinese-central-bank-warns-bitcoin-exchanges/.

[21] L. SHEN, «Bitcoin Just Became More Valuable Than Gold,» Fortune, 3 Μάρτιος 2017. [Ηλεκτρονικό]. Available: http://fortune.com/2017/03/03/gold-bitcoin-price-value/.

[22] J. Garber, «Bitcoin dives after the SEC shoots down plans for another bitcoin ETF,» 28 Μάρτιος 2017. [Ηλεκτρονικό]. Available: https://www.businessinsider.com/bitcoin-price-dives-after-sec-rejects-plans-for-etf-2017-3.

[23] E. Ou, «Someone Wants to Stick a Fork in Bitcoin,» bloomberg, 16 Μάρτιος 2017. [Ηλεκτρονικό]. Available: https://www.bloomberg.com/opinion/articles/2017-03-16/someone-wants-to-stick-a-fork-in-bitcoin.

[24] «Japan Accepts Bitcoin as Legal Payment Method. What’s Next?,» CCN, 5 ΑΠΡΙΛΙΟΣ 2017. [Ηλεκτρονικό]. Available: https://www.ccn.com/japan-accepts-bitcoin-as-legal-payment-method-whats-next/.

[25] R. Bova, «BREAKING: Bitcoin Scaling Consensus Reached. Commentary From Industry Leaders,» Cointelegraph, 23 Μάιος 2017. [Ηλεκτρονικό]. Available: https://cointelegraph.com/news/bitcoin-scaling-consensus-reached-commentary-from-industry-leaders.

[26] W. Suberg, «Japanese Budget Airlines Will Accept Bitcoin In 2017, Install BTMs At Airports,» Cointelegraph, 23 Μάιος 2017. [Ηλεκτρονικό]. Available: https://cointelegraph.com/news/japanese-budget-airlines-will-accept-bitcoin-in-2017-install-btms-at-airports.

[27] J. Young, «Breaking: Chinese Exchanges Resume Withdrawals, Bitcoin Likely to Surge,» Cointelegraph, 31 Μάιος 2017. [Ηλεκτρονικό]. Available: https://cointelegraph.com/news/breaking-chinese-exchanges-resume-withdrawals-bitcoin-likely-to-surge.

[28] «Morgan Stanley says investors shouldn't buy Bitcoin.,» Fortune, 13 Ιούλιος 2017. [Ηλεκτρονικό]. Available:

https://twitter.com/FortuneMagazine/status/885300525683683329/photo/1.

[29] C. KANTOURIS, «Russian Wanted by U.S. Over $4 Billion Bitcoin Crime Caught in Greece,» Bloomberg, 26 Ιούλιο 2017. [Ηλεκτρονικό]. Available: https://www.bloomberg.com/news/articles/2017-07-26/russian-wanted-in-us-caught-in-greece-for-money-laundering?cmpid=socialflow-twitter-business&utm_content=business&utm_campaign=socialflow-organic&utm_source=twitter&utm_medium=social.

[30] F. Chaparro, «Bitcoin splits in 2,» Business Insider, 1 Αύγουστος 2017. [Ηλεκτρονικό]. Available: https://www.businessinsider.com/bitcoin-price-fork-happens-2017-8.

[31] J. Young, «Institutional Investors Can No Longer Ignore Bitcoin: Goldman Sachs,» Cointelegraph, 11 Αύγουστος 2017. [Ηλεκτρονικό]. Available: https://cointelegraph.com/news/institutional-investors-can-no-longer-ignore-bitcoin-goldman-sachs.

[32] O. Williams-Grut, «Here's why China's crypto crackdown is 'bigger than most people think',» Business Insider, 4 Σεπτέμβριος 2017. [Ηλεκτρονικό]. Available: https://www.businessinsider.com/initial-coin-offering-china-bitcoin-ethereum-peoples-bank-of-china-law-all-crypto-illegal-etoro-2017-9.

[33] F. Imbert, «JPMorgan CEO Jamie Dimon says bitcoin is a 'fraud' that will eventually blow up,» CNBC, 12 Σεπτέμβριος 2017. [Ηλεκτρονικό]. Available: https://www.cnbc.com/2017/09/12/jpmorgan-ceo-jamie-dimon-raises-flag-on-trading-revenue-sees-20-percent-fall-for-the-third-quarter.html.

[34] W. Suberg, «Breaking: Bitcoin Exchange BTCChina will Close by October,» Cointelegraph, 14 Σεπτέμβριος 2017. [Ηλεκτρονικό]. Available: https://cointelegraph.com/news/breaking-bitcoin-exchange-btcchina-will-close-by-october.

[35] T. C. &. R.-R. O'Leary, «China's Bitcoin Exchanges Receive Shutdown Orders and Closure Timeline,» Coindesk, 15 Σεπτέμβριος 2017. [Ηλεκτρονικό]. Available: https://www.coindesk.com/document-lists-closure-steps-for-chinas-bitcoin-exchanges/?utm_content=bufferd5f92&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer.

[36] J. Young, «While China Bans Bitcoin Exchanges, Japanese Government Embraces Them,» Cointelegraph, 27 Σεπτέμβριος 2017. [Ηλεκτρονικό]. Available: https://cointelegraph.com/news/while-china-bans-bitcoin-exchanges-japanese-government-embraces-them.

[37] T. D. a. L. H. Paul Vigna, «Goldman Sachs Explores a New World: Trading Bitcoin,» Wall Street Journal, 2 Οκτώβριος 2017. [Ηλεκτρονικό]. Available: https://www.wsj.com/articles/goldman-sachs-explores-a-new-world-trading-bitcoin-

1506959128.

[38] C. KANTOURIS, «Greece backs extradition of Russian to US over bitcoin fraud,» Associated Press, 4 Οκτώβριος 2017. [Ηλεκτρονικό]. Available: https://apnews.com/3a5056c0c01f45fbafb9e59f456fc3bc?utm_campaign=SocialFlow&utm_source=Twitter&utm_medium=AP.

[39] W. Suberg, «Breaking: Russia Rejects Cryptocurrency as Authorities Block Access to Exchanges,» Cointelegraph, 10 Οκτώβριος 2017. [Ηλεκτρονικό]. Available: https://cointelegraph.com/news/breaking-russia-rejects-cryptocurrency-as-authorities-block-access-to-exchanges.

[40] F. Chaparro, «Bitcoin is skyrocketing because 2 of the biggest exchange groups in the world are launching bitcoin futures — here's what that means,» Business Insider, 1 Νοέμβριος 2017. [Ηλεκτρονικό]. Available: https://www.businessinsider.com/bitcoin-future-explained.

[41] L. Froelings, «South Korea to Regulate Bitcoin as Commodity, Says Bank of Korea Governor,» Cointelegraph, 29 Οκτώβριος 2017. [Ηλεκτρονικό]. Available: https://cointelegraph.com/news/south-korea-to-regulate-bitcoin-as-commodity-says-bank-of-korea-governor.

[42] S. Haig, «Breaking News: Segwit2x Fork Cancelled,» Bitcoin.com, 8 Νοέμβριος 2107. [Ηλεκτρονικό]. Available: https://news.bitcoin.com/breaking-news-segwit2x-fork-cancelled/.

[43] J. Redman, «CME Group Plans to Launch Bitcoin Futures December 10,» bitcoin.com, 20 Νοέμβριος 2017. [Ηλεκτρονικό]. Available: https://news.bitcoin.com/cme-group-plans-to-launch-bitcoin-futures-december-10/.

[44] J. Buck, «Breaking: Tether Allegedly Hacked For $30 Mln,» Cointelegraph, 21 Νοέμβριος 2017. [Ηλεκτρονικό]. Available: https://cointelegraph.com/news/breaking-tether-allegedly-hacked-for-30-mln.

[45] G. C.-D. Jemima Kelly, «Bubble trouble? Bitcoin tops $11,000, but fades after sharp rally,» Reuters, 29 Νοέμβριος 2017. [Ηλεκτρονικό]. Available: https://www.reuters.com/article/uk-markets-bitcoin/bubble-trouble-bitcoin-tops-11000-but-fades-after-sharp-rally-idUSKBN1DS2XK.

[46] A. Samson, «Exchange operator Nasdaq aims to launch bitcoin futures in 2018,» Financial Times, 29 Νοέμβριος 2017. [Ηλεκτρονικό]. Available: https://www.ft.com/content/751d690a-b611-3d38-8cc7-da8bbcee81d3.

[47] A. I. John McCrank, «Bitcoin to start futures trading, stoking Wild West worries,» Reuters, 7 Δεκέμβριος 2017. [Ηλεκτρονικό]. Available: https://www.reuters.com/article/us-bitcoin-futures-analysis/bitcoin-to-start-futures-trading-stoking-wild-west-worries-

idUSKBN1E10J7?feedType=RSS&feedName=businessNews&utm_source=Twitter&utm_medium=Social&utm_campaign=Feed%3A+reuters%2FbusinessNews+%28Bus.

[48] D. Palmer, «CME, CBOE to Begin Bitcoin Futures Trading,» Coindesk, 1 Δεκέμβριος 2017. [Ηλεκτρονικό]. Available: https://www.coindesk.com/cme-begin-trading-bitcoin-futures-december-18/?utm_content=bufferb7511&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer.

[49] A. I. John McCrank, «Bitcoin to start futures trading, stoking Wild West worries,» Reuters, 7 Δεκέμβριος 2017. [Ηλεκτρονικό]. Available: https://www.reuters.com/article/us-bitcoin-futures-analysis/bitcoin-to-start-futures-trading-stoking-wild-west-worries-idUSKBN1E10J7?feedType=RSS&feedName=businessNews&utm_source=Twitter&utm_medium=Social&utm_campaign=Feed%3A+reuters%2FbusinessNews+%28Bus.

[50] O. Kharif, «Bitcoin Mining Service NiceHash Says Hackers Emptied Its Wallet,» Bloomberg, 7 Δεκέμβριος 2017. [Ηλεκτρονικό]. Available: https://www.bloomberg.com/news/articles/2017-12-06/bitcoin-mining-service-nicehash-says-hackers-emptied-its-wallet?cmpid=socialflow-twitter-business&utm_content=business&utm_campaign=socialflow-organic&utm_source=twitter&utm_medium=social.

[51] S. Jacobs, «Bitcoin climbs past $14,000 as demand continues to skyrocket,» Business Insider, 6 Δεκέμβριος 2017. [Ηλεκτρονικό]. Available: https://www.businessinsider.com/bitcoin-climbs-past-14000-2017-12.

[52] F. Chaparro, «Bitcoin futures are about to go live, and they could change the game for cryptocurrencies,» Business Insider, 10 Δεκέμβριος 2017. [Ηλεκτρονικό]. Available: https://www.businessinsider.com/what-is-a-bitcoin-future-2017-12.

[53] D. Dinkins, «Bitcoin Markets Really Like CBOE Futures, Prices Spike Sharply,» Cointelegraph, 11 Δεκέμβριος 2017. [Ηλεκτρονικό]. Available: https://cointelegraph.com/news/bitcoin-markets-really-like-cboe-futures-prices-spike-sharply.

[54] G. B. Alexander Osipovich, «Exchange Giant CME Launches Bitcoin Futures,» Wall Street Journal, 15 Δεκέμβριος 2107. [Ηλεκτρονικό]. Available: https://www.wsj.com/articles/exchange-giant-set-to-launch-bitcoin-futures-after-rival-stumbles-1513432800?reflink=e2twmkts.

[55] R. Perper, «North Korea may be behind a massive cyber attack on a South Korean bitcoin exchange that caused it to collapse,» Business Insider, 21 Δεκέμβριος 2017. [Ηλεκτρονικό]. Available: https://www.businessinsider.com/north-korea-south-korea-bitcoin-heist-2017-12.

[56] BLOOMBERG, «The Sharks Are Beginning to Circle Bitcoin and It's Down 30%,» Fortune, 22 Δεκέμβριος 2017. [Ηλεκτρονικό]. Available: http://fortune.com/2017/12/22/bitcoin-value-loss/.

[57] G. Wearden, «FTSE 100 hits record high; Bitcoin falls as South Korea announces crackdown - as it happened,» The Guardian, 28 Δεκέμβριος 2017. [Ηλεκτρονικό]. Available: https://www.theguardian.com/business/live/2017/dec/28/bitcoin-south-korea-crackdown-stock-markets-ftse-100-business-live.

[58] «Opinion: Nigeria Needs Bitcoin Regulation,» CCN, 31 Δεκέμβριος 2017. [Ηλεκτρονικό]. Available: https://www.ccn.com/nigeria-needs-bitcoin-regulation-than-others/.

[59] D. Dinkins, «Egypt's Top Cleric Declares Bitcoin Trading 'Unlawful',» Cointelegraph, 2017. [Ηλεκτρονικό]. Available: https://cointelegraph.com/news/bitcoin-violates-sharia-law-says-egypts-highest-religious-official-issues-fatwa.

[60] D. M. Blei, A. Y. Ng και M. I. Jordan, «Latent Dirichlet Allocation,» 2003. [Ηλεκτρονικό]. Available: http://jmlr.csail.mit.edu/papers/v3/blei03a.html.

[61] C. L. S. DAVID J. KETCHEN, «THE APPLICATION OF CLUSTER ANALYSIS IN STRATEGIC MANAGEMENT RESEARCH: AN ANALYSIS AND CRITIQUE,» Strategic Management Society, 1996.

[62] D. MacKay, «Chapter 20. An Example Inference Task: Clustering,» 2003. [Ηλεκτρονικό]. Available: http://www.inference.org.uk/mackay/itprnn/ps/284.292.pdf.

[63] A. raskov, H. Stögbauer, R. G. Andrzejak και P. Grassberger, «Hierarchical Clustering Based on Mutual Information,» 2003.

[64] Y. Sasaki, «The truth of the F-measure,» 2007. [Ηλεκτρονικό]. Available: https://www.toyota-ti.ac.jp/Lab/Denshi/COIN/people/yutaka.sasaki/F-measure-YS-26Oct07.pdf.

[65] D. M. W. Powers, «Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation,» 2011. [Ηλεκτρονικό]. Available: http://www.flinders.edu.au/science_engineering/fms/School-CSEM/publications/tech_reps-research_artfcts/TRRA_2007.pdf.

Παράρτημα

Παρατίθεται ο κώδικας και τα Queries που χρειάστηκαν για την δημιουργία της βάσης δεδομένων.

```
/****** Object:  Table [dbo].[CryptocurrenciesPrices]    Script Date: 29/10/2018 09:31:46 ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[CryptocurrenciesPrices](
	[Type] [nvarchar](10) NOT NULL,
	[Coin] [nvarchar](10) NOT NULL,
	[time] [bigint] NOT NULL,
	[closeValue] [decimal](18, 2) NOT NULL,
	[high] [decimal](18, 2) NOT NULL,
	[low] [decimal](18, 2) NOT NULL,
	[openValue] [decimal](18, 2) NOT NULL,
	[volumefrom] [decimal](18, 2) NOT NULL,
	[volumeto] [decimal](18, 2) NOT NULL,
	[DateTimeValue] [datetime] NOT NULL,
	[DateInt] [int] NOT NULL,
 CONSTRAINT [PK_CryptocurrencyHourly_1] PRIMARY KEY CLUSTERED
(
	[Type] ASC,
	[Coin] ASC,
	[time] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON,
ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]
GO
/****** Object:  View [dbo].[CryptocurrenciesDayVolatility]    Script Date: 29/10/2018 09:31:46 ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE VIEW [dbo].[CryptocurrenciesDayVolatility]
AS
SELECT      Type, Coin, openValue, DateTimeValue, DateInt, closeValue, maxValue, minValue, (closeValue - openValue) /
openValue * 100 AS Volatility
FROM        (SELECT      Type, Coin, openValue, DateTimeValue, DateInt,
                             (SELECT      MAX(high) AS Expr1
                              FROM        dbo.CryptocurrenciesPrices
                              WHERE       (Coin = N'btc') AND (CryptocurrenciesPrices_1.DateInt = DateInt)) AS
maxValue,
                             (SELECT      MIN(low) AS Expr1
                              FROM        dbo.CryptocurrenciesPrices AS CryptocurrenciesPrices_3
                              WHERE       (Coin = N'btc') AND (CryptocurrenciesPrices_1.DateInt = DateInt)) AS minValue,
                             (SELECT      TOP (1) closeValue
                              FROM        dbo.CryptocurrenciesPrices AS CryptocurrenciesPrices_2
                              WHERE       (Coin = N'btc') AND (DATEADD(hour, 23,
CryptocurrenciesPrices_1.DateTimeValue) = DateTimeValue)) AS closeValue
                 FROM        dbo.CryptocurrenciesPrices AS CryptocurrenciesPrices_1
                 WHERE       (DateTimeValue >= CONVERT(DATETIME, '2017-01-01 00:00:00', 102)) AND (Coin = N'btc') AND
(DateTimeValue < CONVERT(DATETIME, '2018-01-01 00:00:00', 102)) AND (DATEPART(hh, DateTimeValue) = 0))
                 AS derivedtbl_1
GO
/****** Object:  Table [dbo].[ApplicationsRequestsLimits]    Script Date: 29/10/2018 09:31:46 ******/
SET ANSI_NULLS ON
```

```
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[ApplicationsRequestsLimits](
	[ApplicationID] [bigint] NOT NULL,
	[RequestTypeID] [bigint] NOT NULL,
	[LimitRemaining] [int] NOT NULL,
	[LimitReset] [datetime] NOT NULL,
 CONSTRAINT [PK_ApplicationsRequests] PRIMARY KEY CLUSTERED
(
	[ApplicationID] ASC,
	[RequestTypeID] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON,
ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]
GO
/****** Object:  Table [dbo].[CryptocurrenciesDatesRetrieved]    Script Date: 29/10/2018 09:31:46 ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[CryptocurrenciesDatesRetrieved](
	[Coin] [nvarchar](10) NOT NULL,
	[DateCompleted] [int] NOT NULL,
	[BorderDate] [bit] NOT NULL,
 CONSTRAINT [PK_CryptocurrenciesDatesRetrieved_1] PRIMARY KEY CLUSTERED
(
	[Coin] ASC,
	[DateCompleted] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON,
ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]
GO
/****** Object:  Table [dbo].[Tweets]    Script Date: 29/10/2018 09:31:46 ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[Tweets](
	[PostID] [bigint] NOT NULL,
	[DatasetID] [bigint] NOT NULL,
	[Downloaded] [bit] NOT NULL,
	[Username] [nvarchar](150) NULL,
	[TweetDate] [datetime] NULL,
	[DateAdded] [datetime] NULL,
	[TwitterApplicationID] [bigint] NULL,
	[SystemID] [bigint] NULL,
	[RequestTypeID] [bigint] NULL,
	[SearchID] [bigint] NULL,
	[Comment] [nvarchar](250) NULL,
	[user_id] [bigint] NULL,
	[text] [nvarchar](280) NULL,
	[in_reply_to_status_id] [bigint] NULL,
	[in_reply_to_screen_name] [nvarchar](150) NULL,
	[in_reply_to_user_id] [bigint] NULL,
	[retweet_count] [int] NULL,
	[favorite_count] [int] NULL,
	[EventID] [bigint] NOT NULL,
 CONSTRAINT [PK_Tweets] PRIMARY KEY CLUSTERED
(
	[PostID] ASC,
	[DatasetID] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON,
```

```
ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]
GO
/****** Object:  Table [dbo].[TweetsEvents]    Script Date: 29/10/2018 09:31:46 ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[TweetsEvents](
	[UniqueID] [bigint] IDENTITY(1,1) NOT NULL,
	[Name] [nvarchar](4000) NOT NULL,
	[Link] [nvarchar](500) NOT NULL,
 CONSTRAINT [PK_TweetsEvents] PRIMARY KEY CLUSTERED
(
	[UniqueID] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON,
ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]
GO
/****** Object:  Table [dbo].[TweetsPreprocessed]    Script Date: 29/10/2018 09:31:46 ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[TweetsPreprocessed](
	[PostID] [bigint] NOT NULL,
	[CaseID] [bigint] NOT NULL,
	[DatasetID] [bigint] NOT NULL,
	[text] [nvarchar](300) NULL,
	[hashtags] [nvarchar](300) NULL,
 CONSTRAINT [PK_TweetsPreprocessed] PRIMARY KEY CLUSTERED
(
	[PostID] ASC,
	[CaseID] ASC,
	[DatasetID] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON,
ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]
GO
/****** Object:  Table [dbo].[TweetsSimilarities]    Script Date: 29/10/2018 09:31:46 ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[TweetsSimilarities](
	[DatasetID] [bigint] NOT NULL,
	[CaseID] [bigint] NOT NULL,
	[TweetID1] [bigint] NOT NULL,
	[TweetID2] [bigint] NOT NULL,
	[SentenceSimilarity] [float] NULL,
	[TimeSimilarity] [float] NULL,
	[HashtagsSimilarity] [float] NULL,
	[AuthorSimilarity] [float] NULL,
 CONSTRAINT [PK_TweetsDistances] PRIMARY KEY CLUSTERED
(
	[DatasetID] ASC,
	[CaseID] ASC,
	[TweetID1] ASC,
	[TweetID2] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON,
ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]
GO
```

```
/****** Object:  Table [dbo].[TweetsWords]    Script Date: 29/10/2018 09:31:46 ******/
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[TweetsWords](
	[UniqueID] [bigint] IDENTITY(1,1) NOT NULL,
	[PostID] [bigint] NOT NULL,
	[DataSetID] [bigint] NOT NULL,
	[CaseID] [bigint] NOT NULL,
	[Word] [nvarchar](150) NOT NULL,
 CONSTRAINT [PK_TweetsWordsCount] PRIMARY KEY CLUSTERED
(
	[UniqueID] ASC
)WITH (PAD_INDEX = OFF, STATISTICS_NORECOMPUTE = OFF, IGNORE_DUP_KEY = OFF, ALLOW_ROW_LOCKS = ON,
ALLOW_PAGE_LOCKS = ON) ON [PRIMARY]
) ON [PRIMARY]
GO
ALTER TABLE [dbo].[CryptocurrenciesDatesRetrieved] ADD  CONSTRAINT
[DF_CryptocurrenciesDatesRetrieved_BorderDate]  DEFAULT ((0)) FOR [BorderDate]
GO
ALTER TABLE [dbo].[Tweets] ADD  CONSTRAINT [DF_Tweets_DatasetID]  DEFAULT ((1)) FOR [DatasetID]
GO
ALTER TABLE [dbo].[Tweets] ADD  CONSTRAINT [DF_Tweets_DateAdded]  DEFAULT (getdate()) FOR [DateAdded]
GO
ALTER TABLE [dbo].[Tweets] ADD  CONSTRAINT [DF_Tweets_EventID]  DEFAULT ((0)) FOR [EventID]
GO
ALTER TABLE [dbo].[TweetsEvents] ADD  CONSTRAINT [DF_TweetsEvents_Link]  DEFAULT ('') FOR [Link]
GO
ALTER TABLE [dbo].[TweetsPreprocessed] ADD  CONSTRAINT [DF_TweetsPreprocessed_DatasetID]  DEFAULT ((1)) FOR
[DatasetID]
GO
```

Κώδικας 1: Κώδικας δημιουργίας της Βάσης Δεδομένων

Παρακάτω παραθέτουμε κομμάτι του κώδικα με το οποίο συλλέξαμε τα Tweets από το Twitter, καθώς επίσης και πραγματοποιούσαμε τους ελέγχους που απαιτούνταν προκειμένου να μην ξεπεράσουμε τα όρια που μας έθετε το API του Twitter.

```
import SQL
import datetime as dt
import tweepy
import Settings as st
import Tweets

class Application:
  """twitter application Settings"""

  def __init__(self, UniqueID, Consumer_Key, Consumer_Secret, Access_Token, Access_Token_Secret):
    self.UniqueID = UniqueID
    self.consumer_key = Consumer_Key
    self.consumer_secret = Consumer_Secret
    self.access_token = Access_Token
    self.access_token_secret = Access_Token_Secret
    self.AuthenticateApplication()

  def AuthenticateApplication(self):
    #authorize twitter, initialize tweepy
    auth = tweepy.OAuthHandler(self.consumer_key, self.consumer_secret)
    auth.set_access_token(self.access_token, self.access_token_secret)
    self.api = tweepy.API(auth)

  def CanMakeRequest(self, RequestTypeID):
```

```
        for rs in SQL.GetData("SELECT COUNT(ApplicationID) AS MyCount FROM ApplicationsRequestsLimits WHERE (ApplicationID = ?) AND (RequestTypeID = ?) AND (LimitRemaining <= ?) AND (LimitReset >= ?)", self.UniqueID, RequestTypeID, st.Limit_Rate_Lower_Bound, dt.datetime.utcnow()): #Check if App has exceeded limit for this type of request
            if(rs.MyCount > 0):
                return False
            else:
                return True

    def LogRequest(self, RequestTypeID, post):
        LimitRemaining = int(post._api.last_response.headers['x-rate-limit-remaining'])
        LimitReset = dt.datetime.utcfromtimestamp(int(post._api.last_response.headers['x-rate-limit-reset']))


        print("AppID: " + str(self.UniqueID) + ", LimitRemaining: " + str(LimitRemaining) + ", LimitReset: " + LimitReset.strftime('%d/%m/%Y %H:%M:%S'))

        rows_Affected = SQL.ExcecuteQuery("UPDATE ApplicationsRequestsLimits SET LimitRemaining = ?, LimitReset = ? WHERE (ApplicationID = ?) AND (RequestTypeID = ?)", LimitRemaining, LimitReset, self.UniqueID, RequestTypeID)
        if(rows_Affected == 0):
            SQL.ExcecuteQuery("INSERT INTO ApplicationsRequestsLimits (ApplicationID, RequestTypeID, LimitRemaining, LimitReset) VALUES (?, ?, ?, ?)", self.UniqueID, RequestTypeID, LimitRemaining, LimitReset)

    def GetPost(self, PostID):
        if(self.CanMakeRequest(st.req_get_status) == True):
            try:
                post = self.api.get_status(id=PostID)
                self.LogRequest(st.req_get_status, post)
                Tweets.AddPost(post, self.UniqueID, st.req_get_status)
            except tweepy.TweepError as twerr:
                Tweets.LogError(twerr, PostID)

            return True
        else:
            return False
```

Κώδικας 2: Κώδικας συλλογής Tweets

Παρακάτω παραθέτουμε κομμάτι κώδικα με τον οποίο αντλήσαμε επιτυχώς τα ιστορικά δεδομένα για την τιμή του Bitcoin:

```
import requests
import datetime
import pandas as pd
import matplotlib.pyplot as plt
import SQL
import General as g
import Settings as st
import time


def hourly_price_historical(symbol, comparison_symbol, limit, aggregate, exchange='',toTs=0):
    LogRequest()

    url = 'https://min-api.cryptocompare.com/data/histohour?fsym={}&tsym={}&limit={}&aggregate={}'\
            .format(symbol.upper(), comparison_symbol.upper(), limit, aggregate)
    if exchange:
        url += '&e={}'.format(exchange)

    if toTs>0:
```

```
        url += '&toTs={}'.format(toTs)

    page = requests.get(url)
    data = page.json()['Data']
    df = pd.DataFrame(data)
    df['timestamp'] = [datetime.datetime.fromtimestamp(d) for d in df.time]
    return df

def GetDayCoinPrices(Coin, DayInt):
    NowDayStartInt = g.GetDateInt(datetime.datetime.now())
    DataFound = False
    for rs in SQL.GetData("SELECT COUNT(1) AS MyCount, (SELECT BorderDate FROM CryptocurrenciesDatesRetrieved
WHERE (Coin = ?) AND (DateCompleted = ?)) AS BorderDate FROM CryptocurrenciesDatesRetrieved WHERE (Coin = ?) AND
(DateCompleted = ?)", Coin, DayInt, Coin, DayInt):
        if((rs.MyCount == 0) or (DayInt >= NowDayStartInt)):
            while CheckIfCanMakeRequest() == False:
                time.sleep(15)

            lastTimeStampDate = g.GetLastTimeStampFromDateInt(DayInt)
            df = hourly_price_historical(Coin, 'USD', 50, 1, toTs=lastTimeStampDate)
            validHours = 0
            for index, row in df.iterrows():
                isValid = AddDayToDB(Coin, row, DayInt)
                if(isValid):
                    validHours = validHours + 1
            if((validHours > 0) and (DayInt < NowDayStartInt)):
                DataFound = True
                SQL.ExcecuteQuery("INSERT INTO CryptocurrenciesDatesRetrieved (Coin, DateCompleted) VALUES (?, ?)", Coin,
DayInt)
            elif (DayInt >= NowDayStartInt):
                DataFound = True
        elif(rs.MyCount > 0):
            if(rs.BorderDate == True):
                DataFound = False
            else:
                DataFound = True

    return DataFound

def AddDayToDB(Coin, row, DayInt):
    isValid = True

    if((row["close"] == 0) and (row["high"] == 0) and (row["low"] == 0) and (row["open"] == 0) and (row["volumefrom"] == 0)
and (row["volumeto"] == 0)):
        isValid = False

    if (g.GetDateInt(row["timestamp"]) != DayInt):
        isValid = False


    if(isValid == True):
        rows_Affected = SQL.ExcecuteQuery("UPDATE CryptocurrenciesPrices SET closeValue = ?, high = ?, low = ?, openValue
= ?, volumefrom = ?, volumeto = ?, DateTimeValue = ?, DateInt = ? WHERE (Type = ?) AND (Coin = ?) AND (time = ?)",
row["close"], row["high"], row["low"], row["open"], row["volumefrom"], row["volumeto"], row["timestamp"],
g.GetDateInt(row["timestamp"]), "hourly", Coin, row["time"])
        if(rows_Affected == 0):
            SQL.ExcecuteQuery("INSERT INTO CryptocurrenciesPrices (closeValue, high, low, openValue, volumefrom, volumeto,
DateTimeValue, DateInt, Type, Coin, time) VALUES (?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?)", row["close"], row["high"], row["low"],
row["open"], row["volumefrom"], row["volumeto"], row["timestamp"], g.GetDateInt(row["timestamp"]), "hourly", Coin,
row["time"])
    return isValid
```

```
def CheckIfCanMakeRequest():
   #Check hour
   for rs in SQL.GetData("SELECT COUNT(UniqueID) AS MyCount FROM CryptocompareRequests WHERE (DATEADD(hour, 1,
DateOccured) > GETDATE())"):
      if(rs.MyCount >= st.MaxRequestsPerHour):
         return False

   #Check Minute
   for rs in SQL.GetData("SELECT COUNT(UniqueID) AS MyCount FROM CryptocompareRequests WHERE (DATEADD(minute,
1, DateOccured) > GETDATE())"):
      if(rs.MyCount >= st.MaxRequestsPerMinute):
         return False


   #Check Second
   for rs in SQL.GetData("SELECT COUNT(UniqueID) AS MyCount FROM CryptocompareRequests WHERE (DATEADD(second,
1, DateOccured) > GETDATE())"):
      if(rs.MyCount >= st.MaxRequestsPerSecond):
         return False


   return True

def LogRequest():
   SQL.ExcecuteQuery("INSERT INTO CryptocompareRequests (DateOccured) VALUES (GETDATE()) DELETE FROM
CryptocompareRequests WHERE (DATEADD(hour, 1, DateOccured) < GETDATE())")

def GetCryptocurrencyPrices(Coin):
   DayStartInt = g.GetDateInt(datetime.datetime.now())
   while GetDayCoinPrices(Coin, DayStartInt):
      DayStartInt = DayStartInt - 1
```

Κώδικας 3: Κώδικας συλλογής τιμής Bitcoin ανά ώρα

Ο κώδικας με τον οποίο πραγματοποιήσαμε την προ-επεξεργασία δεδομένων παρατίθεται παρακάτω:

```
import Settings as st
import SQL
import DBSimilarities
import os

st.InitConnectionString()

# Run in python console
import nltk;

nltk.download('stopwords')


from decimal import *
import re
import numpy as np
import pandas as pd
from pprint import pprint
from collections import defaultdict

# Gensim
import gensim
import gensim.corpora as corpora
from gensim.utils import simple_preprocess
```

```
from gensim.models import CoherenceModel

# spacy for lemmatization
import spacy


# Plotting tools
import pyLDAvis
import pyLDAvis.gensim  # don't skip this
import matplotlib.pyplot as plt
#%matplotlib inline


# NLTK Stop words
from nltk.corpus import stopwords
stop_words = stopwords.words('english')
stop_words.extend(['from', 'subject', 're', 'edu', 'use'])


from nltk.stem import PorterStemmer
from nltk.tokenize import sent_tokenize, word_tokenize

ps = PorterStemmer()

def sent_to_words(sentences):
    for sentence in sentences:
        yield(gensim.utils.simple_preprocess(str(sentence), deacc=True))  # deacc=True removes punctuations


# Define functions for stopwords, bigrams, trigrams and lemmatization
def remove_stopwords(texts):
    return [[word for word in simple_preprocess(str(doc)) if word not in stop_words] for doc in texts]

def make_bigrams(texts):
    return [bigram_mod[doc] for doc in texts]

def make_trigrams(texts):
    return [trigram_mod[bigram_mod[doc]] for doc in texts]

def lemmatization(texts, allowed_postags=['NOUN', 'ADJ', 'VERB', 'ADV']):
    """https://spacy.io/api/annotation"""
    texts_out = []
    counter = 0
    for sent in texts:
        doc = nlp(" ".join(sent))
        texts_out.append([token.lemma_ for token in doc if token.pos_ in allowed_postags])
        counter = counter + 1
        if (counter%100 == 0): print("lemmatization counter: " + str(counter))
    return texts_out

def find_hash_tags(s):
    return set(part[1:] for part in s.split() if part.startswith('#'))


def GetProprocessedSentence(review):
    review = re.sub(r"http\S+", "", review) #remove links


    # Remove new line characters
    review = re.sub('\s+', " ", review)

    # Remove distracting single quotes
    review = re.sub("\'", "", review)
```

```
    review = re.sub('@[^\s]+','',review)#remove usernames
    #review = review.replace("@", "")#remove tag symbol (@ keep tag value)
    review = re.sub('[^a-zA-Z0-9]', ' ', review) #keeping only letters, removing !,;' etc.
    review = review.lower()  #mikra grammata oxi kefalaia

    review = gensim.utils.simple_preprocess(str(review), deacc=True)

    review = [word for word in review if word not in stop_words]

    #review = [ps.stem(word) for word in review if word not in stop_words]
    #review = ' '.join(review) #connect words seperated by a space

    return review

getcontext().prec = 2

def SavePreprocessedTweet(PostID, text, hashtags):
    SQL.ExcecuteQuery("INSERT INTO TweetsPreprocessed (PostID, CaseID, DatasetID, text, hashtags) VALUES (?, ?, ?, ?, ?)",
PostID, st.caseID, st.datasetID, ' '.join(text), hashtags)

def SaveWordsOfPostToDB(postID, review):
    for w in review:
        SQL.ExcecuteQuery("INSERT INTO TweetsWords (PostID, DataSetID, CaseID, Word) VALUES (?, ?, ?, ?)", postID,
st.datasetID, st.caseID, w)



data = []
dataPostIDs = []
PostsHashtags = []

print("preprocess Tweets")
counter = 0
for rs in SQL.GetData("SELECT PostID, text FROM Tweets WHERE (Downloaded = 1) AND (DatasetID = ?) AND (PostID NOT
IN (SELECT PostID FROM TweetsPreprocessed WHERE (CaseID = ?) AND (DatasetID = ?)))", st.datasetID, st.caseID,
st.datasetID):
    review = rs.text.lower()

    #get hashtags
    hashtags = list(find_hash_tags(review))#hashtags not processed
    for c in range(0, len(hashtags)):
        hashtags[c] = re.sub('[^a-zA-Z0-9]', '', hashtags[c])
    hashtags = ' '.join(hashtags)
    hashtags = hashtags.lower()
    hashtags = ' '.join([word for word in hashtags.split() if word not in stop_words])
    PostsHashtags.append(hashtags)

    review = GetProprocessedSentence(review)
    data.append(review)
    dataPostIDs.append(rs.PostID)

    #SaveWordsOfPostToDB(rs.PostID, review)
    #SavePreprocessedTweet(rs.PostID, review, hashtags)
    counter = counter + 1
    if (counter%100 == 0): print("read DB counter: " + str(counter))


if (counter > 0):
    SQL.ExcecuteQuery("DELETE FROM TweetsWords WHERE (DataSetID = ?) AND (CaseID = ?)", st.datasetID, st.caseID)
    data_words = data;#list(sent_to_words(data))
```

```
# Build the bigram and trigram models
bigram = gensim.models.Phrases(data_words, min_count=5, threshold=100) # higher threshold fewer phrases.
trigram = gensim.models.Phrases(bigram[data_words], threshold=100)

# Faster way to get a sentence clubbed as a trigram/bigram
bigram_mod = gensim.models.phrases.Phraser(bigram)
trigram_mod = gensim.models.phrases.Phraser(trigram)

# See trigram example
print(trigram_mod[bigram_mod[data_words[0]]])


## Remove Stop Words
#data_words_nostops = remove_stopwords(data_words)
data_words_nostops = data_words

# Form Bigrams
data_words_bigrams = make_bigrams(data_words_nostops)

# Initialize spacy 'en' model, keeping only tagger component (for efficiency)
# python3 -m spacy download en
nlp = spacy.load('en', disable=['parser', 'ner'])

# Do lemmatization keeping only noun, adj, vb, adv
data_lemmatized = lemmatization(data_words_bigrams, allowed_postags=['NOUN', 'ADJ', 'VERB', 'ADV'])

print(data_lemmatized[:1])


# remove words that appear only once
frequency = defaultdict(int)
for text in data_lemmatized:
    for token in text:
        frequency[token] += 1

data_lemmatized = [[ps.stem(token) for token in text if frequency[token] > 1]
        for text in data_lemmatized]

counter = 0
for s in range(0, len(data_lemmatized), 1):
    SaveWordsOfPostToDB(dataPostIDs[s], data_lemmatized[s])
    SavePreprocessedTweet(dataPostIDs[s], data_lemmatized[s], PostsHashtags[s])
    counter = counter + 1
    if (counter%100 == 0): print("Add DB counter: " + str(counter))
```

Κώδικας 4: Κώδικας προ-επεξεργασίας δεδομένων

Ο κώδικας που χρησιμοποιήσαμε για τον υπολογισμό της ομοιότητας κειμένου παρουσιάζεται παρακάτω:

```
import Settings as st
import SQL
import DBSimilarities
import os

st.InitConnectionString()

# Run in python console
import nltk;

nltk.download('stopwords')
```

```
from decimal import *
import re
import numpy as np
import pandas as pd
from pprint import pprint
from collections import defaultdict

# Gensim
import gensim
import gensim.corpora as corpora
from gensim.utils import simple_preprocess
from gensim.models import CoherenceModel

# spacy for lemmatization
import spacy


# Plotting tools
import pyLDAvis
import pyLDAvis.gensim  # don't skip this
import matplotlib.pyplot as plt
#%matplotlib inline

# Enable logging for gensim - optional
import logging
logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s', level=logging.ERROR)

import warnings
warnings.filterwarnings("ignore",category=DeprecationWarning)



# NLTK Stop words
from nltk.corpus import stopwords
stop_words = stopwords.words('english')
stop_words.extend(['from', 'subject', 're', 'edu', 'use'])


from nltk.stem import PorterStemmer
from nltk.tokenize import sent_tokenize, word_tokenize



data = []
data_lemmatized = []
dataPostIDs = []
for rs in SQL.GetData("SELECT PostID, text, hashtags FROM TweetsPreprocessed WHERE (CaseID = ?) AND (DatasetID = ?)
ORDER BY PostID", st.caseID, st.datasetID):
   dataPostIDs.append(rs.PostID)
   rs.text = rs.text;
   data_lemmatized.append(rs.text.split())
   data.append(rs.text)



# Create Dictionary
id2word = corpora.Dictionary(data_lemmatized)

# Create Corpus
texts = data_lemmatized

# Term Document Frequency
corpus = [id2word.doc2bow(text) for text in texts]
```

```
# View
print(corpus[:1])

mallet_path = r"C:\mallet-2.0.8\bin\mallet" # update this path

#mallet-2.0.8

def compute_coherence_values(dictionary, corpus, texts, limit, start=2, step=3):
    """
    Compute c_v coherence for various number of topics
    https://markroxor.github.io/gensim/static/notebooks/topic_coherence_tutorial.html#topic=0&lambda=1&term=
    https://radimrehurek.com/gensim/models/coherencemodel.html
    Parameters:
    ----------
    dictionary : Gensim dictionary
    corpus : Gensim corpus
    texts : List of input texts
    limit : Max num of topics

    Returns:
    -------
    model_list : List of LDA topic models
    coherence_values : Coherence values corresponding to the LDA model with respective number of topics
    """
    coherence_values = []
    model_list = []
    for num_topics in range(start, limit, step):
        print("compute_coherence_values: num_topics: " + str(num_topics) + " start: " + str(start) + ", limit: " + str(limit) + ",
step: " + str(step))
        model = gensim.models.wrappers.LdaMallet(mallet_path, corpus=corpus, num_topics=num_topics, id2word=id2word)

        model_list.append(model)

        #texts = [[dictionary[word_id] for word_id, freq in doc] for doc in corpus]#edw


        coherencemodel = CoherenceModel(model=model, corpus=corpus, dictionary=dictionary, coherence='u_mass',
processes = 1)
        #coherencemodel = CoherenceModel(model=model, texts=texts, coherence='c_v', processes = 1)
        #coherencemodel = CoherenceModel(model=model, texts=texts, coherence='c_v', processes =
1)#dictionary=dictionary,
        coherence_values.append(coherencemodel.get_coherence())

    return model_list, coherence_values


calculated_topics = 160;

if (calculated_topics == 0):
    limit=1000; start=10; step=50;
    #limit=70; start=50; step=2;
    #24
    # Can take a long time to run.
    model_list, coherence_values = compute_coherence_values(dictionary=id2word, corpus=corpus,
texts=data_lemmatized, start=start, limit=limit, step=step)

    # Show graph
    x = range(start, limit, step)
    plt.plot(x, coherence_values)
    plt.xlabel("Num Topics")
    plt.ylabel("Coherence score")
    plt.legend(("coherence_values"), loc='best')
```

```
    plt.show()

    # Print the coherence scores
    for m, cv in zip(x, coherence_values):
        print("Num Topics =", m, " has Coherence Value of", round(cv, 4))
    input()



optimal_model = gensim.models.wrappers.LdaMallet(mallet_path, corpus=corpus, num_topics=calculated_topics, id2word=id2word, iterations=1000)

model_topics = optimal_model.show_topics(formatted=False)
pprint(optimal_model.print_topics(num_words=10))


def format_topics_sentences(ldamodel, corpus, texts):
    # Init output
    sent_topics_df = pd.DataFrame()

    # Get main topic in each document
    for i, row in enumerate(ldamodel[corpus]):
        row = sorted(row, key=lambda x: (x[1]), reverse=True)
        # Get the Dominant topic, Perc Contribution and Keywords for each document
        for j, (topic_num, prop_topic) in enumerate(row):
            if j == 0:  # => dominant topic
                wp = ldamodel.show_topic(topic_num)
                topic_keywords = ", ".join([word for word, prop in wp])
                sent_topics_df = sent_topics_df.append(pd.Series([int(topic_num), round(prop_topic,4), topic_keywords]), ignore_index=True)
            else:
                break
    sent_topics_df.columns = ['Dominant_Topic', 'Perc_Contribution', 'Topic_Keywords']

    # Add original text to the end of the output
    contents = pd.Series(texts)
    sent_topics_df = pd.concat([sent_topics_df, contents], axis=1)
    return(sent_topics_df)


df_topic_sents_keywords = format_topics_sentences(ldamodel=optimal_model, corpus=corpus, texts=data)

# Format
df_dominant_topic = df_topic_sents_keywords.reset_index()
df_dominant_topic.columns = ['Document_No', 'Dominant_Topic', 'Topic_Perc_Contrib', 'Keywords', 'Text']

# Insert Results To DB
print("Insert dominant topic for each doc")


Events_Calculated_ClusterIDs_File = 'Events_Calculated_ClusterIDs.json'
if ((os.path.exists(Events_Calculated_ClusterIDs_File))):
    os.remove(Events_Calculated_ClusterIDs_File)

text_file = open(Events_Calculated_ClusterIDs_File, "w")

SQL.ExcecuteQuery("DELETE FROM TweetsDominantTopicPerTweet WHERE (DatasetID = ?) AND (CaseID = ?)", st.datasetID, st.caseID)

for x in range(0, df_dominant_topic.shape[0], 1):
    PostID = dataPostIDs[x]
    Document_No = int(df_dominant_topic.loc[x,'Document_No'])
    Dominant_Topic = df_dominant_topic.loc[x,'Dominant_Topic']
```

```
   Topic_Perc_Contrib = df_dominant_topic.loc[x,'Topic_Perc_Contrib']
   Keywords = df_dominant_topic.loc[x,'Keywords']
   Text = df_dominant_topic.loc[x,'Text']

   SQL.ExcecuteQuery("INSERT INTO TweetsDominantTopicPerTweet (PostID, DatasetID, CaseID, Document_No, Dominant_Topic, Topic_Perc_Contrib, Keywords, Text) VALUES (?, ?, ?, ?, ?, ?, ?, ?)", PostID, st.datasetID, st.caseID, Document_No, Dominant_Topic, Topic_Perc_Contrib, Keywords, Text)
   text_file.write('{"postID":' + str(PostID) + ', "clusterID":' + str(int(Dominant_Topic)) +  '}\n')

text_file.close()



print("Insert dominant topic for each doc ENDED")



# Find the most representative document for each topic
print("Insert the most representative document for each topic")
# Group top 5 sentences under each topic
sent_topics_sorteddf_mallet = pd.DataFrame()

sent_topics_outdf_grpd = df_topic_sents_keywords.groupby('Dominant_Topic')

for i, grp in sent_topics_outdf_grpd:
   sent_topics_sorteddf_mallet = pd.concat([sent_topics_sorteddf_mallet,
                                            grp.sort_values(['Perc_Contribution'], ascending=[0]).head(1)],
                                           axis=0)

# Reset Index
sent_topics_sorteddf_mallet.reset_index(drop=True, inplace=True)

# Format
sent_topics_sorteddf_mallet.columns = ['Topic_Num', "Topic_Perc_Contrib", "Keywords", "Text"]



SQL.ExcecuteQuery("DELETE FROM TweetsMostRepresentativeTweetPerTopic WHERE (DatasetID = ?) AND (CaseID = ?)", st.datasetID, st.caseID)

for x in range(0, sent_topics_sorteddf_mallet.shape[0], 1):
   Topic_Num = sent_topics_sorteddf_mallet.loc[x,'Topic_Num']
   Topic_Perc_Contrib = sent_topics_sorteddf_mallet.loc[x,'Topic_Perc_Contrib']
   Keywords = sent_topics_sorteddf_mallet.loc[x,'Keywords']
   Text = sent_topics_sorteddf_mallet.loc[x,'Text']

   SQL.ExcecuteQuery("INSERT INTO TweetsMostRepresentativeTweetPerTopic (DatasetID, CaseID, Topic_Num, Topic_Perc_Contrib, Keywords, Text) VALUES (?, ?, ?, ?, ?, ?)", st.datasetID, st.caseID, Topic_Num, Topic_Perc_Contrib, Keywords, Text)



print("Insert the most representative document for each topic ENDED")



print("Insert Number of Documents for Each Topic")
# Number of Documents for Each Topic
topic_counts = df_topic_sents_keywords['Dominant_Topic'].value_counts()

# Percentage of Documents for Each Topic
topic_contribution = round(topic_counts/topic_counts.sum(), 4)
```

```
# Topic Number and Keywords
topic_num_keywords = df_topic_sents_keywords[['Dominant_Topic', 'Topic_Keywords']]

# Concatenate Column wise
df_dominant_topics = pd.concat([topic_num_keywords, topic_counts, topic_contribution], axis=1)

# Change Column names
df_dominant_topics.columns = ['Dominant_Topic', 'Topic_Keywords', 'Num_Documents', 'Perc_Documents']


SQL.ExcecuteQuery("DELETE FROM TweetsNumberOfDocumentsPerTopic WHERE (DatasetID = ?) AND (CaseID = ?)", st.datasetID, st.caseID)

for x in range(0, df_dominant_topics.shape[0], 1):
   PostID = dataPostIDs[x]
   Dominant_Topic = df_dominant_topics.loc[x,'Dominant_Topic']
   Topic_Keywords = df_dominant_topics.loc[x,'Topic_Keywords']
   Num_Documents = df_dominant_topics.loc[x,'Num_Documents']
   Perc_Documents = df_dominant_topics.loc[x,'Perc_Documents']

   SQL.ExcecuteQuery("INSERT INTO TweetsNumberOfDocumentsPerTopic (PostID, DatasetID, CaseID, Dominant_Topic, Topic_Keywords, Num_Documents, Perc_Documents) VALUES (?, ?, ?, ?, ?, ?, ?)", PostID, st.datasetID, st.caseID, Dominant_Topic, Topic_Keywords, Num_Documents, Perc_Documents)

print("Insert Number of Documents for Each Topic ENDED")


print("Calculate distances")


lsi = gensim.models.LsiModel(corpus, id2word=id2word, num_topics=calculated_topics)# initialize an LSI transformation
corpus_lsi = lsi[corpus] # create a double wrapper over the original corpus: bow->tfidf->fold-in-lsi

index = gensim.similarities.MatrixSimilarity(lsi[corpus])

def SimQuery(myword, index, documents, lsi, dictionary):
   print("------------" + myword + "-----------")
   doc = myword
   vec_bow = dictionary.doc2bow(doc.lower().split())
   vec_lsi = lsi[vec_bow] # convert the query to LSI space

   sims = index[vec_lsi]
   return sorted(enumerate(sims), key=lambda item: -item[1])


SQL.ExcecuteQuery("UPDATE TweetsSimilarities SET SentenceSimilarity = NULL WHERE (DatasetID = ?) AND (CaseID = ?)", st.datasetID, st.caseID)

for i in range(0, len(data), 1):
   sims = SimQuery(data[i], index, data, lsi, id2word)

   tweetID1 = dataPostIDs[i]
   for res in sims:
      tweetID2 = dataPostIDs[res[0]]
      distance = res[1]

      DBSimilarities.InformDBSentenceSimilarity(tweetID1, tweetID2, distance)

   counter = counter + 1
   if (counter%100 == 0): print("counter: " + str(counter))
```

Κώδικας 5: Κώδικας υπολογισμού ομοιότητας κειμένου

Το Query που συλλέγει τα απαραίτητα δεδομένα και δημιουργεί το View της ημερήσιας μεταβλητότητας απεικονίζεται παρακάτω:

```
CREATE VIEW [dbo].[CryptocurrenciesDayVolatility]
AS
SELECT      Type, Coin, openValue, DateTimeValue, DateInt, closeValue, maxValue, minValue, (closeValue - openValue) /
openValue * 100 AS Volatility
FROM        (SELECT      Type, Coin, openValue, DateTimeValue, DateInt,
                         (SELECT      MAX(high) AS Expr1
                          FROM        dbo.CryptocurrenciesPrices
                          WHERE       (Coin = N'btc') AND (CryptocurrenciesPrices_1.DateInt = DateInt)) AS
maxValue,
                         (SELECT      MIN(low) AS Expr1
                          FROM        dbo.CryptocurrenciesPrices AS CryptocurrenciesPrices_3
                          WHERE       (Coin = N'btc') AND (CryptocurrenciesPrices_1.DateInt = DateInt)) AS minValue,
                         (SELECT      TOP (1) closeValue
                          FROM        dbo.CryptocurrenciesPrices AS CryptocurrenciesPrices_2
                          WHERE       (Coin = N'btc') AND (DATEADD(hour, 23,
CryptocurrenciesPrices_1.DateTimeValue) = DateTimeValue)) AS closeValue
               FROM        dbo.CryptocurrenciesPrices AS CryptocurrenciesPrices_1
               WHERE       (DateTimeValue >= CONVERT(DATETIME, '2017-01-01 00:00:00', 102)) AND (Coin = N'btc') AND
(DateTimeValue < CONVERT(DATETIME, '2018-01-01 00:00:00', 102)) AND (DATEPART(hh, DateTimeValue) = 0))
               AS derivedtbl_1
```

Κώδικας 6: Query δημιουργίας Πίνακα ημερήσιας μεταβλητότητας

Ο κώδικας με τον οποίο εκτελέσαμε τον αλγόριθμο KMeans παρατίθεται παρακάτω:

```
import numpy as np

import pandas as pd
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

# -*- coding: utf-8 -*-
"""
Created on Fri Feb 16 20:09:58 2018

@author: Geo
"""
import SQL
import Settings as st

import tweetsInputFunctions
import csv
import os
import numpy as np
import tweetsDistance
import math, random, tweetsCRP
import DBSimilarities

import test

import evaluate_clustering

def InitApp():
   print("Application Started")
   st.InitConnectionString()
   print("ConnectionString: " + st.ConnectionString)
```

```
def CalculateDistances():
    distance = np.zeros((N_tweets, N_tweets))

    for i in range(N_tweets):
        for j in range(N_tweets):
            if(distance[i,j] == 0):
                if(i!=j):
                    if(distance[j,i] == 0):
                        distance[i,j] = tweetsDistance.CalculateDistances(dataset, i, j, Similarities)
        if ((i % 100) == 0):
            print(i, j)


def Kmeans(calculatedClusters, sentenceWeight, timeWeight, authorWeight, hashtagWeight, daysWeight):


    distance = np.zeros((N_tweets, N_tweets))
    for i in range(N_tweets):
        for j in range(N_tweets):
            if(distance[i,j] == 0):
                if(i!=j):
                    if(distance[j,i] == 0):
                        distance[i,j] = tweetsDistance.getDistance(dataset, i, j, Similarities, sentenceWeight, timeWeight,
authorWeight, hashtagWeight, daysWeight)
                    else:
                        distance[i,j] = distance[j,i]
                else:
                    distance[i,j] = 0 #gia tin diagwnio bazw miden
        if ((i % 100) == 0):
            print(i, j)

    mms = MinMaxScaler()
    mms.fit(distance)
    data_transformed = mms.transform(distance)


    if (calculatedClusters == 0):
        Sum_of_squared_distances = []
        K = range(1,205,20)
        for k in K:
            km = KMeans(n_clusters=k)
            km = km.fit(data_transformed)
            Sum_of_squared_distances.append(km.inertia_)
            print(str(k))

        plt.plot(K, Sum_of_squared_distances, 'bx-')
        plt.xlabel('k')
        plt.ylabel('Sum_of_squared_distances')
        plt.title('Elbow Method For Optimal k')
        plt.show()


    km = KMeans(n_clusters=calculatedClusters)
    km = km.fit(data_transformed)
    labels = km.labels_

    Events_Calculated_ClusterIDs_File = 'Events_Calculated_ClusterIDs.json'
    if ((os.path.exists(Events_Calculated_ClusterIDs_File))):
        os.remove(Events_Calculated_ClusterIDs_File)

    text_file = open(Events_Calculated_ClusterIDs_File, "w")

    for c in range(N_tweets):
```

```
        text_file.write('{"postID":' + str(dataset[c]['tweetid']) + ', "clusterID":' + str(labels[c]) +  '}\n')


    text_file.close()

    scores, cluster_count, document_count = evaluate_clustering.Evaluate(["", "--challenge1",
Events_Calculated_ClusterIDs_File, "Events_Annotated_ClusterIDs.txt"])
    LogDDCRPResults(0, sentenceWeight, timeWeight, authorWeight, hashtagWeight, daysWeight, scores, cluster_count,
document_count)



def LogDDCRPResults(Alpha, sentenceWeight, timeWeight, authorWeight, hashtagWeight, daysWeight, scores,
cluster_count, document_count):
    F1_Main_Score = scores[0][1]
    NMI = scores[1][1]
    F1_Div = scores[2][1]
    Random_Baseline_F1 = 0
    Divergence_F1 = 0



    rowsAffected = SQL.ExcecuteQuery("UPDATE TweetsDDCPR SET F1_Main_Score = ?, NMI = ?, F1_Div = ?,
Random_Baseline_F1 = ?, Divergence_F1 = ? WHERE (CaseID = ?) AND (DatasetID = ?) AND (SentenceWeight = ?) AND
(TimeWeight = ?) AND (AuthorWeight = ?) AND (HashtagWeight = ?) AND (Alpha = ?) AND (daysWeight = ?)",
F1_Main_Score, NMI, F1_Div, Random_Baseline_F1, Divergence_F1, 5, st.datasetID, sentenceWeight, timeWeight,
authorWeight, hashtagWeight, Alpha, daysWeight)
    if(rowsAffected == 0):
        SQL.ExcecuteQuery("INSERT INTO TweetsDDCPR (DatasetID, CaseID, SentenceWeight, TimeWeight, AuthorWeight,
HashtagWeight, F1_Main_Score, NMI, F1_Div, Random_Baseline_F1, Divergence_F1, Alpha, daysWeight) VALUES (?, ?, ?, ?,
?, ?, ?, ?, ?, ?, ?, ?, ?, ?)", st.datasetID, 5, sentenceWeight, timeWeight, authorWeight, hashtagWeight, F1_Main_Score, NMI,
F1_Div, Random_Baseline_F1, Divergence_F1, Alpha, daysWeight)



def FixEventsAnnotatedClusterIDs():
    print("FixEventsAnnotatedClusterIDs")
    Events_Annotated_ClusterIDs = 'Events_Annotated_ClusterIDs.txt'
    if ((os.path.exists(Events_Annotated_ClusterIDs))):
        os.remove(Events_Annotated_ClusterIDs)

    text_file = open(Events_Annotated_ClusterIDs, "w")

    i = 0
    for rs in SQL.GetData("SELECT PostID, EventID FROM Tweets WHERE (DatasetID = ?) ", st.datasetID):
        mystr = str(rs.PostID) + '\t' + str(rs.EventID)
        if(i>0):
            mystr = '\n' + str(rs.PostID) + '\t' + str(rs.EventID)
        text_file.write(mystr)
        i = i + 1
        if ((i % 100) == 0):
            print(i)

    print("FixEventsAnnotatedClusterIDs, i: " + str(i))

    text_file.close()


InitApp()

FixEventsAnnotatedClusterIDs()

#calculate diastances
```

```
print("Calculating Distances")
dataset = tweetsInputFunctions.getTweets()

N_tweets = len(dataset)

Similarities = DBSimilarities.GetSimilaritiesTable(N_tweets, dataset)

print("Do you want to calculate Distances (y/n) ? ")
CalculateDistancesQ = input()
if(CalculateDistancesQ == 'y'):
   CalculateDistances()
   Similarities = DBSimilarities.GetSimilaritiesTable(N_tweets, dataset)


continueExperiment = 1

while (continueExperiment==1):
   calculatedClusters = 0
   sentenceWeight = 0
   timeWeight = 0
   authorWeight = 0
   hashtagWeight = 0
   daysWeight = 0

   print("Give calculatedClusters:")
   calculatedClusters = int(input())

   print("Give days Weight:")
   daysWeight = float(input())

   print("Give sentence Weight:")
   sentenceWeight = float(input())

   print("Give time Weight:")
   timeWeight = float(input())

   print("Give author Weight:")
   authorWeight = float(input())

   print("Give hashtag Weight:")
   hashtagWeight = float(input())

   Kmeans(calculatedClusters, sentenceWeight, timeWeight, authorWeight, hashtagWeight, daysWeight)

   print("continueExperiment (0 / 1)? ")
   continueExperiment = int(input())
```

Κώδικας 7: Κώδικας υλοποίησης αλγορίθμου K-Means

Παρακάτω παρατίθεται το κομμάτι του κώδικα με τον οποίο έγινε η υλοποίηση ολόκληρου του νευρωνικού δικτύου, το οποίο περιλαμβάνει την δημιουργία των πινάκων που περιγράψαμε καθώς επίσης την εκπαίδευση αλλά και την αξιολόγηση του αλγορίθμου:

```
from pprint import pprint
import math
import datetime
import Settings as st
import SQL
import pandas as pd
import os
from sklearn.model_selection import train_test_split
```

```
from sklearn.preprocessing import StandardScaler
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import classification_report, confusion_matrix

st.InitConnectionString()

DayDataXFileName = "daydataX.h"
DayDataYFileName = "daydataY.h"

HourDataXFileName = "hourdataX.h"
HourDataYFileName = "hourdataY.h"

#Day Volatility
def GetDayVolatilityLabel(val):
   if(val > 0):
      return 1
   else:
      return 0

def GetDayDiff(TweetDate, Date, dayWindow):
   elapsedTime = Date - TweetDate
   Days = divmod(elapsedTime.total_seconds(), 86400)[0] # elapsedTime.total_seconds() / 86400 #
   return math.exp(-(Days/dayWindow))
   #return Days/(dayWindow + 0.5)

def GetDayTuple(Words, Date, dayWindow):
   Date = Date + datetime.timedelta(days=1)
   d = {}
   tweetsFound = False
   for x in Words:
      TweetDate = x["TweetDate"]
      Word = x["Word"]
      Username = "username_" + x["Username"]

      if(TweetDate < Date):
         if not Username in d:
            d[Username] = 0

         if not Word in d:
            d[Word] = 0
            tweetsFound = True

         dateDiff = GetDayDiff(TweetDate, Date, dayWindow)
         d[Word] = d[Word] + dateDiff
         d[Username] = d[Username] + dateDiff
      else:
         break

   return d, tweetsFound

def GetDayDataFromDB(dayWindow):
   print("Get Words")
   Words = []
   for rs in SQL.GetData("SELECT TweetsWords.Word, Tweets.TweetDate, dbo.Tweets.Username FROM TweetsWords
INNER JOIN Tweets ON TweetsWords.PostID = Tweets.PostID WHERE (TweetsWords.DataSetID = ?) AND
(TweetsWords.CaseID = ?) ORDER BY Tweets.TweetDate", st.datasetID, st.caseID):
      Words.append({'Word': rs.Word, 'TweetDate' : rs.TweetDate, 'Username' : rs.Username.lower()})

   #Train Day
   DayY = []
   DayX = []

   print("Fill Day Table")
```

```
    cnt = 0
    for rs in SQL.GetData("SELECT DateTimeValue, Volatility FROM CryptocurrenciesDayVolatility ORDER BY
DateTimeValue"):
        d, tweetsFound = GetDayTuple(Words, rs.DateTimeValue, dayWindow)
        if(tweetsFound):
            DayY.append(GetDayVolatilityLabel(rs.Volatility))
            DayX.append(d)
        cnt = cnt + 1
        if (cnt%20 == 0): print("Fill Day Table counter: " + str(cnt))


    pdDayX = pd.DataFrame(DayX)
    pdDayY = pd.DataFrame(DayY)
    #fill N/A with zeros
    pdDayX = pdDayX.fillna(0)


    pdDayX.to_pickle(DayDataXFileName)
    pdDayY.to_pickle(DayDataYFileName)

    return pdDayX, pdDayY


def GetDayDataFromCSV():
    pdDayX = pd.read_pickle(DayDataXFileName)
    pdDayY = pd.read_pickle(DayDataYFileName)

    return pdDayX, pdDayY

def GetDayData(dayWindow):
    if (os.path.exists(DayDataXFileName)):
        return GetDayDataFromCSV()
    else:
        return GetDayDataFromDB(dayWindow)

pdDayX, pdDayY = GetDayData(1)

print("Start Training")

X_train, X_test, y_train, y_test = train_test_split(pdDayX, pdDayY, test_size = 0.1, shuffle=False)

scaler = StandardScaler()
scaler.fit(X_train)

X_train = scaler.transform(X_train)
X_test = scaler.transform(X_test)

mlp = MLPClassifier(activation='logistic', solver='lbfgs', alpha=0.001, batch_size='auto', beta_1=0.9,
    beta_2=0.999, early_stopping=False, epsilon=1e-08,
    hidden_layer_sizes=(layer1Size,layer2Size), learning_rate='constant',
    learning_rate_init=0.001, max_iter=50000, momentum=0.9,
    nesterovs_momentum=True, power_t=0.5, random_state=None,
    shuffle=False, tol=0.0001, validation_fraction=0.1,
    verbose=False, warm_start=False)

mlp.fit(X_train, y_train.values.ravel())

predictions = mlp.predict(X_test)
print(confusion_matrix(y_test,predictions))
print(classification_report(y_test,predictions))
print("Days Train Finished!")
```

Κώδικας 8: Κώδικας υλοποίησης Νευρωνικό Δικτύου