ΠΑΝΕΠΙΣΤΗΜΙΟ ΣΤΕΡΕΑΣ ΕΛΛΑΔΟΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗΝ ΒΙΟΪΑΤΡΙΚΗ

Σύγκριση τεχνικών συσταδοποίησης (clustering) σε δεδομένα από το Twitter

---

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

του

ΚΡΙΚΩΝΗ ΠΑΝΑΓΙΩΤΗ

Επιβλέπων Καθηγητής: ΑΝΑΓΝΩΣΤΟΠΟΥΛΟΣ ΙΩΑΝΝΗΣ

Λαμία, Μάρτιος 2013

Κρικώνης Χ. Παναγιώτης

Τελειόφοιτος Διπλωματούχος Τμήματος Πληροφορικής με Εφαρμογές στην Βιοϊατρική

Δηλώνω υπεύθυνα ότι δεν έχω υποπέσει σε περιπτώσεις λογοκλοπής ή αντιγραφής, όπως αυτές διασαφηνίζονται παρακάτω και κατόπιν ρητών οδηγιών που έλαβα από τον επιβλέποντά μου.

Κρικώνης Χ. Παναγιώτης

Οδηγίες αποφυγής Λογοκλοπής και Αντιγραφής *[απόσπασμα από http://www.samos.aegean.gr/actuar/dlekkas/reports/OdigiesEPO12.pdf]*

1. *Μην παραθέτετε κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί χωρίς να τα περικλείετε σε εισαγωγικά και χωρίς να αναφέρετε το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών σας, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Πρέπει να αναφέρετε πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*
2. *Η αυτολεξεί παράθεση χωρίς εισαγωγικά, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική σας εργασία. Αυτό τιμωρείται ως αντιγραφή.*
3. *Υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορείτε να εντάξετε στην εργασία σας εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια. Εσείς μπορείτε να χρησιμοποιείτε τέτοιο υλικό, με μέτρο, γιατί οι εργασίες είναι μικρού μεγέθους και πρέπει πάντα να κυριαρχούν οι δικές σας ιδέες.*
4. *Αυστηρά τιμωρείται επίσης η παρουσίαση έργου άλλων ως προσωπικής εργασίας.*

Ευχαριστίες

Η παρούσα εργασία αποτελεί τη Πτυχιακή μου Εργασία στα πλαίσια των σπουδών μου στο τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική του Πανεπιστημίου Στερεάς Ελλάδος, υπό την επίβλεψη του επίκουρου καθηγητή στον τομέα των τεχνολογιών και εφαρμογών διαδικτύου Ιωάννη Αναγνωστόπουλο, στον οποίο οφείλω ιδιαίτερες ευχαριστίες τόσο για την ανάθεση της εργασίας όσο και για τη γενικότερη βοήθεια, υποστήριξη και τις πολύτιμες συμβουλές που μου παρείχε καθ' όλη τη διάρκεια εκπόνησης της εργασίας.

Τέλος θα ήθελα να ευχαριστήσω την οικογένεια μου που με στήριζε και με συμβούλευε σε κάθε βήμα της φοιτητικής μου ζωής και να την αφιερώσω εις μνήμην της γιαγιάς μου Ελένης.

Περιεχόμενα

Πίνακας Εικόνων

Πίνακας Σχημάτων – Γραφημάτων

Περίληψη

Ζούμε μια εποχή τεχνολογικών εξελίξεων και τεχνολογικών αλμάτων με το Διαδίκτυο (Internet) να γίνεται ένα από τους βασικότερους εκφραστές των νέων τεχνολογικών τάσεων. Ωστόσο, ο τρόπος λειτουργίας του και δόμησής του παρουσιάζει εξαιρετικά μεγάλη ανομοιογένεια και αταξία με αποτέλεσμα οι χρήστες να βρίσκονται συχνά μπροστά από αδιέξοδο στην προσπάθεια αναζήτησης και αντίληψης του περιεχομένου των παρεχόμενων πληροφοριών. Καθημερινά κατακλυζόμαστε από μηνύματα κάθε μορφής, όπως ειδήσεις και feeds τόσο από τα διάφορα μέσα ενημέρωσης, όσο, πιο πρόσφατα και από τα διάφορα κοινωνικά δίκτυα όπως το Twitter, το Facebook και το Linkedin.

Τα διαδικτυακά κοινωνικά δίκτυα (OSN) είναι ένας καινούριος τομέας παροχής υπηρεσιών ο οποίος λόγω της αλματώδους ανάπτυξής του, αποτελεί ένα δυναμικό παράγοντα με ισχυρή επίδραση στη παγκόσμια οικονομία, τις κατευθύνσεις της τεχνολογικής ανάπτυξης καθώς και τις κοινωνικές εξελίξεις. Για αυτό το λόγο τα OSN παρέχουν μια μοναδική ευκαιρία για ένα ευρύ φάσμα έρευνας που οπωσδήποτε πρέπει να περιλαμβάνει μεθοδολογίες για τη συλλογή αυτών των δεδομένων καθώς και την επεξεργασία τους. Αυτή την περιοχή προσπαθεί να ενισχύσει η συγκεκριμένη πτυχιακή εργασία, η οποία επικεντρώνεται στο Twitter , ένα από τα μεγαλύτερα OSN, που λόγω των ιδιαιτεροτήτων του έχει αναδειχτεί σε μια πολύ σημαντική πηγή πληροφοριών.

Στα πλαίσια λοιπόν της παρούσας πτυχιακής εργασίας πραγματοποιήθηκε συλλογή ενός συνόλου tweets, τα οποία αντλήθηκαν από προκαθορισμένες κατηγορίες hashtag του κοινωνικού δικτύου Twitter. Ακολουθώντας τα βήματα μιας διαδικασίας εξόρυξης δεδομένων, πραγματοποιήθηκε προεπεξεργασία και έπειτα συσταδοποίηση των tweets αυτών, βασιζόμενοι στο κειμενικό τους περιεχόμενο. Η εκτέλεση των αλγορίθμων και η αναπαράσταση των αποτελεσμάτων της έρευνας μας έγιναν με τη χρήση της εξειδικευμένης πλατφόρμας του WEKA.

Abstract

We live an era of technology advances and huge technological steps where the Internet becomes a basic place of demonstration of the technology trends. Nevertheless, the way of operation and construction of the WWW is extremely uneven and this results in dead-ends when the users are trying to locate or understand the content of the provided information. We are daily inundated with any kind of messages, such as new and feeds, by the various media and most recently from the various social networks like Twitter, Facebook and Linkedin.

The online social networks (OSN) is an new service sector that due to its rapid development has a strong impact on the global economy, the direction of technological development and social change. These are the reasons why the OSN provide a unique opportunity for research that should include methodologies for the collection of data and its processing.

In this diploma thesis we collected a data set of tweets that contained at predefined hashtag categories. After following certain steps of a data mining procedure, those tweets had been pre-processed and clustered in different teams, according to their textual content. By using WEKA platform we managed to execute the clustering algorithms and present the results of the whole procedure.

1. Εισαγωγή

Το Internet στις μέρες μας αποτελεί ένα χαώδες δίκτυο στο οποίο υπάρχει κάθε λογής πληροφορία σε διάφορους ιστοτόπους, ιστοσελίδες ή ιστολόγια, κάθε ένα από τα οποία μπορεί να είναι περισσότερο ή λιγότερο αξιόπιστο και αναγνωρισμένο. Παρόλα αυτά ο χρήστης διαθέτει τη δυνατότητα να αναζητεί την πληροφορία που τον ενδιαφέρει από το τεράστιο αυτό σύνολο δεδομένων. Βεβαίως τα παραπάνω προϋποθέτουν ότι ο χρήστης γνωρίζει πώς να αξιολογεί και να συγκρίνει το σύνολο των πηγών αυτών που του παρέχουν την πληροφορία που αναζητά.

Στη σύγχρονη, βέβαια, εποχή της τεχνολογίας και της διαδικτυακής κοινωνικότητας του ατόμου, δεν μπορεί να παραλειφθεί η συνεισφορά των κοινωνικών δικτύων στην παροχή διαφόρων τύπων πληροφοριών προς το χρήστη. Οι δυνατότητες και η ανάπτυξη των κοινωνικών δικτύων έχουν εξελιχθεί σε τέτοια έκταση και ποιότητα που διεισδύουν ταχύτατα σε όλους τους τομείς της κοινωνικής ζωής και εξέλιξης του ατόμου. Τα κοινωνικά δίκτυα χρησιμοποιούν τις πολλαπλασιαστικές δυνατότητες των μελών τους και γίνονται ιστότοποι συγκέντρωσης τεράστιου αριθμού πληροφοριών αλλά και ταυτόχρονα της διανομής τους.

Κάτι αντίστοιχο συναντάται και στο Twitter. Το Twitter, ως ένα μεγάλο και ιδιαίτερα αναγνωρισμένο micro-blogging κοινωνικό δίκτυο προβάλει στο χρήστη ένα σύνολο πληροφοριών που ανέρχεται περίπου στο ποσό των 400 εκατομμυρίων tweets καθημερινά [1]. Ειδικότερα στην περίπτωση που ο χρήστης «ακολουθεί» ένα μεγάλο σύνολο άλλων χρηστών στο Twitter, το προσωπικό home-line του προφίλ του αυξάνεται με υπερβολικά μεγάλους ρυθμούς από το σύνολο των tweets που προβάλλονται.

Ξεχωριστή σημασία παρουσιάζει το γεγονός ότι οι χρήστες πολλές φορές περιλαμβάνουν στα tweets που δημοσιεύουν και hashtag. Τα hashtag αποτελούν λέξεις κλειδιά που ξεκινούν με το πρόθεμα «#» και χρησιμοποιούνται για να

ομαδοποιήσουν και να κατηγοριοποιήσουν, με τη μορφή θεματικών ενοτήτων, τα tweets που δημοσιεύουν οι χρήστες. Με την δημιουργία ενός hashtag παρέχεται η δυνατότητα στους χρήστες να δουν το σύνολο των tweets που αποτελούν το hashtag αυτό.

1.1. Πρόβλημα

Το πρόβλημα έγκειται στο γεγονός πως το σύνολο της πληροφορίας αυτής που προσφέρεται από το Twitter στο χρήστη, με οποιαδήποτε μορφή, όπως εικόνων, βίντεο, κειμένου κ.α, μπορεί να είναι παντελώς «διάσπαρτο». Έτσι ο χρήστης ως επί το πλείστον αδυνατεί να αντιληφθεί το σκοπό ή το περιεχόμενο της δημοσίευσης ενός tweet. Αν ο χρήστης έχει τη δυνατότητα να διαβάσει ένα προς ένα τα περιεχόμενα του κάθε tweet που εμφανίζεται στο προφίλ του, τότε το πρόβλημα τίθεται υπό έλεγχο, καθώς ο ίδιος ο χρήστης με λογικά κριτήρια αντιλαμβάνεται και καθορίζει το θεματικό περιεχόμενο της κάθε δημοσίευσης.

Με βάση την ύπαρξη και τη διαιώνιση του παραπάνω προβλήματος κρίνεται επιτακτική η ανάγκη άμεσης οργάνωσης και διαχωρισμού της πληροφορίας που παρέχεται στους χρήστες μέσω των social media. Μέσω της οργάνωσης αυτής ο χρήστης μπορεί να επιλέγει το σύνολο των πληροφοριών που επιθυμεί μεταξύ διαφορετικών θεματικών ενοτήτων. Αυτό επιτυγχάνεται με τη χρήση μιας διαδικασίας ομαδοποίησης, της συσταδοποίησης.

1.2. Σκοπός Πτυχιακής Εργασίας

Η παρούσα διπλωματική εργασία πραγματοποιήθηκε με στόχο την έρευνα και τον διαχωρισμό της πληροφορίας αυτής που παρέχεται στο σύνολο των χρηστών του Twitter. Η άντληση των tweet, που όπως αναφέρθηκε συγκεντρώνουν το σύνολο των πληροφοριών, πραγματοποιήθηκε με βάση συγκεκριμένα hashtag, τα οποία ορίστηκαν εξαρχής και αφορούσαν ένα επιθυμητό εύρος θεματικών

ενοτήτων. Με την κατάλληλη χρήση συγκεκριμένων τεχνικών επεξεργασίας και αλγορίθμων συσταδοποίησης πραγματοποιήθηκε διαχωρισμός σε κατηγορίες των tweets, βασιζόμενοι στο κειμενικό τους περιεχόμενο. Η υλοποίηση και η αναπαράσταση των κατηγοριών που προέκυψαν, έγιναν με τη χρήση του περιβάλλοντος του WEKA.

1.3. Βασικές Λειτουργίες

Αρχικά, στα πλαίσια της παρακάτω πτυχιακής εργασίας, μελετήθηκε διεξοδικά το API που προσφέρει το κοινωνικό δίκτυο Twitter και αναπτύχθηκε ένα πρόγραμμα το οποίο πραγματοποιούσε την άντληση των απαραίτητων δεδομένων. Τα δεδομένα αυτά αποτελούσαν tweets που συλλέγονταν από προκαθορισμένες θεματικές ενότητες - hashtag. Το API που χρησιμοποιήθηκε για την εφαρμογή μας ονομάζεται Twitter Search API και επιλέχθηκε η χρήση του (έναντι του REST API), εξαιτίας της πιο ακριβής αναζήτησης του για πρόσφατα tweets.

Οι θεματικές ενότητες, βάση των οποίων αντλήθηκε και το σύνολο των tweets, χωρίζονταν σε 5 μεγάλες κατηγορίες. Κάθε μία από τις κατηγορίες αυτές διαχωρίζονται περαιτέρω σε άλλες 5 υποκατηγορίες θεματικών ενοτήτων. Έτσι ο συνολικός αριθμός των υποκατηγοριών hashtag ανέρχονταν στις 25. Για κάθε μια υποκατηγορία από τα hashtag, που είχαμε ορίσει εξαρχής, η εφαρμογή αντλούσε τα 200 πιο πρόσφατα tweets που περιλάμβανε το hashtag αυτό, με αποτέλεσμα να πραγματοποιηθεί συλλογή 5000 tweets συνολικά.

Επόμενη κατά σειρά διαδικασία ήταν η κατάλληλη επεξεργασία των δεδομένων που είχαμε λάβει. Το κειμενικό περιεχόμενο των tweet που αντλήσαμε περιείχε αρκετά περιττά στοιχεία, τα οποία απαλείφτηκαν με κατάλληλες διαδικασίες. Επιπλέον για την ορθή αναπαράσταση των δεδομένων πραγματοποιήθηκε διαχωρισμός του περιεχομένου σε tokens – λέξεις. Γενικότερα η προεπεξεργασία του κειμενικού περιεχομένου πραγματοποιεί τη βασική λειτουργία της εξαγωγής των χαρακτηριστικών όρων των tweet κάθε hashtag, που ονομάζονται όροι δεικτοδότησης (index terms) και είναι οι κατάλληλοι όροι, εφόσον

ολοκληρωθεί η διαδικασία, να αναπαραστήσουν και να αντιπροσωπεύσουν το κειμενικό περιεχόμενο των tweet.

Προκειμένου να πραγματοποιηθεί η επιθυμητή κατηγοριοποίηση των tweets, αναπαραστήσαμε τα term, που προέκυψαν από κάθε tweet, σε μια μορφή επεξεργάσιμη και κατανοητή από τους αλγορίθμους συσταδοποίησης. Για την αναπαράσταση αυτή απαιτήθηκε η δεικτοδότηση του κάθε term ξεχωριστά με βάση της τεχνικής απόδοσης βάρους TF – IDF. Με τον τρόπο αυτό κάθε term αποτελούσε και ένα διάνυσμα με συγκεκριμένη διάσταση και συγκεκριμένο βάρος για κάθε έναν πεδίο. Τα πεδία ουσιαστικά αποτελούσαν τις 25 διαφορετικές υποκατηγορίες hashtag, εκ των οποίων αντλήθηκαν τα δεδομένα.

Εφόσον ολοκληρώσαμε την επεξεργασία και το μετασχηματισμό των term σε κατάλληλα διανύσματα, σειρά έχει η διαδικασία της επιλογής των αλγορίθμων συσταδοποίησης για την εξόρυξη των δεδομένων. Οι αλγόριθμοι εφαρμόστηκαν με τη χρήση του περιβάλλοντος του WEKA. Από το σύνολο των αποτελεσμάτων που παρήγαγε ο κάθε αλγόριθμος, επιλέχθηκαν τα καταλληλότερα (με βάση συγκεκριμένα κριτήρια) και αναπαραστάθηκαν με τη μορφή γραφημάτων.

2. Κοινωνικά Δίκτυα και Social Media

2.1. Εισαγωγή στην Κοινωνική Δικτύωση

2.1.1. Ορισμός Κοινωνικών Δικτύων και Social Media

Ο άνθρωπος ως κοινωνικό ον από τα αρχαία χρόνια αναζητούσε τρόπους επικοινωνίας με τους κοντινούς και μη ανθρώπους. Ανεξάρτητα του μέσου, είτε αυτό ήταν σήματα καπνού, αλληλογραφία, τηλέφωνο, e-mail, βιντεοκλήση ο απώτερος σκοπός ήταν πάντα η εδραίωση και συνέχιση ενός κοινωνικού δικτύου μεταξύ ατόμων ή οργανισμών.

Κοινωνικό δίκτυο αποτελεί η συγκέντρωση ή συμμετοχή των ατόμων σε συγκεκριμένες ομάδες, όπως για παράδειγμα οι αγροτικές κοινότητες, οι χώροι εργασίας, τα πανεπιστήμια και τα σχολεία. Στη βιβλιογραφία εμφανίζονται διάφοροι ορισμοί για το τι είναι ένα κοινωνικό δίκτυο:

- Ο Χτούρης (Χτούρης 2004) ορίζει ως κοινωνικά δίκτυα τα «πολυδιάστατα συστήματα επικοινωνίας και διαμόρφωσης της ανθρώπινης πρακτικής και της κοινωνικής ταυτότητας».

- Οι Walker, MacBride και Vachon (1977), όρισαν ως κοινωνικό δίκτυο το άθροισμα των προσωπικών επαφών μέσω των οποίων το άτομο διατηρεί την κοινωνική του ταυτότητα,λαμβάνει συναισθηματική υποστήριξη, υλική ενίσχυση και συμμετοχή στις υπηρεσίες, έχει πρόσβαση στις πληροφορίες και δημιουργεί νέες κοινωνικές επαφές.

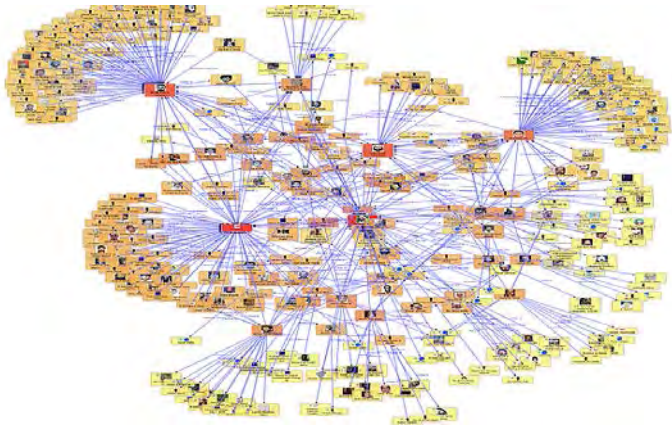
- Τα κοινωνικά δίκτυα συνήθως αποτελούνται από τα μέλη της οικογένειας, τους φίλους και τους γνωστούς και περιλαμβάνουν τρεις κρίσιμες έννοιες: α) το μέγεθος ή το εύρος, το οποίο αναφέρεται στον αριθμό των ατόμων που συμμετέχουν στο δίκτυο, β) τη σύνθεση, δηλαδή το ποσοστό συμμετοχής μελών της ευρύτερης οικογένειας ή φίλων στο δίκτυο, γ) τη συχνότητα, που δηλώνει το πόσο συχνά τα μέλη ενός κοινωνικού δικτύου αλληλεπιδρούν μεταξύ τους (Χτούρης, Παπάνης, Ρόντος, 2004).

Ένα κοινωνικό δίκτυο επομένως, ιδιαίτερα όταν αυτό πραγματοποιείται μέσω της online κοινωνικής δικτύωσης, είναι μια κοινωνική δομή που αποτελείται από τα άτομα ή οργανισμούς που ονομάζονται «κόμβοι», οι οποίοι συνδέονται με έναν ή περισσότερους συγκεκριμένους τύπους σχέσεων, όπως η φιλία, συγγένεια, γνώσεις, κοινά ενδιαφέροντα, οικονομικές ανταλλαγές, αντιπάθειες, σεξουαλικές σχέσεις ή τις κοινές πεποιθήσεις.

Η ανάλυση των κοινωνικών δικτύων παρατηρεί τις κοινωνικές σχέσεις από την άποψη της θεωρίας του δικτύου αποτελούμενες από κόμβους και δεσμούς (ονομάζονται επίσης ακμές ή συνδέσεις). Οι κόμβοι είναι τα άτομα στο πλαίσιο των δικτύων, και οι δεσμοί είναι οι σχέσεις μεταξύ των ατόμων αυτών. Το γράφημα που προκύπτει με βάση τις δομές συχνά είναι ιδιαίτερα πολύπλοκο. Μπορεί να υπάρχουν πολλά είδη των δεσμών μεταξύ των κόμβων.

Πολλές έρευνες σε μια σειρά διαφορετικών επιστημονικών πεδίων έχουν δείξει ότι τα κοινωνικά δίκτυα εφαρμόζονται σε πολλά επίπεδα της ζωής του ανθρώπου, από οικογένειες μέχρι το επίπεδο των εθνών, και διαδραματίζουν κρίσιμο ρόλο στον καθορισμό του τρόπου επίλυσης προβλημάτων, στην ενημέρωση των πωλητών, καθώς και στο τρόπο ατομικής και συλλογικής συμπεριφοράς των ατόμων.

Στην απλούστερη μορφή του, ένα κοινωνικό δίκτυο είναι ένας χάρτης των ειδικών δεσμών (Εικόνα 1), όπως η φιλία, μεταξύ των κόμβων υπό μελέτη. Οι κόμβοι στους οποίους ένα άτομο επομένως συνδέεται είναι και οι κοινωνικές επαφές του εν λόγω ατόμου. Το δίκτυο μπορεί επίσης να χρησιμοποιηθεί για τη μέτρηση του κοινωνικού κεφαλαίου- την αξία που ένα άτομο παίρνει από το κοινωνικό δίκτυο. Οι έννοιες αυτές συχνά εμφανίζονται σε ένα κοινωνικό διάγραμμα δικτύου, όπου οι κόμβοι είναι τα σημεία και οι δεσμοί τους οι γραμμές.



Εικόνα 1 Γράφος Κοινωνικού Δικτύου

Για την εδραίωση ενός κοινωνικού δικτύου αναγκαία συνθήκη είναι το κοινωνικό μέσον. Το κοινωνικό μέσον μπορεί να έχει την μορφή οποιασδήποτε αλληλεπιδραστικής επικοινωνίας. Σήμερα με τον όρο κοινωνικό μέσον ή social media αναφερόμαστε στην χρήση web-based τεχνολογιών ώστε να μετατρέψουν την επικοινωνία σε διαδραστικό διάλογο.

Αναλυτικότερα ο ορισμός των social media, όπως δίνεται από τη Wikipedia [2] είναι ο ακόλουθος: «μέσα σχεδιασμένα για να διαχέονται μέσω της κοινωνικής αλληλεπίδρασης, με υψηλή προσβασιμότητα και με τεχνικές δημοσίευσης που μπορούν να μεταβάλλονται. Τα social media χρησιμοποιούν το Internet και τεχνολογίες που βασίζονται στο web για να μετασχηματίζουν μονολόγους που εκπέμπονται στα media (one to many) σε social media διαλόγους (many to many).

Οι πληροφορίες που διακινούνται μέσω των κοινωνικών δικτύων έχουν τύπο κειμένου, φωτογραφιών και βίντεο. Οι χρήστες μοιράζονται με μια ομάδα ή όλους πληροφορίες όπως ημερομηνία γέννησης, όνομα κι επώνυμο, διεύθυνση, e-mail, φωτογραφίες και βίντεό τους ή και τρίτων προσώπων, εφαρμογές και πολλά άλλα. Τα γνωστότερα social media σήμερα είναι το facebook,το YouTube (εξειδικευμένο στη χρήση βίντεο), το twitter και το flickr (εξειδικευμένο στη χρήση φωτογραφιών. Τα μέσα αυτά μπορούν να χρησιμοποιηθούν για ψυχαγωγία, ενημέρωση, οικονομικούς και διαφημιστικούς λόγους, παράθεση απόψεων ,διάλογο ακόμη και για οργάνωση. Επομένως τα social media αναφέρονται σε μια ποικιλία υπηρεσιών πληροφορίας που χρησιμοποιούνται συνεργατικά από πολλούς ανθρώπους και διακρίνονται στις εξής κατηγορίες:

ΚΑΤΗΓΟΡΙΕΣ	ΠΑΡΑΔΕΙΓΜΑΤΑ
Blogs	Blogger, LiveJournal, WordPress
MicroBlogs	Twitter, GoogleBuzz
Opinion mining	Epinions, Yelp

Photo and Video Sharing	Flickr, YouTube
Social bokkmarking	Delicious, StumbleUpon
Social networking sites	Facebook, LinkedIn, MySpace, Orkut
Social news	Digg, Slashdot
Wikis	Scholarpedia, Wikihow, Wikipedia, Eventmaps

Τα κριτήρια που ξεχωρίζουν τα κοινωνικά μέσα από παραδοσιακά ή μέσα μαζικής ενημέρωσης, όπως οι εφημερίδες, η τηλεόραση και το ραδιόφωνο είναι η προσβασιμότητα από όλους, το γεγονός ότι είναι δωρεάν ή πολύ φτηνά, η δυνατότητα όλων να κοινοποιήσουν πληροφορίες, η ευκολία στην χρήση όπως και η χρονική ελευθερία, για παράδειγμα μια εφημερίδα μπορεί να τυπώνεται μια φορά την μέρα ή την εβδομάδα, ενώ οι ανανεώσεις στα σύγχρονα κοινωνικά μέσα μπορεί να φτάνουν και το ένα δευτερόλεπτο.

2.1.2. Ιστορική Αναδρομή Κοινωνικών Δικτύων

Οι πρώτες υπηρεσίες κοινωνικής δικτύωσης στο Διαδίκτυο ήταν το CBBS (Computerized Bulletin Board System | 1978) και το Usenet (1980) τα οποία συστάθηκαν από ομοϊδεάτες προγραμματιστές με σκοπό την επικοινωνία σχετικά με συγκεκριμένα θέματα. Οι δημιουργοί του CBBS, Ward Christensen και Randy Suess προγραμματιστές της IBM ανέπτυξαν ένα εικονικό σύστημα, όπου οι χρήστες μπορούσαν να αναρτήσουν δημόσια μηνύματα προσομοιώνοντας τον πίνακα ανακοινώσεων γραφείου. Λόγω όμως της δημιουργίας και της χρήσης τους από προγραμματιστές, το ευρύ κοινό θεώρησε ότι αφορούσε άτομα με μεγάλο ενδιαφέρον και γνώσεις στην τεχνολογία, με αποτέλεσμα να μην είναι πολύ

δημοφιλής στο κοινό.

Η ιστορία των κοινωνικών δικτύων, ξεκινάει ουσιαστικά από τα μέσα της δεκαετίας του ’90 όπου τα πρώτα κοινωνικά δίκτυα κάνουν την εμφάνισή τους με τη μορφή γενικών κοινοτήτων. Παραδείγματα αυτών είναι το WELL (Whole Earth 'Lectronic Link | 1985), το Tripod (1992), το Theglobe (1994), και το GeoCities (1994). Στην ουσία αυτό που προσπάθησαν να κάνουν οι κοινότητες αυτές ήταν να φέρουν κοντά τους χρήστες, να μοιραστούν προσωπικές πληροφορίες και ιδέες μέσω εργαλείων και προσωπικών δημοσιεύσεων. Ουσιαστικά αποτελούσαν προγόνους των σημερινών ιστολογίων.

Κοινό χαρακτηριστικό αυτών των Ιστοσελίδων Κοινωνικής Δικτύωσης ήταν το chat room, ένας ειδικός δικτυακός χώρος με δυνατότητα παρουσίας πολλαπλών χρηστών (multi-user) που επέτρεπε στους ανθρώπους να πληκτρολογούν μηνύματα ο ένας στον άλλο και να λαμβάνουν απαντήσεις σε πραγματικό χρόνο .

Οι σύγχρονες Ιστοσελίδες Κοινωνικής Δικτύωσης άρχισαν να διακρίνονται από τις προγενέστερες, απλούστερες κοινότητες αποστολής μηνυμάτων από τον τρόπο με τον οποίο το δίκτυο του χρήστη γίνεται ορατό στους άλλους. Η πρώτη Ιστοσελίδα Κοινωνικής Δικτύωσης που συνδύασε αυτά τα χαρακτηριστικά λειτουργίας ήταν το Six Degrees.com (1997). Οι κύριες υπηρεσίες που παρεχόταν στα μέλη του, ήταν η δημιουργία προφίλ, η αποστολή μηνυμάτων σε διαδικτυακούς φίλους και η ομαδοποίηση τους σε λίστα, καθώς και την πλοήγηση στις λίστες των φίλων. Παρόλο που η ιστοσελίδα ήταν εξαιρετικά δημοφιλής με πολλά εκατομμύρια χρήστες έκλεισε δύο χρόνια αργότερα καθώς δεν απέφερε κέρδη. Ο δημιουργός της ιστοσελίδας μάλιστα ανέφερε ότι ο κόσμος δεν ήταν έτοιμος να αποδεχτεί μια τέτοια πρωτοπορία.

Από το 1997 μέχρι το 2001, αναπτύχθηκαν web-καινοτομίες που έδιναν στους χρήστες όχι μόνο τη δυνατότητα να βλέπουν ποιος είναι φίλος με ποιόν, αλλά τους επέτρεπε να έχουν καλύτερο έλεγχο στην συνδεσιμότητα τους με άλλους χρήστες. Τα εργαλεία της κοινότητας που παρέχοντας οδήγησαν στην δυνατότητα

δημιουργίας διαφόρων συνδυασμών προφίλ και δημόσιου βαθμού φιλίας.

Ιστοσελίδες όπως το AsianAvenue (1999),το BlackPlanet (1999), και το MiGente (2000) επέτρεπαν στους χρήστες τους να δημιουργούν συνδυασμό από προφίλ, όπως προσωπικό, επαγγελματικό, αισθηματικό και να δημιουργούν φίλους χωρίς να απαιτείται η έγκριση της σύνδεσης.

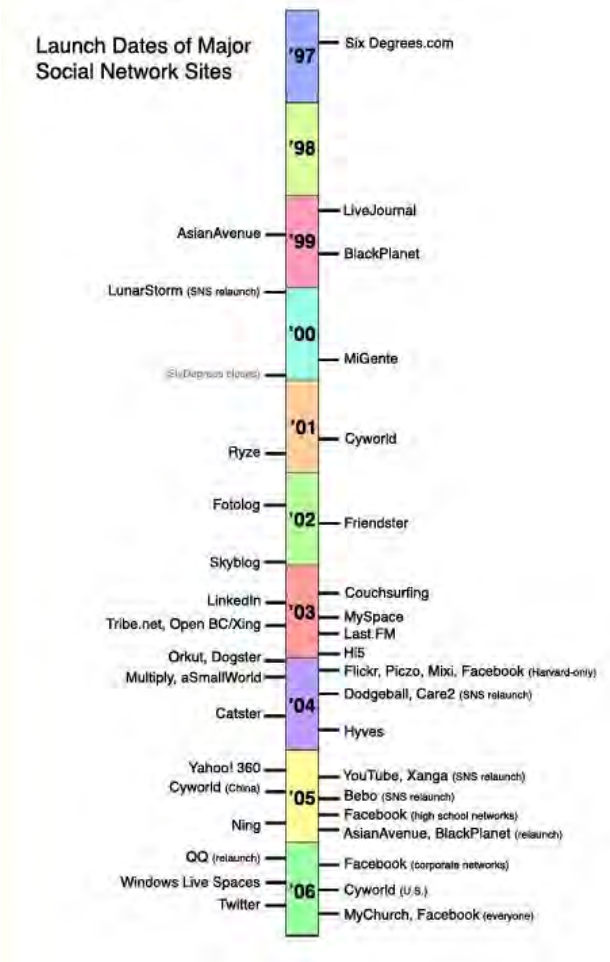
Η νέα γενιά κοινωνικών δικτύων εμφανίστηκε το 2001 με το Ryze (2001) που είχε σκοπό να βοηθήσει τα άτομα να αξιοποιούν τα επιχειρηματικά τους δίκτυα το οποίο ποτέ δεν απέκτησε μεγάλη δημοσιότητα, ενώ από το 2003, αναπτύχθηκαν πολλές νέες υπηρεσίες κοινωνικής δικτύωσης και εμφανίστηκε ο όρος YASNS: «Yet Another Social Networking Service».

Χαρακτηριστικά μπορούμε να αναφέρουμε τα δίκτυα LinkedIn, Visible Path, και Xing τα οποία αποτάθηκαν στον επιχειρηματικό κόσμο, ενώ κάποια άλλα όπως τα: Couchsurfing(2003 | συνδέσεις ταξιδιωτών), Dogster (2004 | φιλίες μεταξύ ατόμων βάσει ενδιαφέροντος για τους σκύλους), Care2 (2004 |συναντήσεις ακτιβιστών), MyChurch (σύνδεση χριστιανικών εκκλησιών και των μελών τους) αποτέλεσαν προσπάθειες για δημιουργία κοινοτήτων κοινών ενδιαφερόντων.

Καθώς όλο και περισσότεροι άνθρωποι αποκτούσαν πρόσβαση και ταχύτερες συνδέσεις στο διαδίκτυο με ταυτόχρονη μείωση του κόστους, κοινωνικά δίκτυα όπως το MySpace (2003), ή το HiFive (2003), άρχισαν να προσελκύουν το παγκόσμιο ενδιαφέρον και να γίνονται δημοφιλή και ευρέως γνωστά και αναγνωρίσιμα. Το γεγονός αυτό αύξησε σημαντικά τον όγκο του περιεχομένου που δημιουργούνταν από τους χρήστες και γινόταν διαθέσιμο στο διαδίκτυο και σαν αποτέλεσμα οι ιστοσελίδες που παρείχαν πλατφόρμες για τη δημοσίευση μουσικής Last.FM (2003) φωτογραφιών Flickr.com (2004), ή video YouTube.com (2005) , καθώς και υπηρεσίες instant messaging, συζητήσεων, ιστολογίων, άρχισαν και αυτές να αποκτούν χαρακτηριστικά Ιστοσελίδες Κοινωνικής Δικτύωσης.

Το 2004 ήρθε στο φως το Facebook, που υπήρξε πολύ καλός ανταγωνιστής, και η ανάπτυξη του ήταν το ίδιο γρήγορη. Το 2006 ήταν η χρόνια που το Facebook

σταμάτησε να απευθύνεται μόνο στην κοινότητα των αμερικανικών κολλεγίων, και άρχισε να χρησιμοποιείται από ανθρώπους σε όλο τον κόσμο. Αυτό που το έκανε τόσο αγαπητό ήταν το γεγονός ότι αναπτύχθηκε μία πληθώρα εφαρμογών καθιστώντας το ευχάριστο στη χρήση και το γεγονός ότι δεν υπήρχε κανένα γεωγραφικό όριο στην επικοινωνία μεταξύ των χρηστών. [3]



Εικόνα 2 Ιστορική Εξέλιξη Ιστοσελίδων Κοινωνικής Δικτύωσης

2.1.3. Οφέλη και Κίνδυνοι μέσω της χρήσης των Κοινωνικών Δικτύων

Η χρήση online κοινωνικών δικτύων προσφέρει στους χρήστες πολλά οφέλη και πλεονεκτήματα. Η γρήγορη και ανεξέλεγκτη διάδοση του Διαδικτύου οδήγησε

στην εύκολη κατάργηση των αποστάσεων μεταξύ των ανθρώπων με αποτέλεσμα τα κοινωνικά δίκτυα να αποκτήσουν παγκόσμια έκταση. Η χρήση των online σελίδων κοινωνικής δικτύωσης επομένως προσφέρει ένα μεγάλο σύνολο από οφέλη στους χρήστες τους, κρύβοντας όμως και ένα σύνολο κινδύνων.

Οφέλη

- Η δυνατότητα δημιουργίας δεσμών με πολύ μεγάλο αριθμό ατόμων, εφόσον το Διαδίκτυο συγκεντρώνει πλήθος άτομα από όλο τον κόσμο.

- Η δυνατότητα δημιουργίας μεγάλης ποικιλίας κοινωνικών δεσμών εφόσον στο Διαδίκτυο συρρέουν άτομα από διαφορετικές χώρες, κοινωνίες, πολιτισμούς και με διαφορετικές συνήθειες και χαρακτηριστικά. Με τον τρόπο αυτό ο κάθε χρήστης έχει τη δυνατότητα να γνωρίσει νέους πολιτισμούς και έθιμα από περιοχές που στην πραγματικότητα δεν θα είχε επισκεφθεί ποτέ.

- Η δυνατότητα επαφής με πολλούς διαφορετικούς πολιτισμούς και επομένως η δυνατότητα διεύρυνσης των γνώσεων και των πνευματικών οριζόντων του ατόμου.

- Η δυνατότητα δημιουργίας δεσμών με άτομα που μπορεί να βρίσκονται σε μεγάλη γεωγραφική απόσταση το ένα από το άλλο, εφόσον το Διαδίκτυο καταργεί τις αποστάσεις. Η δυνατότητα αυτή αποτελεί ένα προνόμιο μόνο των online κοινωνικών δικτύων καθώς η δημιουργία των στενών κοινωνικών δικτύων του άμεσου κοινωνικού περίγυρου ενός ατόμου δεν επιτυγχάνεται σε μεγάλες γεωγραφικές αποστάσεις.

- Η δυνατότητα ελεύθερης επιλογής ανάμεσα σε μεγάλο πλήθος κοινωνικών ομάδων και η αναζήτηση της ομάδας που εκφράζει και ωφελεί το άτομο με τον καλύτερο δυνατό τρόπο.

- Η δυνατότητα αναζήτησης και εύρεσης περιεχομένου (φωτογραφιών, βίντεο κλπ) στο οποίο οι χρήστες θα αδυνατούσαν να έχουν πρόσβαση με διαφορετικό τρόπο.

- Η δυνατότητα άμεσης, γρήγορης και ταυτόχρονα έγκυρης ενημέρωσης για οτιδήποτε συμβαίνει τόσο γύρω από το χρήστη όσο και σε οποιοδήποτε σημείο του κόσμου.

- Η δυνατότητα δραστηριοποίησης και σύμπραξης για κοινούς σκοπούς και στόχους με άτομα που μπορεί να βρίσκονται οπουδήποτε στη γη.

- Η δυνατότητα πρόσβασης σε ψυχαγωγικό περιεχόμενο, μέσω της ενασχόλησης με εφαρμογές ψυχαγωγικού χαρακτήρα (όπως online παιχνίδια), καθώς επίσης και με την παρακολούθηση βίντεο.

Κίνδυνοι

- Η έκθεση σε πολύ μεγαλύτερο αριθμό κινδύνων κοινωνικού χαρακτήρα. Ενώ θα περίμενε κανείς ότι το Διαδίκτυο λόγω της απόστασης που εμπεριέχει, θα προφύλασσε τα άτομα από την επαφή με πραγματικούς κινδύνους, είναι πολλές οι περιπτώσεις όπου ιδιαίτερα οι νέοι βιώνουν έντονα καταστάσεις μέσα από την χρήση του κυβερνοχώρου, οι οποίες είναι τραυματικές για την ψυχική τους υγεία. Χαρακτηριστικό παράδειγμα αποτελεί η έκφραση λεκτικού κυρίως ρατσισμού από ένα σύνολο χρηστών σε κάποιο άλλο. Η έκφραση αυτή συνοδεύεται ως επί το πλείστον και με προσβλητικές και χυδαίες εκφράσεις ή δημοσιεύσεις στους ιστότοπους κοινωνικής δικτύωσης.

- Ο κίνδυνος εξαπάτησης, δηλ. η σύνδεση με άτομα που ισχυρίζονται ότι είναι κάποιοι που δεν είναι στην πραγματικότητα. Οι χρήστες του Διαδικτύου μπορούν να διατηρήσουν την ανωνυμία τους όταν έρχονται σε επαφή με άλλους χρήστες, αλλά μπορούν εύκολα και να

εξαπατήσουν ή να εξαπατηθούν από τους υπόλοιπους χρήστες εξαιτίας αυτής της δυνατότητας.

- Η επιβλαβής έκθεση της προσωπικής ζωής του ατόμου. Τα online κοινωνικά δίκτυα είναι χώροι όπου μπορεί να συγκεντρώνεται πλήθος διαφορετικών και άγνωστων ατόμων. Η δημοσίευση προσωπικών στοιχείων σε αυτά, καθιστά τους χρήστες τους ευάλωτους σε πολύ μεγαλύτερο αριθμό κακόβουλων ατόμων. Η παράνομη φυσικά απόκτηση και χρήση των στοιχείων ενός χρήστη αποδεδειγμένα στο παρελθόν οδήγησε σε ανεξέλεγκτες συνέπειες.

- Η παρενόχληση από άτομα εντελώς άγνωστα μέλη του κοινωνικού δικτύου χωρίς τη δυνατότητα προστασίας ή αντιμετώπισης τέτοιων ενεργειών. Ειδικότερα το φαινόμενο αυτό εντείνεται σε νεαρές ηλικίες χρηστών και φτάνει στα όρια της σεξουαλικής παρενόχλησης πολλές φορές.

- Η εύκολη μετάδοση και επαφή με ηλεκτρονικό περιεχόμενο που δεν είναι ασφαλές για τους χρήστες του, όπως κακόβουλο λογισμικό, πορνογραφικό υλικό, υλικό προσηλυτισμού, υλικό εθισμού σε επικίνδυνες ουσίες, υλικό με θέμα τη βία, ρατσιστικό περιεχόμενο κ.ά.

- Ο βομβαρδισμός με διαφημιστικά μηνύματα από τις σελίδες του ιστότοπου.

2.2. Twitter

2.2.1. Εισαγωγή

Το Twitter αποτελεί μια διαδικτυακή ιστοσελίδα κοινωνικής δικτύωσης, που προσφέρει υπηρεσίες microblogging και επιτρέπει στους χρήστες του να στέλνουν

και να διαβάζουν μηνύματα κειμένου έως 140 χαρακτήρων, γνωστά ως «tweets». Ως υπηρεσία κοινωνικής δικτύωσης κέρδισε γρήγορα την αναγνωρισιμότητα του σε όλο τον κόσμο, με πάνω από 500 εκατομμύρια εγγεγραμμένους έως το 2012, οι οποίοι δημιουργούν πάνω από 340 εκατομμύρια tweets την ημέρα. Κατά μέσο όρο το Twitter διαχειρίζεται καθημερινά πάνω από 1,6 δισεκατομμύρια αιτήματα αναζήτησης στους διακομιστές του (servers). Όλα τα παραπάνω οδήγησαν την ιστοσελίδα του Twitter στο να γίνει μια από τις 10 «συχνά-επισκεπτόμενες» ιστοσελίδες του Παγκόσμιου Ιστού, με αποτέλεσμα να του δοθεί και το όνομα «the SMS of Internet» [4].

Οι εγγεγραμμένοι χρήστες του Twitter, δημιουργώντας ένα λογαριασμό (profile), έχουν τη δυνατότητα να δημοσιεύουν (post) tweets τα οποία γίνονται ορατά από τους υπόλοιπους χρήστες αν επισκεφθούν τη σελίδα του προφίλ του συντάκτη που δημοσίευσε το προφίλ. Παρόλα αυτά οι συντάκτες των tweets μπορούν να επιλέξουν αν τα tweets αυτά θα είναι ορατά μόνο στην λίστα των φίλων τους, ή θα είναι δημοσίως προσβάσιμα και από το υπόλοιπο σύνολο χρηστών. Οι φίλοι που έχει ο κάθε χρήστης της ιστοσελίδας ονομάζονται followers.

Η ανταλλαγή των tweets μεταξύ των χρηστών μπορεί να γίνει είτε μέσω της ίδια της ιστοσελίδας του Twitter, είτε μέσω της υπηρεσίας των SMS και άλλων εφαρμογών για κινητά τηλέφωνα που έχουν δημιουργηθεί. Οι υπηρεσίες κοινωνικής δικτύωσης που παρέχονται από το Twitter είναι εντελώς δωρεάν, εφόσον φυσικά γίνει η χρήση τους μέσω της ίδιας της ιστοσελίδας. Η αποστολή SMS για τη δημοσίευση ενός tweet χρεώνει τον τηλεφωνικό λογαριασμό του χρήστη σαν να έστελνε ένα κανονικό μήνυμα κειμένου μέσω της τηλεφωνικής του συσκευής.

Το Twitter παρέχει τη δυνατότητα στους χρήστες του, να δημιουργούν ομάδες από tweets με σχετικό μεταξύ τους περιεχόμενο. Με τη χρήση ετικετών (tags), οι χρήστες δημιουργούν τα λεγόμενα «hashtag», ομαδοποιώντας έτσι tweets με παρόμοιο θέμα. Η δημιουργία ενός hashtag γίνεται με την εισαγωγή του prefix «#» πριν τη λέξη που ορίζει το θέμα του hashtag. Παράδειγμα αποτελεί το hashtag #tv, που συγκεντρώνει το σύνολο των tweet που αναφέρονται σε οποιοδήποτε

θέμα σχετίζεται με την τηλεόραση (tv).

Ένα χαρακτηριστικό που ξεχωρίζει το Twitter από τα υπόλοιπα social networks είναι η πολύ λιτή διεπαφή του. Η βασική ιδέα γύρω από το συγκεκριμένο social network είναι αυτή των followers. Όταν επιλέγει ο χρήστης να κάνει follow έναν άλλον, τότε τα tweets του δεύτερου εμφανίζονται με αντίστροφη χρονολογική σειρά στην κεντρική σελίδα του πρώτου. Τα tweets αυτά και ένα πεδίο στο οποίο αναγράφονται τα πιο πολυσυζητημένα θέματα της παρούσας χρονικής περιόδου καλύπτουν το μεγαλύτερο μέρος της σελίδας του χρήστη.

Το σύνολο των δυνατοτήτων που προσφέρει το Twitter, σε σχέση με τα υπόλοιπα κοινωνικά δίκτυα, και συγκεκριμένα η δυνατότητα γρήγορης και εύκολης δημοσίευσης ενός tweet, έχει οδηγήσει την ιστοσελίδα αυτή να αποτελεί, εκτός από ένα κοινωνικό δίκτυο, ένα ιστόχωρο όπου οι χρήστες μπορούν να ενημερωθούν γρήγορα και έγκυρα για τα σημαντικότερα θέματα που συμβαίνουν στον κόσμο, την στιγμή που συμβαίνουν. Βάση αυτού το Twitter κατατάσσεται σε πολύ υψηλή βαθμολογία στην προτίμηση των χρηστών για την γενική τους ενημέρωση σε ευρύ σύνολο θεμάτων [5], [6].

Η χρήση της ιστοσελίδας του Twitter χρησιμοποιείται κυρίως από άτομα μέσης ηλικίας τα οποία δεν έχουν χρησιμοποιήσει κάποια άλλη σελίδα κοινωνικής δικτύωσης [7]. Μόνο το 11% του συνόλου των χρηστών αποτελείται από ανήλικους χρήστες της ηλικίας των 12-18 χρονών. Επίσης οι γυναίκες αποτελούν το 53% των χρηστών, υπερτερώντας έναντι των ανδρών χρηστών με το ποσοστό τους να φτάνει το 47% [8]. Τέλος ένα αποθαρρυντικό στατιστικό για την ιστοσελίδα του Twitter είναι πως μόλις το 5% των εγγεγραμμένων χρηστών αντιπροσωπεύουν το 80% της συνολικής δραστηριότητας της ιστοσελίδας, με αποτέλεσμα να αντιλαμβανόμαστε ότι μεγάλο μέρος των χρηστών παραμένουν ανενεργοί.

Το Twitter έχει γίνει διεθνώς αναγνωρίσιμο από το λογότυπο «μπλε πουλί» που διαθέτει ως υπογραφή. Το αρχικό λογότυπο της εταιρίας που δημιουργήθηκε και με τη δημιουργία της ιστοσελίδας ανέφερε με μπλε γράμματα τη λέξη «twitter» (Εικόνα 3). Το λογότυπο εκείνο διατηρήθηκε ως το Σεπτέμβριο του 2010. Το 2010 στο λογότυπο του Twitter προστέθηκε ένα μπλε πουλί που ονομάστηκε «Larry the Bird». Εκτός όμως από την προσθήκη του Larry άλλαξε και ο χρωματισμός των γραμμάτων σε μαύρα (Εικόνα 4). Το τελευταίο λογότυπο το οποίο διατηρείται έως και σήμερα έγινε 5 Ιουνίου του 2012 όπου ο Larry the Bird αντικαταστάθηκε από ένα μεγαλύτερο μπλε πουλί με το όνομα «Twitter Bird» (Εικόνα 5).

Εικόνα 3 Αρχικό λογότυπο του Twitter

Εικόνα 4 Δεύτερο κατά σειρά λογότυπο

Εικόνα 5 Το τωρινό λογότυπο του Twitter

2.2.2. Ιστορία του Twitter

Η ιδέα της δημιουργίας της ιστοσελίδας κοινωνικής δικτύωσης εμφανίστηκε σε ένα συνέδριο των μελών του διοικητικού συμβουλίου της εταιρίας Odeo. Στο συνέδριο αυτό παρευρισκόταν ο τότε προπτυχιακός φοιτητής του Πανεπιστημίου της Νέας Υόρκης, Jack Dorsey. Ο Dorsey παρουσίασε μια ιδέα στα στελέχη της εταιρίας, σύμφωνα με την οποία παρείχε τη δυνατότητα επικοινωνίας ενός ατόμου με μια ομάδα με τη χρήση των υπηρεσιών SMS. Βασιζόμενος στην ήδη υπάρχουσα υπηρεσία TXTMob, ο Dorsey με μια ομάδα προγραμματιστών δημιούργησε το πρώτο κώδικα του έργου του, μήκους 10958 γραμμών, που εξελίχθηκε στη συνέχεια σε 40404. Το έργο του αρχικά είχε το κωδικό όνομα twttr, επηρεασμένο από το Flickr.

Το πρώτο πρωτότυπο του Twitter χρησιμοποιήθηκε ως μια εσωτερική υπηρεσία επικοινωνίας για τους εργαζόμενος της εταιρίας Odeo και η πλήρης εκδοχή του εισήχθη δημοσίως στο διαδίκτυο στις 15 Ιουλίου του 2006. Τον Οκτώβριο του 2006 τα κάποια μέλη της εταιρίας Odeo, δημιουργούν την εταιρία Obvious την εταιρία Obvious Corporation και εξαγοράζουν την εταιρία Odeon για την οποία εργάζονταν. Από την άνοιξη του 2007 το Twitter θα αποσχιστεί και θα αποτελέσει πλέον αυτόνομη εταιρία.

Η καθοριστική κίνηση για την εκρηκτική αύξηση της δημοτικότητας του Twitter θα γίνει το 2007 , όταν στην διάρκεια ενός συνεδρίου στο Texas (South by Southwest Interactive), με στρατηγικές κινήσεις από την πλευρά των δημιουργών, ο αριθμός των Tweets από 20.000 ανά μέρα θα ανέλθει σε 60.000.

Όσον αφορά το όνομα του καινούριου αυτού κοινωνικού δικτύου , είναι κάτι που απασχόλησε αρκετά τους δημιουργούς του, οι οποίοι αρχικά σκέφτονταν να το ονομάσουν Twitch (συσπώμαι), λόγω της δόνησης του κινητού τηλεφώνου που χρησιμοποιούν οι χρήστες του. Αναζητώντας όμως περαιτέρω επιλογές ανακάλυψαν την λέξη Twitter ,ένας ορισμός της οποίας ήταν : μια μικρή έκρηξη επουσιωδών πληροφοριών. Αυτό φάνηκε πιο κατάλληλο στους υπεύθυνους από την αρχική ιδέα και έγινε τελικά η ονομασία του social network.

Όσον αφορά τα οικονομικά της εταιρίας, ο πρώτος γύρος χρηματοδότησης της εταιρίας εικάζεται ότι ανέρχεται σε ποσό ανάμεσα στα 1 έως 5 εκατομμύρια δολάρια. Ο δεύτερος γύρος (2008) σκαρφαλώνει στο ποσό των 22 εκατομμυρίων, ενώ ο τρίτος φτάνει σε αυτό των 35 εκατομμυρίων δολαρίων με πολλές πλέον εταιρίες όπως οι Union Square Ventures, Digital Garage, Spark Capital και Bezos Expeditions να στηρίζουν το Twitter.

Κάποιοι στον επιχειρηματικό κόσμο ισχυρίζονται ότι η μακροβιότητα του Twitter δεν είναι εγγυημένη καθώς η εταιρία δεν έχει στη διάθεσή της τα απαραίτητα εισοδήματα. Για αυτό το λόγο το μέλος του συμβουλίου της εταιρίας Todd Chaffee πρότεινε το Twitter να αποκτήσει τη δυνατότητα να χρησιμοποιείται για ηλεκτρονικό εμπόριο καθώς ούτως ή άλλως ένας μεγάλος αριθμός πωλητών το

χρησιμοποιεί για να προωθήσει τα προϊόντα του και πολλοί χρήστες για να συλλέξουν συστάσεις για κάποιο προϊόν που τους ενδιαφέρει.

2.2.3. Περιεχόμενο των tweets

Η εταιρία ερευνών Pear Analytics, με βλαση το San Antonio αφού πήρε ένα τυχαίο δείγμα από tweets, της τάξης των 2000, για μία περίοδο δύο εβδομάδων (Αύγουστος 2009 από τις 11:00π.μ έως τις 05:00μ.μ), χώρισε τα tweets σε έξι κατηγορίες:

1. News: Ο τύπος των μηνυμάτων με περιεχόμενο ανάλογο με αυτό που θα συναντήσει κανείς σε κανάλια της τηλεόρασης ή και άλλα μέσα ενημέρωσης

2. Spam: Μηνύματα χωρίς ιδιαίτερη χρηστική αξία όπως “Μάθετε πως να κερδίσετε ένα εκατομμύριο δολάρια σε λίγες ώρες”

3. Self-Promotion: Μηνύματα από πάσης φύσεως εμπόρους που διαφημίζουν τα προϊόντα τους

4. Pointless Babble (Άσκοπη Φλυαρία): Μηνύματα που στέλνει ο κάθε χρήστης με ανούσιες πληροφορίες όπως “Αυτή τη στιγμή πίνω καφέ”

5. Conversational (Ομιλιτικά): Μηνύματα που αποτελούν μέρος ενός διαλόγου με μορφή ανάλογη με αυτήν του Instant Messaging.

6. Pass-along value (Διάδοσης Αξιών): Μηνύματα που περνάν απαράλλακτα από χρήστη σε χρήστη. Τα μηνύματα αυτά ονομάζονται retweet.

Η Pear Analytics στη συνέχεια ανέλυσε τη συχνότητα με την οποία εμφανίζεται κάθε κατηγορία μηνυμάτων και κατέληξε στα παρακάτω ποσοστά : 3,60% για την κατηγορία news, 3,75% για τα spam, 5,85% για τα self-promotion, 40,55% για τα Pointless Babble, 37,55% για τα Conversational και 8,70% για τα Pass-Along value. Όπως μπορεί να δει κανείς, βάση της έρευνας που πραγματοποίησε η εν λόγω εταιρία, η επικρατέστερη κατηγορία είναι αυτή των μηνυμάτων που περιέχουν ασήμαντες πληροφορίες, κάτι στο οποίο στρέφονται και οι περισσότεροι που κατακρίνουν το συγκεκριμένο social network.

Στην έρευνα της Pear ανταποκρίθηκε άμεσα η ερευνήτριας κοινωνικής δικτύωσης Danah Boyd, διαφωνώντας με την ετικέτα που δόθηκε σε ένα σύνολο μηνυμάτων ως «Άσκοπη Φλυαρία». Πρότεινε μάλιστα η κατηγορία αυτών των μηνυμάτων να χαρακτηριστεί ως «Κοινωνικός Καλλωπισμός» ή «Περιφερειακή Συνείδηση», εξηγώντας πως υπάρχουν άτομα που θέλουν να ξέρουν τι κάνει, τι πιστεύει και τι νιώθει ο κόσμος γύρω τους, ακόμα και αν η συνύπαρξη δεν είναι εφικτή σε πραγματικό επίπεδο, και αυτός είναι και ο λόγος που οι ίδιοι δημοσιεύουν τέτοιας μορφής tweet [9].

3.Εξόρυξη Δεδομένων σε Κοινωνικά Δίκτυα

3.1. Εισαγωγή

Η εξόρυξη γνώσης από μεγάλες αποθήκες δεδομένων έχει εξελιχθεί σε ένα από τα βασικότερα ερευνητικά ζητήματα στον τομέα της ανάλυσης των κοινωνικών δικτύων και αποτελεί αντικείμενο μελέτης από πολλούς ερευνητές και μηχανικούς, ιδιαίτερα τα τελευταία χρόνια με τη ραγδαία αύξηση του όγκου της πληροφορίας. Η έρευνα στον τομέα αυτόν έχει προχωρήσει θεαματικά και έχουν εξαχθεί πολλά και σημαντικά αποτελέσματα.

Την τελευταία δεκαετία έχει παρατηρηθεί μια αλματώδης αύξηση στην παραγωγή και στη συλλογή δεδομένων. Ωστόσο η πρόοδος στην τεχνολογία των βάσεων δεδομένων μας παρέχει νέες τεχνικές για την αποδοτικότερη και αποτελεσματικότερη συλλογή, αποθήκευση και διαχείριση των δεδομένων. Είναι κοινώς αποδεκτό ότι κάθε χρόνο τα δεδομένα διπλασιάζονται, ενώ η χρήσιμη πληροφορία δείχνει να μειώνεται. Αυτό είναι και το κύριο πρόβλημα που προσπαθεί να λύσει ο τομέας της εξόρυξης γνώσης.

Οι δικτυακές εφαρμογές που διαχειρίζονται μεγάλες αποθήκες δεδομένων έχουν αρχίσει να κάνουν χρήση μεθόδων και τεχνικών της εξόρυξης γνώσης με σκοπό τη βελτίωση της ποιότητας των παρεχόμενων υπηρεσιών μέσω της μελέτης της συμπεριφοράς των πελατών και της εξαγωγής συμπερασμάτων από αυτή.

Κάθε χρόνο παράγονται τεράστιοι όγκοι δεδομένων (της τάξης των petta-bytes και exa-bytes) από το σύνολο των χρηστών που χρησιμοποιεί τα social media, τα οποία αποθηκεύονται σε τεράστιες βάσεις δεδομένων. Η δυνατότητα ανάλυσης και ερμηνείας των δεδομένων καθώς και η εξαγωγή «χρήσιμης» γνώσης από αυτά έχει ξεπεράσει κάθε όριο. Έτσι, φαίνεται επιτακτική η ανάγκη για μια νέα γενιά εργαλείων, μεθόδων και τεχνικών για ευφυή ανάλυση των βάσεων δεδομένων. Αυτή η ανάγκη έχει προσελκύσει την προσοχή πολλών ερευνητών από διάφορες ερευνητικές περιοχές όπως τεχνητή νοημοσύνη, στατιστική, αποθήκες δεδομένων, διαδραστική ανάλυση και επεξεργασία, έμπειρα συστήματα και οπτικοποίηση δεδομένων, με αποτέλεσμα ένας νέος ερευνητικός τομέας να δημιουργείται, γνωστός ως εξόρυξη δεδομένων και γνώσης (Data and Knowledge Mining).

Επομένως εξαιτίας της ευρέως διαδεδομένης χρήσης των social media μέσω του internet, υπάρχει ένα πρωτοφανές ποσό δεδομένων διαθέσιμο ως αντικείμενο μελέτης πολλών πεδίων, όπως η κοινωνιολογία, οι επιχειρήσεις, η ψυχολογία, η διασκέδαση, η πολιτική κοκ. Εφαρμόζοντας τις τεχνικές Εξόρυξης Γνώσης στα social media μπορούμε να ανακαλύψουμε ενδιαφέρουσες πλευρές της ανθρώπινης συμπεριφοράς και της ανθρώπινης αλληλεπίδρασης. Το data mining μπορεί να χρησιμοποιηθεί σε σύνδεση με τα social media για να βελτιωθεί η αντίληψη που έχουν οι άνθρωποι σχετικά με ένα θέμα, για να προσδιορισθούν ομάδες ανθρώπων

ανάμεσα στις μάζες του πληθυσμού, για να μελετηθούν ομάδες που αλλάζουν με το χρόνο, για να βρεθούν άνθρωποι με επιρροή, να γίνει η σύσταση ενός προϊόντος ή μιας δραστηριότητας σε ένα άτομο, ή ακόμα και να κατατάξουμε το σύνολο των πληροφοριών που παρέχεται στα κοινωνικά δίκτυα σε κατηγορίες με βάση το περιεχόμενο τους.

3.2. Εξόρυξη Δεδομένων

Όλα τα παραπάνω ζητήματα οδήγησαν την επιστημονική κοινότητα στη δημιουργία ενός νέου τομέα, που καλείται σήμερα εξόρυξη δεδομένων. Με τον όρο αυτό καλούμε τον ορθό προσδιορισμό ενδιαφερουσών δομών σε δεδομένα.

Ένας τυπικός ορισμός για την Εξόρυξη Γνώσης βρέθηκε στο WordNet του Πανεπιστημίου του Princeton [10] όπου ορίζεται ως: «η επεξεργασία των δεδομένων που χρησιμοποιεί ικανότητες αναζήτησης δεδομένων και στατιστικούς αλγορίθμους για να ανακαλύψει πρότυπα και συσχετισμούς σε μεγάλες προϋπάρχουσες βάσεις δεδομένων. Ένας τρόπος για να ανακαλύψουμε νέα σημασία στα δεδομένα». Πιο απλά, η ιδέα κλειδί πίσω από την Εξόρυξη Δεδομένων είναι η εύρεση καινούργιας πληροφορίας σε ένα σύνολο δεδομένων, η οποία είναι κρυμμένη ή λανθάνουσα.

Πρέπει σε αυτό το σημείο να τονιστεί -επειδή δεν αναφέρεται στον ορισμό - πως για να εξάγουμε πραγματικά χρήσιμη πληροφορία συνήθως απαιτείται να έχουμε όσο το δυνατό πιο πολλά δεδομένα. Αυτό έχει να κάνει περισσότερο με την ακρίβεια και την λεπτομέρεια της πληροφορίας αυτής. Έτσι μπορούμε να πούμε γενικά πως η σπουδαιότητα ενός αλγόριθμου εξόρυξης δεδομένων μπορεί να περιγραφεί από τον τύπο ***ΑΠΟΔΟΣΗ + ΠΟΙΟΤΗΤΑ ΠΛΗΡΟΦΟΡΙΑΣ = ΣΠΟΥΔΑΙΟΤΗΤΑ***.

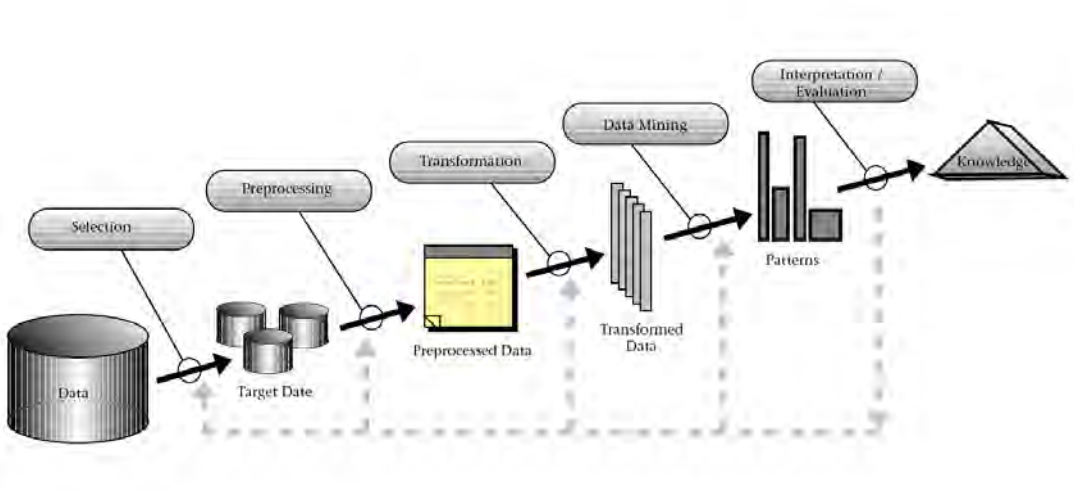
Από το παραπάνω συμπεραίνουμε πως στις περισσότερες εφαρμογές είναι άσκοπο να έχουμε υψηλή απόδοση αποτελεσμάτων με υπολογισμούς που απαιτούν απροσδιόριστα πολύ χρόνο ή να έχουμε πολύ γρήγορα αποτελέσματα

πολύ χαμηλής ποιότητας.

Ο όρος «Εξόρυξη Δεδομένων» είναι σχετικά καινούργιος και εμφανίστηκε στη δεκαετία του '90. Ωστόσο, παρεμφερείς τάσεις και έρευνες είναι ακόμα πιο παλιές. Ο τομέας της εξόρυξης δεδομένων σχετίζεται με πολλούς άλλους τομείς όπως την στατιστική (statistics), την τεχνητή νοημοσύνη (artificial intelligence), τη μηχανική μάθησης (machine learning), τις βάσεις δεδομένων (data bases), τις μηχανές αναζήτησης (search engines), τα συστήματα υποστήριξης αποφάσεων (decision support systems), τα συστήματα άμεσης ανάλυσης δεδομένων (OLAP) καθώς επίσης με τις τεχνικές Ανάκτησης Πληροφορίας, όπως θα δούμε και παρακάτω, ειδικότερα όταν τα δεδομένα που αντλούνται από τα Κοινωνικά Δίκτυα αναφέρονται σε απλό κείμενο.

3.2.1. Διαδικασία Εξόρυξης Δεδομένων

Η διαδικασία της εξόρυξης γνώσης από βάσεις δεδομένων είναι μια διαλογική και επαναληπτική διαδικασία που αποτελείται από μια σειρά από τα ακόλουθα βήματα (Εικόνα 6):



Εικόνα 6 Διαδικασία Εξόρυξης Δεδομένων

- Την επιλογή των δεδομένων. Υπάρχουν διαφορετικά είδη αποθηκών πληροφοριών που μπορούν να χρησιμοποιηθούν στη διαδικασία εξόρυξης γνώσης. Κατά συνέπεια, οι πολλαπλές πηγές δεδομένων μπορούν να συνδυαστούν καθορίζοντας το σύνολο στο οποίο τελικά η διαδικασία εξόρυξης πρόκειται να εφαρμοστεί.
- Τη δημιουργία του στόχου-συνόλου δεδομένων. Επιλογή του συνόλου δεδομένων (μεταβλητές, δείγματα δεδομένων) στο οποίο η διαδικασία εξόρυξης πρόκειται να εκτελεσθεί.
- Τον καθορισμό και την προ-επεξεργασία δεδομένων. Αυτό το βήμα περιλαμβάνει βασικές διαδικασίες όπως η αφαίρεση του θορύβου, η συλλογή των απαραίτητων πληροφοριών για τη διαμόρφωση ή τη μέτρηση του θορύβου, η απόφαση σχετικά με τις στρατηγικές διαχείρισης των ελλειπόντων πεδίων δεδομένων
- Το μετασχηματισμό των δεδομένων. Τα δεδομένα μετασχηματίζονται ή παγιώνονται σε μορφές κατάλληλες για εξόρυξη . Χρήση των μεθόδων μείωσης διαστάσεων ή μετασχηματισμού για τη μείωση του αριθμού των υπό εξέταση μεταβλητών ή την εύρεση κατάλληλης αντιπροσώπευσης των δεδομένων χωρίς μεταβλητές. Σε αυτό το στάδιο εμπλέκονται οι τεχνικές Ανάκτησης Πληροφοριών, που αναλύονται διεξοδικότερα σε επόμενη ενότητα.
- Την επιλογή των στόχων και των αλγορίθμων εξόρυξης δεδομένων. Σε αυτό το βήμα αποφασίζουμε το στόχο της διαδικασίας εξόρυξης γνώσης, επιλέγοντας τους στόχους εξόρυξης δεδομένων που θέλουμε να επιτύχουμε. Επίσης, επιλέγονται οι μέθοδοι που θα χρησιμοποιηθούν. Αυτό περιλαμβάνει την επιλογή του κατάλληλου μοντέλου και παραμέτρων. Επίσης η μέθοδος εξόρυξης δεδομένων πρέπει να αντιστοιχηθεί με τις απαιτήσεις και τα γενικά κριτήρια της διαδικασίας εξόρυξης γνώσης.
- Την εξόρυξη δεδομένων. Εφαρμόζοντας ευφυείς μεθόδους, ψάχνουμε για ενδιαφέροντα πρότυπα γνώσης. Τα πρότυπα θα μπορούσαν να είναι μιας συγκεκριμένης αντιπροσωπευτικής μορφής ή ενός συνόλου τέτοιων

αντιπροσωπευτικών, όπως κανόνες κατηγοριοποίησης, δένδρα, παλινδρόμηση, συσταδοποίηση κλπ. Η απόδοση και τα αποτελέσματα της μεθόδου εξόρυξης δεδομένων εξαρτώνται από τα προηγούμενα βήματα.

- Την αξιολόγηση των προτύπων. Τα εξαγόμενα πρότυπα αξιολογούνται με κάποια μέτρα, προκειμένου να προσδιοριστούν τα πρότυπα τα οποία αντιπροσωπεύουν τη γνώση, δηλαδή τα αληθινά ενδιαφέροντα πρότυπα.

- Τη σταθεροποίηση και παρουσίαση της γνώσης. Σε αυτό το βήμα, η εξορυγμένη γνώση ενσωματώνεται στο σύστημα και κάποιες τεχνικές αντιπροσώπευσης γνώσης χρησιμοποιούνται για να παρουσιάσουν την εξορυγμένη γνώση στο χρήστη. Επίσης, ελέγχουμε για επίλυση τυχών συγκρούσεων με προηγούμενη εξορυγμένη γνώση.

3.3. Ανάκτηση Πληροφοριών

Η Ανάκτηση Πληροφορίας (Informations Retrieval) είναι η επιστημονική περιοχή που μελετά τα προβλήματα που σχετίζονται με την αναπαράσταση, την οργάνωση και την επεξεργασία στοιχείων πληροφορίας, με στόχο την αποτελεσματική και αποδοτική πρόσβαση των χρηστών σε αυτά.

Αν και η γνωστική περιοχή της Ανάκτησης Πληροφορίας ξεκίνησε με τη μελέτη εγγράφων κειμένου (text), στη συνέχεια επεκτάθηκε και στη μελέτη άλλων τύπων δεδομένων, κάτι που επιβλήθηκε από τις ανάγκες των σύγχρονων εφαρμογών. Έτσι, σήμερα μπορούμε να χρησιμοποιούμε μεθόδους ανάκτησης πληροφορίας για την πρόσβαση σε πολυμεσικά δεδομένα (όπως εικόνα, ήχο, βίντεο) καθώς και σε δεδομένα διαθέσιμα μέσω του παγκόσμιου ιστού (world wide web).

Είναι εύκολα αντιληπτό επομένως, πως η δυνατότητα εξόρυξης γνώσης με βάση το περιεχόμενο του Παγκόσμιου των ιστοσελίδων κοινωνικής δικτύωσης εκτός από τα πολυμεσικά δεδομένα, σχετίζεται στενά με το πεδίο της εξόρυξης γνώσης

από κείμενο (text mining), καθώς οι ιστοσελίδες περιέχουν και κείμενο, το οποίο χρησιμοποιείται μεταγενέστερα για να αναπαραστήσουμε έγγραφα και να εφαρμόσουμε διαδικασίες συσταδοποίησης και κατηγοριοποίησης με βάση τις τεχνικές που προσφέρει ο κλάδος της Ανάκτησης Πληροφοριών.

3.3.1. Ορισμός Μοντέλων Ανάκτησης Πληροφορίας

Πριν εξεταστούν τα επί μέρους μοντέλα, στην υπό-ενότητα αυτή θα δοθεί ένας τυπικός και ακριβής ορισμός για το τι είναι ένα μοντέλο Ανάκτησης Πληροφορίας (ΑΠ) κατά τον Baeza-Yates [Baeza-Yates,99].

Ένα μοντέλο ανάκτησης πληροφορίας είναι η τετράδα $[D, Q, F, R(q_i, d_j)]$ όπου το D είναι ένα σύνολο από λογικές αναπαραστάσεις για τα κείμενα της συλλογής, το Q αντιπροσωπεύει ένα σύνολο από λογικές αναπαραστάσεις για τις πληροφοριακές ανάγκες (ερωτήσεις) του χρήστη, το F αποτελεί το υπόβαθρο για την μοντελοποίηση της αναπαράστασης των κειμένων, των ερωτημάτων και των σχέσεων μεταξύ τους και το $R(q_i, d_j)$είναι μια συνάρτηση κατάταξης, η οποία συνδέει έναν πραγματικό αριθμό με ένα ερώτημα $q_i \in Q$ και μια αναπαράσταση κειμένου $d_j \in D$. Μια τέτοια κατάταξη ορίζει μια διάταξη πάνω στα κείμενα πάντα με βάση το ερώτημα q_i.

Ο παραπάνω ορισμός περιγράφει τη διαδικασία καθορισμού ενός μοντέλου ΑΠ. Η διαδικασία ορισμού ενός μοντέλου είναι η ακόλουθη. Αρχικά επινοείται ένας τρόπος αναπαράστασης για τα κείμενα και την πληροφοριακή ανάγκη του χρήστη. Έπειτα καθορίζεται ένα υπόβαθρο στο οποίο θα μπορούν αυτές οι αναπαραστάσεις να μοντελοποιηθούν. Το υπόβαθρο αυτό, με την σειρά του πρέπει να παρέχει και τον μηχανισμό κατάταξης. Για παράδειγμα στο Boolean μοντέλο, το υπόβαθρο αυτό αποτελείται από αναπαραστάσεις κειμένων και ερωτήσεων ως σύνολα, και τις κλασσικές πράξεις τους. Αντίστοιχα στο Χώρο-Διανυσματικό μοντέλο, το υπόβαθρο αποτελείται από τις διανυσματικές αναπαραστάσεις κειμένων σε έναν πολυδιάστατο διανυσματικό χώρο και τις επιτρεπτές αλγεβρικές πράξεις πάνω σε

διανύσματα.

3.3.2. Κλασσικά μοντέλα Ανάκτησης Πληροφορίας

Στην ενότητα αυτή θα παρουσιαστούν εν συντομία το Boolean και το Χώρο-Διανυσματικό μοντέλο ανάκτησης πληροφοριών [Μακρής].

Τα κλασσικά μοντέλα στην ανάκτηση πληροφορίας θεωρούν ότι κάθε κείμενο περιγράφεται από ένα σύνολο από αντιπροσωπευτικές λέξεις κλειδιά, που ονομάζονται όροι δεικτοδότησης. Ένας όρος δεικτοδότησης, είναι μια λέξη το σημασιολογικό περιεχόμενο της οποίας, περικλείει ένα μέρος του θέματος με το οποίο ασχολείται το κείμενο. Έτσι τα κείμενα μπορούν να αναπαρασταθούν ως σύνολα όρων, που συνοψίζουν το περιεχόμενο τους. Γενικά οι όροι δεικτοδότησης είναι συνήθως ουσιαστικά γιατί τα ουσιαστικά αναπαριστούν μια έννοια χωρίς την ανάγκη να εμφανίζονται δίπλα σε άλλο μέρος του λόγου και η σημασιολογία τους είναι εύκολα αντιληπτή. Σύνδεσμοι και επιρρήματα, θεωρούνται ότι έχουν κυρίως συμπληρωματικό χαρακτήρα. Συχνά όμως χρειάζεται να χρησιμοποιηθούν και αυτά τα μέρη του λόγου στο ευρετήριο.

Με δεδομένη την αναπαράσταση των κειμένων ως συλλογές όρων, δεν έχουν όλοι οι όροι την ίδια ισχύ ως προς την περιγραφή ενός κειμένου. Με άλλα λόγια η ερμηνεία ενός όρου συχνά μπορεί να δίνει μια γενικευμένη ή και ασαφή περιγραφή. Τέτοιοι όροι είναι αυτοί που εμφανίζονται με μεγάλη συχνότητα στην πλειονότητα των κειμένων μιας συλλογής. Για να προσομοιωθεί το γεγονός ότι διαφορετικοί όροι μπορούν να έχουν διαφορετική βαρύτητα ως προς στην δεικτοδότηση των κειμένων, σε κάθε όρο δεικτοδότησης ανατίθεται ένα αριθμητικό βάρος.

Συγκεκριμένα έστω k_i, ένας όρος δεικτοδότησης, και d_j ένα κείμενο. Ο αριθμός $w_{ij} \geq 0$ είναι το βάρος που αντιστοιχεί στο ζεύγος (k_i, d_j) και αντιστοιχεί στο πόσο αντιπροσωπευτικός είναι ο όρος k_iγια το κείμενο d_j.

3.3.2.1. Δεικτοδότηση βάρους όρου

Έστω t είναι ο αριθμός των όρων δεικτοδότησης στο σύστημα και k_i, είναι ένας γενικός όρος δεικτοδότησης. Το σύνολο $K = \{k_1, k_2, \ldots, k_t\}$ είναι το σύνολο όλων των όρων δεικτοδότησης. Ένα βάρος $w_{ij} > 0$ συνδέεται με κάθε όρο k_i, που εμφανίζεται στο κείμενο dj. Για κάποιον όρο δεικτοδότησης που δεν εμφανίζεται στο κείμενο, ισχύει $w_{ij} = 0$. Κάθε κείμενο d_j έχει ένα αντιπροσωπευτικό διάνυσμα d_j, το οποίο αναπαρίσταται ως $d_j = (w_{1j}, w_{2j}, \ldots, w_{tj})$. Επιπλέον έστω g_i μια συνάρτηση που επιστρέφει το βάρος που συνδέεται με τον όρο, σε κάθε $t-$ διάστατο διάνυσμα $g_i(d_j) = w_{ij}$.

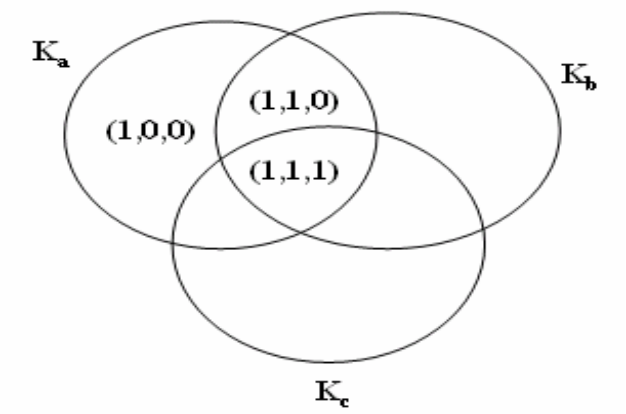
Τα παραπάνω βάρη είναι μεταξύ τους ανεξάρτητα, δηλαδή η τιμή του wij δεν επηρεάζει την τιμή του wi+1j. Αυτή η υπόθεση είναι απλουστευτική δεδομένου ότι συχνά υπάρχουν συμπλέγματα όρων που εμφανίζονται μαζί. Ένα τέτοιο παράδειγμα είναι οι όροι “credit” και “card”. Σε μια συλλογή με θέμα εμπορικών συναλλαγών, αναμένεται αυτοί οι δύο όροι να έχουν σχεδόν ταυτόσημες συχνότητες εμφάνισης. Κατά συνέπεια οι δύο αυτοί όροι είναι συσχετισμένοι μεταξύ τους και ο υπολογισμός της ανάθεσης βαρών θα πρέπει να λαμβάνει υπόψη του αυτή τη συσχέτιση. Λαμβάνοντας υπόψη μας τις συσχετίσεις των όρων μεταξύ τους, η πολυπλοκότητα υπολογισμού των βαρών αυξάνει, συμπαρασύροντας και τον υπολογισμό της κατάταξης. Για το λόγο αυτό, οι διακριτοί όροι δεικτοδότησης θεωρούνται ότι είναι μεταξύ τους ανεξάρτητοι.

3.3.3. Το Boolean μοντέλο

Το Boolean μοντέλο, είναι ένα απλό μοντέλο ανάκτησης πληροφορίας που το υπόβαθρό του και τα ερωτήματα που υποβάλει ο χρήστης βασίζονται στη θεωρία συνόλων και στη άλγεβρα Boole. Συγκεκριμένα στο Boolean μοντέλο, κάθε όρος δεικτοδότησης θεωρείται ότι είτε ανήκει εξ ολοκλήρου σε ένα κείμενο είτε όχι. Κατά συνέπεια τα βάρη θεωρούνται δυαδικά, ως $w_{ij} \in \{0,1\}$. Το κάθε ερώτημα θεωρείται ότι αποτελείται από όρους δεικτοδότησης οι οποίο συνδέονται με έναν από τους τελεστές and, or, not. Δηλαδή κάθε ερώτημα είναι μια Boolean έκφραση που μπορεί

να γραφεί σε Διαζευκτική Κανονική Μορφή (ΔΚΜ). Για παράδειγμα το ερώτημα $[q = k_a \cap (k_b \cup -k_c)]$ μπορεί να γραφεί στην μορφή ΔΚΜ ως $[q_{ΔΚΜ} = (k_a \cap k_b) \cup (k_a \cup -k_c)$. Έστω τώρα ένα διάνυσμα με δυαδικά βάρη που αντιστοιχεί σε ανάθεση αλήθειας σε συζευκτικές εκφράσεις της τριάδας (k_a, k_b, k_c). Για παράδειγμα στην έκφραση $k_a \cap k_b$ μια ανάθεση αλήθειας είναι η $(1,1,0)$.

Άρα το αρχικό ερώτημα μπορεί να αναλυθεί σε διάζευξη τέτοιων διανυσμάτων ως $\bar{q}_{ΔΚΜ} = (1,1,1) \cup (1,1,0) \cup (1,0,0)$. Τα δυαδικά αυτά διανύσματα εισήχθησαν επειδή υπάρχει απευθείας αντιστοιχία του ερωτήματος $\bar{q}_{ΔΚΜ}$, όπως φαίνεται και στο Σχήμα 1.



Σχήμα 1 Συζευτικές συνιστώσες του ερωτήματος $[q = k_a \cap (k_b \cup -k_c)]$

3.3.3.1. Ανάθεση βαρών δεικτοδότησης

Στο Boolean μοντέλο, τα βάρη που ανατίθενται στους όρους δεικτοδότησης είναι δυαδικά δηλαδή, $w_{ij} \in (0,1)$. Ένα ερώτημα q είναι μια συνήθης Boolean έκφραση. Έστω $\bar{q}_{ΔΚΜ}$ η διαζευκτική κανονική μορφή του ερωτήματος και $\bar{q}_{cc}$ καθεμία από τις συζευκτικές συνιστώσες του $\bar{q}_{ΔΚΜ}$. Η ομοιότητα του κειμένου d_j προς το ερώτημα q ορίζεται από την παρακάτω Σχέση.

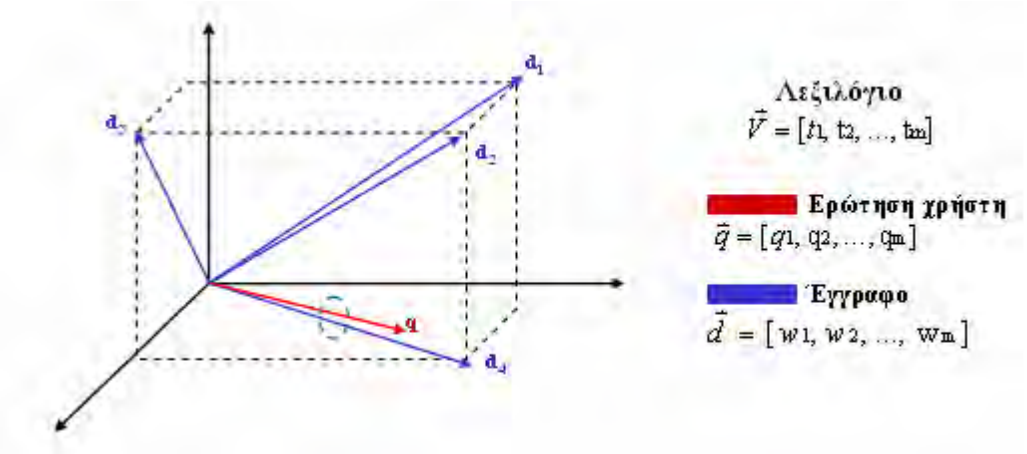
$$sim(d_j, q) = \begin{cases} 1, \text{εάν } \exists\, \bar{q}_{cc} \text{ ώστε } (\bar{q}_{cc} \in \bar{q}_{ΔΚΜ}) \cap (\forall\, k_i, g_i(d_j) = g_i(\bar{q}_{cc})) \\ 0 \qquad\qquad \text{σε άλλη περίπτωση} \end{cases}$$

Αν $\mathrm{sim}(d_j, q) = 1$, τότε το Boolean μοντέλο προβλέπει ότι το κείμενο d_jείναι σχετικό με το ερώτημα q, ενώ σε οποιαδήποτε άλλη περίπτωση είναι άσχετο. Με άλλα λόγια στο μοντέλο αυτό δεν υπάρχει η έννοια της μερικής ικανοποίησης των συνθηκών του ερωτήματος. Για παράδειγμα έστω d_jτέτοιο ώστε να είναι $d_j = (0,1,0)$. Το κείμενο αυτό περιέχει τον όρο k_b, αλλά θεωρείται άσχετο ως προς το ερώτημα $[q = k_a \cap (k_b \cup -k_c)]$. Λόγω αυτής της έλλειψης, το Boolean μοντέλο στην ουσία εκτελεί περισσότερο ανάκτηση δεδομένων παρά πληροφορίας.

Το κύριο πλεονέκτημα του Boolean μοντέλου είναι η απλότητά του. Το κύριο μειονέκτημά του είναι ότι δεν υπάρχει διαβάθμιση σχετικότητας ως προς το ερώτημα κάτι που μπορεί να οδηγήσει σε χαμηλής ποιότητας ανάκτηση πληροφορίας. Ένα δεύτερο μειονέκτημά του είναι ότι συχνά δεν είναι εύκολη η έκφραση της πληροφοριακής ανάγκης του χρήστη με την τυποποίηση που επιβάλλει το μοντέλο αυτό. Λόγω αυτών των χαρακτηριστικών του, το Boolean μοντέλο έχει βρει εφαρμογή σε κυρίως εμπορικά συστήματα βιβλιοθηκών.

3.3.4. Το Χωρο-Διανυσματικό μοντέλο

Το Χώρο-Διανυσματικό μοντέλο (Vector-Space Model), αντιμετωπίζει την ανεπάρκεια της ανάθεσης δυαδικών βαρών και εισάγει ένα υπόβαθρο, στο οποίο επιτρέπεται το προσεγγιστικό ταίριασμα [Salton,68], [Salton,71]. Τα βάρη που ανατίθενται στους όρους δεικτοδότησης, τόσο για τα κείμενα όσο και για τα ερωτήματα είναι μη δυαδικά και χρησιμοποιούνται για τον υπολογισμό του βαθμού ομοιότητας μεταξύ του ερωτήματος και κάθε αποθηκευμένου κειμένου. Κατόπιν, τα κείμενα διατάσσονται με φθίνουσα σειρά, με κριτήριο τον βαθμό ομοιότητάς τους με το ερώτημα του χρήστη. Έτσι στο μοντέλο αυτό λαμβάνονται υπόψη και κείμενα που ικανοποιούν μερικώς τις συνθήκες του ερωτήματος και το τελικό αποτέλεσμα είναι πολύ πιο ακριβές σε σχέση με την ανάκτηση από το Boolean μοντέλο.



Σχήμα 2 Αναπαράσταση Χωρο-Διανυσματικού μοντέλου

3.3.4.1. Ανάθεση Βαρών Δεικτοδότησης

Στο Χώρο-Διανυσματικό μοντέλο το βάρος w_{ij} που αντιστοιχεί στο ζεύγος (k_i, d_i) είναι φυσικός αριθμός και όχι δυαδικός. Επιπλέον ανατίθενται βάρη και στους όρους δεικτοδότησης του ερωτήματος που υποβάλλει ο χρήστης. Έστω λοιπόν ότι (w_i, q)το βάρος που αντιστοιχεί στο ζεύγος (k_i, q), όπου $w_{i,q} \geq 0$. Τότε το διάνυσμα του ερωτήματος ορίζεται $q = (w_{1q}, w_{2q,} \ldots \ldots, w_{tq})$ ως όπου t είναι ο συνολικός αριθμός των όρων δεικτοδότησης στο σύστημα. Όπως και πριν το διάνυσμα του d_j είναι $d_j = (w_{1p}, w_{2j}, \ldots \ldots, w_{tj})$.

Μ' αυτόν τον τρόπο το κείμενο d_j και το ερώτημα χρήστη q αναπαρίστανται ως διανύσματα διαστάσεως t. Στο μοντέλο αυτό προτείνεται ο βαθμός της ομοιότητας μεταξύ του κειμένου d_j και του ερωτήματος q να υπολογιστεί ως ο βαθμός συσχέτισης μεταξύ των δύο διανυσμάτων. Στο παραπάνω Σχήμα 2, παρουσιάζεται μια αναπαράσταση του Χώρο-Διανυσματικού μοντέλου στις τρεις διαστάσεις που αντιλαμβάνεται ο άνθρωπος. Μέτρο του βαθμού συσχέτισης αποτελεί το συνημίτονο της γωνίας που περιέχεται μεταξύ των δύο διανυσμάτων, που παρέχεται από την ακόλουθη Σχέση, όπου $|d_j|$και $|q|$ οι νόρμες των διανυσμάτων.

$$\mathrm{sim}(d_j, q) = \frac{d_j \times q}{|d_j| \times |q|} = \frac{\sum_{i=1}^{t}(w_{i,j} \times w_{i,q})}{\sqrt{\sum_{i=1}^{t} w^2{}_{i,j}} \times \sqrt{\sum_{i=1}^{t} w^2{}_{i,q}}}$$

Εφόσον $w_{i,j} \geq 0$ και $w_{i,q} \geq 0$, η τιμή παίρνει τιμές από 0, όπου δεν υπάρχει καθόλου ταύτιση έως 1 και τα διανύσματα ταυτίζονται πλήρως. Έτσι το μοντέλο αυτό, αντί να προσπαθήσει να προσδιορίσει αν ένα κείμενο είναι ή όχι σχετικό, διατάσσει τα κείμενα με κριτήριο τον βαθμό ομοιότητας τους προς το ερώτημα. Με αυτή τη στρατηγική ένα κείμενο μπορεί να ανακτηθεί ακόμα και αν ταιριάζει κατά προσέγγιση με το ερώτημα. Επειδή δεν είναι επιθυμητή η ανάκτηση όλων των κειμένων που έχουν μη μηδενικό βαθμό σχετικότητας με το ερώτημα, αλλά αυτά που ταιριάζουν περισσότερο, ορίζεται ένα κατώφλι ελέγχου για την τιμή που λαμβάνει το μέγεθος $sim(d_j, q)$. Κείμενα με βαθμό ομοιότητας μεγαλύτερο απ' αυτό το κατώφλι επιστρέφονται στο χρήστη ως σχετικά. Πριν όμως ερμηνευτεί ο μηχανισμός κατάταξης των κειμένων πρέπει να εξεταστεί ο τρόπος υπολογισμού των βαρών.

Το πρόβλημα υπολογισμού των βαρών ανάγεται θεωρητικά στο εξής πρόβλημα ομαδοποίησης. Έστω μια συλλογή κειμένων *C* και ένα σύνολο *A* από κείμενα της συλλογής. Στο πρόβλημα της ΑΠ, το *A* είναι το σύνολο εκείνο των κειμένων που απαντούν σε μια πληροφοριακή ανάγκη. Η διατύπωση της πληροφοριακής ανάγκης που καθορίζει το *A*, μπορεί να είναι σχετικά ασαφής, οπότε τα θέματα που πρέπει να αντιμετωπιστούν είναι δυο ειδών. Πρώτον, πρέπει να καθοριστεί ποια χαρακτηριστικά χαρακτηρίζουν τα κείμενα του *A*, ενώ δεύτερον πρέπει να καθοριστεί ποια χαρακτηριστικά διαχωρίζουν τα κείμενα του συνόλου *A* από τα κείμενα του *C*. Η εξισορρόπηση της επίδρασης αυτών των δύο ομάδων χαρακτηριστικών είναι το αντικείμενο ενός καλού σχήματος ανάθεσης βαρών.

Ένα καλό μέτρο για τον χαρακτηρισμό των στοιχείων εντός του συνόλου *A* είναι η συχνότητα εμφάνισης του όρου k_i, σε κάθε κείμενο d_j. Διαισθητικά όσο πιο συχνά εμφανίζεται ένας όρος k_i, σε ένα κείμενο d_j, τόσο πιο καλή περιγραφή του d_j αποτελεί ο όρος k_i. Η συχνότητα εμφάνισης του όρου, ονομάζεται παράγοντας tf (στην Αγγλική term frequency). Επίσης ένα μέτρο για τον διαχωρισμό των συνόλων *A* και *C* αποτελεί η αντίστροφη συχνότητα εμφάνισης του k_i, στα κείμενα της

συλλογής. Διαισθητικά αν ο όρος k_i, έχει μεγάλη συχνότητα εμφάνισης στη συλλογή, δεν είναι πολύ χρήσιμος για να χαρακτηρίσει ένα κείμενο και άρα να διαχωρίσει μια ομάδα κειμένων μέσα στην συλλογή. Η αντίστροφη συχνότητα εμφάνισης αναφέρεται συνήθως ως παράγοντας idf (στην Αγγλική inverse document frequency). Συνδυάζοντας αυτούς τους δύο παράγοντες προκύπτει το σχήμα υπολογισμού tf - idf, όπως ορίζεται παρακάτω.

3.3.4.2. Ανάθεση βάρους tf - idf

Έστω N ο συνολικός αριθμός των κειμένων και $n_ι$, ο αριθμός των κειμένων στα οποία εμφανίζεται ο όρος k_i. Έστω fr_{ij} η συχνότητα εμφάνισης του όρου k_iστο d_j. Τότε η κανονικοποιημένη συχνότητα f_{ij} του όρου k_i, στο d_j δίνεται από την Σχέση 3, όπου η μέγιστη τιμή max υπολογίζεται πάνω σε κάθε όρο που αναφέρεται στο κείμενο d_j. Αν ο όρος k_iδεν εμφανίζεται στο d_jτότε $f_{ij} = 0$. Επιπλέον, έστω idf_i η αντίστροφη συχνότητα εμφάνισης για τον όρο k_iπου δίνεται από την Σχέση 4 όπου N είναι το σύνολο των εγγράφων της συλλογής και $n_ι$τα έγγραφα που περιέχουν τον όρο k_i. Ο συνδυασμός των δύο αυτών μεγεθών ορίζει το σύστημα ανάθεσης βαρών tf - idf, σύμφωνα με την Σχέση 5. Αντίστοιχα, για τα βάρη των όρων στα ερωτήματα ισχύει η Σχέση 6 [Salton,88].

$$f_{i,j} = \frac{fr_{ij}}{\max_l(fr_{l,j})} \qquad \text{Σχέση 1}$$

$$idf_i = log\frac{N}{n_i} \qquad \text{Σχέση 2}$$

$$w_{i,j} = f_{i,j} \times idf_i = f_{i,j} \times log\frac{N}{n_i} \qquad \text{Σχέση 3}$$

$$w_{i,q} = \left(0.5 + \frac{0.5 \times fr_{i,q}}{max_l(fr_{l,q})}\right) \times log\frac{N}{n_i} \qquad \text{Σχέση 4}$$

Στην παραπάνω Σχέση 1, $fr_{i,q}$ είναι η συχνότητα εμφάνισης του όρου k_i, στο κείμενο που αντιπροσωπεύει την πληροφοριακή ανάγκη q. Ο αθροιστικός παράγοντας 0.5, έχει προκύψει πειραματικά και εξισορροπεί το γεγονός ότι το ερώτημα απαρτίζεται συνήθως από πολύ λίγους όρους.

Το Χώρο-Διανυσματικό μοντέλο πλεονεκτεί διότι το σχήμα υπολογισμού των βαρών που χρησιμοποιεί, βελτιώνει την απόδοση της ανάκτησης. Επιπλέον, η στρατηγική προσεγγιστικού ταιριάσματος επιτρέπει την ανάκτηση κειμένων που προσεγγίζουν τις συνθήκες του ερωτήματος που υποβάλλει ο χρήστης. Ακόμα, ο τρόπος του υπολογισμού της κατάταξης με βάση το συνημίτονο επιτρέπει την ταξινόμηση των κειμένων βάσει του βαθμού ομοιότητας τους με την ερώτηση, ενώ παράλληλα υλοποιείται εύκολα με τις υπάρχουσες δομές δεικτοδότησης. Ένα μειονέκτημα είναι ότι οι όροι δεικτοδότησης θεωρούνται ανεξάρτητοι μεταξύ τους.

Το Χώρο-Διανυσματικό μοντέλο, παρά την απλότητα της σύλληψης και της υλοποίησης του είναι ένα στιβαρό μοντέλο. Η δυνατότητα της εφαρμογής προσεγγιστικού ταιριάσματος, δίνει αποτελέσματα που είναι δύσκολο να βελτιωθούν χωρίς επέκταση του ερωτήματος ή εφαρμογή ανάδρασης του χρήστη. Τα αλγεβρικά μοντέλα που ακολούθησαν το Χώρο-Διανυσματικό μοντέλο αν και έχουν κατά σημεία καλύτερη απόδοση, είναι πιο δύσκολα στην υλοποίηση τους. Πάντως το μοντέλο αυτό δεν αντιμετωπίζει επαρκώς τα προβλήματα της συνωνυμίας και της πολυσημίας. Παρόλα αυτά, λόγω της ευκολίας στην υλοποίηση του, παραμένει το πιο δημοφιλές μοντέλο ΑΠ.

3.4. Αλγόριθμοι Εξόρυξης Δεδομένων

Οι αλγόριθμοι εξόρυξης δεδομένων είναι πολλοί και σε αυτή την ενότητα θα παρουσιάσουμε σε κατηγορίες τους πιο σημαντικούς από αυτούς. Οι κατηγορίες τις οποίες θα αναφέρουμε είναι οι εξής: Κατηγοριοποίηση, Συσταδοποίηση, Κανόνες Συσχέτισης, Πρότυπα Ακολουθιών, Παλινδρόμηση και Δέντρα Απόφασης.

Οι παραπάνω κατηγορίες χωρίς αμφιβολία αναπαριστούν όλη την περιοχή των αλγορίθμων που χρησιμοποιούνται στον τομέα αυτό. Τα τελευταία χρόνια η

ερευνητική κοινότητα δίνει πολύ βάση στη βελτίωση υπαρχόντων τεχνικών και δημιουργία νέων για να αντιμετωπιστούν τα προβλήματα που τίθενται σε αυτές τις κατηγορίες οι οποίες θα αναλυθούν παρακάτω:

- **Κατηγοριοποίηση:** Η κατηγοριοποίηση (classification) αποτελεί μια από τις βασικές εργασίες (tasks) εξόρυξης δεδομένων. Βασίζεται στην εξέταση των χαρακτηριστικών ενός νέου αντικειμένου το οποίο με βάση τα χαρακτηριστικά αυτά αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων. Τα αντικείμενα που πρόκειται να κατηγοριοποιηθούν αναπαριστάνονται γενικά από τις εγγραφές της βάσης δεδομένων και η διαδικασία της κατηγοριοποίησης αποτελείται από την ανάθεση κάθε εγγραφής σε κάποιες από τις προκαθορισμένες κατηγορίες. Η εργασία της κατηγοριοποίησης χαρακτηρίζεται από έναν καλά καθορισμένο ορισμό των κατηγοριών και το σύνολο που χρησιμοποιείται για την εκπαίδευση του μοντέλου αποτελείται από προ-κατηγοριοποιημένα παραδείγματα. Η βασική εργασία είναι να δημιουργηθεί ένα μοντέλο το οποίο θα μπορούσε να εφαρμοστεί για να κατηγοριοποιήσει δεδομένα που δεν έχουν ακόμα κατηγοριοποιηθεί (να ανατεθεί σε κάποια από τις κατηγορίες). Στις περισσότερες περιπτώσεις, υπάρχει ένα περιορισμένος αριθμός κατηγοριών και εμείς θα πρέπει να αναθέσουμε κάθε εγγραφή στην κατάλληλη κατηγορία. Για αυτό το σκοπό χρησιμοποιούνται κάποιες τεχνικές, τις οποίες μπορούμε να κατατάξουμε σε δύο κατηγορίες. Η πρώτη χρησιμοποιεί δέντρα απόφασης (decision trees) και η δεύτερη νευρωνικά δίκτυα (neural networks).

- **Συσταδοποίηση:** Η συσταδοποίηση (clustering) είναι η εργασία του καταμερισμού ενός ετερογενούς πληθυσμού σε ένα σύνολο περισσότερων ετερογενών συστάδων (clusters). Αυτό που διαφοροποιεί τη συσταδοποίηση από την κατηγοριοποίηση είναι ότι η συσταδοποίηση δε βασίζεται σε προκαθορισμένες κατηγορίες. Στην κατηγοριοποίηση, ο πληθυσμός διαιρείται σε κατηγορίες αναθέτοντας κάθε στοιχείο ή εγγραφή σε μια προκαθορισμένη κατηγορία με βάση ένα μοντέλο που αναπτύσσεται μέσω της εκπαίδευσης του με παραδείγματα που έχουν κατηγοριοποιηθεί εκ των προτέρων. Όπως και στην κατηγοριοποίηση έτσι και στη συσταδοποίηση υπάρχουν πολλές εφαρμογές. Για παράδειγμα, ας θεωρήσουμε πως έχουμε διαθέσιμα τα

δεδομένα πελατών μιας εταιρίας πωλήσεων. Χρησιμοποιώντας τεχνικές συσταδοποίησης, μπορούμε να βρούμε τον καταμερισμό των πελατών και της αγοράς, π.χ. μπορούμε να δούμε ποιοι πελάτες αγοράζουν για την οικογένεια τους και ποιοι για τον εαυτό τους ή ποιοι έχουν μεγάλο εισόδημα και ποιοι όχι.

- **Κανόνες Συσχέτισης:** Η εξαγωγή κανόνων συσχέτισης (association rules) θεωρείται μια από τις σημαντικότερες διεργασίες εξόρυξης δεδομένων. Έχει προσελκύσει μεγάλο ενδιαφέρον γιατί παρέχουν έναν συνοπτικό τρόπο για να εκφραστούν οι ενδεχομένως χρήσιμες πληροφορίες που γίνονται εύκολα κατανοητές από τους τελικούς χρήστες. Οι κανόνες συσχέτισης ανακαλύπτουν κρυμμένες «συσχετίσεις» μεταξύ των γνωρισμάτων ενός συνόλου των δεδομένων. Αυτοί οι συσχετισμοί παρουσιάζονται στην ακόλουθη μορφή A→B όπου το A και το B αναφέρονται στα σύνολα γνωρισμάτων που υπάρχουν στα υπό ανάλυση δεδομένα.

- **Πρότυπα Ακολουθιών:** Η εξόρυξη πρότυπων ακολουθιών (sequential patterns) είναι η εξόρυξη των συχνά εμφανιζόμενων προτύπων σχετικών με το χρόνο ή άλλες ακολουθίες. Οι περισσότερες μελέτες στα πρότυπα ακολουθιών επικεντρώνονται στα συμβολικά πρότυπα. Ο χρήστης εδώ μπορεί να προσδιορίσει τους περιορισμούς στα είδη των προτύπων ακολουθιών που εξάγονται με την παροχή των προσχεδίων προτύπων (template patterns) υπό μορφή σειριακών επεισοδίων, παράλληλων επεισοδίων ή κανονικών εκφράσεων. Παραδείγματα προτύπων ακολουθιών έχουμε στην καθημερινή μας ζωή όπως τα κείμενα, οι μουσικές νότες, τα δεδομένα του καιρού και οι ακολουθίες του DNA.

- **Παλινδρόμηση:** Η παλινδρόμηση (regression) είναι θέμα το οποίο έχει μελετηθεί πολύ στην στατιστική και στα νευρωνικά δίκτυα. Κύριος σκοπός εδώ είναι η πρόβλεψη της τιμής μιας μεταβλητής μελετώντας τις τιμές που είχε στο παρελθόν. Συνήθως χρησιμοποιούμε ένα μοντέλο για την μεταβλητή. Η παλινδρόμηση καλύπτει ένα μεγάλο τμήμα του τομέα της εξόρυξης δεδομένων που έχει να κάνει με προβλέψεις.

- **Δέντρα Απόφασης:** Τα δέντρα απόφασης (decision trees) έχουν μελετηθεί

αρκετά σαν ένα ζήτημα μηχανικής μάθησης. Για να γίνει κατανοητό, ας υποθέσουμε ότι έχουμε ένα σύνολο εγγραφών και καθεμία από αυτές έχει μια λίστα χαρακτηριστικών. Ένα δέντρο απόφασης στο σύνολο των εγγραφών είναι ένα δέντρο όπου σε κάθε κόμβο του (που δεν είναι φύλλο) υπάρχει ένα ερώτημα που αναφέρεται στα χαρακτηριστικά των εγγραφών και κάθε ερώτημα καταλήγει σε ένα συγκεκριμένο παιδί ενός κόμβου. Τα φύλλα του δηλώνουν τις κλάσεις. Έτσι ένα δέντρο απόφασης εκτελεί κατηγοριοποίηση χρησιμοποιώντας ερωτήματα σχετικά με τα χαρακτηριστικά των εγγραφών. Οι εφαρμογές που χρησιμοποιούν δέντρα απόφασης είναι παρόμοιες με αυτές που κάνουν κατηγοριοποίηση.

3.5. Σχετικές έρευνες για την Εξόρυξη Δεδομένων στα Κοινωνικά Δίκτυα

Στις μέρες, με τη συνεχή ανάπτυξη των διαδικτυακών τεχνολογιών καθώς και με την ενασχόληση ολοένα και περισσότερων χρηστών στο τομέα του Internet, τα κοινωνικά δίκτυα αποτελούν πλέον τη νέα «τάση» στο διαδίκτυο. Με περισσότερους από 40 εκατομμύρια χρήστες να χρησιμοποιούν την microblogging πλατφόρμα των 140-χαρακτήρων, το Twitter έχει εξελιχθεί σε ένα δίκτυο με τη δικιά του δυναμική.

Λαμβάνοντας, επομένως, υπόψη την επικρατούσα κατάσταση στον τομέα αυτό, ολοένα και περισσότερες προσπάθειες και έρευνες πραγματοποιούνται για τη λεγόμενη «Ανάλυση» του Twitter, με απώτερο σκοπό την εξαγωγή συμπερασμάτων τόσο για το περιεχόμενο που συναντάται στο σύνολο του, όσο και για το γενικότερο τρόπο χρήσης του δικτύου, και των υπηρεσιών που προσφέρει στους χρήστες.

Οι Dolan Antenucci, Gregory Handy, Askhay Modi και Miller Tinkerhess στην έρευνα που πραγματοποίησαν το 2011 με τίτλο «CLASSIFICATION OF TWEETS VIA CLUSTERING OF HASHTAGS», παρουσιάζουν ένα αλγόριθμο για να αντιληφθούν τις σχέσεις που υπάρχουν μεταξύ του περιεχομένου ενός tweet και του συνόλου των hashtags που περιλαμβάνουν αυτό το tweet. Αναλυτικότερα κάθε tweet αναπαραστάθηκε ως μια λίστα συχνότητας των λέξεων (non-stopword και non-hashtag) που εμπεριέχονταν στο tweet αυτό. Αναλυτικότερα απέδειξαν ότι η

κατηγοριοποίηση των hashtags καθώς και η μείωση της διάστασης των δεδομένων, μπορεί να οδηγήσει σε ορθή ταξινόμηση των tweet σε κατηγορίες, χωρίς να επηρεαστεί η ακρίβεια των αλγορίθμων ταξινόμησης. Επίσης έδειξαν ότι η πολύ-διάσταση ταξινόμηση επιφέρει καλύτερα αποτελέσματα απ' ότι οι προσεγγίσεις Naïve, ειδικότερα αν το σύνολο των δεδομένων είναι συμπιεσμένα. Τέλος χρησιμοποίησαν την κατηγοριοποίηση των hashtag και από τις κλάσεις που δημιουργήθηκαν πρότειναν στο χρήστη, που δημοσίευε ένα tweet, μια ετικέτα, που αντιστοιχούσε σε μία κλάση.

Στο έργο του, σχετικά με τη συσταδοποίηση των hashtag, με τίτλο «EXPLORING TWITTER HASHTAGS», ο Poschko υποστηρίζει ότι δύο hashtag είναι παρόμοια αν συνυπάρχουν σε ένα tweet ταυτόχρονα. Στην έρευνα αυτή, που πραγματοποιήθηκε εξίσου το 2011, ο Poschko δημιουργεί ένα συγκεντρωτικό πίνακα, όπου αναπαριστά τις κλάσεις που προκύπτουν από την κατηγοριοποίηση των hashtags. Ως μέτρο ομοιότητας μεταξύ των κλάσεων χρησιμοποιεί την έννοια της «συχνότητας συνύπαρξης» (co-occurrence frequency). Η «συχνότητα συνύπαρξης» για την έρευνα του, αποτελεί μια μετρική με βάση της οποίας μετρά το πόσες φορές εμφανίζεται το ίδιο σύνολο hashtag σε παρόμοια tweet.

Ένας άλλος κλάδος στην ανάλυση των κοινωνικών δικτύων ονομάζεται «sentiment analysis». Με τον όρο «sentiment analysis» αναφερόμαστε στην εφαρμογή μεθόδων επεξεργασίας φυσικής γλώσσας, υπολογιστικής γλωσσολογίας και ανάλυση κειμένου, τόσο σε γραμματικό όσο και σε ερμηνευτικό επίπεδο, για την εξαγωγή συμπεράσματος της στάσης ή της διάθεσης του ατόμου που κάνει κάποια δημοσίευση σε ένα κοινωνικό δίκτυο. Οι Davidov, Tsur και Rappoport στην έρευνα που πραγματοποίησαν και δημοσίευσαν με τίτλο «ENHANCED SENTIMENT LEARNING USING TWITTER HASHTAGS AND SMILEYS» προσπάθησαν να χαρακτηρίσουν ένα σύνολο από tweets, που είχαν συλλέξει ως δεδομένα, όσο αναφορά την άποψης της διάθεσης του ατόμου που έκανε τη δημοσίευση.

Η εργασία με τίτλο «MINING TWITTER FEEDS FOR TOP STORIES» που πραγματοποιήθηκε και δημοσιεύτηκε από τον Bhattarai έχει ως στόχο την κατηγοριοποίηση των tweet σε κατηγορίες βάση του θέματος στο οποίο αναφέρονται στα περιεχόμενα τους. Ο Bhattarai γνωρίζοντας τη δυσκολία που προκύπτει στην

προσπάθεια ανάλυσης του περιεχομένου των tweets, εξαιτίας της έλλειψης χαρακτηριστικών γνωρισμάτων, της φτωχής χρήσης της γλώσσας στην οποία δημοσιεύονται καθώς και στην ύπαρξη αρκετών «άσκοπων» στοιχείων στο περιεχόμενο τους, χρησιμοποίησε μια εξειδικευμένη τεχνική συσταδοποίησης (μη-επιβλεπόμενης μηχανικής μάθησης), που ονομάζεται Latent Dirichlet Allocation (LDA). Σκοπός της έρευνας αυτής ήταν η κατηγοριοποίηση των tweets σε «topic» κλάσεις (με βάση τη θεματική ενότητα δηλαδή), από όπου θα αντλούνταν ως εκπρόσωποι του κάθε θέματος ένα σύνολο από tweets. Το τελικό στάδιο της εργασίας του Bhattarai συγκρίνει την αποτελεσματικότητα μεταξύ της μεθόδου που χρησιμοποίησε ο ίδιος, για τον διαχωρισμό των θεματικών ενοτήτων, και άλλων μεθόδων που κάνουν χρήση των ετικετών που βρίσκονται στα hashtags του εκάστοτε tweet.

4. Συσταδοποίηση Δεδομένων

4.1. Βασικές Έννοιες της Συσταδοποίησης

4.1.1. Ορισμός Συσταδοποίησης

Η Συσταδοποίηση ή αλλιώς η Ανάλυση Συστάδων είναι η διαδικασία ομαδοποίησης ενός συνόλου αντικειμένων κατά τέτοιο τρόπο ώστε τα αντικείμενα στην ίδια ομάδα (cluster) να είναι περισσότερο όμοια μεταξύ τους και όσο το δυνατόν πιο διαφορετικά με τα αντικείμενα που ανήκουν σε διαφορετικές ομάδες (clusters) [11].

Η συσταδοποίηση δεν αποτελεί ένα μοναδικό ειδικό αλγόριθμο που κατηγοριοποιεί τα δεδομένα, αλλά το γενικό πλαίσιο το οποίο ακολουθείτε και πραγματοποιείται για την κατηγοριοποίηση αυτή. Η διαδικασία αυτή επιτυγχάνεται από ένα σύνολο διαφορετικών αλγορίθμων που διαφέρουν σημαντικά μεταξύ τους ως προς τον τρόπο που ορίζουν ο καθένας την έννοια της κλάσης, καθώς και στην

συνολική διαδικασία που οδηγούνται προκειμένου να τις εντοπίσουν.

Ο μαθηματικός ορισμός της συσταδοποίησης παρατίθεται παρακάτω[11]:

> ***ΟΡΙΣΜΟΣ:*** Δοθέντος ενός συνόλου δεδομένων με τη μορφή διανυσμάτων $D = \{t_1, t_2, \dots \dots, t_n\}$ και μιας ακέραιης τιμής k (αναμενόμενος αριθμός συστάδων), το πρόβλημα της συσταδοποίησης είναι να οριστεί μια αντιστοιχιση $f: D \to \{1, \dots \dots, k\}$, όπου κάθε t_i ανατίθεται σε μια συστάδα K_j, $1 \leq j \leq k$. Μια συστάδα K_j περιέχει ακριβώς εκείνα τα διανύσματα που τις ανατέθηκαν, δηλαδή $K_j = \{t_i | f(t_i) = K_j\ , 1 \leq i \leq n\ και\ t_i \in D\}$

Η έννοια του όρου cluster περιλαμβάνει ομάδες με μικρές αποστάσεις μεταξύ των στοιχείων που τις αποτελούν, πυκνές περιοχές που εντοπίζονται στο σύνολο των δεδομένων του χρήστη που θα συσταδοποιηθεί, καθώς και ειδικές στατιστικές κατανομές. Οι έννοιες αυτές βέβαια ποικίλουν αναλόγως του προβλήματος συσταδοποίησης, των τύπων του συνόλου δεδομένων καθώς και τον αλγόριθμο που θα χρησιμοποιηθεί.

Η διαδικασία της ανάλυσης των συστάδων, βάση των παραπάνω, αντιλαμβανόμαστε ότι μπορεί να θεωρηθεί ένα πολυδιάστατο πρόβλημα βελτιστοποίησης μίας κατηγοριοποίησης. Η κατάλληλη επιλογή του αλγορίθμου συσταδοποίησης καθώς και η σωστή επεξεργασία και αναπαράσταση των δεδομένων μπορεί να οδηγήσει σε διαφορετικά αποτελέσματα κάθε φορά με βάση τα κριτήρια που θέτουμε στο πρόβλημα.

Να επισημάνουμε στο σημείο αυτό πως η συσταδοποίηση δεν είναι μια αυτόματη διαδικασία κατηγοριοποίησης που εφαρμόζεται σε ένα σύνολο δεδομένων για την παραγωγή ενός συνόλου ομάδων. Αντίθετα αποτελεί μια επαναληπτική διαδικασία για την ανακάλυψη γνώσης, που χρησιμοποιεί την έννοια των συνεχόμενων δοκιμών και μπορεί να οδηγήσει σε ένα επιτυχές ή και εντελώς άστοχο αποτέλεσμα.

Εκτός από τον όρο συσταδοποίηση, υπάρχουν μια σειρά από όρους με παρόμοια σημασία, όπως αυτόματη ταξινόμηση, αριθμητική ταξινόμηση,

botryology (από την ελληνική λέξη βότρυς που σημαίνει σταφύλι), καθώς και τυπολογική ανάλυση. Οι λεπτές διαφορές στους όρους αυτούς απαντώνται και σε διαφορετικούς τύπους προβλημάτων, όπου γίνεται και διαφορετική χρήση αλγορίθμων και επιλογή χαρακτηριστικών στα δεδομένα.

Η βασική, όμως, διαφορά που εντοπίζεται μεταξύ της ταξινόμησης και της συσταδοποίησης είναι ότι η πρώτη τεχνική αποτελεί μια εποπτευόμενη από το χρήστη κατηγοριοποίηση των νέων, χωρίς ετικέτα ομάδας, αντικειμένων, στα οποία αποδίδεται μια ετικέτα κατηγορίας χρησιμοποιώντας ένα μοντέλο που αναπτύχθηκε από αντικείμενα με γνωστές ετικέτες κατηγορίας. Αντίθετα η συσταδοποίηση λαμβάνει τις ετικέτες αυτές από το σύνολο των δεδομένων, χωρίς την ανάγκη προγενέστερης δημιουργίας των ετικετών αυτών.

4.1.2. Εφαρμογές Συσταδοποίησης

Οι κατηγορίες ή οι εννοιολογικά σημαντικές ομάδες αντικειμένων που αποτελούνται από κοινά χαρακτηριστικά, παίζουν ένα σημαντικό ρόλο στον τρόπο με τον οποίο οι άνθρωποι αναλύουν, αντιλαμβάνονται και περιγράφουν τον κόσμο. Η διαδικασία της συσταδοποίησης επομένως αποτελεί αναπόσπαστο κομμάτι σε διάφορους τομείς της ζωής των ανθρώπων και των επιστημών, λαμβάνοντας χώρα άλλοτε θεμιτά και άλλοτε αθέμιτα ή ασυνείδητα στην καθημερινότητα του ατόμου. Παρακάτω παρουσιάζονται ορισμένα παραδείγματα της εφαρμογής τεχνικών συσταδοποίησης.

- ***Ανάκτηση Πληροφοριών.*** Το σύνολο των περιεχομένων που συναντάται στον Παγκόσμιο Ιστό αποτελείται από δισεκατομμύρια ιστοσελίδων. Η προσπάθεια αναζήτησης επομένως πληροφοριών ή άντλησης δεδομένων πολλές φορές κρίνεται αδύνατη εξαιτίας του τεράστιου αυτού όγκου πληροφοριών. Η χρήση τεχνικών συσταδοποίησης στον τομέα αυτό, θα μπορούσε να ομαδοποιήσει το σύνολο των ιστοσελίδων αυτών σε μικρότερες ομάδες συστάδων, κάθε μια από της οποίες θα

περιλαμβάνει μια συγκεκριμένη πτυχή του συνόλου των διαφορετικών θεμάτων που συναντώνται. Επίσης η δυνατότητα διαχωρισμού και των ομάδων σε ακόμα μικρότερες ομάδες ιστοσελίδων, θα δημιουργούσε μια ιεραρχική δομή στα δεδομένα, κάνοντας την οποιαδήποτε προσπάθεια αναζήτησης ευκολότερη.

- ***Ιατρική – Φαρμακευτική.*** Βασιζόμενοι στο γεγονός ότι κάθε ασθένεια συχνά εμφανίζει ένα πλήθος παραλλαγών της γενικότερης μορφής της, μπορούμε να δημιουργήσουμε ένα σύνολο ομάδων (συστάδων) που θα χωρίζει τις παραλλαγές αυτές σε υποκατηγορίες. Με τον τρόπο αυτό επιτυγχάνεται η ευκολότερη διάγνωση, όσο αναφορά την ιατρική επιστήμη, καθώς και η μετέπειτα θεραπεία της νόσου από φαρμακευτικής απόψεως.

- ***Κλιματολογικές και Γεωλογικές Συνθήκες***. Η κατανόηση τόσο του κλίματος στη Γή όσο και της γεωλογικής συμπεριφοράς που συναντάται, απαιτεί την εύρεση υποδειγμάτων στην ατμόσφαιρα, στους ωκεανούς, στη θερμοκρασία και γενικότερα σε ένα μεγάλο σύνολο παραγόντων. Η εφαρμογή της ανάλυσης των συστάδων μπορεί να επιφέρει αποτελέσματα τόσο για την πρόγνωση γεγονότων όσο και για την πρόληψη μεγάλων ανατρέψιμων καταστροφών.

- ***Επιχειρήσεις.*** Το σύνολο των δεδομένων που συναντάται στις βάσεις δεδομένων μιας εταιρίας, αναλόγως φυσικά και με το μέγεθος της, απαιτεί τμηματοποίηση. Η τμηματοποίηση αυτή αποφέρει αύξηση της εταιρικής απόδοσης, αποκτώντας τη δυνατότητα επέκτασης των αγοραστικών της δραστηριοτήτων.

4.1.3. Διαδικασία Συσταδοποίησης

Η διαδικασία της συσταδοποίησης ακολουθεί τα τέσσερα βασικά βήματα που παρατίθενται παρακάτω [11]:

- ***Επιλογή Χαρακτηριστικών Γνωρισμάτων.*** Σκοπός του βήματος αυτού είναι η κατάλληλη επιλογή, από τα δεδομένα, του συνόλου των χαρακτηριστικών τους, που πιστεύουμε ότι θα αποφέρει το καλύτερο αποτέλεσμα μετά την πραγματοποίηση της συσταδοποίησης. Η διαδικασία της προεπεξεργασία των δεδομένων κρίνεται απαραίτητη σε αυτό το βήμα, ώστε τα δεδομένα να αναπαριστούνται με τη μορφή διανυσμάτων.

- ***Επιλογή Αλγορίθμου Συσταδοποίησης.*** Σε αυτό το βήμα γίνεται η επιλογή του κατάλληλου αλγορίθμου συσταδοποίησης. Η επιλογή αυτή εξαρτάται από τα δεδομένα που πρόκειται να συσταδοποιηθούν και τις ανάγκες τις εκάστοτε εφαρμογής. Το μέτρο γειτνίασης και το κριτήριο συσταδοποίησης είναι αυτά που κυρίως χαρακτηρίζουν έναν αλγόριθμο συσταδοποίησης.

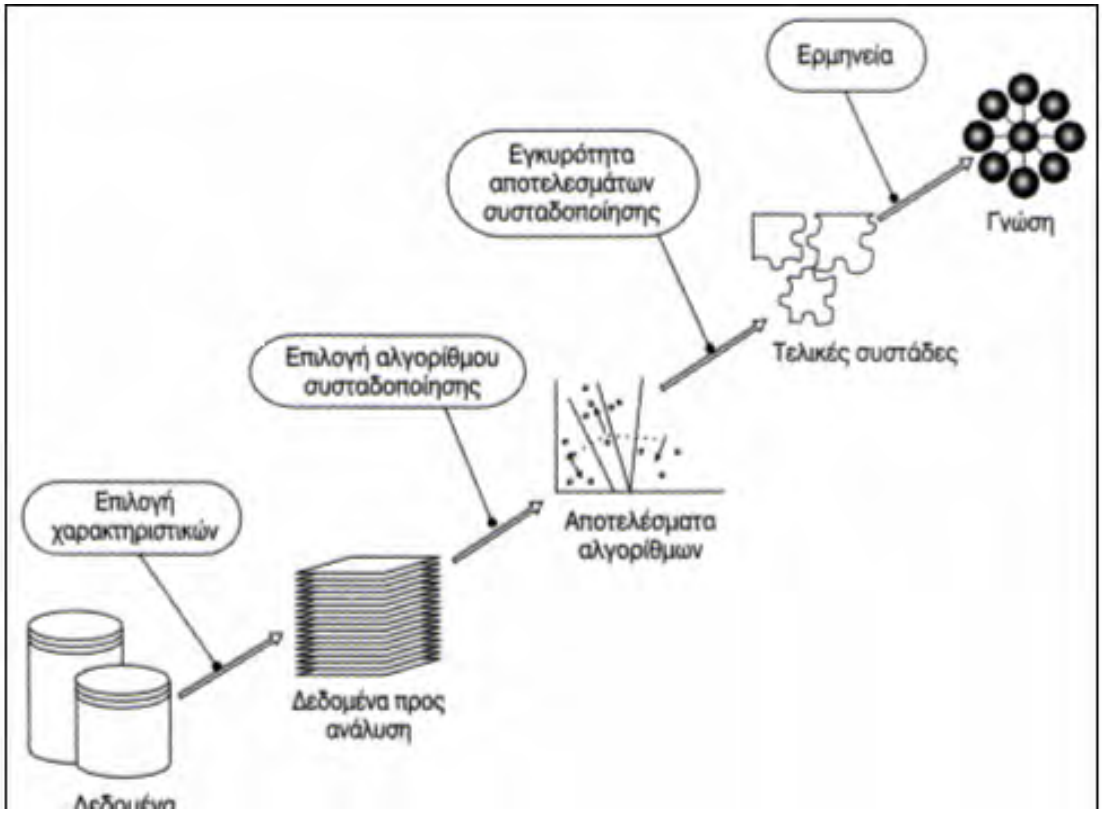
    1. Με το μέτρο γειτνίασης, υπολογίζεται η ομοιότητα μεταξύ των στοιχείων.
    2. Το Κριτήριο Συσταδοποίησης, εκφράζεται συνήθως μέσω μιας συνάρτησης κόστους ή κάποιου άλλου τύπου κανόνων.

- ***Επικύρωση Αποτελεσμάτων.*** Η επικύρωση των αποτελεσμάτων αποτελεί ίσως το κρισιμότερο βήμα σε όλη τη διαδικασία της συσταδοποίησης, καθώς το σύνολο των συστάδων που θα παραχθούν με τη χρήση οποιουδήποτε αλγορίθμου συσταδοποίησης δεν είναι γνωστό. Η ακρίβεια των αποτελεσμάτων μπορεί να επικυρωθεί με κριτήρια και τεχνικές που αναπτύσσονται σε επόμενη υποενότητα της εργασίας.

- ***Ερμηνεία Αποτελεσμάτων***. Τα αποτελέσματα της συσταδοποίησης θα πρέπει να συνδυαστούν με άλλα πειραματικά στοιχεία και αποτελέσματα προηγούμενων αναλύσεων ίδιων στοιχείων προκειμένου να εξαχθούν τα σωστά αποτελέσματα.

Με το πέρας της ερμηνείας των αποτελεσμάτων και της κατάληξης του συμπεράσματος ακολουθεί η απεικόνιση των αποτελεσμάτων συνήθως με τη μορφή γραφημάτων.

Τα βήματα που περιγράφονται παραπάνω απεικονίζονται παραστατικά και με λογική σειρά, όπως αναφέρθηκαν στην Εικόνα 7.



Εικόνα 7 Διαδικασία Συσταδοποίησης

4.2. Είδη Συσταδοποίησης και Αξιολόγηση

Όπως έχει αναφερθεί και παραπάνω η συσταδοποίηση αποτελεί μια πολυδιάστατη διαδικασία για την τμηματοποίηση δεδομένων σε κλάσεις - ομάδες. Πολλοί αλγόριθμοι έχουν προταθεί και δημιουργηθεί κατά καιρούς για την επίλυση διαφόρων προβλημάτων συσταδοποίησης. Κάθε ένας από αυτούς τους αλγορίθμους διαφέρει τόσο στο σκοπό χρήσης του αναφορικά με το πρόβλημα, όσο

στη διαδικασία που ακολουθεί για να ολοκληρώσει τη συσταδοποίηση. Η προσπάθεια, επομένως, κατηγοριοποίησης των αλγορίθμων αυτών δεν αποτελεί μια ξεκάθαρη πάντα διαδικασία, εξαιτίας κυρίως της πολυδιάστατης φύσης του προβλήματος. Παρόλα αυτά στην ενότητα αυτή θα γίνει μια προσπάθεια διαχωρισμού των αλγορίθμων σε ξεχωριστές κατηγορίες.

Ο πρώτος διαχωρισμός της συσταδοποίησης, με βάση το σύνολο των συστάδων αν είναι φωλιασμένο ή μη, αναφέρεται από το Murty and Flynn το 1999 [12], όπου κατηγοριοποιούν τους αλγορίθμους συσταδοποίησης σε **Ιεραρχικούς** (Hierarchical) και **Μη-Ιεραρχικούς** (Non- Hierarchical) ή αλλιώς **Διαχωριστικούς** (Partitional).

Στους ιεραρχικούς αλγορίθμους συσταδοποίησης [13] δημιουργείται ένα εμφωλιασμένο σύνολο από συστάδες. Κάθε επίπεδο της ιεραρχίας έχει ένα ξεχωριστό σύνολο συστάδων. Στο κατώτατο επίπεδο, κάθε αντικείμενο βρίσκεται στη δική του συστάδα. Στο ανώτατο επίπεδο, όλα τα αντικείμενα ανήκουν στην ίδια και μοναδική συστάδα. Η ιεραρχική συστοδοποίηση επομένως επιτρέπει στην κάθε κλάση τη δυνατότητα ύπαρξης υποκλάσεων αυτής.

Αντίθετα στη διαχωριστική συσταδοποίηση ο αλγόριθμος δημιουργεί μόνο ένα σύνολο συστάδων. Οι διαχωριστικοί αλγόριθμοι, δηλαδή, διαιρούν το σύνολο των δεδομένων σε μη επικαλυπτόμενα υποσύνολα (κλάσεις) τέτοια ώστε κάθε αντικείμενο να ανήκει ακριβώς σε ένα υποσύνολο. Ο συνολικός αριθμός των επιθυμητών παραγόμενων συστάδων αποτελεί είσοδο για ένα διαχωριστικό αλγόριθμο σε αντίθεση με οποιονδήποτε ιεραρχικό.

Οι ιεραρχικοί αλγόριθμοι διαχωρίζονται περαιτέρω με βάση την υπόθεση αν κάθε αντικείμενο του συνόλου δεδομένων αποτελεί μια ξεχωριστή συστάδα και φτάνοντας στο αποτέλεσμα ανήκουν όλα τα αντικείμενα σε μία κλάση, ή το αντίστροφο. Στην πρώτη περίπτωση οι αλγόριθμοι αυτοί καλούνται **Ιεραρχικοί Συσσωρευτικοί (Agglomerative ή bottom-up),** ενώ στη δεύτερη περίπτωση οι αλγόριθμοι ονομάζονται **Ιεραρχικοί Διαιρετικοί (Divisive ή top-down).**

Αναλυτικότερα στους **ιεραρχικούς συσσωρευτικούς αλγόριθμους συσταδοποίησης** [14] **(Hierarchical Agglomerative Clustering - HAC)** αρχικά υπολογίζεται η ομοιότητα μεταξύ των συστάδων, εφόσον πριν έχουν μετρηθεί οι αποστάσεις των δεδομένων και έχουν κατηγοριοποιηθεί σε κλάσεις. Σε κάθε βήμα του αλγορίθμου γίνεται συγχώνευση των δύο πλησιέστερων συστάδων, με βάση την μεγαλύτερη ομοιότητα. Το αποτέλεσμα λοιπόν είναι η δημιουργία μια δενδροειδής δομής, που καλείται «dendrogram», με μια συστάδα στην κορυφή να αποτελείται από όλα τα δεδομένα, και τόσες συστάδες στη βάση του δέντρου όσο είναι το σύνολο το δεδομένων συσταδοποίησης. Τα ενδιάμεσα επίπεδα αποτελούν τα βήματα συγχώνευσης των πλησιέστερων κλάσεων.

Με βάση λοιπόν την παραπάνω περιγραφή ο αλγόριθμος που παρουσιάζει τη διαδικασία συσταδοποίησης ενός συσσωρευτικού αλγόριθμου [15] παρατίθεται παρακάτω:

1. Θέσε τα N δεδομένα σε N διαφορετικές κλάσεις και υπολόγισε τις ομοιότητες μεταξύ όλως των ζευγαριών των κλάσεων (αρχικά των N δεδομένων).
2. Αναζήτησε μεταξύ των αποτελεσμάτων το ζευγάρι με τη μεγαλύτερη ομοιότητα και συγχώνευσε τις αντίστοιχες συστάδες σε μια νέα
3. Υπολόγισε εκ νέου τις ομοιότητες της νέας συστάδας με όλες τις υπόλοιπες.
4. Επανέλαβε τα βήματα 2,3 μέχρι να απομείνει 1 κλάση.

Η εύρεση της ομοιότητα μεταξύ δύο συστάδων πραγματοποιείται με διάφορες μεθόδους, με διαφορετικό προσανατολισμό η καθεμία. Από τις μεθόδους αυτές θα αναφερθούμε στην απλή σύνδεση, στην πλήρη σύνδεση και στη μέση απόσταση, οι ποίες είναι και στηριζόμενες σε γράφο τεχνικές (graph-based).

- **Απλή σύνδεση ή Απόσταση απλού συνδέσμου (single link):** αποτελεί την μικρότερη απόσταση μεταξύ ενός στοιχείου μιας συστάδας και ενός

στοιχείου μιας άλλης. Έτσι $dis(K_i, K_j) = \min\left(dis(t_{il}, t_{jm})\right) \forall t_{il} \in K_i \nexists K_j \; και \; t_{jm} \in K_j \nexists K_i$.

- **Πλήρης σύνδεση ή Απόσταση πλήρους συνδέσμου (complete link):** αποτελεί τη μεγαλύτερη απόσταση μεταξύ ενός στοιχείου μιας συστάδας και ενός στοιχείου μιας άλλης. Έτσι $dis(K_i, K_j) = max\left(dis(t_{il}, t_{jm})\right) \forall t_{il} \in K_i \nexists K_j \; και \; t_{jm} \in K_j \nexists K_i$.

- **Μέση Απόσταση (average link):** είναι η μέση απόσταση μεταξύ ενός στοιχείου της μιας συστάδας και ενός της άλλης. Στην απόσταση αυτή, δηλαδή, χρησιμοποιούνται όλα τα μέλη των συστάδων για την εύρεση ομοιότητας, υπολογίζοντας το μέσο όρο. Έτσι $dis(K_i, K_j) = mean\left(dis(t_{il}, t_{jm})\right) \forall t_{il} \in K_i \nexists K_j \; και \; t_{jm} \in K_j \nexists K_i$.

Το βασικό μειονέκτημα των ιεραρχικών μεθόδων είναι ότι έχουν μεγάλη χρονική πολυπλοκότητα της τάξης O(n3), όπου με n συμβολίζεται ο συνολικός αριθμός των αντικειμένων. Η πολυπλοκότητα αυτή μειώνεται σε O(n2) με τη χρήση τεχνικών βελτιστοποίηση των αλγορίθμων [16]. Άλλο ένα μειονέκτημα αυτών των μεθόδων είναι ότι δεν μπορούν να αναιρέσουν καμία προηγούμενη ενέργεια τους.

Οι **διαιρετικοί ιεραρχικοί αλγόριθμοι (Hierarchical Divisive Algorithms – HAD)** συσταδοποίησης, όπως αναφέρθηκε, ακολουθούν αντίστροφη διαδικασία από τους HAC. Η διαφορά έγκειται στο γεγονός ότι ο αλγόριθμος ξεκινά συγκεντρώνοντας το σύνολο των δεδομένων σε μια κλάση και έπειτα γίνεται ο διαχωρισμός. Επομένως στο Βήμα 1 ο αλγόριθμος ξεκινά με όλα τα N δεδομένα σε μια συστάδα και στο Βήμα 2 γίνεται ο χωρισμός σε δύο μέρη. Ο διαχωρισμός όμως των δεδομένων είναι ιδιαίτερα περίπλοκος όσο προχωράει ο αλγόριθμος. Αυτό οφείλεται στο γεγονός των $2^{(N-1)} - 1$ διαφορετικών διαχωρισμών που μπορεί να πραγματοποιηθούν για ένα σύνολο N δεδομένων. Η επιλογή του «καλύτερου» διαχωρισμού κρίνεται ιδιαίτερα δύσκολη διαδικασία, οπότε η χρήση των αλγορίθμων αυτή δεν είναι ιδιαίτερα πρακτική [15].

Η διαδικασία, τέλος, της συσταδοποίησης στους διαχωριστικούς αλγόριθμους περιλαμβάνει ένα σύνολο των N δεδομένων τα οποία εκχωρούνται σε ένα προκαθορισμένο αριθμό K κλάσεων. Σκοπός αυτού του είδους της συσταδοποίησης είναι η εύρεση της βέλτιστης λύσης στην παραγωγή όσο το δυνατόν πιο συνεκτικών και ορθών (ως προς τα δεδομένα) κλάσεων. Αυτό επιτυγχάνεται μέσω της δοκιμής ενός μεγάλου συνόλου δυνατών περιπτώσεων και παραγόντων, που αφορά τόσο το συνολικό αριθμό κλάσεων που θέλουμε να επιτύχουμε, όσο και την χρήση κάθε φορά του σωστού μέτρου ομοιότητας των αντικειμένων.

Ο μέχρι τώρα διαχωρισμός των αλγορίθμων βασίζονταν στο αν το σύνολο των συστάδων αποτελούνταν από εμφωλιασμένες ή όχι. Ένας ακόμα διαχωρισμός που πραγματοποιείται με το κριτήριο της «επικάλυψης» χωρίζει τη συσταδοποίηση σε **Επικαλυπτόμενη** και **Αποκλειστική**. Η εκχώρηση ενός αντικειμένου αποκλειστικά σε μια συστάδα χαρακτηρίζει τη διαδικασία της συσταδοποίησης ως αποκλειστική. Αντίθετα η ταυτόχρονη ύπαρξη ενός αντικειμένου σε δύο ή περισσότερες κλάσεις χαρακτηρίζει τον τύπο της συσταδοποίησης ως επικαλυπτόμενο.

Ο τρόπος εκχώρησης των δεδομένων σε συστάδες οδηγεί σε ένα τύπο συσταδοποίησης που ονομάζεται **Ασαφής** (fuzzy) [17]. Στην ασαφή συσταδοποίηση κάθε αντικείμενο – δεδομένο ανήκει σε κάθε κατηγορία, με μια στάθμιση ιδιότητας μέλους (βάρος) ανάμεσα στο 0 και στο 1. Με απλά λόγια οι συστάδες στη συσταδοποίηση αυτή αντιμετωπίζονται ως ασαφή σύνολα, των οποίων τα αντικείμενα δεν είναι προκαθορισμένα. Το βάρος αυτό που έχει κάθε αντικείμενο για κάθε κλάση, εκφράζει την πιθανότητα του κάθε αντικειμένου να ανήκει σε μία κλάση.

Μια άλλη κατηγορία αλγορίθμων συσταδοποίησης αποτελούν οι **Βασισμένοι σε Γράφους**. Σύμφωνα με τη συσταδοποίηση αυτή τα δεδομένα είναι κόμβοι ενός γράφου, ενώ οι ακμές που τα συνδέουν αντιπροσωπεύουν τη σχέση μεταξύ των δεδομένων και φέρουν κάποια τιμή – βάρος (weight), που καθορίζεται με βάση την ομοιότητα των κόμβων. Το πρόβλημα επίλυσης των αλγορίθμων αυτών

ανάγεται στο διαχωρισμό του γράφου σε υπογράφους επιδιώκοντας το βέλτιστο δυνατό αποτέλεσμα. Ο διαχωρισμός αυτός πραγματοποιείται συνήθως με την αφαίρεση ακμών ώστε το άθροισμα των ακμών που κόβονται να ελαχιστοποιείται, μένοντας έτσι στο γράφο ακμές με μεγάλη βαρύτητα [17].

Ο τελευταίος τύπος συσταδοποίησης ονομάζεται **συσταδοποίηση βάση πυκνότητας.** Μια συστάδα είναι μια πυκνή περιοχή αντικειμένων, η οποία περιβάλλεται από άλλες περιοχές χαμηλότερης πυκνότητας. Αυτός ο τύπος συσταδοποίησης χρησιμοποιείται σε δεδομένα τα οποία εμπεριέχουν πολύ θόρυβο ή ακραίες τιμές με αποτέλεσμα να παράγουν ακανόνιστες ή συμπλεκόμενες μεταξύ τους συστάδες κατά τη διαδικασία της συσταδοποίησης. Η κατηγοριοποίηση των δεδομένων αυτών θα κρίνονταν αδύνατη αν επιλέγαμε αλγορίθμους που βασίζονται στην έννοια της γειτνίασης, καθώς η ύπαρξη του θορύβου τείνει να δημιουργήσει γέφυρες μεταξύ των διαφορετικών κλάσεων, με αποτέλεσμα οι αποστάσεις να μην αποτελούν αξιόπιστη τεχνική υπολογισμού και κατηγοριοποίησης των δεδομένων.

Η εκτίμηση της ποιότητας των συστάδων, που παράγονται από τους αλγόριθμους συσταδοποίησης, είναι πρωταρχικής σημασίας, προκειμένου να ελεγθεί η απόδοση των αλγορίθμων αυτών. Υπάρχουν 2 γενικές προσεγγίσεις για τη μελέτη της ποιότητας των αλγορίθμων.

Η πρώτη προσέγγιση επιτρέπει τη σύγκριση διαφορετικών συνόλων συστάδων, χωρίς να χρησιμοποιείται οποιαδήποτε εξωτερική γνώση. Η τεχνική αυτή ονομάζεται **Internal Quality Measure**.

Η Συνεκτικότητα (Cohesion) των συστάδων αποτελεί χαρακτηριστικό παράδειγμα μετρικής για την ομοιότητα των δεδομένων των κλάσεων, στην περίπτωση έλλειψης οποιασδήποτε εξωτερικής πληροφορίας. Η συνηθέστερη μέθοδος υπολογισμού της συνεκτικότητας είναι να χρησιμοποιήσουμε την σταθμισμένη ομοιότητα της εσωτερικής ομοιότητας των συστάδων. Ο υπολογισμός γίνεται βάση του παρακάτω τύπου $cohesion\ (C_i) = \sum_{\forall x \in C_i}^{n} proximity(x, center\ of\ class)$. Για τον ορθότερο

υπολογισμό της απόδοσης ενός αλγορίθμου με βάση τη συνεκτικότητα θα πρέπει να συμπεριλαμβάνει ταυτόχρονα και την έννοια του διαχωρισμού των κλάσεων. Ο διαχωρισμός αποτελεί μετρική ανομοιότητας διαφορετικών κλάσεων. Επομένως ο συνδυασμός υψηλού βαθμού συνεκτικότητας των στοιχείων της ίδιας κλάσης και υψηλό ποσοστό διαχωρισμού της κλάσης με οποιαδήποτε άλλη κλάση ενός προβλήματος συσταδοποίησης οδηγεί στο συμπέρασμα πως πιθανότητα ο αλγόριθμος που χρησιμοποιήθηκε ήταν αποδοτικός και ορθός.

Η δεύτερη προσέγγιση της ποιοτικής αξιολόγησης των αλγορίθμων, μας επιτρέπει να εκτιμήσουμε την απόδοση της συσταδοποίησης, συγκρίνοντας τις ομάδες που παράγονται από μια τεχνική συσταδοποίησης με γνωστές κατηγορίες. Οι εξωτερικές πληροφορίες που παρέχονται για τα δεδομένα κατηγοριοποίησης με αυτήν την τεχνική ελέγχου, είναι συνήθως με τη μορφή ετικετών κατηγοριών. Οι μέθοδοι αυτή που εξετάζουν την εγκυρότητα των συστάδων ονομάζονται **Με-Επίβλεψη**. Οι μετρικές που θα εξετάσουμε είναι η εντροπία, η ακρίβεια, η ανάκληση και η F-μέτρο. Στις μετρικές αυτές η συνήθης διαδικασία που ακολουθείτε είναι ο υπολογισμός του βαθμού αντιστοιχίας ανάμεσα στις ετικέτες των κατηγοριών, που δίδονται ως δεδομένο, και στις πραγματικές ετικέτες των συστάδων που προκύπτουν από τους αλγορίθμους.

Η **Εντροπία (entropy)** αποτελεί το βαθμό τον οποίο κάθε συστάδα αποτελείται από αντικείμενα μίας μοναδικής κατηγορίας. Για κάθε συστάδα υπολογίζεται αρχικά η κατανομή κατηγοριών των δεδομένων, δηλαδή για τη συστάδα j υπολογίζεται το p_{ij}. Το p_{ij} είναι η πιθανότητα ένα μέλος της συστάδας i να ανήκει στην κατηγορία j με τη σχέση $p_{ij} = \frac{m_{ij}}{m_i}$, όπου m_i είναι το πλήθος των αντικειμένων στην κατηγορίας i και m_{ij} είναι το πλήθος των αντικειμένων της κατηγορίας j στην συστάδα i. Κάνοντας χρήση του παραπάνω υπολογισμού πιθανοτήτων, η εντροπία κάθε συστάδα i, υπολογίζεται χρησιμοποιώντας τη σχέση $e_i = \sum_{j=1}^{L} p_{ij} log_2 p_{ij}$, όπου L είναι το πλήθος των κατηγοριών που δίδονται ως δεδομένο. Η συνολική εντροπία του συστήματος συσταδοποίησης υπολογίζεται ως το άθροισμα των επιμέρους εντροπιών.

Η **Ανάκληση** αποτελεί την έκταση στην οποία μια συστάδα περιέχει το σύνολο των αντικειμένων μια συγκεκριμένης κατηγορίας. Η ανάκληση της συστάδας i, σε σχέση με την κατηγορία j δίδεται από τον τύπο $recall(i,j) = \frac{m_{ij}}{m_j}$.

Η αναλογία μιας συστάδας, η οποία αποτελείται από αντικείμενα μιας συγκεκριμένης κατηγορίας, ονομάζεται **Ακρίβεια.** Η ακρίβεια μια συστάδας i, σε σχέση με την κατηγορία j δίνεται από τον τύπο $precision\ (i,j) = p_{ij}$.

Η **F-μέτρο** μετρική αποτελεί ένα συνδυασμό μεταξύ της ακρίβειας και της ανάκλασης, που μετρά την έκταση στην οποία μια συστάδα περιέχει *μόνο* αντικείμενα μίας συγκεκριμένης κατηγορίας και *όλα* τα αντικείμενα της κατηγορίας. Η F-μέτρο υπολογίζεται από τη σχέση $F(i,j) = (2 \times precision(i,j) \times recall(i,j)/ precision(i,j) + recall(i,j))$.

Υπάρχει ένα μεγάλο σύνολο διαφορετικών μετρικών ποιότητας μιας διαδικασίας συσταδοποίησης. Το αποτέλεσμα της απόδοσης που δίδεται από κάθε μετρική μπορεί να διαφέρει για τον ίδιο αλγόριθμο, αναλόγως με το ποια χρησιμοποιείται. Ωστόσο αν κάποιος αλγόριθμος αποδίδει θετικά στην πλειονότητα των μετρικών, μπορούμε να ισχυριστούμε ότι πράγματι αποτελεί ένα καλό αλγόριθμο για τη λύση του συγκεκριμένου προβλήματος που εξετάζουμε. Στην επόμενη υποενότητα θα παρουσιαστούν 2 αλγόριθμοι που χαρακτηρίστηκαν ως η καταλληλότερη για το πρόβλημα της συσταδοποίησης της παρούσας εργασίας.

4.3. Αλγόριθμος Προσδοκίας-Μεγιστοποίησης (Expectation-Maximization)

Ο αλγόριθμος προσδοκίας – μεγιστοποίησης (Expectation – Maximazation ή EM) αποτελεί έναν αλγόριθμο συνδυαστικών μοντέλων. Τα συνδυαστικά μοντέλο θεωρούν τα δεδομένα ως ένα σύνολο παρατηρήσεων, από ένα συνδυασμό διαφορετικών κατανομών πιθανοτήτων. Η χρήση του αλγορίθμου EM εντοπίζεται κυρίως σε προβλήματα όπου παρατηρούνται κρυφά ή ελλειπή δεδομένα

συσταδοποίησης.

Ο αλγόριθμος ΕΜ επαναλαμβάνει με διαδοχικό τρόπο δύο βασικά βήματα, της πρόβλεψης (E steps) και της μεγιστοποίησης (M steps). Σε κάθε Ε-βήμα υπολογίζεται η προσδοκώμενη τιμή του λογαρίθμου της πιθανοφάνειας ολόκληρου του σετ των δεδομένων, λαμβάνοντας υπόψη τα παρατηρηθέντα στοιχεία και τις κατ' εκτίμηση παραμέτρους από την προηγούμενη επανάληψη. Στο βήμα-Μ η δεσμευμένη προσδοκώμενη τιμή του λογαρίθμου της πιθανοφάνειας ολόκληρου του σετ δεδομένων μεγιστοποιείται. Η διαδικασία αυτή των βημάτων ακολουθείτε επαναληπτικά έως ότου η παρατητηθείσα πιθανοφάνεια να μην αλλάζει από βήμα σε βήμα ή σταματά σε ένα συγκεκριμένο κατώφλι ορισμένο από το χρήστη [18].

Παρακάτω παρατίθεται σε απλά βήματα ο αλγόριθμος ΕΜ:

1. Επίλεξε ένα αρχικό σύνολο παραμέτρων του μοντέλου (με τυχαίο ή μη τρόπο)

2. **Βήμα Προσδοκίας (E-step):** Για κάθε αντικείμενο, υπολόγισε την πιθανότητα να ανήκει σε κάθε κατανομή, δηλαδή υπολόγισε την ποσότητα $prob(κατανομή\ j|x_i, \Theta)$.

3. **Βήμα Μεγιστοποίησης (M-step):** Δοθέντων των πιθανοτήτων από το βήμα προσδοκίας, βρες τις νέες εκτιμήσεις των παραμέτρων που μεγιστοποιούν την αναμενόμενη πιθανοφάνεια.

4. Επανέλαβε Βήμα 2,3 μέχρι να μην αλλάζουν οι παράμετροι (εναλλακτικά, σταμάτα όταν φτάσεις στο κατώφλι).

Προτού συνεχίσουμε στην ανάλυση του αλγορίθμου ΕΜ, κρίνεται απαραίτητο η διευκρίνιση για το τι είναι η πιθανότητα και τι συνάρτηση πιθανοφάνειας.

Αρχικά υποθέτουμε ότι υπάρχουν *K* κατανομές και *m* αντικείμενα τα οποία θα εισαχθούν στον αλγόριθμο ΕΜ, με $X = \{x_1, \dots \dots, x_m\}$. Έστω επίσης ότι η *j*-οστή κατανομή έχει παραμέτρους θ_j και ότι Θ είναι το σύνολο όλως των παραμέτρων,

δηλαδή $\Theta = \{\theta_1, \dots\dots, \theta_K\}$. Τότε, $prob(x_i|\theta_i)$ είναι η **πιθανότητα** του *i*-οστού αντικειμένου να προέρχεται από την *j*-οστή κατανομή. Η πιθανότητα ότι έχει επιλεχθεί η j-οστή κατανομή για να παράγει ένα αντικείμενο δίνεται από το βάρος $w_j, 1 \leq j \leq K$. Τότε η πιθανότητα ενός αντικειμένου x δίνεται από τη σχέση $prob(x|\Theta) = \sum_{j=1}^{K} w_j p_j(x|\theta_j)$.

Έστω επίσης $X = (X_1, \dots\dots, X_n)$ το τυχαίο δείγμα από το σύνολο των δεδομένων. Η συνάρτηση πιθανότητας του X είναι $\prod_{i=1}^{n} f(x_i, \theta)$. Η συγκεκριμένη συνάρτηση όταν οι τιμές x_i είναι γνωστές μπορεί να θεωρηθεί ως συνάρτηση της παραμέτρου ϑ. Έτσι μπορούμε να θέσουμε $L(\theta) = \prod_{i=1}^{n} f(x_i, \theta)\ \forall\ \theta \in \Theta$. Η L(θ) ονομάζεται συνάρτηση πιθανοφάνειας καθότι εκφράζει πόσο πιθανοφανείς, η διαφορετικά πόσο σύμφωνες με το συγκεκριμένο δείγμα του συνόλου δεδομένων είναι οι διάφορες τιμές τις παραμέτρου θ.

Συνοψίζοντας μπορούμε να πούμε ότι ο αλγόριθμος ΕΜ χειρίζεται τα προβλήματα αυτού του είδους με τον εξής τρόπο :

1. Αντικαθιστά τις ελλείπουσες τιμές με τις κατ' εκτίμηση τιμές
2. Εκτιμά τις παραμέτρους
3. Επανεκτιμά τις προσδοκώμενες τιμές των πιθανοτήτων υποθέτοντας ότι οι νέες εκτιμήσεις των παραμέτρων είναι σωστές
4. Επανεκτιμά τις παραμέτρους κ.ο.κ μέχρι να επιτευχθεί σύγκλιση

Συμπερασματικά τα πλεονεκτήματα της χρήσης του αλγορίθμου ΕΜ είναι τα εξής:

- Σε κάθε επανάληψη αυξάνεται η πιθανότητα με αποτέλεσμα να παρουσιάζει αξιόπιστη ολική σύγκλιση
- Είναι γρήγορος και εύκολα εφαρμόσιμος, με ελάχιστες υπολογιστικές απαιτήσεις σε χώρο εικονικής μνήμης.
- Το κόστος επανάληψης του είναι ιδιαίτερα μικρό, καθώς ολοκληρώνεται

σε μικρό χρονικό διάστημα.

- Εκτιμά ελλειπή δεδομένα με αποτέλεσμα να παρακάμπτει και λάθη που πραγματοποιήθηκαν κατά τη συλλογή των δεδομένων, στο βήμα της Ανάκτησης Πληροφορίας.

4.4. Αλγόριθμος DBScan

Η συσταδοποίηση βάση πυκνότητας, όπως αναφέρεται και σε προηγούμενη υποενότητα του κεφαλαίου, εντοπίζει περιοχές με υψηλή πυκνότητα οι οποίες διαχωρίζονται μεταξύ τους με τη βοήθεια περιοχών χαμηλής πυκνότητας. Η πλήρης ονομασία του αλγορίθμου είναι Destiny Based Spatial Clustering of Applications with Noise. Ο αλγόριθμος DBScan αποτελεί μια απλή και αποτελεσματική τεχνική συσταδοποίησης ενός συνόλου χωρικών δεδομένων βάση της πυκνότητας που έχουν μεταξύ τους. Τα δεδομένα, εφόσον είναι χωρικά μπορούν να αναπαρασταθούν στο χώρο ως σημεία.

Ως πυκνότητα ορίζεται το ελάχιστο πλήθος σημείων που απέχουν συγκεκριμένη απόσταση μεταξύ τους. Παρόλα αυτά θα πρέπει να διευκρινιστεί ο όρος «πυκνότητα», καθώς ο αλγόριθμος αυτός χρησιμοποιεί μια νέα έννοια για την πυκνότητα. Αναλυτικότερα ο αλγόριθμος DBScan βασίζεται στη λεγόμενη «προσέγγιση βάση κέντρου» για τον υπολογισμό της πυκνότητας μίας περιοχής δεδομένων. [19] Στην προσέγγιση βάση κέντρου, η πυκνότητα εκτιμάται για ένα συγκεκριμένο σημείο στο σύνολο δεδομένων, μετρώντας το πλήθος των σημείων εντός μιας συγκεκριμένης ακτίνας *Eps,* αυτού του σημείου. Αυτή συμπεριλαμβάνει και το ίδιο το σημείο.

Για τον αλγόριθμο DBScan επομένως ως πυκνότητα για τα συγκεκριμένα σημεία, που αναφέρουμε και παραπάνω και θα εξηγήσουμε εκτενέστερα έπειτα, θεωρεί το σύνολο των υπόλοιπων σημείων που εμπεριέχονται εντός αυτής της ακτίνας. Αντιλαμβανόμαστε πως αν η ακτίνα *Eps* είναι μεγάλη, τότε όλα τα σημεία θα έχουν πυκνότητα ίση με το σύνολο όλων των δεδομένων προς συσταδοποίηση.

Ομοίως αν η ακτίνα είναι πολύ μικρή, τότε όλα τα σημεία θα έχουν πυκνότητα ίση με 1, αφού θα περιλαμβάνουν μόνο τον εαυτό τους.

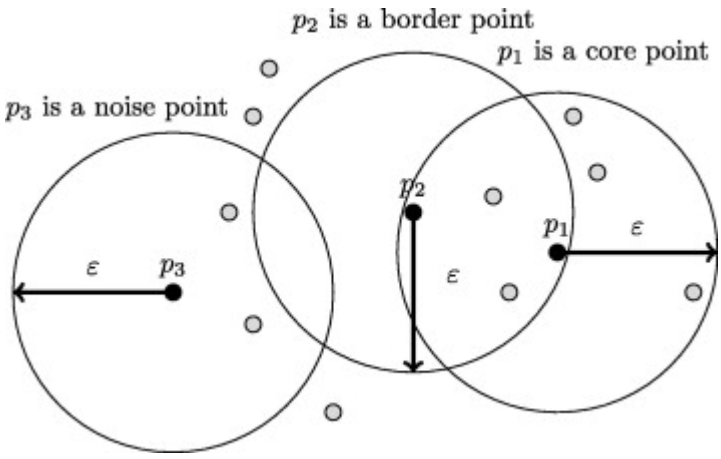
Η προσέγγιση της πυκνότητα βάσει κέντρου επιτρέπει τον διαχωρισμό των σημείων ως:

1. Εσωτερικό σημείο μιας πυκνής περιοχής (**σημείο πυρήνα**). Τα σημεία αυτά βρίσκονται στο εσωτερικό μιας συστάδας που δημιουργείται βάσης της πυκνότητας. Ένα σημείο αποτελεί σημείο πυρήνα αν το πλήθος των σημείων εντός μιας συγκεκριμένης γειτονιάς γύρω από το σημείο, η οποία υπολογίζεται από τις αποστάσεις του σημείου αυτού από τα υπόλοιπα, ξεπερνά ένα συγκεκριμένο κατώφλι (*MinPts*) που ορίζει εξαρχής ο χρήστης.

2. Σημείο στην άκρη μιας πυκνής περιοχής (**οριακό σημείο**). Σημείο ορίου αποτελεί ένα σημείο που βρίσκεται εντός της γειτονιάς ενός σημείου πυρήνα αλλά δεν αποτελεί το ίδιο σημείο πυρήνα.

3. Σημείο μιας οριακής περιοχής (**σημείου θορύβου**). Τα σημεία θορύβου είναι τα σημεία τα οποία δεν είναι ούτε ορίου, ούτε πυρήνα.

Ο παραπάνω διαχωρισμός των σημείων που αναφέρουμε γίνεται αντιληπτός αν κοιτάξουμε την Εικόνα 8. Το σημείο p1 της εικόνας αποτελεί ένα σημείο πυρήνα, αφού περιέχει στη γειτονιά του 6 άλλα σημεία. Το σημείο p2 αποτελεί ένα σημείο ορίου καθώς βρίσκεται ακριβώς στην ακτίνα Eps, που στο συγκεκριμένο παράδειγμα αναπαριστάται με ε. Τέλος το σημείο p3 αποτελεί ένα σημείο θορύβου.



Εικόνα 8 Χωρικά σημεία του αλγορίθμου DBScan

Αρχικά ο αλγόριθμος DBScan ξεκινά υπολογίζοντας τις ακτίνες Eps για κάθε χωρικό σημείο, δημιουργώντας με τον τρόπο αυτό γειτονιές για το κάθε σημείο. Στο εσωτερικό της γειτονίας του κάθε σημείου περικλείονται άλλα σημεία. Μετρά λοιπόν το πλήθος των σημείων που συναντάται σε κάθε γειτονιά. Αν το πλήθος αυτό ξεπερνά το κατώφλι MinPts που δίνει ο χρήστης, τότε το σημείο με τη γειτονιά αυτή επιλέγεται ως σημείο πυρήνα. Έπειτα, γνωρίζοντας πλέον τα σημεία τα οποία αποτελούν τους πυρήνες, καταχωρεί τα σημεία που βρίσκονται εντός των γειτονιών αυτών ως οριακά σημεία και τα υπόλοιπα ως θορύβου.

Απλοποιώντας τον αλγόριθμο που παρουσιάζει η Wikipedia [20] για την λειτουργία του DBScan, καταλήγουμε στον παρακάτω:

1. Χαρακτήρισε κάθε σημείο ως πυρήνα, συνοριακό ή θόρυβο.
2. Αγνόησε όλα τα σημεία θορύβου.
3. Δημιούργησε ένα γράφο με μια κορυφή για κάθε σημείο.
4. Τοποθέτησε μια ακμή μεταξύ όλων των κεντρικών σημείων που είναι σε απόσταση έως Eps μεταξύ τους.
5. Θέσε κάθε ομάδα συνδεδεμένων κεντρικών σημείων ως μια διαφορετική συστάδα.

6. Ανάθεσε κάθε συνοριακό σημείο σε μία από τις συστάδες που περιέχει πιο κοντινό του κεντρικό σημείο.

Η βασική πολυπλοκότητα χρόνου του αλγορίθμου DBScan είναι O(m × χρόνος εύρεσης των σημείων στη γειτονιά – Eps), όπου m είναι το πλήθος των σημείων. Η χείριστη περίπτωση αυξάνει την πολυπλοκότητα σε O(m2). Παρόλα αυτά σε προβλήματα μικρών διαστάσεων υπάρχουν δομές δεδομένων που ονομάζονται, kd-δέντρα, οι οποίες επιτρέπουν την αποτελεσματική ανάκτηση όλων των σημείων εντός μια συγκεκριμένης και δεδομένης απόστασης από κάποιο άλλο σημείο. Η χρήση τέτοιων δομών ελαττώνει την πολυπλοκότητα του αλγορίθμου σε O(m logm). Τέλος οι απαιτήσεις σε χώρο του αλγορίθμου είναι O(m), ακόμα και για πολύ μεγάλα σύνολα δεδομένων, επειδή τα χαρακτηριστικά που αποθηκεύονται για κάθε σημείο είναι πολύ λίγα.

Επομένως τα οφέλη της χρήσης του αλγορίθμου DBScan, βάση του συνόλου των χαρακτηριστικών του, που περιγράψαμε παραπάνω, είναι τα εξής:

1. Δυνατότητα συσταδοποίησης και δεδομένων που εμπεριέχουν πολύ «θόρυβο» εξαιτίας της ανοχής σε αυτόν, με βάση την προσέγγιση της πυκνότητας.

2. Κατηγοριοποίηση οποιοδήποτε τύπου αυθαίρετων σχημάτων και μεγεθών, όσο αναφορά το σύνολο των δεδομένων.

3. Ικανότητα γρήγορης κατηγοριοποίησης, αναλογιζόμενοι και τη διαδικασία που επιτελεί για την επίλυση ενός προβλήματος συσταδοποίησης.

5. Τεχνική Περιγραφή Υλοποίησης

Η παρούσα διπλωματική εργασία αποτελεί εφαρμογή μιας διαδικασίας Εξόρυξης Δεδομένων στο Κοινωνικό Δίκτυο που ονομάζεται Twitter. Όπως έχει αναφερθεί και σε προηγούμενο κεφάλαιο μια διαδικασία εξόρυξης δεδομένων, για την υλοποίηση της, ακολουθεί ένα προκαθορισμένο σύνολο βημάτων. Αρχικά πρέπει να αντληθεί το σύνολο των πληροφοριών που θα χρησιμοποιηθεί αργότερα για την ανάλυση και την εξόρυξη γνώσης. Έπειτα απαιτείται η ορθή προεπεξεργασία των δεδομένων αυτών, ώστε να μετασχηματιστούν στη μορφή που απαιτείται για το επόμενο βήμα. Η επιλογή και η χρήση του κατάλληλου αλγορίθμου εξόρυξης στα μετασχηματισμένα πλέον δεδομένα, αποτελούν το βήμα αυτό. Τέλος με την εξαγωγή συμπερασμάτων και την αναπαράσταση αυτών ολοκληρώνουμε τη διαδικασία της εξόρυξης δεδομένων. Στις παρακάτω υποενότητες του κεφαλαίου αυτού θα παρουσιαστούν αναλυτικά κάθε βήμα της διαδικασίας αυτής.

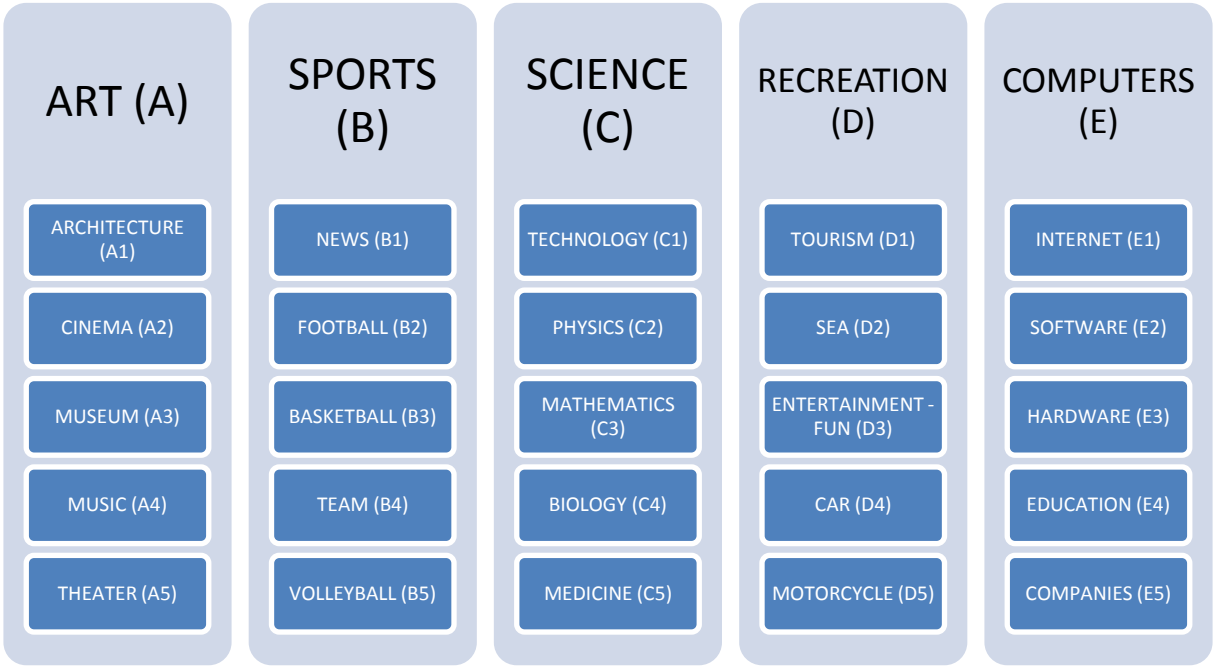
5.1. Συλλογή Δεδομένων

Η άντληση των πληροφοριών από ένα κοινωνικό δίκτυο, όπως το Twitter, αποτελεί το πρώτο και το σημαντικότερο βήμα για την μετέπειτα ανάλυση αυτών. Τα δεδομένα τα οποία αντλούνται από το Twitter στην συγκεκριμένη διπλωματική αποτελούν tweets που δημοσιεύονται στο public timeline. Η επιλογή βέβαια των tweets αυτών δεν αποτελούσε μια τυχαία, αυθαίρετη διαδικασία. Τα tweets που συλλέξαμε ικανοποιούν κάποια συγκεκριμένα κριτήρια, τα οποία ορίστηκαν από εμάς εξαρχής. Τα κριτήρια αυτά είναι:

- **Θεματική Ενότητα tweets:** Πραγματοποιήθηκε άντληση ενός συνόλου tweet από συγκεκριμένες θεματικές ενότητες που ονομάζονται hashtags. Τα hashtag από τα οποία συλλέξαμε τα δεδομένα διακρίνονται σε 5 βασικές κατηγορίες, οι οποίες είναι:

- ART
- SPORTS
- SCIENCE
- RECREATION
- COMPUTERS

Κάθε ένα από τα παραπάνω hashtag χωριζόταν περεταίρω σε 5 υποκατηγορίες δημιουργώντας την παρακάτω ιεραρχική δομή για το κάθε hashtag. (Εικόνα 9). Οι ετικέτες που εμφανίζονται για κάθε κατηγορία hashtag (π.χ A1, A2 …), δόθηκαν για τον εύκολο διαχωρισμό μεταξύ τους, έτσι ώστε να μην χρειάζεται η συνεχής αναφορά στο όνομα της κάθε κατηγορίας, όποτε ήταν απαραίτητο.



Εικόνα 9 Κατηγορίες των Hashtag

Η επιλογή των παραπάνω κατηγοριών έγινε με βάση την κατηγοριοποίηση που παρέχει στο site του το Dmoz ή αλλιώς το Open Directory Project (ODP). Το ODP είναι ένα project ανοικτού κώδικα με σκοπό την ταξινόμηση των περιεχομένων του Παγκόσμιου Ιστού (World Wide Web) με την μορφή καταλόγων. Χρησιμοποιεί ένα ιεραρχικό σχήμα οντολογιών για την οργάνωση των περιεχομένων μιας ιστοσελίδας [21]. Η αναζήτηση επομένως των tweets που επιθυμούσαμε να συλλέξουμε ήταν εξαρχής στοχευμένη. Επικεντρωνόταν εξ' ολοκλήρου στον εντοπισμό των 25 υποκατηγοριών hashtag που συναντάμε παραπάνω (A1 – E5) και έπειτα στη συλλογή των tweets που εμπεριέχονταν σε αυτές.

Στο περιεχόμενο των tweets που συγκεντρώσαμε εντοπίσαμε βέβαια και αναφορές σε άλλα hashtag, στα οποία δώσαμε και την ετικέτα «σχετικά hashtag» ή «related hashtag». Κάθε μια από τις 25 υποκατηγορίες συγκέντρωνε και ένα δικό της αριθμό από related hashtags. Η επαλήθευση των σχετικών αυτών ετικετών πραγματοποιήθηκε και με τη χρήση ενός online εργαλείου του Twitter που ονομάζεται Hashonomy. Το Hashonomy αποτελεί μια κοινωνική bookmarking υπηρεσία του Twitter, όπως το ίδιο αναφέρει στην ιστοσελίδα του, το οποίο παρέχει τη δυνατότητα προβολής των related hashtag για κάθε hashtag που του δίδεται να αναζητήσει.

- **Συνολικός αριθμός tweets:** Επόμενο κατά σειρά κριτήριο της επιλογής των tweets αποτέλεσε ο συνολικός αριθμός που θα επιλέγονταν από κάθε hashtag. Ως ικανοποιητικός αριθμός δείγματος δεδομένων κρίθηκε ο αριθμός των 200 tweet ανά υποκατηγορία hashtag. Συνολικά αντλήθηκαν περίπου 5000 tweets από το Twitter.

- **Περιεχόμενο tweets:** Ως προϋπόθεση για την υλοποίηση της διπλωματικής ορίστηκε πως η ανάλυση των περιεχομένων από τα δεδομένα που αντλήθηκαν θα περιοριζόταν μόνο στο κείμενο το οποίο περιείχαν. Επομένως τα tweets που ελέχθησαν ως αντιπροσωπευτικό δείγμα δεδομένων αποτελούνταν μόνο από αλφαριθμητικούς χαρακτήρες και υπερσυνδέσμους (urls). Tweet τα οποία εμπεριείχαν οποιασδήποτε μορφής πολυμέσα,

απορρίφθηκαν εξαρχής και δεν αντλήθηκαν καν.

- **Επικαιρότητα των tweets:** Το Twitter θεωρείτε εκτός από ένα κοινωνικό δίκτυο επικοινωνίας μεταξύ των χρηστών και ένα δίκτυο άμεσης και επίκαιρης ενημέρωσης των γεγονότων που διαδραματίζονται στον κόσμο. Οι χρήστες επιλέγουν το Twitter για να ενημερωθούν και να δημοσιεύσουν γεγονότα που γίνονται την κάθε στιγμή. Δεδομένης λοιπόν της παραπάνω κατάστασης θεωρήθηκε ως ορθό από μέρους μας να γίνει άντληση των tweets που ανταποκρίνονταν στην επικαιρότητα των γεγονότων ή των πληροφοριών που ανέφεραν. Συλλέχθηκαν έτσι tweets που ο χρόνος δημοσίευσης τους δεν ξεπερνούσε τις 2 ημέρες από τη στιγμή της συλλογής τους. Τα δεδομένα που συλλέξαμε επομένως επαληθεύουν τα γεγονότα και τα δρώμενα εκείνης της χρονικής περιόδου.

- **Μοναδικότητα των tweets:** Βασικό κριτήριο για την επιλογή και άντληση των tweet από το Twitter, αποτέλεσε η μοναδικότητα τους. Το περιεχόμενο σαν σύνολο για κάθε tweet ήταν διαφορετικό από οποιοδήποτε άλλο tweet συλλέγονταν από το ίδιο hashtag. Βάση αυτού του κριτηρίου και το σύνολο των αποτελεσμάτων μας κρίνεται με περισσότερη εγκυρότητα και ρεαλιστικότητα.

- **Γλώσσα του περιεχομένου των tweets:** Το τελικό κριτήριο για την επιλογή των δεδομένων (tweets) αποτέλεσε η χρήση της αγγλικής ή ελληνικής γλώσσας από τους χρήστες που δημοσίευσαν το tweet. Πιο απλά, γινόταν συλλογή μόνο των tweet που εκπλήρωναν τα παραπάνω κριτήρια και ήταν γραμμένα στα Ελληνικά ή στα Αγγλικά.

5.1.1. Τρόπος Συλλογής Δεδομένων

Η συλλογή των tweets πραγματοποιήθηκε με δυο τρόπους από το Twitter, αυτοματοποιημένα αρχικά και έπειτα χειροκίνητα. Η αυτοματοποιημένη συλλογή γινόταν με την υλοποίηση και χρήση ενός προγράμματος - λογισμικού που κατασκευάστηκε από εμάς. Αντίθετα η χειροκίνητη άντληση των δεδομένων γινόταν με την χρήση ενός online tool (δικτυακό εργαλείου) που προσέφερε η ιστοσελίδα του Twitter και ονομάζεται Twitter Advanced Search.

5.1.1.1. Αυτοματοποιημένος τρόπος άντλησης tweets μέσω υλοποίησης λογισμικού

Η υλοποίηση του προγράμματος για την άντληση των tweet που ικανοποιούσαν τα κριτήρια που θέσαμε, στηρίχτηκε στη δυνατότητα, που προσφέρει το Twitter, αλληλεπίδρασης με το περιβάλλον του, επομένως και με το σύνολο των περιεχομένων του. Η δυνατότητα αυτή παρέχεται με τη βοήθεια του API (Application programming interface).

Με τη χρήση του API υπάρχει η δυνατότητα δημιουργίας εφαρμογών, websites, ή άλλα projects τα οποία είναι σε θέση να αλληλεπιδρούν με το Twitter. Αναλυτικότερα κάποια εφαρμογή μέσω του API μπορεί να «ποστάρει» νέα tweets, να σβήνει κάποια tweets, να χαρακτηρίζει ένα tweet ως «αγαπημένο» (favorite), να βλέπει tweets από οποιαδήποτε timeline (home ή public), να κάνει σύνθετες αναζητήσεις tweet, users και hashtags με βάση συγκεκριμένα κριτήρια. Ουσιαστικά η δημιουργίας μιας τέτοιας εφαρμογής κάνει ότι ακριβώς μπορεί να κάνει ένα χρήστης απλά πιθανότερα γρηγορότερα λόγω του άμεσης επικοινωνίας με τα δεδομένα του Twitter και του αυτοματοποιημένου τρόπου λειτουργίας.

Στην πραγματικότητα το Twitter παρέχει τρία APIs., το Representational State Transfer (**REST**) API, το **Search** API και το **Streaming** API, καθένα από τα οποία βρίσκει και διαφορετική χρησιμότητα αναλόγως της διαδικασίας που θέλεις

να πραγματοποιήσεις.

- Το REST API μπορεί να υλοποιήσει όλες τις βασικές λειτουργίες που μπορεί να κάνει κανείς στο Twitter. Μέσω του REST API μπορεί δηλαδή, να δημοσιεύσει κάποιο μήνυμα στο status του, να κάνει retweet κάποιο μήνυμα, να κάνει follow και unfollow άλλα άτομα, να στείλει direct messages , ή και να οργανώσει τις λίστες του [22].

- Το Search API, από την άλλη μεριά, δίνει τη δυνατότητα στον προγραμματιστή, να κάνει ότι μπορεί να κάνει και το Twitter Advanced Search, που αποτελεί online εργαλείο της ιστοσελίδας του δικτύου. Μπορεί δηλαδή, να αναζητήσει tweets με βάση κάποιες λέξεις-κλειδιά ή με βάση κάποιους συνδυασμούς λέξεων. Μπορεί ακόμα να κάνει αναζήτηση με βάση το hashtag, το σύνολο των tweets που εμπεριέχονται σε αυτό, το χρήστη που τα δημοσίευσε ή χρήστες που αναφέρονται στα tweets αυτά και άλλα [23].

- Το Streaming API απευθύνεται σε όσους θέλουν να συλλέγουν μεγάλη ποσότητα πληροφοριών σε πραγματικό χρόνο, σε αντίθεση με το REST API που τα δεδομένα δεν είναι τόσο πρόσφατα. Συνήθως χρησιμοποιείται για εφαρμογές που συλλέγουν δεδομένα για αναλυτική έρευνα. Ο λόγος που δεν το χρησιμοποιήσαμε στην εφαρμογή μας είναι γιατί ψάχνει μόνο στο 1% των tweets που δημιουργούνται και γιατί δε μπορεί να ψάξει μόνο για hashtag με βάση τα ερωτήματα [24].

Το Twitter API επικοινωνεί με το λογισμικό που δημιουργήσαμε μέσω του πρωτοκόλλου HTTP. Μέσω λοιπόν του πρωτόκολλου επικοινωνίας HTTP είναι δυνατή η δημιουργία ενός ερωτήματος προς το Server του Twitter API, με τη χρήση των μεθόδων GET και POST [25], επιστρέφοντας ένα σύνολο αποτελεσμάτων με βάση τα κριτήρια που χρησιμοποιήθηκαν στο ερώτημα αυτό. Πιο συγκεκριμένα το πρόγραμμα μας έκανε χρήση κυρίως της εντολής ¨GET search/tweets¨ με την

version 1.0 του Twitter Search API.

Αξίζει να σημειωθεί πως στο Search API και συγκεκριμένα στην εντολή ¨GET search/tweets¨ θέσαμε κάποιους περιορισμούς στην άντληση των tweets. Οι κυριότεροι από αυτούς είναι οι εξής:

- Αναζητούσαμε tweets που συναντώνται μόνο σε συγκεκριμένες hashtag, οι οποίες αναφέρονται στην ενότητα 5.1. Ο περιορισμός αυτός τίθεται με τη χρήση ενός ειδικού χαρακτήρα «%23», που αντιπροσωπεύει τα hashtags για το Twitter API, και έπειτα το όνομα του hashtag (πχ. %23art).

- Επιτρέπαμε στο πρόγραμμα να πραγματοποιεί 150 requests ανά ώρα, καθώς τα αιτήματα που στέλναμε στο Server του Twitter API αποτελούνταν από μη αυθεντικοποιημένη πηγή. Οι κλήσεις αυτές καταμετρούνταν μέσω της δημόσιας IP του υπολογιστή που έκανα τις κλήσεις.

- Ο συνολικός αριθμός tweet που συγκεντρωνόταν ανά κατηγορία hashtag δεν ξεπερνούσε σε σύνολο τα 200 ανά σελίδα εμφάνισης του ερωτήματος. Ο περιορισμός αυτός τίθονταν εξίσου στο ερώτημα με τη μορφή παραμέτρου με το όνομα “rpp=200”. Εφόσον εμείς θέλαμε να συλλέξουμε τελικά 200 tweets για κάθε hashtag συλλέγαμε μόνο την πρώτη σελίδα κάθε ερωτήματος που στελλόταν στον Server του API.

- Η ημερομηνία των δημοσιευμένων tweets ενός hashtag αποτελούσε ένα ακόμα περιορισμό του ερωτήματος, που εξασφαλιζόταν με τη χρήση της εντολής “until” και έπειτα την ημερομηνία που επιθυμούσαμε. Αναλόγως με την ημερομηνία που γινόταν η αναζήτηση και η άντληση των tweets, θέταμε ως περιορισμό έως και 2 ημέρες πριν.

Η version 1.0 του Twtter API παρείχε τη δυνατότητα άντλησης δεδομένων,

ικανοποιώντας βέβαια τον περιορισμό των 150 requests ανά ώρα. Παρόλα αυτά διδόταν και η δυνατότητα ελέγχου της ταυτότητας του χρήστη και εξουσιοδότησης από τη μεριά του REST API, αν όριζες την εφαρμογή σου ως web application και έκανες χρήση εκείνου του API, αντί του Search. Η εξουσιοδότηση σε αυτή την περίπτωση γίνεται μέσω του πρωτοκόλλου OAuth. Το OAuth είναι ένα ανοιχτό πρότυπο για εξουσιοδότηση. Επιτρέπει στους χρήστες να μοιράζονται προσωπικά τους δεδομένα (φωτογραφίες, βίντεο, λίστες επαφών, tweets), που είναι αποθηκευμένα στο REST API του Twitter, μέσω εφαρμογών λογισμικού που δημιουργούν οι χρήστες αυτοί. Η αλληλεπίδραση αυτή δεν απαιτεί την χρήση κωδικών (passwords), αλλά tokens. Τα tokens αυτά μπορεί να παρέχουν συγκεκριμένα δικαιώματα πρόσβασης και για περιορισμένη χρονική διάρκεια. Η χρήση της εξουσιοδότησης παρείχε υψηλότερο ρυθμό άντλησης δεδομένων (rate limit), αλλά αρκετούς αρνητικούς περιορισμούς.

Η αρχική προσπάθεια άντλησης των δεδομένων από το Search API έγινε τη χρήση της γλώσσας PHP. Χρησιμοποιήθηκε η βιβλιοθήκη «tmhOAuth» που προσφέρεται από το Twitter API. Η βασική ιδέα ήταν η άντληση των απαιτούμενων tweet και η αποθήκευση τους σε μια βάση δεδομένων (SQL κατά προτίμηση). Η προσπάθεια αυτή εγκαταλείφθηκε αρκετά γρήγορα εξαιτίας τεσσάρων κυρίως παραγόντων. Αρχικά ο κώδικας που απαιτούνταν για την άντληση των δεδομένων ήταν εξαιρετικά πολύπλοκος και δύσκολα προσαρμόσιμος σε οποιεσδήποτε αλλαγές απαιτούνταν. Επίσης απαιτούνταν η εξουσιοδότηση μέσω του πρωτοκόλλου OAuth, το οποίο προϋπέθετε και τη χρήση στατικού (static) IP για την ενεργή διατήρηση της σύνδεσης μεταξύ του λογισμικού και του Twitter API. Επιπλέον τα ερωτήματα που γινόταν στο Server του Twitter API επιστρέφονταν σε αρχεία JSON, όπου η επεξεργασία που απαιτούνταν για την αποθήκευση τους ήταν εξαιρετικά δύσκολη και χρονοβόρα. Τέλος οι υπολογιστικές απαιτήσεις που χρειαζόταν ήταν ιδιαίτερα υψηλές, καθώς τα δεδομένα αποθηκεύονταν στην μνήμη RAM του υπολογιστή μέχρι την καθολική τους αποθήκευση στη βάση δεδομένων.

Η επόμενη προσπάθεια άντλησης των δεδομένων επιτεύχθηκε με τη χρήση της γλώσσας Python. Πρόκειται για μια ιδιαίτερα νέα *αντικειμενοστραφή γλώσσα*

υψηλού επιπέδου, η οποία δημιουργήθηκε το 1990 και άρχισε να χρησιμοποιείται ευρύτατα την τελευταία δεκαετία. Χαρακτηρίζεται ως ιδιαίτερα εύκολη γλώσσα στην αναγνωσιμότητα του κώδικα της και στην ευκολία εκμάθησης και χρήσης από το χρήστη. Το μεγάλο σύνολο των έτοιμων βιβλιοθηκών που προσφέρει, διευκολύνει την πραγματοποίηση αρκετά συνηθισμένων εργασιών. Η γενικότερη ευκολία που τη χαρακτηρίζει έγκειται στο γεγονός της χρήσης των κενών διαστημάτων (whitespaces) για τον διαχωρισμό των συντακτικών δομών του προγράμματος, καθώς και στη χρήση πλήρων αγγλικών λέξεων στη θέση συμβόλων.

Η χρήση της βιβλιοθήκης Twython διευκόλυνε ακόμα περισσότερο την όλη διαδικασία συλλογής των δεδομένων. Το Twython αποτελεί μια βιβλιοθήκη που προσφέρει έναν εύκολο και συνεχώς ενημερωμένο τρόπο για την επικοινωνία με τα δεδομένα του Twitter API με τη χρήσης της γλώσσας προγραμματισμού Python. Η εγκατάσταση της βιβλιοθήκης αυτής γίνεται με τη χρήση μόνο μίας εντολής ενώ η άντληση των tweets, από συγκεκριμένα hashtag, και η εμφάνιση τους δεν ξεπερνά σε προγραμματιστικό κώδικα τις 6-10 γραμμές. Η παρουσίαση και η ανάλυση του κώδικα παρουσιάζεται στο Παράρτημα Ι της πτυχιακής εργασίας.

Το πρόγραμμα το οποίο δημιουργήσαμε αρχικά αλληλεπιδρούσε με το Twitter Search API version 1.0 και αντλούσε 200 μοναδικά tweets από κάθε hashtag που ορίζαμε ως παράμετρο του προγράμματος μας. Τηρώντας τον περιορισμό των 150 request ανά ώρα, η λειτουργία του προγράμματος κρίνονταν βέλτιστη στην λήψη και αποθήκευση των δεδομένων. Τα δεδομένα που λαμβάνονταν ήταν με την μορφή αρχείων JSON, τα οποία έπειτα αποθηκεύονταν με τη μορφή απλού αρχείου κειμένου (txt format), για κάθε ένα hashtag του οποίου ολοκληρώνονταν η λήψη των tweets. Η επικοινωνία με το Search API πραγματοποιούνταν εντελώς ανώνυμα, χωρίς τη ανάγκη εξουσιοδότησης ή επικύρωσης (authentication) του χρήστη.

Τα αποτελέσματα τα οποία επέστρεφε ένα αρχείο JSON για κάθε request αποτελούνταν από:

- Το περιεχόμενο του tweet. Αν το περιεχόμενο συμπεριλάμβανε

κάποιον υπερσύνδεσμο (hyperlink), εμφανιζόταν ξεχωριστά.

- Το χρήστη που το δημοσίευσε.
- Την ημερομηνία που το δημοσίευσαν.
- Το hashtag στο οποίο υπάρχει αναφορά.

Όλα τα παραπάνω συνέβαιναν μέχρι τον Σεπτέμβριο του 2012 όπου το Twitter άλλαξε πολιτική και δημιούργησε το Twitter API version 1.1. Η αλλαγή της πολιτικής αυτής πραγματοποιήθηκε εξαιτίας της αλόγιστης χρήσης των API από τους χρήστες, προκειμένου να αντλήσουν δεδομένα, που οδήγησε πολύ φορές μέχρι και στην «κατάρρευση» του Twitter Server. Επομένως η εταιρία αποφάσισε να δημιουργήσει το νέο API, με το οποίο για να αποκτήσεις πρόσβαση απαιτείται η εξουσιοδότηση μέσω του πρωτοκόλλου OAuth καθώς και η δημιουργία λογαριασμού Twitter από τον οποίο θα πραγματοποιείς κανονικά login στο API με το username και το password που διαθέτεις στο λογαριασμό σου.

Από μέρους μας έγινε προσπάθεια αλλαγής του κώδικα του προγράμματος μας για να λειτουργεί με βάση τα νέα δεδομένα που προέκυψαν. Αρχικά δημιουργήθηκε ένας λογαριασμός στο Twitter και μας δόθηκε η απαραίτητη εξουσιοδότηση που απαιτούνταν μέσω του πρωτοκόλλου OAuth, δηλώνοντας την εφαρμογή μας ως web application. Συναντήθηκαν αρκετά προβλήματα και δαπανήθηκε αρκετός χρόνος έως ότου μπορέσουμε να αποκτήσουμε πρόσβαση και πάλι στους Server του Twitter API, αλλά τα βασικότερα και σημαντικότερα προβλήματα εντοπίστηκαν στην άντληση των tweets.

Τα request προς τον Twitter API Server που πραγματοποιούσαμε προηγουμένως με την έκδοση 1.0 περιοριζόταν στα 150. Εισάγοντας μια μεγάλη χρονική καθυστέρηση στο πρόγραμμα μας μεταξύ των συνεχόμενων request καταφέρναμε να λαμβάνουμε όλο το σύνολο των tweet χωρίς να χρειαστεί επανασύνδεση με το Twitter Search API, για την λήψη των υπολοίπων. Με την δεύτερη έκδοση του API όμως, η χρονική αυτή καθυστέρηση δεν ήταν αρκετή για να διατηρήσουμε ενεργή τη σύνδεση μας με το API του Twitter, καθώς σε **διαρκή**

επικοινωνία πάνω από 4 περίπου ώρες η σύνδεση διακοπτόταν από τον ίδιο το Server του API. Στην προσπάθεια επανασύνδεσης με το API, τα δεδομένα που είχαν αντληθεί στα πρώτα 150 request χανόταν με αποτέλεσμα να αντλούμε και πάλι τα ίδια στην επόμενη επανασύνδεση.

Προσπαθώντας να δώσουμε λύση στο παραπάνω πρόβλημα, αποθηκεύαμε τα δεδομένα που λαμβάναμε, προτού πραγματοποιήσουμε εκ νέου σύνδεση με το API, με το μορφή αρχείων txt κειμένου. Αμέσως μετά την αποθήκευση αλλάζαμε την IP στην σύνδεση μας στο Internet και αποκτούσαμε εκ νέου πρόσβαση στο API. Στο ενδιάμεσο διάστημα όμως της διακοπής του προγράμματος, της αποθήκευσης των tweets και της επανασύνδεσης με αλλαγμένη πλέον IP, υπήρχαν αρκετοί χρήστες οι οποίες πόσταραν καινούργια tweets, που αναφερόταν στο hashtag που ορίζαμε στην αναζήτηση μας. Αυτό είχε ως αποτέλεσμα τα νέα αυτά tweet να εμφανίζονται πρώτα στη λίστα των αποτελεσμάτων του προγράμματος μας, κάνοντας αδύνατο τον εντοπισμό του προηγούμενου σημείου που σταματήσαμε τη λειτουργία του προγράμματος μας.

Έγιναν αρκετές προσπάθειες επίλυσης των παραπάνω προβλημάτων και διαφορετικές δοκιμές του προγράμματος μας. Οι προσπάθειες μας σταμάτησαν όταν το πρόγραμμα μας κρίθηκε ως κακόβουλο λογισμικό, βάση του OAuth πρωτόκολλου, και απαγορεύτηκε η περαιτέρω είσοδος, επικοινωνία και χρήση του Twitter Search API. Το θετικό βέβαια στην παραπάνω διαδικασία ήταν ότι μέχρι το Σεπτέμβριο του 2012, όπου και ενεργοποιήθηκε το νέο version του Twitter API, από μέρους μας είχε πραγματοποιηθεί η συλλογή των tweets σε 22 από τις συνολικές 25 υποκατηγορίες hashtag, που παρουσιάζονται και στην Εικόνα 9.

Ο συνολικός αριθμός των tweets που είχε συγκεντρωθεί, πριν την εμφάνιση του API version 1.1, ανερχόταν στο ποσό των 4400 με 4800 μοναδικών tweets περίπου. Ο συνολικός χρόνος που απαιτήθηκε για την λήψη των tweets από τα 22 αυτά hashtag ήταν 7-8 εβδομάδες. Κατά μέσο όρο για κάθε hashtag απαιτούνταν η λειτουργία του προγράμματος 14 με 16 ώρες, συμπεριλαμβάνοντας και στο χρόνο αυτό την αποθήκευση που γινόταν σε αρχεία txt. Βέβαια ο χρόνος αυτός δεν ήταν

πάντα σταθερός, καθώς εξαρτιόταν τόσο από την ταχύτητα της σύνδεσης του Internet που διαθέταμε, όσο και από την «κίνηση» που επικρατούσε στο Server του Twitter API.

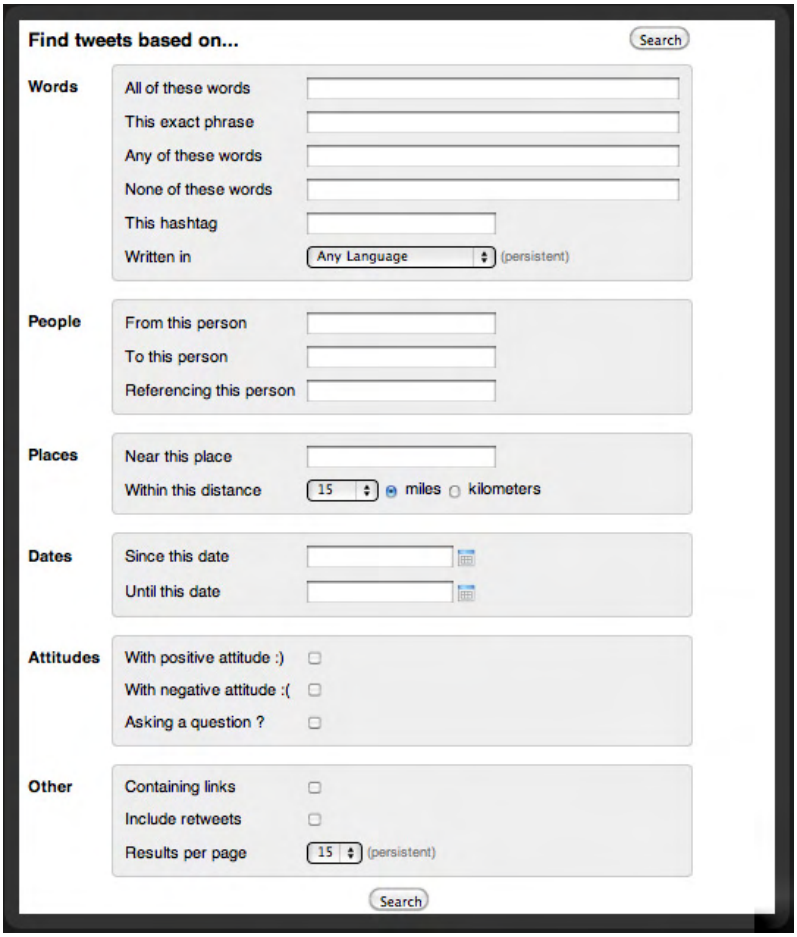
Συλλογιζόμενοι τα παραπάνω ζητήματα αποφασίσαμε πως θα ήταν καλύτερο να χρησιμοποιήσουμε ένα online εργαλείο που παρέχεται από την ιστοσελίδα του Twitter, για την άντληση των tweets των επόμενων τριών hashtag που είχαν απομείνει. Κάναμε επομένως χρήση του online tool που ονομάζεται Twitter Advanced Search.

5.1.1.2. Συλλογή tweets με το Twitter Advanced Search

Το Twitter Advanced Search αποτελεί ένα εργαλείο το οποίο παρέχει τη δυνατότητα αναζήτησης ενός μεγάλου συνόλου δεδομένων, χρησιμοποιώντας διαφορετικά κριτήρια. Μέσω του εργαλείου αυτού, που παρέχεται από την επίσημη ιστοσελίδα του Twitter, μπορείς να αναζητήσεις tweets που εμπεριέχουν ένα keyword, που θέτεις ως κριτήριο, tweets που περιέχουν πολυμέσα, tweets που δημοσιεύτηκαν από συγκεκριμένους users ή σε συγκεκριμένη ημερομηνία, tweet που εμπεριέχουν κάποιο hashtag και άλλα πολλά. Το Twitter Advanced Search επικοινωνεί και αντλεί τα δεδομένα, που εμφανίζει στο χρήστη που κάνει την αναζήτηση, με το Streaming API του Twitter. Χρησιμοποιώντας ως χρήστης το κριτήριο που κρίνεις ως ορθό για τα δεδομένα που αναζητάς, το Twitter Advanced Search μετατρέπει την αναζήτηση σε ένα ερώτημα μεθόδου GET με τα αντίστοιχα κριτήρια ως χαρακτηριστικά της μεθόδου.

Βασιζόμενοι λοιπόν στις παραπάνω δυνατότητες που παρέχονται από το διαδυκτιακό αυτό tool αναζήτησης, καθώς και στα προβλήματα που αντιμετωπίσαμε με την εμφάνιση του Twitter API version 1.1, αποφασίσαμε να χρησιμοποιήσουμε το Twitter Advanced Search για να αντλήσουμε τα tweets των υπόλοιπων τριών hashtag που απέμεναν. Η αναζήτηση πραγματοποιούνταν με κριτήρια την πληκτρολόγηση του hashtag, του οποίου αναζητούσαμε τα tweets του,

στο πεδίο «this hashtag:» και επιλογή του πεδίου της ημερομηνίας «since this day:» στο οποίο βάζαμε την ημερομηνία 2 ημερών πριν (Εικόνα 10).



Εικόνα 10 Twitter Advanced Search

Η αναζήτηση ήταν ιδιαίτερα γρήγορη και αποδοτική, όσο αναφορά τα αποτελέσματα που επέστρεφε το Twitter Advanced Search. Το δυσκολότερο όμως κομμάτι ήταν η συλλογή των 200 πρώτων μοναδικών tweet που θα εμφανίζονταν για κάθε hashtag. Η επιλογή και η αποθήκευση τους γινόταν με χειροκίνητο τρόπο σε αρχεία του Microsoft Office Excel (.xls), καθώς η στοίχιση των tweet κρινόταν αδύνατη σε αρχεία απλού κειμένου txt. Εκτός από το κειμενικό περιεχόμενο των tweet, συλλέγονταν επίσης και ο user που έκανε τη δημοσίευση καθώς και η ώρα που έγινε το post.

Συνολικά με αυτή τη χειροκίνητη διαδικασία συλλέχθηκαν 600 – 650 περίπου tweets που αφορούσαν τις τελευταίες τρεις κατηγορίες hashtag που είχαν αρχικά οριστεί. Η συνολική διάρκεια συλλογής αυτών των tweets διήρκησε 1 εβδομάδα περίπου και κάθε μέρα συλλέγονταν περίπου 100 μοναδικά tweet. Εφόσον οδηγηθήκαμε στη δημιουργία αρχείων xls για την αποθήκευση των tweet, αποφασίσαμε να μεταφέρουμε και τα υπόλοιπα δεδομένα (4400 – 4800 tweets), που συλλέξαμε με τη χρήση του προγράμματος λογισμικού, σε ίδιας μορφής αρχείων.

Μετά την ολοκλήρωση της λήψης των tweet και των τελευταίων τριών hashtag, είχαμε στη διάθεση μας ένα σύνολο 5000-5500 μοναδικά tweet που χρειάστηκαν συνολικά 9 βδομάδες περίπου για την απόκτηση σου. Η άντληση των δεδομένων ξεκίνησε το μήνα Ιούλιο του 2012 και ολοκληρώθηκε επιτυχώς στις αρχές του μήνα Σεπτεμβρίου. Επόμενο μέλημα της πτυχιακής μας έρευνας ήταν η επεξεργασία των δεδομένων αυτών, ώστε να κριθούν κατάλληλα για τη χρησιμοποίηση τεχνικών και αλγορίθμων συσταδοποίησης.

5.2. Επεξεργασία Δεδομένων

Εφόσον είχε πραγματοποιηθεί ή λήψη των επιθυμητών tweets, είχε ουσιαστικά ολοκληρωθεί και το πρώτο βήμα της πτυχιακής μας εργασίας που αφορούσε την άντληση δεδομένων. Το επόμενο κατά σειρά βήμα σε μια διαδικασία Εξόρυξης Δεδομένων, που πραγματοποιείται και στη δικιά μας πτυχιακή εργασία, είναι η κατάλληλη επεξεργασία των δεδομένων και ο ορθός μετασχηματισμός τους, προκειμένου να οδηγηθούμε στο τελικό στάδιο. Η επεξεργασία αυτή αποτελείται από τρείς βασικές διεργασίες:

- Η **Προεπεξεργασία** του κειμενικού περιεχομένου των tweet
- Η **Δεικτοδότηση** των term που προκύπτουν από την παραπάνω διαδικασία
- Η κατάλληλη **Αναπαράσταση** των term αυτών

5.2.1. Προεπεξεργασία κειμενικού περιεχομένου των tweets

Η Προεπεξεργασία του κειμενικού περιεχομένου των δεδομένων (tweets) αποτελεί μια πολύ βασική διεργασία σε ένα σύστημα Ανάκτησης Πληροφορίας, η οποία προηγείται της δεικτοδότησης των στοιχείων και της διανυσματικής τους αναπαράστασης. Η Προεπεξεργασία του περιεχομένου πραγματοποιεί τη βασική λειτουργία της εξαγωγής των χαρακτηριστικών όρων των tweet κάθε hashtag, που ονομάζονται όροι δεικτοδότησης (index terms) και είναι οι κατάλληλοι όροι, εφόσον ολοκληρωθεί η διαδικασία, να αναπαραστήσουν και να αντιπροσωπεύσουν το κειμενικό περιεχόμενο των tweet.

Τα στάδια από τα οποία αποτελείται η διαδικασία Προεπεξεργασίας των συλλεγμένων tweet είναι και πραγματοποιούνται με την σειρά που αναφέρονται:

- **Αναγνώριση και αφαίρεση της οποιασδήποτε δομής των περιεχομένων.** Στην περίπτωση που το κειμενικό περιεχόμενο των αρχείων, που βρίσκονται αποθηκευμένα το σύνολο των 5000 περίπου tweets, αποτελείται από οποιαδήποτε δομικά στοιχεία, απαιτείται η αφαίρεση των δομών αυτών. Τα αρχικά αρχεία απλού κειμένου (txt), όπου αποθηκεύαμε τα tweet που αντλούσαμε μέσω του αυτοματοποιημένου προγράμματος, εμπεριείχαν κάποια δομικά στοιχεία εξαιτίας της άμεσης αποθήκευσης των tweet από JSON μορφή. Η μετέπειτα όμως αποθήκευση των tweet αυτών σε αρχεία του Office Excel έλυσε το θέμα της δόμησης των στοιχείων. Ένας διαχειριστής υπολογιστικών φύλλων, όπως είναι το Excel, δεν διαθέτει στοιχεία δόμησης, παρά μόνο γραμμές και στήλες.

- **Λημματοποίηση (Tokenization).** Στο βήμα αυτό κάθε tweet διαχωρίζεται σε λήμματα (tokens). Τα λήμματα μπορεί να είναι λέξεις, αριθμοί ή σημεία στίξης. Στην παρούσα πτυχιακή εργασία ο διαχωρισμός των tweet έγινε βάση λέξεων και αριθμών, χωρίς να

λαμβάνονται υπόψη τα σημεία στίξης, ως ξεχωριστά tokens – terms.

- **Αφαίρεση των common words.** Ως common word ορίζεται ένας όρος ο οποίος έχει μεγάλη συχνότητα εμφάνισης στα tweet και ο οποίος δεν αποδίδει καμία βοήθεια στον χαρακτηρισμό του κειμενικού περιεχομένου, επομένως και στην εξαγωγή συμπερασμάτων και στατιστικών μετρήσεων. Οι όροι αυτοί είναι συνήθως προθέσεις, άρθρα, σύνδεσμοι προτάσεων, καθώς και κάποια ρήματα. Εάν συμπεριληφθούν οι όροι αυτοί στην ανάλυση των δεδομένων το πιθανότερο είναι να αποτελέσουν μια αιτία θορύβου στα δεδομένα, με αποτέλεσμα την μείωση της εγκυρότητας των αποτελεσμάτων.

5.2.1.1. Χωρισμός των tweets σε terms

Η άντληση των tweet από το Twitter Search API είχαν ως αποτέλεσμα την δημιουργία 25 συνολικά αρχείων του Excel (xls), καθένα από τα οποία εμπεριείχε ένα σύνολο 200 tweet. Τα αρχεία αυτά δομούνταν με την εξής λογική. Στην πρώτη στήλη (A) αποθηκεύονταν το κειμενικό περιεχόμενο των tweet, στην επόμενη στήλη (B) ο user που πραγματοποίησε τη δημοσίευση του συγκεκριμένου tweet και στην αμέσως επόμενη (C) η ώρα και η ημερομηνία δημοσίευσης. Κάθε γραμμή του Excel αποτελούσε και ένα διαφορετικό tweet με τα τρία διαφορετικά πεδία που αναφέρονται παραπάνω (κείμενο, χρήστης, ημερομηνία). Απεικονιστικό παράδειγμα της δομής που ακολουθήθηκε στα αρχεία παρουσιάζεται στην παρακάτω εικόνα (Εικόνα 11).

Ο διαχωρισμός των tweets σε terms, για κάθε αρχείο, έγινε βάση της στήλης A, δηλαδή βάση του κειμενικού τους περιεχομένου. Οι χρήστες (στήλη B) και οι ώρες δημοσίευσης (στήλη C) αποθηκεύτηκαν στα αρχεία αυτά για μεταγενέστερη χρήση και μελλοντική ίσως εξέλιξη της ανάλυσης πάνω στην πτυχιακή αυτή.

	A	B	C
1	Tweet	User	Date
2			
3	Στήλη1	Στήλη2	Στήλη3
4	I'm really enjoying studying #Greek #Architecture! VERY impresssiiiivee!	Hiba ElGihani @PearlHiba	4/7/2012 - 10:20
5	20 Modern Chic Living Room Designs for a Charming Lookhttp://su.pr/2LIVgq #design #architecture	Tony Fafa @Reeph	4/7/2012 - 10:23
6	Museum of Contemporary Art Cleveland by Farshid Moussavi: An glossy, augmented hexagon is Cleveland's... http://dlvr.it/2JmVlV #architecture	Mark Magazine @markmagazine	4/7/2012 - 10:24
7	Amazing Properties - http://www.homeadore.com/2012/09/18/amazing-houses/ ...#architecture	Home Adore @HomeAdore	4/7/2012 - 10:24
8	Via @Dezeen - Vito Acconci interview: "Architecture is not about space but about time" http://dlvr.it/2Jm5zJ #architecture	Architecture Feeds @archfeeds	4/7/2012 - 10:24
9	Vito Acconci interview: "Architecture is not about space but about time": Architecture magazines... http://bit.ly/X3v6Eo #architecture	Archiburner @archiburner	4/7/2012 - 10:25
10	Wander the streets of #Sydney to admire the #architecture - tips here from a local architect http://bit.ly/P6HVuW #weekend #favourites	Lara & Terence @gran_tourismo	4/7/2012 - 10:26
11	Via @DesignBoom - bellposition furniture collection by pos1t1onhttp://dlvr.it/2JID70 #architecture #design	Architecture Feeds @archfeeds	4/7/2012 - 10:27
12	It is in dialogue with pain that many beautiful things acquire their value. #art #architecture	Carlie Walters @carliewalters	4/7/2012 - 10:31
13	As I love pyramids, I think this photograph is amazing!http://www.flickr.com/photos/73230975@N03/6893326896/ ... | #architecture	Colorcoat @Colorcoat	4/7/2012 - 10:31
14	Housing in Taipei / Chin Architects http://archdai.ly/WdvMaJ #architecture	ArchDaily @ArchDaily	4/7/2012 - 10:38
15	Two Homes in Luque by BAUEN:http://www.architectureoflife.net/Blog/2213/Two-Homes-in-Luque-by-BAUEN.aspx ... #architecture #sustainability#residential #house #Paraguay #nature pic.twitter.com/DVm1JM1S	Architecture of Life @ArchofLife	4/7/2012 - 10:38
16	VIDEO___Toblerone House by studio mk27 http://shar.es/5FCOp #ARCHITECTURE	architecturalvideos* @einstainmc2	4/7/2012 - 10:38
17	this is just some amazing work]School in Balsiai / Sigitas Kucevičius Architecture Studio[http://www.archdaily.com/278514/school-in-balsiai-sigitas-kucevicius-architecture-studio/ ... #architecture#LIETUVA	Gintas Reisgys @GintasReisgys	4/7/2012 - 10:40
18	RT @AndisKakeli: Emotional #architecture http://bit.ly/PsV70k /via@WenataBabkowski @egoubano @rohitmal @terrinakamura@keithalink	Terri Nakamura @terrinaka	4/7/2012 - 10:44
19	El Campanil Theater, Antioch, CA. #architecture #theater #retro#Hipstamatic #HelgaViking #Pistil pic.twitter.com/Su4CONry	Mary @msmaryb	4/7/2012 - 10:45
20	My 1st rumah idaman :3 #perspective #house #architecturehttp://instagr.am/p/Qtrtxmknvv/	Andreina.A @boyaang	4/7/2012 - 10:45
21	The count down begins! The @RIBA Stirling Prize for #Architecture#stirling2012 http://bit.ly/Wh01d - winner is announced tonight!	B Irlam-Mowbray @irlammowbray	4/7/2012 - 10:47
22	#liege #liège #belgium #architecture #town #city #ville #urban#instagood #instamod #europe #w @ En Neuvicehttp://instagr.am/p/Qtrlk3K86A/	E-liege @E_liege	4/7/2012 - 10:48
	#instagram #instagrammer #instadaily #skyscraper #Tokyo#architecture #Japan #park #blue #summer #beautiful #s	JapanNews24 @JapanNewsTwo4	4/7/2012 - 10:49

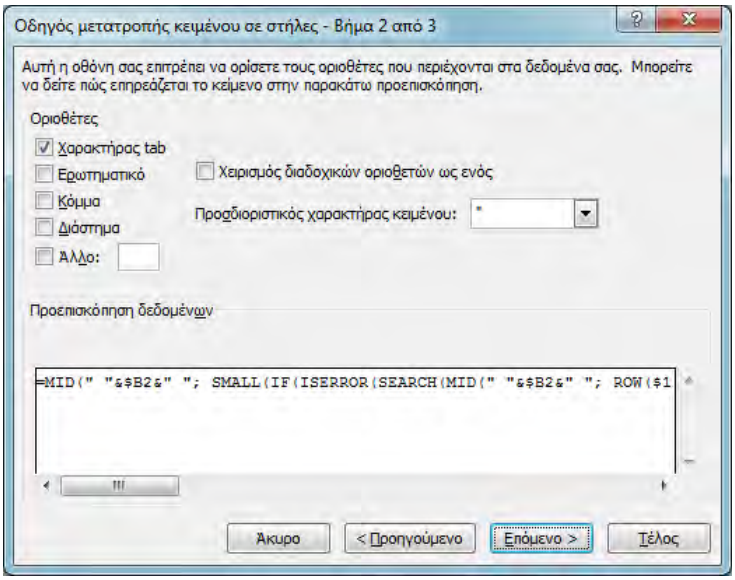
Εικόνα 11 Απεικονιστικό Παράδειγμα Tweet αποθηκευμένα στο Excel

Η αρχική προσπάθεια διαχωρισμού του κειμενικού περιεχομένου των tweet σε λέξεις (terms) πραγματοποιήθηκε με τη λειτουργία που παρέχεται από το ίδιο το Excel και ονομάζεται «Κείμενο σε Στήλες» (Data to Columns). Η επιλογή αυτή στο Microsoft Office 2010, που χρησιμοποιήθηκε, βρίσκεται στο μενού Δεδομένα της γραμμής μενού. Η επιλογή «Κείμενο σε Στήλες» παρέχει τη δυνατότητα της διαίρεσης των περιεχομένων ενός κελιού σε δεξιότερες στήλες, επιλέγοντας ένα σύνολο κριτηρίων. Επιλέγοντας πρώτα τα δεδομένα στα οποία επιθυμείς να εφαρμοστεί η επιλογή αυτή και έπειτα πατώντας το κουμπί «Δεδομένα σε Στήλες» εμφανίζεται ένα παράθυρο, στο οποίο επιλέγεις τα κριτήρια διαχωρισμού, όπως φαίνεται και στην παρακάτω εικόνα (Εικόνα 12). Προτού, όμως, πραγματοποιήσουμε τη διαδικασία διαχωρισμού σε terms, δημιουργήσαμε αντίγραφα των αρχείων αυτών, εισάγοντας μόνο στη στήλη Α το κειμενικό περιεχόμενο των tweet και δεν συμπεριλάβαμε καθόλου τους χρήστες και τις ημερομηνίες δημοσίευσης.

Με βάση τα νέα πλέον αντίγραφα που δημιουργήθηκαν, τα κριτήρια που επιλέξαμε ήταν να πραγματοποιείται διαχωρισμός όταν συναντάται:

- χαρακτήρας κενού (Spacebar),
- κόμμα,
- τελεία
- και ερωτηματικό μεταξύ λέξεων,
- καθώς επίσης και απόστροφος μεταξύ αυτών (π.χ στη λέξη «I'm»).

Επιπλέον χρησιμοποιήθηκε ως κριτήριο διαχωρισμό και η χρήση της « **/** » καθέτου, για να μπορέσουμε, μαζί με τον περιορισμό του εντοπισμού του σημείου στίξης της τελείας, να διαχωρίσουμε και τους υπερσυνδέσμους που εμπεριέχονταν στα tweet.



Εικόνα 12 Επιλογή κριτηρίων διαχωρισμού στο Excel

Με τον τρόπο αυτό κάθε μια λέξη των κελιών της στήλης Α, που αποτελούσε ουσιαστικά το κειμενικό περιεχόμενο των tweet, διαχωρίστηκε στις δεξιότερες στήλες χωρίζοντας έτσι τα tweets σε terms ανά γραμμή. Κάθε γραμμή των αρχείων Excel αποτελεί το σύνολο των λέξεων (terms) του κάθε tweet. Η αποτελεσματικότητα του διαχωρισμού μέσω της χρήσης της επιλογή «Δεδομένα σε στήλες» ήταν ιδιαίτερα ενθαρρυντική, αν και εμφάνιζε κάποια λάθη στην

προσπάθεια διαχωρισμού των υπερσυνδέσμων.

Τα μικρά προβλήματα που αντιμετώπιζε η παραπάνω τεχνική, λύθηκαν άμεσα με τη χρήση ενός έτοιμου προγράμματος, με το όνομα «Parsing Excel». Το συγκεκριμένο πρόγραμμα αναζητήθηκε και βρέθηκε στο διαδίκτυο ανάμεσα σε ένα σύνολο έτοιμων βιβλιοθηκών – προγραμμάτων που παρέχονται για την εφαρμογή του Excel. Το Parsing Excel πραγματοποιεί tokenization για κάθε λέξη που εντοπίζει σε ένα αρχείο του Excel. Δέχεται ως είσοδο ένα αρχείο xls και σε ένα χρονικό διάστημα των 10-15 λεπτών επιστρέφει ως αποτέλεσμα ένα αρχείο txt με την κάθε λέξη που συναντάται στο Excel, την μια κάτω από την άλλη και αλφαβητικά ταξινομημένες. Επίσης το πρόγραμμα επιτρέπει την επιλογή για το αν θα συμπεριληφθούν τα σημεία στίξης στην διαδικασία της δημιουργίας terms, την οποία εμείς απενεργοποιήσαμε. Το πρόγραμμα αυτό αποτέλεσε για εμάς ένα τύπο επικύρωσης και αύξησης της αποτελεσματικότητας της τεχνικής που χρησιμοποιήσαμε εμείς με τη βοήθεια του Excel.

Δυστυχώς δεν είναι γνωστή η γλώσσα προγραμματισμού με την οποία κατασκευάστηκε το πρόγραμμα αυτό, με αποτέλεσμα να μην γνωρίζουμε και τον ακριβή τρόπο που πραγματοποιεί την όλη διαδικασία διαχωρισμού. Εικάζουμε κάποια σενάρια για τον τρόπο λειτουργίας της εφαρμογής αυτής, που βασίζονται στην περιγραφή που παρέχεται για τη βιβλιοθήκη που χρησιμοποιεί. Η λειτουργία του προγράμματος και των αποτελεσμάτων που παρείχε, ελέγχθηκαν διεξοδικά από τη δικιά μας μεριά καθώς επίσης συγκρίθηκαν και με τα terms που παρήχθησαν με τη χρήση της δικιάς μας τεχνικής.

Το αποτέλεσμα του διαχωρισμού των tweets σε terms που τα αντιπροσώπευαν είχε ως αποτέλεσμα την δημιουργία 25 αρχείων απλού κειμένου (txt). Τα αρχεία αυτά περιείχαν τα σύνολο των terms, για κάθε διαφορετικό hashtag, αλφαβητικά ταξινομημένα.

5.2.1.2. Υπολογισμός συχνότητας εμφάνισης των terms και απαλοιφή περιττών λέξεων

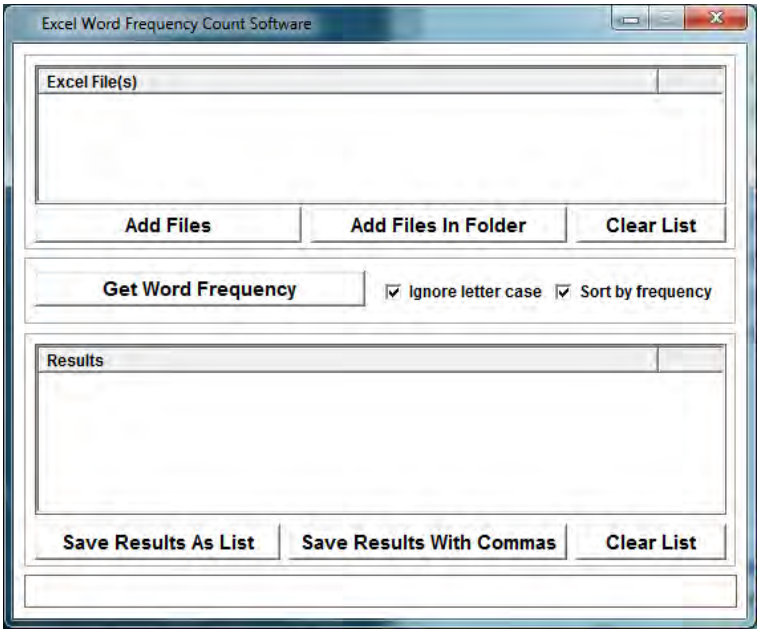
Η επόμενη διαδικασία που ακολουθήθηκε ήταν της απαλοιφής όλων των περιττών terms που υπήρχαν στο σύνολο των δεδομένων μας. Με τον όρο περιττό term αναφερόμαστε στην απαλοιφή των common words, που αναφέρονται σε προηγούμενη ενότητα, καθώς επίσης και των διπλότυπων terms που εντοπίζονταν στα αρχεία txt που δημιουργήθηκαν. Για την επίτευξη της παραπάνω διαδικασίας θεωρήσαμε πως θα ήταν ορθότερη η πραγματοποίηση της σε ένα ενιαίο αρχείο το οποίο θα δημιουργούνταν από την ενσωμάτωση των 25 διαφορετικών αρχείων σε 1 που θα περιείχε όλα τα terms. Η δημιουργία, όμως, ενός ενιαίου αρχείου θα έκανε αδύνατο τον υπολογισμό της συχνότητας εμφάνισης του κάθε term που απαιτείται στο βήμα της δεικτοδότησης των term.

Με τον όρο συχνότητα εμφάνισης (Term Frequency - TF) εννοούμε τη συχνότητα με την οποία εμφανίζεται ένα συγκεκριμένο term σε κάθε αρχείο από τα 25 που είχαμε δημιουργήσει. Τα 25 αυτά αρχεία ουσιαστικά αποτελούν τα terms τα οποία εξήχθησαν από τα tweets, που αντλήθηκαν για κάθε ένα hashtag. Επομένως υπολογίζοντας τη συχνότητα εμφάνισης των terms για κάθε αρχείο, μετράμε τη συχνότητα με την οποία εμφανίζεται κάθε term στο συγκεκριμένο hashtag που συναντάται.

Για να γίνει ο υπολογισμός της συχνότητας εμφάνισης των terms, τα 25 αρχεία απλού κειμένου (txt) που είχαν δημιουργηθεί στο προηγούμενο βήμα της Λημματοποίησης, μετατράπηκαν σε 25, αντίστοιχα, αρχεία xls. Τα terms αποτελούσαν τις γραμμές της στήλης Α του Excel, για κάθε αρχείο, ενώ στη στήλη Β γινόταν ο υπολογισμός της συχνότητας εμφάνισης με τη χρήση της συνάρτησης COUNTIF που παρέχεται από το Excel. Η συνάρτηση COUNTIF επιστρέφει το σύνολο των κελιών που επαληθεύουν μια συγκεκριμένη συνθήκη. Αρχικά ταξινομώντας την στήλη Α με αλφαβητική σειρά, είχαμε συγκεντρώσει τα διπλότυπα terms στις αρχικές γραμμές της στήλης Α. Έπειτα χρησιμοποιώντας κατάλληλα τη συνάρτηση COUTIF υπολογίζαμε πόσες φορές συναντάται το κάθε term ανάμεσα στο συνολικό

πλήθος των terms του ίδιου αρχείου. Σε περίπτωση που είχε πραγματοποιηθεί η ενοποίηση όλων των αρχείων txt σε ένα ενιαίο αρχείο, η συνάρτηση COUNTIF θα επέστρεφε ως αποτέλεσμα τη συχνότητα εμφάνισης ενός term ανάμεσα σε όλα τα terms που συναντώνται σε όλα τα hashtag, που αποτελεί λάθος αποτέλεσμα βάση του ορισμού της TF.

Η επαλήθευση των αποτελεσμάτων της συχνότητας εμφάνισης έγινε τόσο με βάση την παρατήρηση και τον έλεγχο που πραγματοποιήσαμε στα αποτελέσματα που εξήχθησαν, όσο και με βάση τη σύγκριση των αποτελεσμάτων που λάβαμε με τη χρησιμοποίηση του προγράμματος «Excel Word Frequency Count Software». Η εφαρμογή αυτή, όπως και το Parsing Excel που χρησιμοποιήθηκε στο tokenization, αποτελεί ένα έτοιμο λογισμικό που παρέχεται με τη μορφή βιβλιοθηκών για το Excel. Ο κώδικας του προγράμματος είναι γραμμένος στη γλώσσα προγραμματισμού JAVA και η λειτουργία του είναι αντίστοιχη της συνάρτησης COUNTIF. Δέχεται ένα αρχείο του Excel και με βάση τα περιεχόμενα σε κείμενο του αρχείου παράγει νέα αρχεία xls όπου στην στήλη Α εισάγει το term, ενώ στη στήλη Β το TF. Το πρόγραμμα Excel Word Frequency Count Software προσφέρει επίσης και περιβάλλον διεπαφής με το χρήστη που φαίνεται στην παρακάτω εικόνα (Εικόνα 13).



Εικόνα 13 Περιβάλλον της εφαρμογής Excel Word Frequency Count Software

Μετά την ολοκλήρωση του υπολογισμού της συχνότητας εμφάνισης των terms, ακολούθησε η απαλοιφή των common words. Η απαλοιφή αυτών, των περιττών για τα δεδομένα μας, λέξεων ολοκληρώθηκε εκμεταλλευόμενοι τη συνάρτηση FIND και του λογικού τελεστή OR που παρέχονται ως εργαλεία του Excel. Δημιουργώντας μια λίστα με το σύνολο των common words, ελέγχαμε κάθε term, ένα προς ένα, αν αντιστοιχούσε στις λέξεις αυτές που εμπεριέχονταν στην λίστα.

Ως κριτήρια της συνάρτησης FIND ορίσαμε το σύνολο των λέξεων που εμπεριέχονταν στην λίστα που προαναφέρθηκε. Για να παρέχεται η δυνατότητα ταυτόχρονου ελέγχου ενός term με όλο το σύνολο των common words, ανάμεσα στα κριτήρια γινόταν η χρήση του λογικού τελεστή OR. Ο έλεγχος πραγματοποιούνταν για όλο το σύνολο των term σε κάθε αρχείο xls ξεχωριστά.

Εφόσον ολοκληρώθηκε η διαδικασία της απαλοιφής των common words, σειρά είχε η απαλοιφή των διπλότυπων term που εντοπίζονταν στο ίδιο αρχείο xls. Εφόσον είχε γίνει ο υπολογισμός της συχνότητας εμφάνισης του κάθε term δεν υπήρχε πλέον λόγος για την διατήρηση των ίδιων ακριβώς term στο ίδιο αρχείο. Πριν την απαλοιφή των διπλότυπων term προηγήθηκε η ταξινόμηση της στήλης Α του κάθε αρχείου, που περιέχει ονομαστικά το κάθε term, κατά αλφαβητικό τρόπο.

Έπειτα, η απαλοιφή των διπλότυπων πραγματοποιήθηκε με την χρήση των συναρτήσεων FIND, που χρησιμοποιήθηκε και στην απαλοιφή των common words, καθώς επίσης και της συνάρτηση EXACT. Η συνάρτηση EXACT πραγματοποιούσε τον έλεγχο αν δύο συνεχόμενα term είναι όμοια μεταξύ τους, ελέγχοντας πάντα το προηγούμενο με το επόμενο. Δεν υπήρχε λόγος να γίνει έλεγχος σε μεγαλύτερο εύρος term από την στιγμή που τα δεδομένα μας ήταν αλφαβητικά ταξινομημένα. Η συνάρτηση FIND επέστρεφε τη θέση του διπλότυπου term, που έβρισκε η συνάρτηση EXACT, το οποίο έπειτα απαλείφονταν. Η διαδικασία αυτή ακολουθήθηκε στο σύνολο των 25 αρχείων που εμπεριέχονταν όλα τα terms για κάθε κατηγορία.

Με την ολοκλήρωση της διαδικασίας της απαλοιφής των common words και

των διπλότυπων term, ενοποιήσαμε όλα τα ξεχωριστά αρχεία xls σε ένα ενιαίο. Το σύνολο των terms, που ενσωματώθηκαν στο στάδιο αυτό, έφτανε σε αριθμό τα 15750.

5.2.1.3. Αντιστοίχηση terms σε κατηγορίες - mapping

Με την δημιουργία ενός ενιαίου αρχείου του Excel τα terms, εκτός από τη συχνότητα εμφάνισης, απέκτησαν και μια ακόμα στήλη (πεδίο), που αντιστοιχούσε στην κατηγορία του hashtag από το οποίο προήλθαν. Η προσθήκη της στήλης αυτής κρίνονταν απαραίτητη για τη δυνατότητα αντιστοίχισης του κάθε term στην κατηγορία του hashtag από το οποίο προήλθε, κυρίως όσο αφορά το στάδιο της δεικτοδότησης και της συσταδοποίησης.

Αποτέλεσμα της παραπάνω διαδικασίας, συμπεριλαμβανομένης και της απαλοιφής των common words, ήταν η δημιουργία ενός μεγάλου σε μέγεθος αρχείου που περιείχε το σύνολο των 15750 terms από όλες τις κατηγορίες. Στην στήλη B του xls υπήρχαν ονομαστικά τα terms που απέμειναν, στη στήλη A υπήρχε η τιμή της συχνότητας εμφάνισης του κάθε term ενώ στη στήλη C υπήρχε μία ετικέτα βάση της κατηγορίας στην οποία άνηκε. Για παράδειγμα αν κάποιο term προέκυψε από ένα tweet του hashtag Technology (C1), θα έχει την ετικέτα C1 στην στήλη C.

Ανάμεσα στο τεράστιο αυτό σύνολο terms εντοπίζονταν και terms με το ίδιο όνομα και αρκετές φορές με το ίδιο term frequency. Τα term αυτά διέφεραν μεταξύ τους στην στήλη C του xls καθώς καθένα από αυτά προερχόταν από διαφορετικό hashtag. Επομένως τα term που υπήρχαν στο συνολικό πλέον αρχείο ήταν το καθένα μοναδικό σε σχέση με τα υπόλοιπα και ως χαρακτηριστικά είχε τρία πεδία, το όνομα, το TF και την ετικέτα της κατηγορίας.

Στην παρακάτω εικόνα (Εικόνα 14) παρουσιάζεται ένα μέρος του συνόλου των δεδομένων του συνολικού αρχείου.

	A	B	C	D	E
1	Term Frequency	Terms	Category		
2	1	aaaa	C2		
3	1	aaaand	D3		
4	3	aac	A4		
5	1	aaron	A2		
6	1	abacus	C3		
7	1	abate	D5		
8	1	abbey	A3		
9	2	abcs	D3		
10	2	abdelaziz	A3		
11	1	abdulaziz	A3		
12	1	abdulrahman	A3		
13	1	abelen	B5		
14	1	aberdeen	A2		
15	1	ability	E3		
16	1	ability	C3		
17	1	abortion	B1		
18	1	aboutfun	D3		
19	1	above	A3		
20	1	above	B1		
21	1	abraham	A1		
22	1	abroad	E4		
23	1	absolute	C2		
24	1	absolutely	D3		
25	1	absolutely	E1		
26	1	absorbed	A3		
27	1	abstinence	E4		
28	1	abstract	A3		
29	1	abuse	E4		
30	2	abuse	B1		
31	1	abuse	D2		
32	1	academic	E4		
33	2	academy	C5		
34	1	acc	D1		
35	1	acceler	C2		
36	1	accelerator	E3		
37	2	accenture	E2		
38	1	accepted	C3		
39	2	access	A1		
40	3	access	E4		
41	1	access	D3		
42	2	access	E1		
43	1	access	E2		
44	1	accessing	E3		

Εικόνα 14 Αρχείο του Excel για αντιστοίχιση term με hashtag

5.2.1.4. Υπολογισμός idf και κατασκευή vectors

Προκειμένου να εφαρμόσουμε τις τεχνικές εξόρυξης δεδομένων στα term στα οποία καταλήξαμε θα πρέπει να αναπαραστήσουμε τα term αυτά σε μια μορφή που είναι επεξεργάσιμη. Η πιο ορθή και αποδοτική μέθοδος αναπαράστασης των

terms είναι η διανυσματική αναπαράσταση, όπου δημιουργείται ένας διανυσματικός χώρος στον οποίο το κάθε term αποτελεί ένα ξεχωριστό διάνυσμα. Ο λόγος για τον οποίο χρησιμοποιείται η διανυσματική αναπαράσταση ως μέθοδος αναπαράστασης των term είναι ότι με τον τρόπο αυτό μπορούμε να αναπαραστήσουμε το κάθε term ως ένα ξεχωριστό όρο που αποτελείται από ένα σύνολο χαρακτηριστικών πεδίων.

Οι δυο βασικοί τρόποι που χρησιμοποιούνται για την διανυσματική αναπαράσταση των κειμένων είναι:

- Boolean Αναπαράσταση
- Αναπαράσταση βασισμένη στη συχνότητα εμφάνισης των όρων στα δεδομένα.

Για την αναπαράσταση με βάση τη συχνότητα εμφάνισης των όρων θα χρησιμοποιήσουμε τη μέθοδο της δεικτοδότησης TF – IDF. Με βάση το μοντέλο τα term αναπαριστώνται ως διανύσματα σε ένα πολυδιάστατο χώρο. Κάθε διάσταση αντιπροσωπεύει έναν μοναδικό όρο της συλλογής κειμένων, ενώ η τιμή που αντιστοιχεί σε κάθε διάσταση είναι ένας πραγματικός αριθμός που προκύπτει ως συνάρτηση της συχνότητας εμφάνισης του όρου αυτού του term. Η χρήση του TF-IDF γίνεται στην πτυχιακής μας εργασίας στα πλαίσια της απόδοσης βάρους σε κάθε term της συλλογής μας.

Το μοντέλο TF – IDF αποτελείται από τις εξής ποσότητες:

- TF
- IDF

Η ποσότητα TF είναι ουσιαστικά η συχνότητα εμφάνισης ενός term σε κάθε κατηγορία hashtag και έχει υπολογιστεί στην ενότητα 5.1.2.1.2. Αποτελεί τη στήλη Α του αρχείου xls που εμπεριέχει το συνολικό αριθμό των terms. Αντίστοιχα η ποσότητα IDF αποτελεί ένας βάρος που δηλώνει τη σημαντικότητα ενός term σε σχέση με ολόκληρη τη συλλογή όλων των term από τις διάφορες κατηγορίες των

hashtag. Το τελικό βάρος TF-IDF που αποδίδεται στον όρο προκύπτει από τον πολλαπλασιασμό των TF και IDF. Για περισσότερες πληροφορίες για το μοντέλο αυτό και τη διανυσματική αναπαράσταση μπορείτε να ανατρέξετε στην ενότητα 3.3.4

Το βάρος IDF υπολογίζεται βάση της Σχέσης 4 που βρίσκεται στο κεφάλαιο 3.3.4.2 και είναι η παρακάτω:

$$\boldsymbol{idf_i = log \frac{N}{n_i}}$$

όπου, στη προκειμένη περίπτωση, N είναι το σύνολο των διαφορετικών κατηγοριών hashtag, βάση των οποίων συλλέξαμε τα δεδομένα, και n_i είναι ο συνολικός αριθμός των φορών που εντοπίστηκε ένα συγκεκριμένο term στο σύνολο των terms των διαφορετικών hashtag.

Στο αρχείο xls που είχε δημιουργηθεί στο προηγούμενο βήμα, της αντιστοίχισης των terms σε κατηγορίες, η στήλη A αποτελούσε τη συχνότητα εμφάνισης (TF) του κάθε term, η στήλη B την ονομασία του term και η στήλη C την κατηγορία στην οποία άνηκε. Για τον υπολογισμό του IDF κάθε όρου προστέθηκε μια στήλη D στην οποία εμφανιζόταν, ως αριθμός, πόσες φορές συναντάται το συγκεκριμένο term στο σύνολο των terms των διαφορετικών hashtag. Ο υπολογισμός αυτός πραγματοποιήθηκε με τη χρήση της συνάρτησης COUNTIF του EXCEL. Η συνάρτηση COUTIF επιστρέφει το σύνολο των κελιών που ικανοποιούν ένα συγκεκριμένο κριτήριο. Το κριτήριο που τέθηκε ήταν η σύγκριση του κάθε term για το αν είναι όμοιο με κάποιο από το σύνολο των υπολοίπων. Με τον τρόπο αυτό εμφανιζόταν στη στήλη D το πόσες φορές συναντιόνταν το κάθε term συνολικά.

Ο υπολογισμός του IDF, εφόσον δημιουργήθηκε η στήλη D, ήταν ένα απλός τύπος του EXCEL της μορφής « = log (D(διπλανό) / 25)», που τοποθετούνταν στη στήλη E. Ο υπολογισμός έγινε βάση της Σχέσης που αναφέρεται παραπάνω. Έπειτα η συμπλήρωση του τύπου προς τα κάτω έκανε τον υπολογισμό του idf για το σύνολο των terms. Αποτέλεσμα της παραπάνω διαδικασίας ήταν η δημιουργία ενός αρχείου xls, με 2 νέες στήλες, που μέρος αυτού παρουσιάζεται στην παρακάτω

εικόνα (Εικόνα 15).

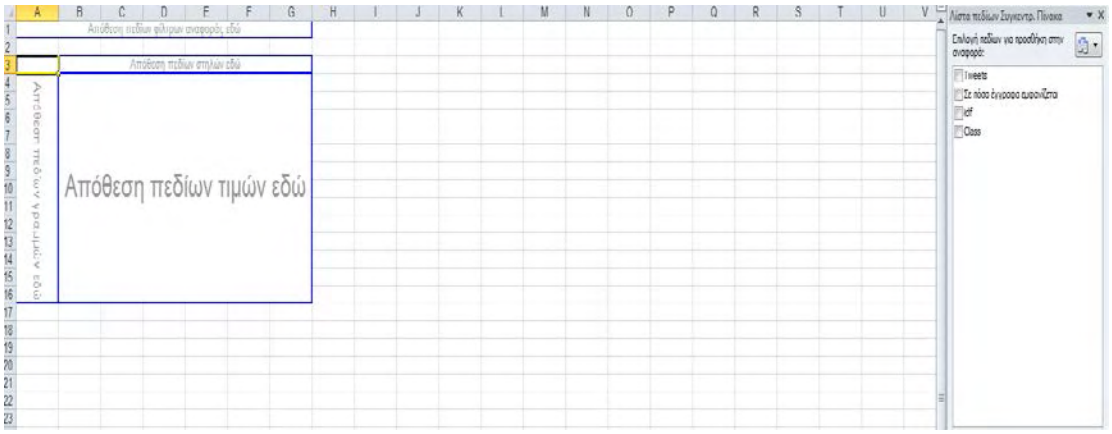
	A	B	C	D	E
4	3	aac	A4	1	1,39794001
5	1	aaron	A2	1	1,39794001
6	1	ability	E3	2	1,09691001
7	1	ability	C3	2	1,09691001
8	1	abortion	B1	1	1,39794001
9	1	aboutfun	D3	1	1,39794001
10	1	above	A3	2	1,09691001
11	1	above	B1	2	1,09691001
12	1	abraham	A1	1	1,39794001
13	1	abroad	E4	1	1,39794001
14	1	absolute	C2	1	1,39794001
15	1	absolutely	D3	2	1,09691001
16	1	absolutely	E1	2	1,09691001
17	1	absorbed	A3	1	1,39794001
18	1	abstinence	E4	1	1,39794001
19	1	abstract	A3	1	1,39794001
20	1	abuse	E4	3	0,92081875
21	2	abuse	B1	3	0,92081875
22	1	abuse	D2	3	0,92081875
23	1	academic	E4	1	1,39794001
24	2	academy	C5	1	1,39794001
25	1	acc	D1	1	1,39794001
26	1	acceler	C2	1	1,39794001
27	1	accelerator	E3	1	1,39794001
28	2	accenture	E2	1	1,39794001
29	1	accepted	C3	1	1,39794001
30	2	access	A1	5	0,69897
31	3	access	E4	5	0,69897
32	1	access	D3	5	0,69897
33	2	access	E1	5	0,69897
34	1	access	E2	5	0,69897
35	1	accessing	E3	1	1,39794001
36	4	accident	D4	1	1,39794001
37	1	accompaniment	E2	1	1,39794001

Εικόνα 15 Αρχείο του Excel με υπολογισμένο το idf

Τελικό βήμα της επεξεργασίας των δεδομένων, προτού ακολουθήσει η συσταδοποίησης τους, αποτελεί η αναπαράσταση τους με τη μορφή διανυσμάτων. Η διανυσματική μορφή αναπαριστά κάθε term ως ένα σύνολο διαφορετικών πεδίων, όπου στο καθένα από αυτά το κάθε term έχει και ένα βάρος. Στην δικιά μας πτυχιακή εργασία η διανυσματική αναπαράσταση των terms έγινε με τον εξής τρόπο. Ως πεδία για κάθε term αποτέλεσαν το σύνολο των διαφορετικών κατηγοριών hashtag. Αντίθετα ως βάρος τοποθετήθηκε η τιμή TF – IDF, που είχαμε υπολογίσει στο τελευταίο αρχείο xls, που εμφανίζει κάθε term για την συγκεκριμένη κατηγορία που ανήκει. Για παράδειγμα το term «accident» έχει τιμή βάρους TF – IDF 1,39794001 για την κατηγορία D4, ενώ το accelerator έχει τιμή βάρους TF – IDF 1,39794001 για την κατηγορία E3.

Το τελικό πλέον, έτοιμο προς χρήση, αρχείο που δημιουργήθηκε αποτελούνταν από 26 στήλες συνολικά και 2500 γραμμές. Στη στήλη Α αναφέρονταν ονομαστικά το κάθε term. Οι επόμενες 25 στήλες δεξιότερα αποτελούσαν τις διαφορετικές κατηγορίες hashtag για τις οποίες αντλήσαμε τα αρχικά tweets. Για κάθε μια από της κατηγορίες αυτές που το αντίστοιχο term εμφάνιζε τιμή βάρους TF – IDF, καταχωρούνταν στην αντίστοιχη στήλη, σε αντίθετη περίπτωση γραφόταν η τιμή 0 για τις κατηγορίες στις οποίες δεν εμφανιζόταν το term.

Η χρήση του εργαλείου «Συγκεντρωτικός Πίνακας» που προσφέρει το EXCEL πραγματοποίησε την παραπάνω διαδικασία. Η εντολή αυτή βρίσκεται στο μενού «Εισαγωγή» της γραμμής μενού για το EXCEL στο Office 2010. Αρχικά επιλέγεις το σύνολο των δεδομένων για τα οποία θα δημιουργήσεις τον συγκεντρωτικό πίνακα. Εμείς επιλέξαμε ως σύνολο δεδομένων τις στήλες Β , C και Ε του παραπάνω αρχείου, που φαίνεται και στην Εικόνα 15, και αποτελούν αντίστοιχα το term, την κατηγορία που ανήκει το term και το βάρος TF – IDF που έχει για την κατηγορία αυτή. Αμέσως μετά εμφανίζεται ένα νέο φύλλο εργασίας του EXCEL που έχει την παρακάτω μορφή (Εικόνα 16).



Εικόνα 16 Δημιουργία Συγκεντρωτικού Πίνακα

Με απλή διαδικασία συρσίματος (drag and drop) επιλέγεις δεξία, που εμφανίζονται τα πεδία, ποιο πεδίο θέλεις να αποτελεί το πάνω μέρος του πίνακα, ποιο το αριστερό μέρος του πίνακα και τέλος με βάση ποιο πεδίο θα δίδονται οι τιμές. Εμείς επιλέξαμε ως πάνω μέρος του πίνακα τις κατηγορίες hashtag, αποτέλεσμα να δημιουργηθούν 25 στήλες απευθείας, ως αριστερό τμήμα το όνομα του term που αποτελούσε την πρώτη στήλη του πίνακα που δημιουργήθηκε, και ως απόθεση πεδίου τιμών το βάρος TF – IDF που είχαμε υπολογίσει. Η επαλήθευση και ο έλεγχος των αποτελεσμάτων του πίνακα αποτέλεσε μια ιδιαίτερα επίπονη διαδικασία από μέρους μας, όμως βάση του αποτελέσματος άξιζε τον κόπο. Ο πίνακας που δημιουργήθηκε από την παραπάνω διαδικασία απεικονίζεται παραδειγματικά παρακάτω (Εικόνα 17).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1945	sisters	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1,09691	1,09691	0	0	0	0	0
1946	sit	0	0	0	0	0	0	0	1,09691	0	0	0	0	0	1,09691	0	0	0	0	0	0	0	0	0	0
1947	site	0	0	0	0	0,49485	0,49485	0	0	0	0	0	0	0,49485	0,49485	0	0	0	0	0,49485	0	0,49485	0,49485	0,49485	0
1948	sites	1,09691	0	0	0	0	0	0	0	0	0	1,09691	0	0	0	0	0	0	0	0	0	0	0	0	0
1949	sitting	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1,09691	1,09691	0	0	0	0	0
1950	six	0	0	0	0	0	0,79588	0	0	0	0	0,79588	0	0	0	0	0	0	0	0	0	0,79588	0	0	0,79588
1951	size	0	0	0	0	0	0	0	0,920819	0	0	0	0	0,920819	0	0	0	0	0	0	0,920819	0	0	0	0
1952	skeleton	0	0	1,09691	0	0	0	0	0	0	0	0	0	0	0	0	0	1,09691	0	0	0	0	0	0	0
1953	skills	0	0	0	0	0	0	0	0	0	1,09691	0	0	1,09691	0	0	0	0	0	0	0	0	0	0	0
1954	skin	0	0	0	0	0,920819	0	0	0	0	0	0	0	0	0	0,920819	0	0	0	0,920819	0	0	0	0	0
1955	sky	0,69897	0,69897	0,69897	0	0	0	0	0	0	0	0	0,69897	0	0	0	0	0,69897	0	0	0	0	0	0	0
1956	skyline	1,09691	0	0	0	0	0	0	0	0	0	1,09691	0	0	0	0	0	0	0	0	0	0	0	0	0
1957	sleep	0	0	0,443697	0	0	0	0	0	0,443697	0,443697	0,443697	0,443697	0	0,443697	0,887395	0	0	0,443697	0	0	0	0	0	0
1958	sleeping	0	0	0	0	0	0	0	0	0	0	0,920819	0	0,920819	0,920819	0	0	0	0	0	0	0	0	0	0
1959	sm	0	0,69897	0	0	0	0,69897	0	0	0	0	0	0	0	0	0	0	0	0	0,69897	0	0,69897	0	0	0	0,
1960	small	0,619789	0,619789	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,619789	0	0	0	0	0,619789	0,619789	0	0,6
1961	smart	0	0	0,552842	0	0	0,552842	0	0,552842	0	0	0,552842	0	0,552842	0	0	0	0	0	0	0	0,552842	0	0	0,552842	
1962	smartphon	0	0	0	0	0	0,69897	0	0	0	0	0,69897	0	0	0	0	0	0	0	0	0	0,69897	0	0,69897	0,69897	
1963	smile	0	0	0	0,79588	0	0	0	0,79588	0	0	0	0	0	0	0	0	0	0,79588	0,79588	0	0	0	0	0	
1964	smith	0	0	0	0	0	0	1,09691	0	1,09691	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1965	smoke	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1,09691	0	0	0	0	1,09691	0	0	0	0	
1966	snow	0	0	0	0	0	0	0,920819	0	0	0	0	0	0	0	0	0	0,920819	0	0	0	0	0	0	0	0,9
1967	soccer	0	0	0	0	0	0	0,79588	0	0,79588	0	0	0	0,79588	0	0,79588	0	0	0	0	0	0	0	0	0	
1968	social	0	0	0	0	0	0,552842	0	0	0	0,552842	0,552842	0	0	0	0	0,552842	0	0	0	0	0,552842	0	0	0,552842	0,5
1969	socialmedi	0	0	0	0	0	0,69897	0	0	0	0	0,69897	0	0	0	0	0,69897	0	0	0	0	0,69897	0	0	0	0,
1970	society	0	0	0,920819	0	0	0	0	0	0	0	0,920819	0	0	0	0	0	0	0	0	0	0	0	0	0,920819	
1971	socks	0	0	0	0	0	1,09691	0	0	0	0	0	0	0	0	0	0	0	0	0	1,09691	0	0	0	0	
1972	software	0	0	0	0	0	0	0	0	0	0	0,69897	0	0	0	0	0	0	0	0	0	0,69897	0,69897	0,69897	0	0,
1973	solar	0	0	0	0	0	0	0	0	0	0	1,09691	0	0	0	0	0	0	0	0	0	0	0	1,09691	0	
1974	sold	0,920819	0	0	0	0	0	0	0	0	0	0,920819	0	0,920819	0	0	0	0	0	0	0	0	0	0	0	
1975	solid	0	0	0,920819	0	0,920819	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,9
1976	solo	0	0	0	1,09691	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1,09691	
1977	solution	0	0	0	0	0	0	0	0	0	0	0	0	1,09691	0	0	0	0	0	0	0	0	0	0	0	1,
1978	solutions	0	0	0	0	0	0	0	0	0	0	0	0	0,69897	0	0	0	0	0	0	0	0,69897	0,69897	0	0,69897	0,
1979	solve	0	0	0	0	0	1,09691	0	0	0	0	0	1,09691	0	0	0	0	0	0	0	0	0	0	0	0	
1980	somebody	0	0	0	0,920819	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,920819	0	0	0	0	0,9
1981	someone	0	0	0	0	0	0	0	0	0	0	0	0	0,920819	0,920819	0	0	0	0	0	0	0	0	0	0,920819	
1982	son	0	0	0	1,09691	0	1,09691	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1983	song	0	0	0	0,79588	0	0	0	0	0,79588	0	0	0	0	0,79588	0	0	0	0,79588	0	0	0	0	0	0	
1984	songs	0	0	0	0	0	0	0	0	0	0	0	0	0,920819	0	0	0	0	0,920819	0	0,920819	0	0	0	0	
1985	soo	0	0,920819	0	0	0	0	0	0	0	0,920819	0	0	0	0	0	0	0	0,920819	0	0	0	0	0	0	
1986	soon	0	0	0	0	0	0	0	0	0,79588	0,79588	0	0	0	0	0,79588	0	0	0	0	0	0,79588	0	0	0	
1987	sore	0	0	0	0	0	0	0,920819	0,920819	0	0,920819	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1988	soul	0	0	0	0,920819	0	0	0	0	0,920819	0	0	0,920819	0	0	0	0	0	0	0	0	0	0	0	0	

Εικόνα 17 Μέρος του Συγκεντρωτικού Πίνακα Διανυσμάτων

Με τη δημιουργία του Συγκεντρωτικού Πίνακα ολοκληρώθηκε η διαδικασία της επεξεργασίας των δεδομένων, με την αναπαράσταση των δεδομένων μας με τη μορφή διανυσμάτων. Στην επόμενη ενότητα ακολουθεί το προτελευταίο βήμα της πτυχιακής μας εργασίας, η συσταδοποίηση των διανυσμάτων αυτών και ο χωρισμός τους σε κλάσεις.

5.3. Μετρήσεις με το πρόγραμμα WEKA

Εφόσον έχει ολοκληρωθεί η επεξεργασία και ο μετασχηματισμός των term σε κατάλληλα διανύσματα, σειρά έχει η διαδικασία της επιλογής των αλγορίθμων εξόρυξης δεδομένων. Οι δυο βασικότερες τεχνικές εξόρυξης γνώσης από τα δεδομένα είναι η ταξινόμηση και η συσταδοποίηση. Στην παρούσα πτυχιακή εργασία θα γίνει χρήση μόνο αλγορίθμων συσταδοποίησης, οι οποίο θα κατηγοριοποιούν τα terms σε ομάδες.

Η συσταδοποίηση μπορεί να ορισθεί, εντελώς περιγραφικά, ως μια διαδικασία ομαδοποίησης όμοιων μεταξύ τους όρων, ανάμεσα σε ένα σύνολο δεδομένων, σε κοινές ομάδες που ονομάζονται κλάσεις. Περισσότερες πληροφορίες για τον ορισμό, καθώς και τη διαδικασία που ακολουθείτε προκειμένου να ολοκληρωθεί η συσταδοποίηση, μπορείτε να ανατρέξετε στο κεφάλαιο 4 της πτυχιακής εργασίας, όπου αναλύεται διεξοδικά. Η διαδικασία της συσταδοποίησης στην προκειμένη πτυχιακή εργασία θα γίνει με τη χρήση μιας εφαρμογής με το όνομα WEKA.

Το WEKA είναι ένα λογισμικό για εξόρυξη δεδομένων που αναπτύχθηκε στη γλώσσα προγραμματισμού JAVA και αποτελείται από μια συλλογή αλγορίθμων μηχανικής μάθησης. Παρέχει δυνατότητα για:

- Προεπεξεργασία δεδομένων με τη χρήση κατάλληλων εργαλείων που ονομάζονται φίλτρα (filters)
- Δημιουργία «μοντέλων» από τα δεδομένα με κάποια διαδικασία εκπαίδευσης
- Συσταδοποίηση δεδομένων
- Ταξινόμηση των δεδομένων
- Την εύρεση κανόνων συσχέτισης
- Χρησιμοποίηση στατιστικών μεγεθών για την αξιολόγηση των διάφορων αλγορίθμων μάθησης
- Απεικόνιση τόσο των αρχικών δεδομένων όσο και των

αποτελεσμάτων μετά της διαδικασία της εκπαίδευσης.

Η προεπεξεργασία των δεδομένων πραγματοποιήθηκε από μέρους μας, με αποτέλεσμα να μην απαιτείται η χρήση φίλτρων του WEKA κατά τη διαδικασία της συσταδοποίησης.

Τα βασικά αρχεία τα οποία δέχεται σαν είσοδο το WEKA έχουν κατάληξη ARFF (Attribute Relation File Format) και πρόκειται για ένα αρχείο κειμένου χαρακτήρων ASCII. Πρωταρχικό μέλημα μας επομένως ήταν η μετατροπή του EXCEL αρχείου που διαθέταμε, με το σύνολο των διανυσμάτων, σε αρχείο ARFF προκειμένου να είναι συμβατό με την εφαρμογή του WEKA. Η μετατροπή αυτή από xls αρχείο σε arff έγινε με τη βοήθεια μιας εφαρμογής με το όνομα «Excel to Arff Converter». Η εφαρμογή αυτή έπαιρνε ως είσοδο ένα αρχείο xls, το οποίο αρχικά μετέτρεπε σε csv και έπειτα σε αρχείο arff.

Τα αρχεία με κατάληξη arff έχουν συγκεκριμένη δομή που αποτελείται από δυο κύρια συστατικά, τα χαρακτηριστικά (attributes) και τα δεδομένα (data). Η πρώτη γραμμή ενός αρχείο arff αποτελεί το όνομα του αρχείου και εμφανίζει την εξής μορφή, @relation (όνομα αρχείου). Ακριβώς στην επόμενη γραμμή ξεκινάει η δήλωση των χαρακτηριστικών πεδίων (attributes) που περιγράφουν το σύνολο των δεδομένων που ακολουθεί. Η δήλωση γίνεται χρησιμοποιώντας την σύμβαση: **@attribute <attribute-name> <datatype>**, όπου **<attribute-name>** το όνομα του χαρακτηριστικού. Το όρισμα **<datatype>** καθορίζει τον τύπο του χαρακτηριστικού. Το WEKA υποστηρίζει τέσσερες διαφορετικούς τύπους:

- Αριθμητικά δεδομένα (numeric)
- Δεδομένα που ορίζουν κατηγορία (ονομαστικά) (nominal)
- Αλφαριθμητικά (string)
- Ημερομηνίες με συγκεκριμένο format (date)

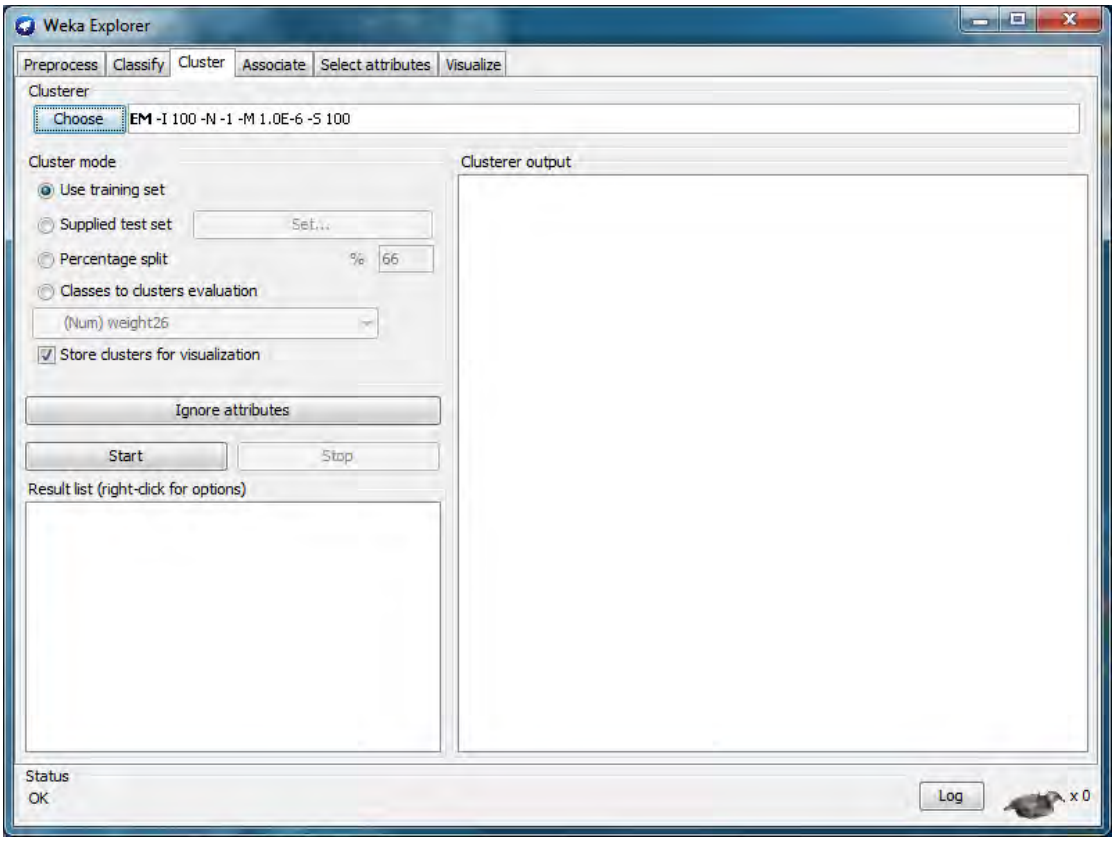
Στην δικιά μας περίπτωση τα δεδομένα αποτελούσαν τα terms που επεξεργαστήκαμε κα ως attributes θεωρήσαμε τα βάρη (μηδενικά η μη), που εμφανίζει κάθε term στην αντίστοιχη από τις 25 κατηγορίες hashtag, καθώς και ένα μοναδικό id που δώσαμε κάθε term, ως αντιστοιχία. Ως τύπος χαρακτηριστικού ορίστηκε για όλα τα δεδομένα το numeric, όπως φαίνεται και στην παρακάτω εικόνα (Εικόνα 18).

```
@RELATION twitter_visualise

@ATTRIBUTE termID NUMERIC
@ATTRIBUTE weight1 NUMERIC
@ATTRIBUTE weight2 NUMERIC
@ATTRIBUTE weight3 NUMERIC
@ATTRIBUTE weight4 NUMERIC
@ATTRIBUTE weight5 NUMERIC
@ATTRIBUTE weight6 NUMERIC
@ATTRIBUTE weight7 NUMERIC
@ATTRIBUTE weight8 NUMERIC
@ATTRIBUTE weight9 NUMERIC
@ATTRIBUTE weight10 NUMERIC
@ATTRIBUTE weight11 NUMERIC
@ATTRIBUTE weight12 NUMERIC
@ATTRIBUTE weight13 NUMERIC
@ATTRIBUTE weight14 NUMERIC
@ATTRIBUTE weight15 NUMERIC
@ATTRIBUTE weight16 NUMERIC
@ATTRIBUTE weight17 NUMERIC
@ATTRIBUTE weight18 NUMERIC
@ATTRIBUTE weight19 NUMERIC
@ATTRIBUTE weight20 NUMERIC
@ATTRIBUTE weight21 NUMERIC
@ATTRIBUTE weight22 NUMERIC
@ATTRIBUTE weight23 NUMERIC
@ATTRIBUTE weight24 NUMERIC
@ATTRIBUTE weight25 NUMERIC


@DATA
1,0,0,0,0,0,0,0,0,0,0,1.096910013,0,0,0,0,0,0,0,0,1.096910013,0,0,0,0,0.5,0,0.5
2,0,0,1.096910013,0,0,1.096910013,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.5,0.5,0,0,0
3,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.096910013,0,0,1.096910013,0,0,0,0,0,0,0.5,0.5
4,0,0,0,0,0,0.920818754,0,0,0,0,0,0,0,0,0,0,0,0.920818754,0,0,0,0,0,0,0,0.920818754,0,0,0.333333333,0,0.333333333,0.333333333
5,0.698970004,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.698970004,0,0,0,0,0.698970004,0.698970004,0,0.698970004,0,0,0.2,0,0,0.2,0.6
6,0,0,0.920818754,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.920818754,0,0,0.920818754,0,0.333333333,0,0,0,0.666666667
7,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.096910013,1.096910013,0,0,0,0,0,0,0,1
8,0,0,0,0,0,0,0,0,0,0,0,0,0,0.920818754,0,0,0,0,0,0,0,0.920818754,0,0,0,0,0.920818754,0,0,0.333333333,0.333333333,0.333333333
9,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.096910013,0,0,0,0,1.096910013,0,0,0,0,0,0,0,0.5,0.5
10,0,0,0,1.096910013,1.096910013,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0
11,1.096910013,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.096910013,0,0,0,0,0,0,0,0,0,0.5,0,0,0.5,0
12,0,0,0,0,0,0.920818754,0.920818754,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.920818754,0,0,0,0,0.333333333,0.333333333,0,0,0.333333333
13,0,0,0.795880017,0,0,0,0,0,0,0,0,0,0.795880017,0,0,0,0,0,0,0.795880017,0,0,0,0,0,0,0,0.795880017,0,0,0.25,0.25,0,0.25,0.25
14,0,0,0.920818754,0,0,0,0,0,0,0,0,0,0,0.920818754,0,0,0,0,0,0,0,0.920818754,0,0,0,0,0.333333333,0,0,0.333333333,0,0,0.333333333
15,0,0,0,0,0,0,0,0,0,0.920818754,0,0,0,0,0,0,0.920818754,0,0,0,0,0,0,0,0,0,0,0,0,0.920818754,0,0,0.333333333,0.333333333,0,0.333333333
16,0,0,0,0,0.698970004,0.698970004,0,0,0,0,0,0,0,0.698970004,0.698970004,0,0,0,0,0,0,0,0.698970004,0,0,0,0,0,0,0,0,0.4,0,0.4,0.2,0
17,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.096910013,0,1.096910013,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0
18,0,0,0.619788758,0,0.619788758,0,0,0,0,0,0,0,0.619788758,0,0,0.619788758,0,0.619788758,0.619788758,0,0,0,0,0,0.619788758,0,0,0,0,0,0,0,0,0.333333333,0.166666667,0.333333333
```

Εικόνα 18 Αναπαράσταση του arff αρχείου

Στη συνέχεια η χρήση του @data υποδηλώνει ότι ακολουθούν τα δεδομένα. Κάθε ένα term αντιπροσωπεύεται από μια γραμμή. Οι τιμές των χαρακτηριστικών για κάθε term διαχωρίζονται μεταξύ τους με κόμμα. Θα πρέπει οι τιμές των δεδομένων να εμφανίζονται κατά σειρά, όπως έχουν δηλωθεί με την αντίστοιχη σειρά τα attributes στην επικεφαλίδα του arff αρχείου. Στην περίπτωση που κάποιες τιμές λείπουν, τότε στην θέση τους αναγράφεται ένα λατινικό ερωτηματικό, γεγονός που ευτυχώς δεν εντοπίσαμε στο σύνολο των δεδομένων μας. Εφόσον είχαμε το αρχείο arff που αποτελεί την είσοδο για την εφαρμογή του WEKA υπολείπονταν η διαδικασία της συσταδοποίησης των δεδομένων που εντοπίζονται στο αρχείο.

Το WEKA δίνει τη δυνατότητα για την εφαρμογή μιας σειρά αλγορίθμων συσταδοποίησης χρησιμοποιώντας της καρτέλα Cluster, όπως απεικονίζεται στην παρακάτω εικόνα (Εικόνα 19)



Εικόνα 19 Καρτέλα Cluster του WEKA

Στο πεδίο «Clusterer» ο χρήστης επιλέγει τον αλγόριθμο που επιθυμεί για χρησιμοποιήσει για την πραγματοποίηση της διαδικασίας της συσταδοποίησης. Οι αλγόριθμοι που προσφέρονται πατώντας το κουμπί «Choose» είναι οι εξής:

- COBWEN
- DBSCAN
- FarthestFirst
- FilteredClusterer
- HierarchicalClusterer
- MakeDestinyBasedClusterer

- OPTICS
- sIB
- SimpleKMeans
- XMeans
- EM

Στο πεδίο «Cluster mode» η επιλογή «Percentage split» παρέχει τη δυνατότητα να επιλέξεις το ποσοστό των δεδομένων που θα αποτελέσουν το σύνολο εκπαίδευσης του αλγορίθμου (training set) και τα υπόλοιπα κατηγοριοποιούνται βάση της εκπαίδευσης που έλαβε ο αλγόριθμος. Οι τιμές που δόθηκαν ως ποσοστό training set από εμάς ήταν 25%, 30%, 66%, 75%.

Πραγματοποιώντας της παραπάνω επιλογές για κάθε διαφορετικό αλγόριθμο, το πάτημα του κουμπιού «Start» της εφαρμογής του WEKA, εκτελούσε τον αλγόριθμο και παρουσίαζε τα αποτελέσματα του στο δεξί παράθυρο με το όνομα «Clusterer output». Κάνωντας δεξι κλικ πάνω στο μοντέλο που δημιουργείτε στο πεδίο «Result list» παρέχεται η δυνατότητα αποθήκευσης των αποτελεσμάτων του παραμετροποιημένου αλγορίθμου (Save Result Buffer) που χρησιμοποίησε το συγκεκριμένο μοντέλο, καθώς επίσης η δυνατότητα εποπτικής παρουσίασης – απεικόνισης των αποτελεσμάτων (Visualize cluster assignments).

Ο συνδυασμός μεταξύ των διαφορετικών αλγορίθμων και της επιλογής διαφορετικού ποσοστού κάθε φόρα συνόλου εκπαίδευσης για τον κάθε αλγόριθμο οδήγησαν σε ένα σύνολο αποτελεσμάτων που εμφανίζεται στον παρακάτω πίνακα.

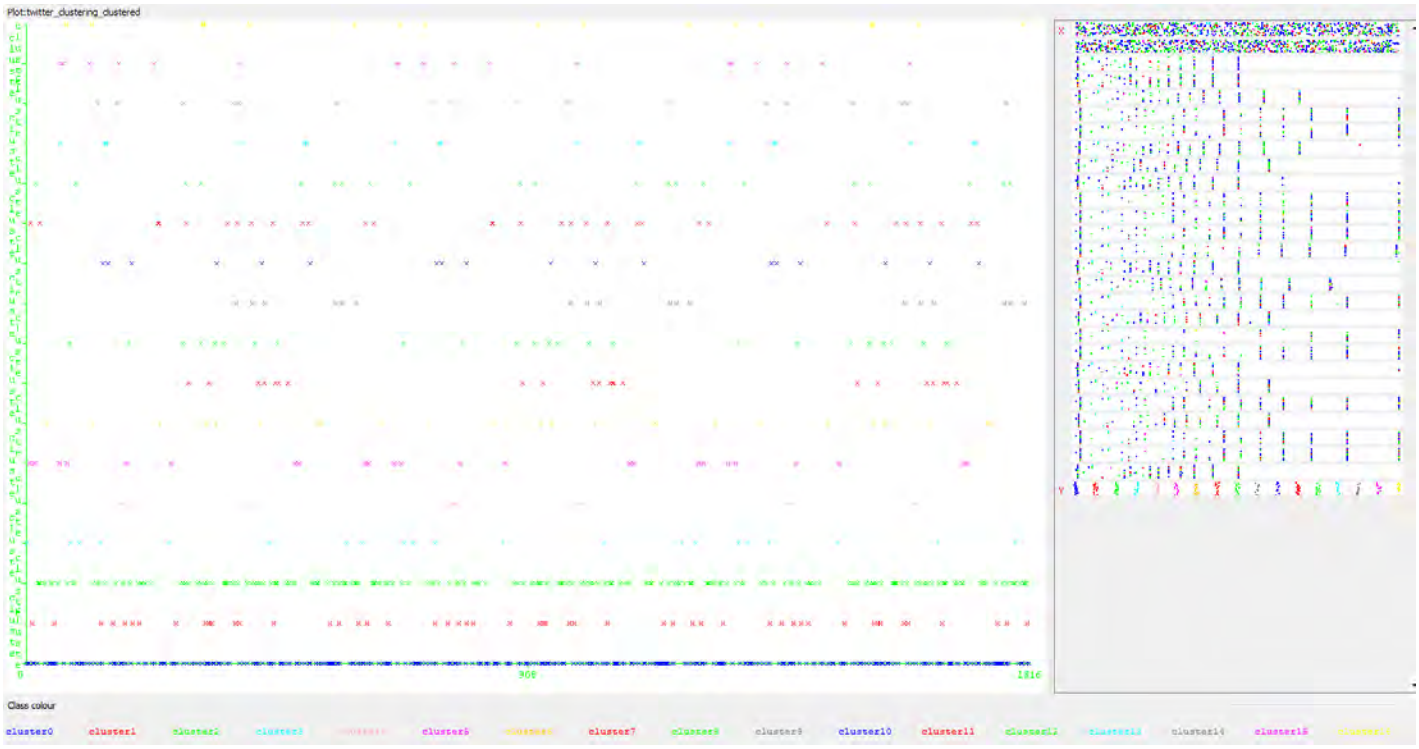
ΑΛΓΟΡΙΘΜΟΣ	% TEST SET	% TRAINING SET	ΑΡΙΘΜΟΣ ΚΛΑΣΕΩΝ	ΑΡΙΘΜΟΣ ΚΑΤΗΓΟΡΙΟΠΟΙΗΜΕΝΩΝ TERMS
COBWEN	66	33	1	1623
COBWEN	33	66	1	824
COBWEN	25	75	1	606
COBWEN	75	25	1	1817
DBSCAN	66	33	4	1250

DBSCAN	33	66	1	823
DBSCAN	25	75	1	606
DBSCAN	75	25	17	1217
FarthestFirst	66	33	2	1623
FarthestFirst	33	66	2	823
FarthestFirst	25	75	2	606
FarthestFirst	75	25	2	1817
FilteredClusterer	66	33	2	1623
FilteredClusterer	33	66	2	823
FilteredClusterer	25	75	2	606
FilteredClusterer	75	25	2	1817
HierarchicalClusterer	66	33	2	1623
HierarchicalClusterer	33	66	2	823
HierarchicalClusterer	25	75	-	0
HierarchicalClusterer	75	25	2	1817
MakeDestinyBasedClusterer	66	33	2	1623
MakeDestinyBasedClusterer	33	66	2	823
MakeDestinyBasedClusterer	25	75	2	606
MakeDestinyBasedClusterer	75	25	2	1817
OPTICS	66	33	1	823
OPTICS	33	66	1	1623
OPTICS	25	75	1	1817
OPTICS	75	25	1	606
sIB	66	33	2	1623
sIB	33	66	2	823
sIB	25	75	2	606
sIB	75	25	2	1817
SimpleKMeans	66	33	2	1623
SimpleKMeans	33	66	2	823
SimpleKMeans	25	75	2	606

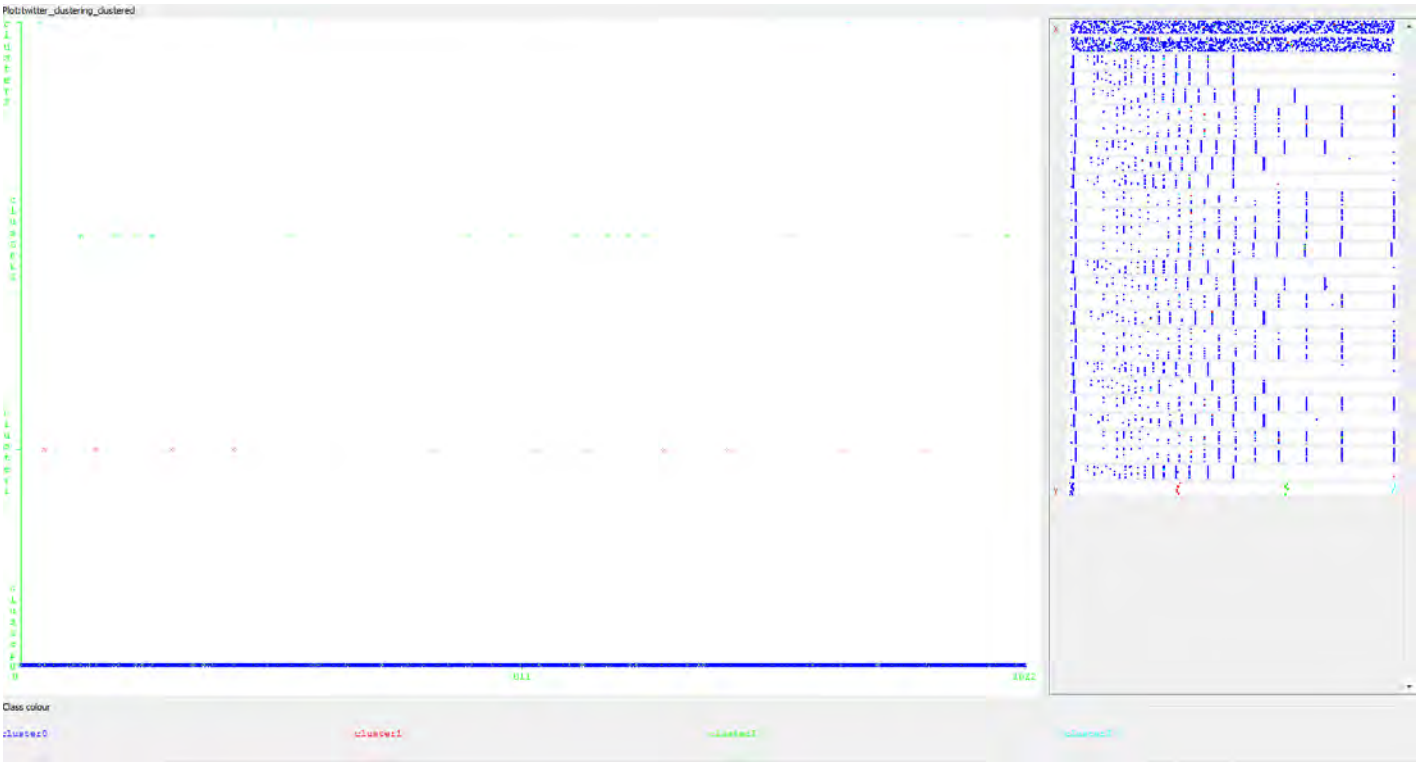
SimpleKMeans	75	25	2	1817
XMeans	66	33	2	1623
XMeans	33	66	3	823
XMeans	25	75	3	606
XMeans	75	25	2	1817
EM	66	33	8	1623
EM	33	66	8	823
EM	25	75	4	606
EM	75	25	8	1817

Εξετάζοντας τα αποτελέσματα που προέκυψαν από την εκτέλεση των αλγορίθμων συσταδοποίησης αντιλαμβανόμαστε πως οι αλγόριθμοι που επιλέχθηκαν για την εξαγωγή συμπερασμάτων βάση των 5 μεγάλων κατηγοριών (A,B,C,D,E) hashtag, που φαίνονται στην ενότητα 5.1, ήταν ο αλγόριθμος EM, καθώς και ο αλγόριθμος DBSCAN με ποσοστό training set 33%. Αντίστοιχα ο επιλεγμένος για την εξαγωγή συμπερασμάτων βάση των 25 υποκατηγοριών hashtag, της ενότητας 5.1, ήταν ο αλγόριθμος DBSCAN με ποσοστό training set 25%.

Στο σημείο αυτό να αναφέρουμε πως για τους παραπάνω αλγορίθμους πραγματοποιήθηκε αποθήκευση των αποτελεσμάτων τους μέσω του WEKA (Save Result Buffer). Το αποθηκευμένο αρχείο περιείχε το σύνολο των terms που άνηκε σε κάθε μια από τις κλάσεις που δημιούργησε ο εκάστοτε αλγόριθμος, με αποτέλεσμα να γνωρίζουμε κάθε ένα από το σύνολο των terms μας σε ποια κλάση κατηγοριοποιήθηκε. Επίσης πραγματοποιήθηκε απεικονιστική αναπαράσταση της κατηγοριοποίησης των δεδομένων αυτών που παρουσιάζεται στις παρακάτω εικόνες. Στις εικόνες ο οριζόντιος άξονας εμφανίζει το TermID του κάθε term, ενώ ο κατακόρυφος την κλάση του WEKA στην οποία ανήκει.

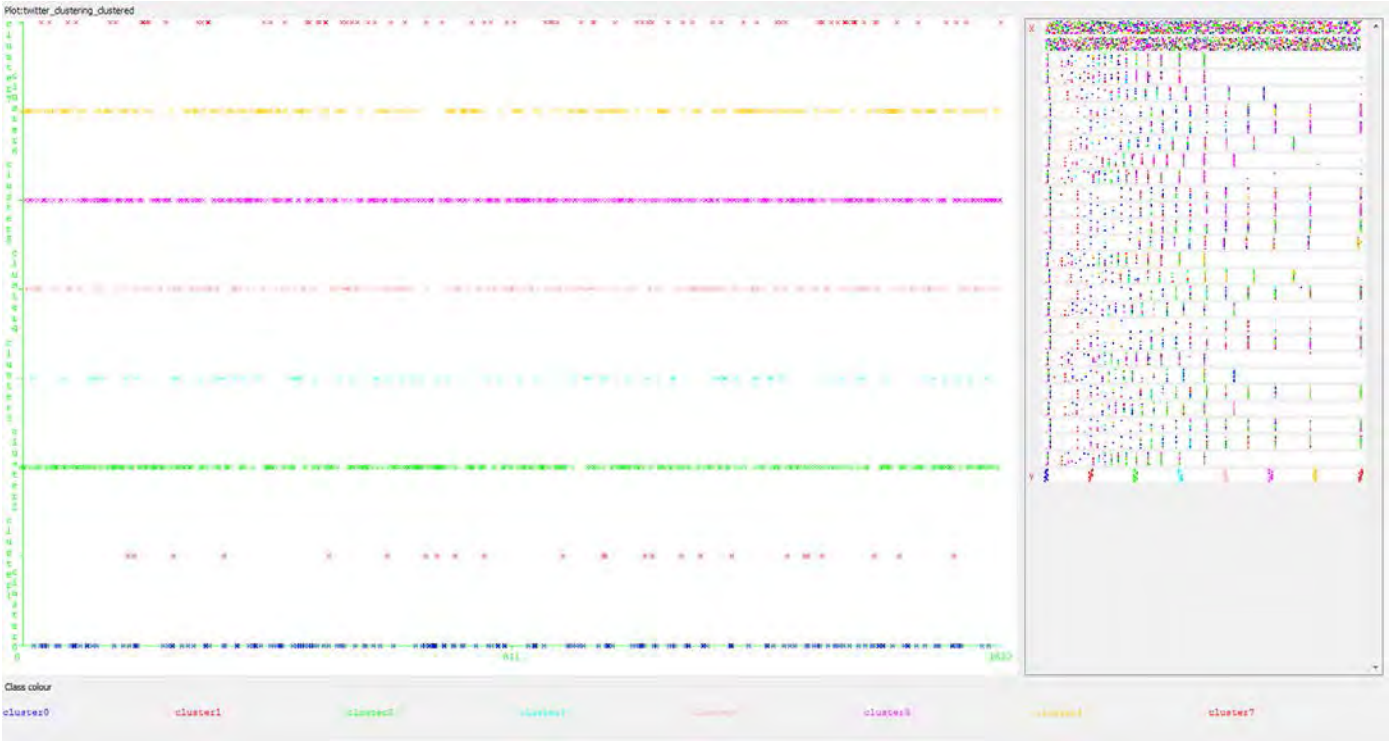


Εικόνα 21 Απεικόνιση κλάσεων αλγορίθμου DBScan με 25% training set

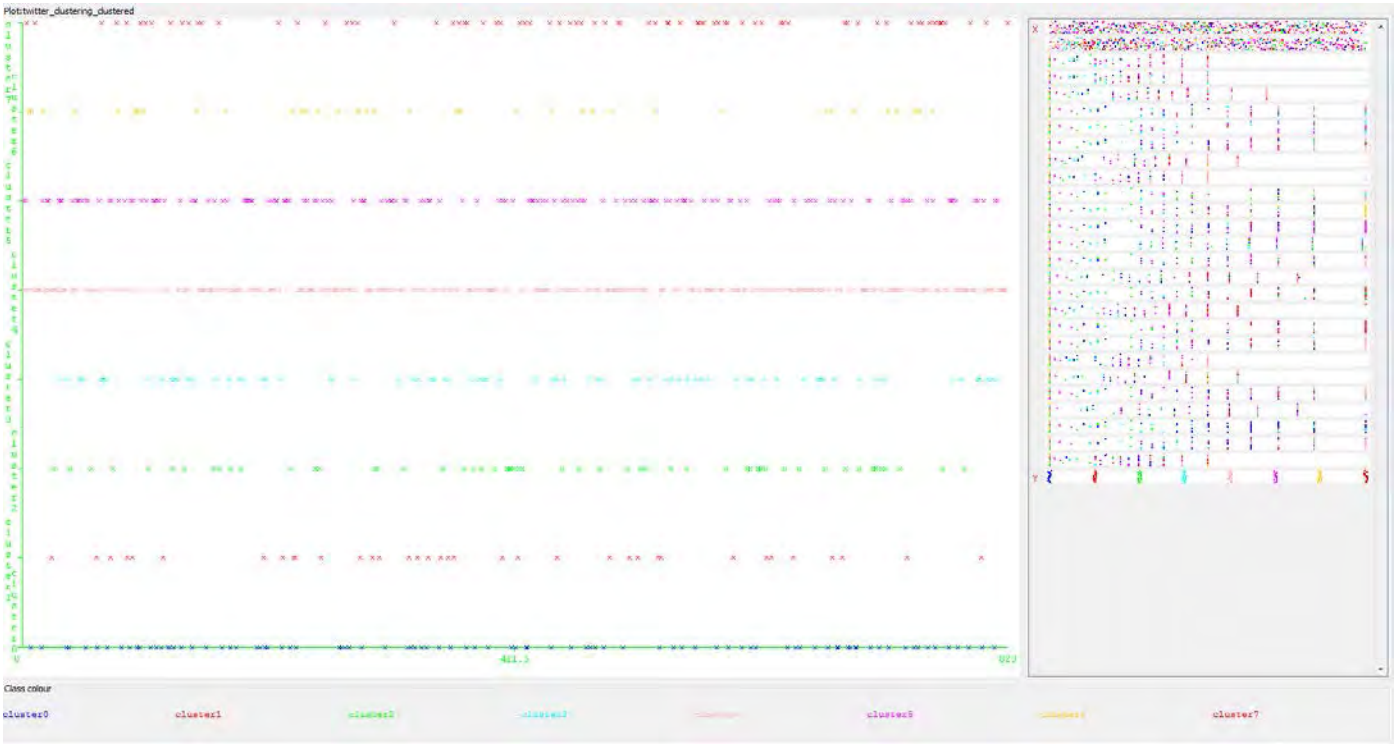


Εικόνα 20 Απεικόνιση κλάσεων αλγορίθμου DBScan με 33% training set

Εικόνα 22 Απεικόνιση κλάσεων αλγορίθμου EM με 25% training set



Εικόνα 23 Απεικόνιση κλάσεων αλγορίθμου EM με 33% training set



Εικόνα 25 Απεικόνιση κλάσεων αλγορίθμου EM με 66% training set

Εικόνα 24 Απεικόνιση κλάσεων αλγορίθμου EM με 75% training set

Τελική εργασία για την ολοκλήρωση της πτυχιακής εργασίας μας αποτέλεσε η εξαγωγή συμπερασμάτων και η εποπτική απεικόνιση αυτών που παρουσιάζεται στην επόμενη ενότητα.

5.4. Εξαγωγή και Αναπαράσταση Αποτελεσμάτων

Πρωταρχικός στόχος για τη δημιουργία της πτυχιακής εργασίας ήταν η άντληση κάποιων tweet, που αναφέρονταν σε συγκεκριμένα hashtag, από το κοινωνικό δίκτυο Twitter και η κατηγοριοποίηση αυτών των tweet, για να εξετάσουμε αν οι νέες κατηγορίες που δημιουργούνται έχουν κάποια συσχέτιση με τις αρχικές από τις οποίες αντλήθηκαν.

Οι κατάλληλες ενέργειες για την κατηγοριοποίηση, ή αλλιώς συσταδοποίηση, αυτών πραγματοποιήθηκαν και περιγράφηκαν σε προηγούμενες ενότητες. Λαμβάνοντας, επομένως, υπόψη τα αποτελέσματα που είχε η κατηγοριοποίηση των term, στα οποία οδηγηθήκαμε, καθώς και τον πρωταρχικό στόχο της πτυχιακής μας εργασίας, κρίνεται απαραίτητο η εξαγωγή συμπερασμάτων για ένα υποσύνολο των αλγορίθμων που χρησιμοποιήθηκε. Για να πραγματοποιηθεί η εξαγωγή συμπερασμάτων θα πρέπει να συγκρίνουμε τις κλάσεις που «παρήγαγε» κάθε αλγόριθμος με τον αριθμό των αρχικών κατηγοριών από τις οποίες συγκεντρώσαμε τα tweet, είτε αυτές είναι οι 5 βασικές κατηγορίες, είτε οι 25 υποκατηγορίες αυτών.

Η επιλογή των καταλληλότερων κλάσεων, βάση του αριθμού τους, εντοπίζεται στους αλγορίθμους EM και DBSCAN. Όλοι οι δυνατοί συνδυασμοί μεταξύ του training set και του test set για τον αλγόριθμο EM δημιουργούν 8 ή 4 κλάσεις με βάση τα δεδομένα που ταξινομούν. Οι κλάσεις αυτές επειδή σαν αριθμός είναι πιο κοντά στον αριθμό 5, που αντιπροσωπεύει τις 5 βασικές κατηγορίες A,B,C,D,E, θα συγκριθούν με τις 5 αυτές βασικές κατηγορίες hashtag. Με βάση την ίδια ακριβώς λογική και οι κλάσεις του αλγορίθμου DBSCAN, του οποίου το 33% των δεδομένων αποτελεί το training set, θα συγκριθούν με τις 5 βασικές κατηγορίες hashtag. Αντίθετα επειδή ο αλγόριθμος DBSCAN, με 25% των συνολικών δεδομένων να

αποτελεί το training set, δημιουργεί 17 κλάσεις θα συγκριθεί με τις 25 σε αριθμό υποκατηγορίες hashtag που είναι κοντινότερα ως αριθμοί μεταξύ τους.

Τα αρχεία που σώσαμε τα αποτελέσματα συσταδοποίησης του WEKA για τους παραπάνω αλγόριθμους, ήταν αρχικά αρχεία με κατάληξη csv, τα οποία μετατράπηκαν με μια απλή αντιγραφή των δεδομένων που εμπεριείχαν σε αρχεία xls του EXCEL. Επομένως εφόσον είχαμε 6 αλγορίθμους που επιλέξαμε ως ορθούς, δημιουργήσαμε και 6 αρχεία xls, τα οποία αποτελούνταν από 27 στήλες. Η πρώτη στήλη ήταν το TermID του κάθε term, οι επόμενες 25 στήλες ήταν τα βάρη TF-IDF για καθεμία από τις 25 κατηγορίες hashtag και η τελευταία στήλη ήταν το όνομα της κλάσης, στην οποία κατέταξε ο εκάστοτε αλγόριθμος του WEKA, το term αυτό.

Ας δούμε αναλυτικότερα τη διαδικασία που ακολουθήθηκε πρώτα για τα αποτελέσματα των αλγορίθμων που συγκρίθηκαν με τις 5 βασικές κατηγορίες hashtag και έπειτα για το μοναδικό αρχείο που συγκρίθηκε με τις 25 υποκατηγορίες των hashtag.

Σύγκριση με βάση τις 5 βασικές κατηγορίες hashtag

Με τον όρο «σύγκριση» αναφέρουμε την πιθανότητα κάθε μια από τις κλάσεις που δημιούργησε ο κάθε αλγόριθμος να αποτελεί μια από τις κατηγορίες hashtag, από τις οποίες συλλέξαμε τα αρχικά tweets. Για να υπολογίσουμε την πιθανότητα για κάθε κλάση ξεχωριστά θα πρέπει να υπολογίζουμε το συνολικό αριθμό terms που κατηγοριοποιήθηκαν στην κλάση του WEKA (για την οποία κάνω τον υπολογισμό της πιθανότητας) και έπειτα να υπολογίσουμε πόσα από αυτά άνηκαν σε κάθε μια κατηγορία των hashtag.

Ξεκινώντας να συγκρίνουμε τα αποτελέσματα των αλγορίθμων με τις 5 βασικές κατηγορίες των hashtag, θα επιλέξουμε τα αρχεία xls των αλγορίθμων που κατασκεύασαν συνολικά 8 ή 4 κλάσεις κατηγοριοποιημένων terms και θα υπολογίσουμε την πιθανότητα κάθε μια από αυτές τις κατασκευασμένες κλάσεις να αποτελεί και μια από τις παραπάνω κατηγορίες hashtag. Η περιγραφή της

διαδικασίας θα γίνει για ένα από τους 5 αλγορίθμους που ικανοποιούν τις συνθήκες αυτές, αφού η διαδικασία που ακολουθείτε στους υπόλοιπους είναι όμοια.

Σε ένα αρχείο αποτελεσμάτων xls τα δεδομένα χωρίζονται σε κλάσεις που δημιούργησε ο αλγόριθμος. Για κάθε μια από αυτές τις κλάσεις αυτές υπολογίζω τον συνολικό αριθμό term από τον οποίο αποτελείται. Γνωρίζοντας τώρα για κάθε term, που απεικονίζεται σε ένα από τα αρχεία xls, από ποια κατηγορία hashtag αντλήθηκε, υπολογίζουμε το συνόλου των term που ανήκουν στην hashtag A, στην B, στην C, στην D και στην E, για κάθε μια από τις κλάσεις που δημιούργησε ο αλγόριθμος που εξετάζουμε.

Οι αλγόριθμοι οι οποίοι θα εξετάσουμε και θα συγκρίνουμε με βάση τις 5 βασικές κατηγορίες hashtag είναι οι εξής:

- DBSCAN με 33% training set
- EM με 25% training set
- EM με 33% training set
- EM με 66% training set
- EM με 75 % training set

Τα αποτελέσματα συσταδοποίησης των παραπάνω αλγορίθμων έχουν αποθηκευτεί με την μορφή αρχείων xls. Ο υπολογισμός επομένως των κατάλληλων μετρικών που απαιτούνται για την ανάλυση των αποτελεσμάτων, έχει πραγματοποιηθεί βάση των δεδομένων που εμπεριέχονται στα αρχεία αυτά. Η γενική διαδικασία που ακολουθήθηκε για τον υπολογισμό των μετρικών στα δεδομένα των αλγορίθμων με την μορφή βημάτων είναι η εξής:

- Αρχικά κάθε ένας από τους παραπάνω αλγόριθμους παράγει ένα σύνολο κλάσεων με τα ονόματα *«Cluster 0», «Cluster 1», «Cluster 2»* και ούτω καθεξής, Τα ονόματα των κλάσεων που δίδει το WEKA ορίζονται ανάλογα με τον αριθμό των συνολικών κλάσεων που δημιουργεί ο κάθε αλγόριθμος. Αν κάποιος αλγόριθμος δεν έχει καταφέρει να κατηγοριοποιήσει όλα τα term, τότε δημιουργείται μια

ξεχωριστή κατηγορία term που ονομάζονται «*Un-clustered»* (μη-κατηγοριοποιημένα).

- Έπειτα πραγματοποιήθηκε ο υπολογισμός του συνόλου των terms που αντιστοιχούν σε κάθε μια κλάση, από εκείνες που δημιούργησε ο εκάστοτε αλγόριθμος. Για παράδειγμα στην κλάση *«Cluster 0»* κατηγοριοποιήθηκαν συνολικά κ terms, στην κλάση *«Cluster 1»* κατηγοριοποιήθηκαν n terms και συνεχίζουμε τη διαδικασία για όλες τις δημιουργημένες κλάσεις.

- Για κάθε μια από τις παραπάνω κλάσεις που δημιούργησε ο κάθε αλγόριθμος, υπολογίσαμε πόσα απ' αυτά τα terms αντλήθηκαν αρχικά από εμάς από την κατηγορία hashtag A, πόσα από τη B, τη C, τη D και την E εξίσου. Ο καθορισμός της κατηγορίας απ' όπου αντλήθηκε το κάθε term έγινε με βάση το βάρος TF-IDF που είχε το κάθε term για κάθε μία από τις κατηγορίες hashtag. **Έγινε η θεώρηση πως αν το βάρος ενός term για κάποια κατηγορία hashtag ήταν μεγαλύτερο του 0 τότε το term αυτό άνηκε στην κατηγορία που εμφάνιζε αυτό το βάρος**. Η θεώρηση αυτή βασίστηκε στον συγκεντρωτικό πίνακα που δημιουργήθηκε στην ενότητα 5.2.1.4. Ανατρέχοντας στον πίνακα εκείνο, βλέπουμε πως κάθε term έχει μια τιμή βάρους TF-IDF για κάθε μια από τις 25 υποκατηγορίες hashtag. Οι κανόνες που ορίσαμε είναι οι εξής:

    1. Αν η μη μηδενική τιμή του βάρους αντιστοιχούσε στην υποκατηγορία hashtag A1 έως A5 το συγκεκριμένο term άνηκε στην βασική κατηγορία hashtag A.
    2. Αν η μη μηδενική τιμή του βάρους αντιστοιχούσε στην υποκατηγορία hashtag B1 έως B5 το συγκεκριμένο term άνηκε στην βασική κατηγορία hashtag B.
    3. Αν η μη μηδενική τιμή του βάρους αντιστοιχούσε στην υποκατηγορία hashtag C1 έως C5 το συγκεκριμένο term

άνηκε στην βασική κατηγορία hashtag C.

4. Αν η μη μηδενική τιμή του βάρους αντιστοιχούσε στην υποκατηγορία hashtag D1 έως D5 το συγκεκριμένο term άνηκε στην βασική κατηγορία hashtag D.
5. Αν η μη μηδενική τιμή του βάρους αντιστοιχούσε στην υποκατηγορία hashtag E1 έως E5 το συγκεκριμένο term άνηκε στην βασική κατηγορία hashtag E.
6. Αν κάποιο term εμφάνιζε κάποια τιμή βάρους TF-IDF σε περισσότερες από μια ομάδα (από τις παραπάνω) το term θεωρούνταν ότι άνηκε σε όλες τις ομάδες που εμφάνιζε την τιμή αυτή.

Βάση των παραπάνω κανόνων και ειδικότερα του κανόνα με τον αριθμό 6, αντιλαμβανόμαστε πως υπήρχε περίπτωση κάποιο term να μην αντιστοιχούσε μονοσήμαντα σε μία μόνο από τις 5 βασικές κατηγορίες των hashtag. Για παράδειγμα ένα term μπορούσε να ανήκει ταυτόχρονα και στην κατηγορία hashtag A, B και C.

Για να γίνει κατανοητή η παραπάνω διαδικασία θα δώσουμε ως παράδειγμα τον υπολογισμό των πιθανοτήτων για τα αποτελέσματα του αλγόριθμου DBSCAN, που χρησιμοποίησε το 33% των συνολικών term ως training set. Ο αλγόριθμος αυτός δημιούργησε συνολικά 4 κλάσεις, με τα ονόματα «*Cluster 0», «Cluster 1», «Cluster 2»* και «*Cluster 3»*. Συνολικά κατηγοριοποιήθηκαν 1250 term σε αυτές τις 4 κλάσεις. Υπήρξε και ένα σύνολο 410 terms τα οποία παρέμειναν μη-κατηγοριοποιημένα, ή αλλιώς unclustered, όπως τα αναφέρει ο αλγόριθμος σαν ξεχωριστή κατηγορία. Υπολογίζοντας πόσα term αντιστοιχούν σε κάθε κλάση που δημιούργησε ο αλγόριθμος προέκυψε ότι στην κλάση «*Cluster 0»* εντάχθηκαν 1175 terms, στην κλάση «*Cluster 1»* 12 terms, στην κλάση *«Cluster 2»* 14 terms και στην κλάση *«Cluster 3»* 12 terms. Από το σύνολο των 1250 terms, όπως αναφέραμε 410 terms παρέμειναν un-clustered. Επόμενο βήμα ήταν ο υπολογισμός των term της κάθε κλάσης που ανήκουν στις κατηγορίες hashtag A,B,C,D,E.

Τα αποτελέσματα που προέκυψαν από τις μετρήσεις, με βάση τους παραπάνω κανόνες, για κάθε αλγόριθμο παρουσιάζονται παρακάτω με την μορφή πινάκων.

Αλγόριθμος EM με 75% training set

Κλάση του WEKA	Σύνολο Terms που ανήκουν στην κλάση	Terms που ανήκουν στην κατηγορία hashtag A	Terms που ανήκουν στην κατηγορία hashtag B	Terms που ανήκουν στην κατηγορία hashtag C	Terms που ανήκουν στην κατηγορία hashtag D	Terms που ανήκουν στην κατηγορία hashtag E
Cluster 0	79	68	34	13	47	16
Cluster 1	95	41	40	90	37	50
Cluster 2	218	110	124	80	107	84
Cluster 3	214	78	95	114	70	173

Αλγόριθμος DBSCAN με 33% training set

Κλάση του WEKA	Σύνολο Terms που ανήκουν στην κλάση	Terms που ανήκουν στην κατηγορία hashtag A	Terms που ανήκουν στην κατηγορία hashtag B	Terms που ανήκουν στην κατηγορία hashtag C	Terms που ανήκουν στην κατηγορία hashtag D	Terms που ανήκουν στην κατηγορία hashtag E
Cluster 0	1175	558	577	614	518	659
Cluster 1	12	9	6	2	7	9
Cluster 2	14	4	4	9	8	6
Cluster 3	12	9	7	5	7	6
Un-clustered	410	203	209	218	173	232

Αλγόριθμος EM με 25% training set

Κλάση του WEKA	Σύνολο Terms που ανήκουν στην κλάση	Terms που ανήκουν στην κατηγορία hashtag A	Terms που ανήκουν στην κατηγορία hashtag B	Terms που ανήκουν στην κατηγορία hashtag C	Terms που ανήκουν στην κατηγορία hashtag D	Terms που ανήκουν στην κατηγορία hashtag E
Cluster 0	113	67	25	33	81	65
Cluster 1	279	147	164	80	90	100
Cluster 2	307	148	135	217	105	271
Cluster 3	119	119	68	55	49	62
Cluster 4	113	50	29	101	8	36
Cluster 5	144	66	50	81	57	119
Cluster 6	506	196	298	230	331	259
Cluster 7	236	88	125	141	76	105

Αλγόριθμος EM με 33% training set

Κλάση του WEKA	Σύνολο Terms που ανήκουν στην κλάση	Terms που ανήκουν στην κατηγορία hashtag A	Terms που ανήκουν στην κατηγορία hashtag B	Terms που ανήκουν στην κατηγορία hashtag C	Terms που ανήκουν στην κατηγορία hashtag D	Terms που ανήκουν στην κατηγορία hashtag E
Cluster 0	175	123	63	93	133	87
Cluster 1	25	16	22	16	20	14
Cluster 2	385	148	140	288	142	269
Cluster 3	68	44	40	32	55	40
Cluster 4	226	82	90	44	102	193
Cluster 5	451	270	325	130	146	150
Cluster 6	210	64	81	169	70	118
Cluster 7	83	36	42	76	45	41

Αλγόριθμος EM με 66% training set

Κλάση του WEKA	Σύνολο Terms που ανήκουν στην κλάση	Terms που ανήκουν στην κατηγορία hashtag A	Terms που ανήκουν στην κατηγορία hashtag B	Terms που ανήκουν στην κατηγορία hashtag C	Terms που ανήκουν στην κατηγορία hashtag D	Terms που ανήκουν στην κατηγορία hashtag E
Cluster 0	90	21	47	65	40	88
Cluster 1	36	22	11	14	35	17
Cluster 2	54	16	23	54	18	27
Cluster 3	73	68	38	26	31	35
Cluster 4	309	138	152	67	97	159
Cluster 5	148	84	66	116	73	77
Cluster 6	37	12	30	6	20	8
Cluster 7	77	32	30	58	43	29

Εφόσον ολοκληρώσαμε τον υπολογισμό των απαραίτητων συνόλων απέμεινε ο υπολογισμός της πιθανότητας για κάθε κλάση. Ο αριθμός με τη μορφή δεκαδικών ψηφιών, που θα υπολογίσουμε, θα αποτελεί την πιθανότητα κάθε μια από τις κλάσεις που δημιούργησε ένας αλγόριθμος του WEKA, να αποτελεί μια βασική κατηγορία hashtag $\{A, B, C, D, E\}$. Ο υπολογισμός της πιθανότητας, βάση ορισμού, προκύπτει από την διαίρεση ενός συνόλου στοιχείων με μια συγκεκριμένη ιδιότητα προς το ολικό σύνολο των στοιχείων με όλες τις ιδιότητες.

Στην προκειμένη περίπτωση, το σύνολο με την συγκεκριμένη ιδιότητα είναι τα terms μιας κλάσης (Cluster 0 ή Cluster 1 κ.λπ.) που ανήκουν σε μια συγκεκριμένη κατηγορία hashtag. Αντίστοιχα το ολικό σύνολο αποτελεί τον αριθμό όλων των terms που υπολογίστηκαν για κάθε μια κλάση. Για παράδειγμα, ανατρέχοντας στον πίνακα των αποτελεσμάτων για τον αλγόριθμο EM με 66% training set, η πιθανότητα να αποτελεί η κλάση Cluster 0, την κατηγορία hashtag A είναι $\frac{21}{90}$, την κατηγορία hashtag B είναι $\frac{47}{90}$, την κατηγορία hashtag C είναι $\frac{65}{90}$, την κατηγορία

hashtag D είναι $\frac{40}{90}$ και την κατηγορία hashtag E είναι $\frac{88}{90}$.

Σύγκριση με βάση τις 25 υποκατηγορίες hashtag

Η παραπάνω μεθοδολογία που παρουσιάζεται αναφέρεται στην σύγκριση των αποτελεσμάτων 5 εκ των 6 συνολικών αλγορίθμων, των οποίων τα αποτελέσματα αποθηκεύσαμε, με βάση τις 5 βασικές κατηγορίες hashtag $\{A, B, C, D, E\}$. Η σύγκριση αυτή έγινε εξαιτίας του πλήθους των κλάσεων που δημιούργησε ο κάθε ένας από τους αλγόριθμους.

Ο αλγόριθμος DBSCAN που χρησιμοποιούσε το 25% του συνόλου των terms ως training set, προκειμένου να πραγματοποιήσει τη διαδικασία της κατηγοριοποίησης – συσταδοποίησης, ως αποτέλεσμα δημιούργησε 17 ξεχωριστές κλάσεις. Επίσης δημιουργήθηκε και μια ξεχωριστή κατηγορία terms, με το όνομα Un-clustered, τα οποία δεν κατηγοριοποιήθηκαν σε κάποια κλάση. Ο αριθμός των 17 κλάσεων είναι ιδιαίτερα μεγάλος για να συγκριθεί με τις 5 βασικές κατηγορίες hashtag, καθώς οι δύο αριθμοί μεταξύ τους απέχουν αρκετά. Παρόλα αυτά κρίνεται εφικτή η σύγκριση των αποτελεσμάτων του αλγορίθμου αυτού με τις 25 υποκατηγορίες hashtag που φαίνονται και στην ενότητα 5.1 $\{ A_1, A_2, A_3, A_4, A_5, B_1, B_2, B_3, B_4, B_5, C_1, C_2, C_3, C_4, C_5, D_1, D_2, D_3, D_4, D_5, E_1, E_2, E_3, E_4, E_5 \}$. Σκοπός και αυτής της σύγκρισης ήταν ο υπολογισμός της πιθανότητας κάθε μια από τις κλάσεις που δημιούργησε ο αλγόριθμος DBSCAN να αποτελεί μία από τις 25 υποκατηγορίες hashtag.

Τα δύο πρώτα βήματα της διαδικασίας του υπολογισμού των απαραίτητων συνόλων ήταν ολόιδια με την αντίστοιχη διαδικασία κατά τη σύγκριση με τις 5 βασικές κατηγορίες. Αναλυτικότερα

- Εφόσον γνωρίζουμε τον αριθμό των κλάσεων, πρέπει να υπολογίσουμε το σύνολο των terms που έχει ενταχθεί σε κάθε μια από αυτές τις 17 κλάσεις που δημιούργησε μέσω του WEKA ο αλγόριθμος. Για

παράδειγμα στην κλάση *«Cluster 0»* κατηγοριοποιήθηκαν συνολικά κ terms, στην κλάση *«Cluster 1»* κατηγοριοποιήθηκαν n terms και συνεχίζουμε τη διαδικασία και για τις 17 δημιουργημένες κλάσεις.

- Στο σημείο αυτό είναι που συναντώνται οι διαφορές στην υπολογιστική διαδικασία. Σε αντίθεση λοιπόν με τις συγκρίσεις που πραγματοποιήθηκαν στους άλλους αλγορίθμους, στον συγκεκριμένο αλγόριθμο DBSCAN **έγινε η θεώρηση πως αν το βάρος ενός term για κάποια από τις 25 υποκατηγορίες hashtag ήταν μεγαλύτερο του 0 τότε το term αυτό άνηκε στην υποκατηγορία που εμφάνιζε αυτό το βάρος.** Με βάση τη θεώρηση αυτή δημιουργήθηκαν και οι παρακάτω κανόνες:

    1. Αν ένα term εμφάνιζε μη μηδενική τιμή βάρους στην υποκατηγορία A1 τότε το term αυτό άνηκε στην υποκατηγορία αυτή. Αντίστοιχη λογική ακολουθούνταν και για τις υπόλοιπες υποκατηγορίες A2 έως E5.

    2. Αν κάποιο term εμφάνιζε κάποια τιμή βάρους TF-IDF σε περισσότερες από μια υποκατηγορία (από τις παραπάνω) το term θεωρούνταν ότι άνηκε ταυτόχρονα σε όλες τις υποκατηγορίες που εμφάνιζε την τιμή αυτή.

Όπως και στην προηγούμενη περίπτωση συγκρίσεων, αντιλαμβανόμαστε πως υπήρχε περίπτωση κάποιο term να μην αντιστοιχούσε μονοσήμαντα σε μία μόνο από τις 25 υποκατηγορίες των hashtag. Για παράδειγμα ένα term μπορούσε να ανήκει ταυτόχρονα και στην υποκατηγορία hashtag A1, στην A2 και στην E2.

Τα αποτελέσματα των υπολογισμών που προκύπτουν παρουσιάζονται στους παρακάτω πίνακες:

Αλγόριθμος DBSCAN με 25% training set

Υποκατηγορίες A1, A2, A3, A4, A5

Κλάση του WEKA	Σύνολο Terms που ανήκουν στην κλάση	Terms που ανήκουν στην κατηγορία hashtag A1	Terms που ανήκουν στην κατηγορία hashtag A2	Terms που ανήκουν στην κατηγορία hashtag A3	Terms που ανήκουν στην κατηγορία hashtag A4	Terms που ανήκουν στην κατηγορία hashtag A5
Un-clustered	880	130	126	139	139	106
Cluster 0	376	48	52	52	44	55
Cluster 1	51	10	12	4	8	10
Cluster 2	207	30	36	39	29	26
Cluster 3	21	3	2	4	4	1
Cluster 4	24	6	4	2	1	3
Cluster 5	24	2	2	2	6	3
Cluster 6	24	4	2	4	2	2
Cluster 7	24	5	6	2	5	6
Cluster 8	24	2	6	4	5	5
Cluster 9	18	5	7	5	5	3
Cluster 10	18	1	2	3	2	4
Cluster 11	33	4	8	5	2	2
Cluster 12	21	2	3	4	4	2
Cluster 13	18	2	2	2	0	2
Cluster 14	18	0	3	0	5	2
Cluster 15	18	1	3	4	3	2
Cluster 16	18	3	2	4	4	2

Υποκατηγορίες B1, B2, B3, B4, B5

Κλάση του WEKA	Σύνολο Terms που ανήκουν στην κλάση	Terms που ανήκουν στην κατηγορία hashtag B1	Terms που ανήκουν στην κατηγορία hashtag B2	Terms που ανήκουν στην κατηγορία hashtag B3	Terms που ανήκουν στην κατηγορία hashtag B4	Terms που ανήκουν στην κατηγορία hashtag B5
Un-clustered	880	203	140	112	75	95
Cluster 0	376	94	70	51	31	44
Cluster 1	51	12	7	0	4	3
Cluster 2	207	55	36	36	18	22
Cluster 3	21	7	5	3	4	7
Cluster 4	24	4	1	4	0	5
Cluster 5	24	3	2	2	2	3
Cluster 6	24	4	5	3	2	3
Cluster 7	24	2	5	3	2	1
Cluster 8	24	8	3	5	2	2
Cluster 9	18	5	6	2	4	4
Cluster 10	18	1	2	1	1	5
Cluster 11	33	8	5	4	2	3
Cluster 12	21	2	3	5	6	2
Cluster 13	18	4	1	2	0	3
Cluster 14	18	6	3	7	1	2
Cluster 15	18	3	1	2	1	2
Cluster 16	18	4	0	1	1	1

Υποκατηγορίες C1, C2, C3, C4, C5

Κλάση του WEKA	Σύνολο Terms που ανήκουν στην κλάση	Terms που ανήκουν στην κατηγορία hashtag C1	Terms που ανήκουν στην κατηγορία hashtag C2	Terms που ανήκουν στην κατηγορία hashtag C3	Terms που ανήκουν στην κατηγορία hashtag C4	Terms που ανήκουν στην κατηγορία hashtag C5
Un-clustered	880	159	135	112	132	117
Cluster 0	376	74	45	63	56	53
Cluster 1	51	7	6	5	6	12
Cluster 2	207	41	34	35	30	34
Cluster 3	21	6	3	7	5	4
Cluster 4	24	5	3	4	3	5
Cluster 5	24	6	3	4	3	4
Cluster 6	24	5	1	2	4	1
Cluster 7	24	7	4	7	3	3
Cluster 8	24	7	5	2	3	4
Cluster 9	18	4	1	5	1	4
Cluster 10	18	3	3	5	4	1
Cluster 11	33	7	4	2	6	4
Cluster 12	21	5	5	2	0	4
Cluster 13	18	1	4	3	3	3
Cluster 14	18	3	3	3	5	5
Cluster 15	18	6	3	3	2	1
Cluster 16	18	6	5	0	3	1

Υποκατηγορίες D1, D2, D3, D4, D5

Κλάση του WEKA	Σύνολο Terms που ανήκουν στην κλάση	Terms που ανήκουν στην κατηγορία hashtag D1	Terms που ανήκουν στην κατηγορία hashtag D2	Terms που ανήκουν στην κατηγορία hashtag D3	Terms που ανήκουν στην κατηγορία hashtag D4	Terms που ανήκουν στην κατηγορία hashtag D5
Un-clustered	880	139	69	144	117	93
Cluster 0	376	45	24	57	42	50
Cluster 1	51	8	2	7	6	0
Cluster 2	207	23	13	30	25	28
Cluster 3	21	3	1	5	3	4
Cluster 4	24	2	2	2	4	5
Cluster 5	24	5	0	3	4	1
Cluster 6	24	2	0	2	1	4
Cluster 7	24	6	1	5	5	1
Cluster 8	24	2	2	5	1	2
Cluster 9	18	2	2	2	4	4
Cluster 10	18	1	2	4	1	1
Cluster 11	33	5	2	3	5	3
Cluster 12	21	7	2	2	2	2
Cluster 13	18	1	0	1	2	2
Cluster 14	18	1	1	2	2	3
Cluster 15	18	1	2	1	2	1
Cluster 16	18	3	3	1	2	3

Υποκατηγορίες E1, E2, E3, E4, E5

Κλάση του WEKA	Σύνολο Terms που ανήκουν στην κλάση	Terms που ανήκουν στην κατηγορία hashtag E1	Terms που ανήκουν στην κατηγορία hashtag E2	Terms που ανήκουν στην κατηγορία hashtag E3	Terms που ανήκουν στην κατηγορία hashtag E4	Terms που ανήκουν στην κατηγορία hashtag E5
Un-clustered	880	156	130	128	182	187
Cluster 0	376	68	43	64	72	77
Cluster 1	51	11	7	5	15	15
Cluster 2	207	40	33	34	42	45
Cluster 3	21	8	8	4	5	4
Cluster 4	24	9	4	4	3	6
Cluster 5	24	5	1	4	2	5
Cluster 6	24	6	5	6	7	5
Cluster 7	24	5	3	4	8	3
Cluster 8	24	7	1	4	1	3
Cluster 9	18	2	4	2	4	2
Cluster 10	18	2	0	4	4	4
Cluster 11	33	3	3	2	4	9
Cluster 12	21	3	1	0	4	3
Cluster 13	18	1	1	4	2	3
Cluster 14	18	4	5	1	2	2
Cluster 15	18	2	2	7	4	6
Cluster 16	18	4	3	2	3	2

Ο υπολογισμός της πιθανότητας κάθε μιας από τις κλάσεις που δημιούργησε ο συγκεκριμένος αλγόριθμος DBSCAN, να αποτελεί μια υποκατηγορία hashtag $\{ A_1, A_2, A_3, A_4, A_5, B_1, B_2, B_3, B_4, B_5, C_1, C_2, C_3, C_4, C_5, D_1, D_2, D_3, D_4, D_5, E_1, E_2, E_3, E_4, E_5 \}$ πραγματοποιείται με τη ίδια τεχνική βάση του ορισμού της πιθανότητας. Ο αριθμητής του κλάσματος της πιθανότητας αποτελεί το σύνολο των terms μίας εκ των 17 κλασεων, που δημιούργησε ο αλγόριθμος, τα οποία ανήκουν σε μια υποκατηγορία hashtag. Ο παρονομαστής αντίθετα αποτελεί το σύνολο όλων των terms τα οποία κατηγοριοποιήθηκαν στη συγκεκριμένη κλάση. Για παράδειγμα

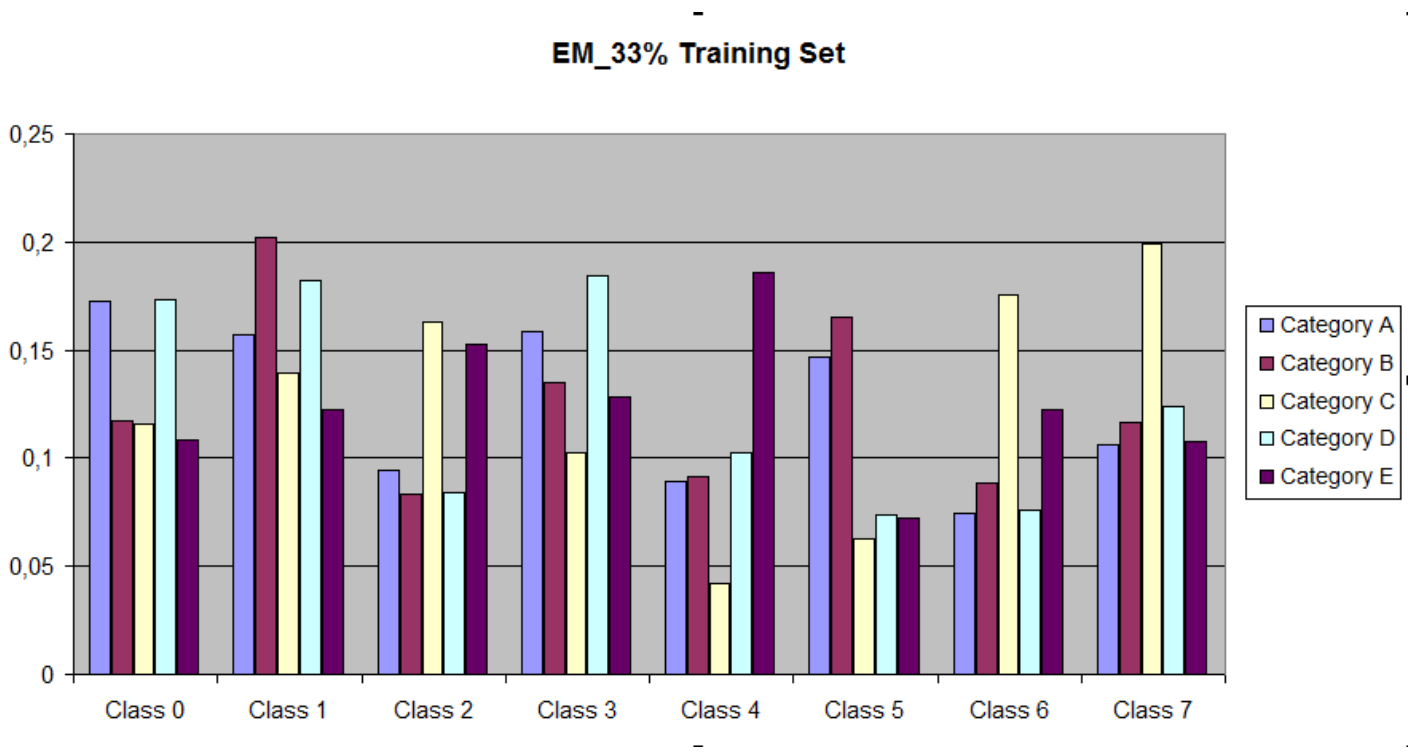
βλέποντας τους παραπάνω πίνακες των συνόλων του αλγορίθμου DBSCAN η πιθανότητα να αποτελεί η κλάση *«Cluster 0»* την υποκατηγορία hashtag A5 είναι $\frac{55}{376}$. Αντίστοιχα η πιθανότητα να αποτελεί η κλάση *«Cluster 11»* την υποκατηγορία hashtag E5 είναι $\frac{9}{33}$.

Να τονίσουμε στο σημείο αυτό πως οι τιμές πιθανότητας, που υπολογίζουμε και στις δυο περιπτώσεις συγκρίσεων (και με τις 5 βασικές κατηγορίες hashtag και με τις 25 υποκατηγορίες αυτών), δεν είναι **κανονικοποιημένες**, δεν δημιουργούν δηλαδή κανονική κατανομή, με αποτέλεσμα να μην μπορούν να αθροιστούν στην μονάδα. Για παράδειγμα αθροίζοντας όλες τις πιθανότητες του να αποτελεί η κλάση *«Cluster 0»* την κατηγορία hashtag A ή B ή C ή D ή E θα έπρεπε να είναι το αποτέλεσμα ίσο με τη μονάδα, γεγονός που δεν συμβαίνει. Επομένως πριν την τελική παρουσίαση των αποτελεσμάτων με τη μορφή γραφημάτων όλες οι πιθανότητες και στις δύο περιπτώσεις συγκρίσεων κανονικοποιήθηκαν Η διαδικασία της κανονικοποίησης, βάση του ορισμού της, προκύπτει διαιρώντας κάθε αριθμό με το άθροισμα των αριθμών σε μια στήλη ή γραμμή, όταν ο αριθμός αυτός αποτελεί ποσοστό πιθανότητας.

Στις επόμενες σελίδες παρουσιάζονται τα αποτελέσματα, με τη μορφή γραφημάτων, για κάθε ένα αλγόριθμο που αναλύσαμε ξεχωριστά. Οι τιμές που απεικονίζονται στις ποσοστιαίες στήλες της περιοχής των γραφημάτων αποτελούν τις κανονικοποιημένες πιθανότητες, σύμφωνα με τις οποίες οι κλάσεις (του κάθε αλγόριθμου) που βρίσκονται στον οριζόντιο άξονα του γραφήματος αποτελούν τις κατηγορίες ή υποκατηγορίες hashtag που εμφανίζονται στο υπόμνημα. Τα γραφήματα συνοδεύονται με τους πίνακες που εμπεριέχουν τις πιθανότητες αυτές για κάθε αλγόριθμο.

EM Algorithm 33% Training Set

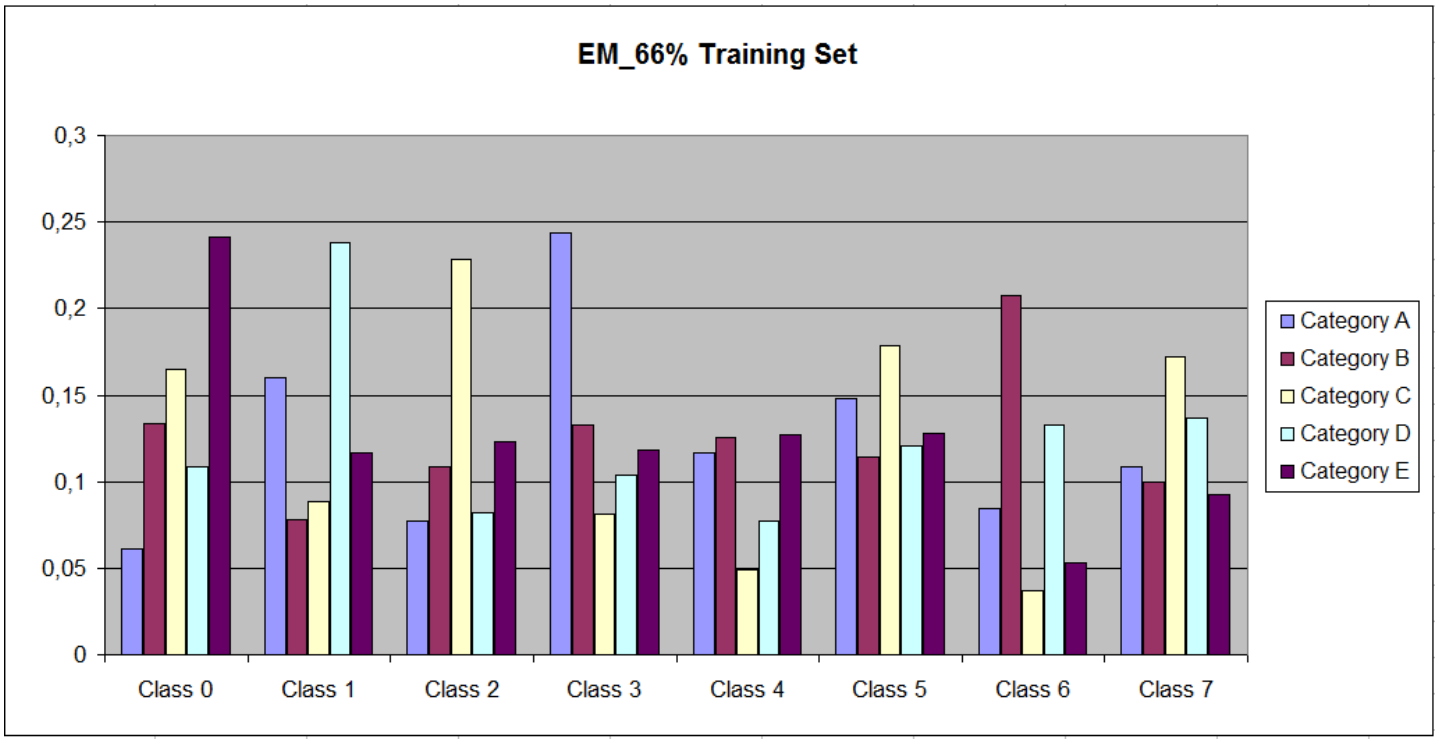
EM_33% Training Set						
	Category A	Category B	Category C	Category D	Category E	Μέσο Ποσοστό Επιτυχίας Ταύτισης παραγόμενων κλάσεων με αρχικές κατηγορίες hashtag
Class 0	0,172509345	0,117621123	0,115693501	0,173191347	0,108391311	0,137481325
Class 1	0,157081042	0,202082367	0,139329434	0,182306681	0,122096661	0,160579237
Class 2	0,094351709	0,083505945	0,162852161	0,084050216	0,152337388	0,115419484
Class 3	0,158813842	0,135082877	0,102448497	0,184316612	0,128253822	0,14178313
Class 4	0,089052679	0,09144916	0,042384449	0,102850593	0,186193048	0,102385986
Class 5	0,146937042	0,165482495	0,062752671	0,073772677	0,072514515	0,10429188
Class 6	0,074800028	0,088574079	0,175198055	0,075960357	0,122510918	0,107408688
Class 7	0,106454313	0,116201954	0,199341233	0,123551516	0,107702337	0,130650271



Γράφημα 1 ΕΜ αλγόριθμος με 33% training set

EM Algorithm 66% Training Set

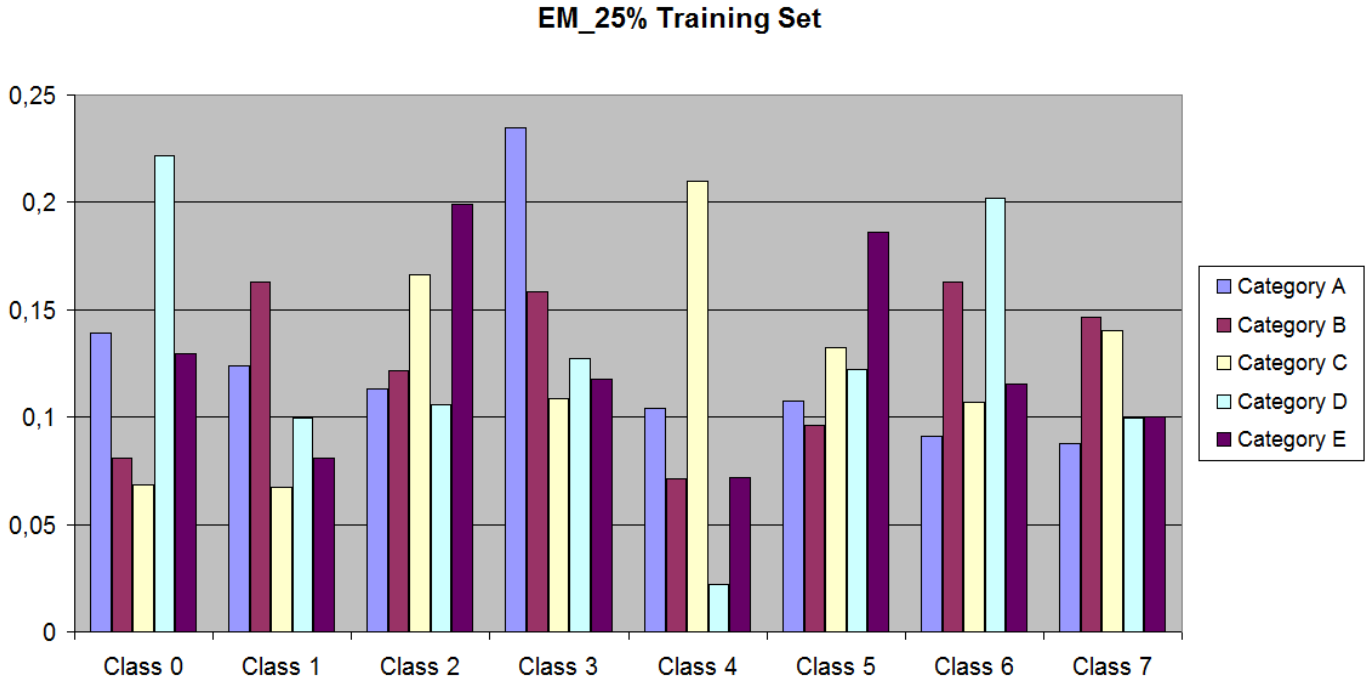
EM_66% Training Set						
	Category A	Category B	Category C	Category D	Category E	Μέσο Ποσοστό Επιτυχίας Ταύτισης παραγόμενων κλάσεων με αρχικές κατηγορίες hashtag
Class 0	0,060972741	0,13347306	0,164761709	0,108905553	0,241003648	0,141823342
Class 1	0,159710425	0,078110623	0,088723108	0,238249277	0,116385685	0,136235824
Class 2	0,077437734	0,108859012	0,228138617	0,081679165	0,123237701	0,123870446
Class 3	0,243446672	0,133038544	0,081262975	0,104077832	0,118184955	0,136002196
Class 4	0,116718501	0,125728453	0,049460452	0,076924962	0,126836242	0,099133722
Class 5	0,14834174	0,113970964	0,178815048	0,120864579	0,128241152	0,138046696
Class 6	0,084755508	0,207238524	0,037004084	0,132456011	0,053287982	0,102948422
Class 7	0,108616679	0,09958082	0,171834006	0,136842621	0,092822636	0,121939353



Γράφημα 2 EM αλγόριθμος με 66% training set

EM Algorithm 25% Training Set

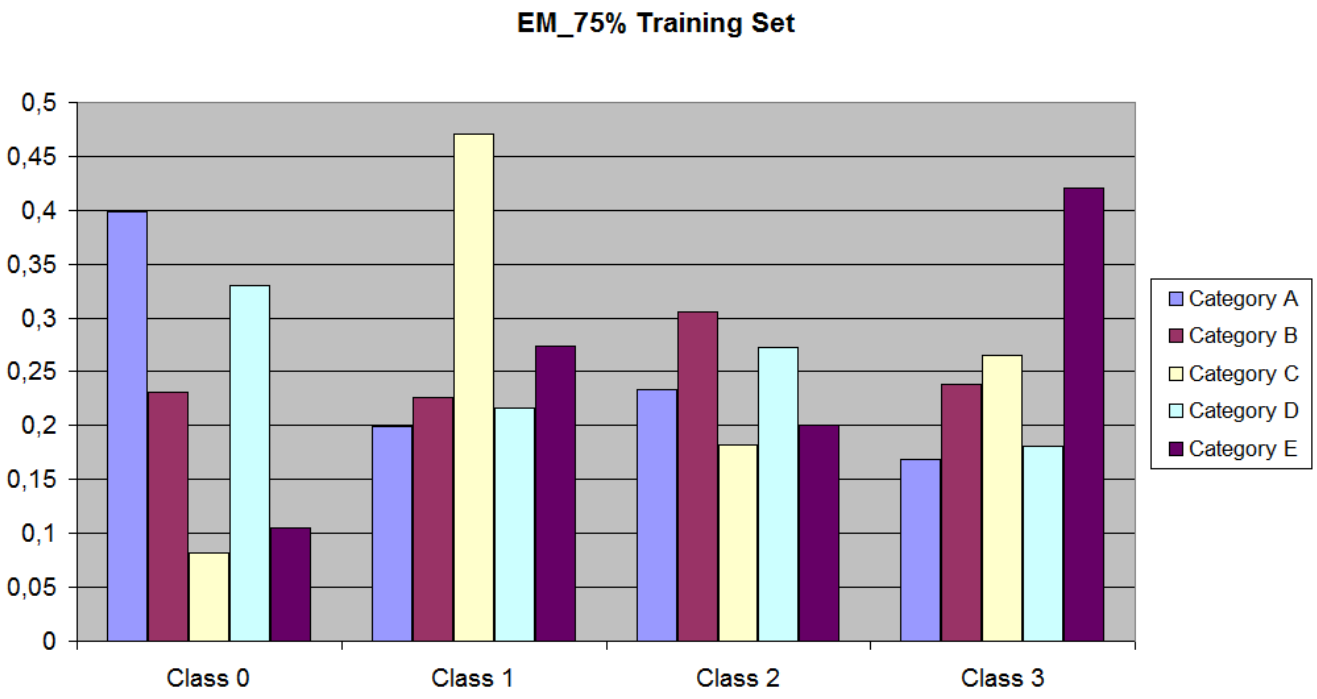
EM_25% Training Set						
	Category A	Category B	Category C	Category D	Category E	Μέσο Ποσοστό Επιτυχίας Ταύτισης παραγόμενων κλάσεων με αρχικές κατηγορίες hashtag
Class 0	0,139087435	0,080616275	0,068616486	0,221512494	0,129579153	0,127882369
Class 1	0,123595743	0,162707914	0,067371221	0,099685412	0,080740864	0,106820231
Class 2	0,113088885	0,121721608	0,166076144	0,105692866	0,198853832	0,141086667
Class 3	0,234580441	0,158173871	0,108591863	0,127244297	0,117367328	0,14919156
Class 4	0,103797154	0,071038872	0,210005827	0,021878998	0,071766153	0,095697401
Class 5	0,107515254	0,09611174	0,132162627	0,122321522	0,186159932	0,128854215
Class 6	0,090864734	0,163017934	0,10679915	0,202148963	0,115306118	0,13562738
Class 7	0,087470355	0,146611786	0,140376681	0,099515448	0,10022662	0,114840178



Γράφημα 3 ΕΜ αλγόριθμος με 25% training set

EM Algorithm 75% Training Set

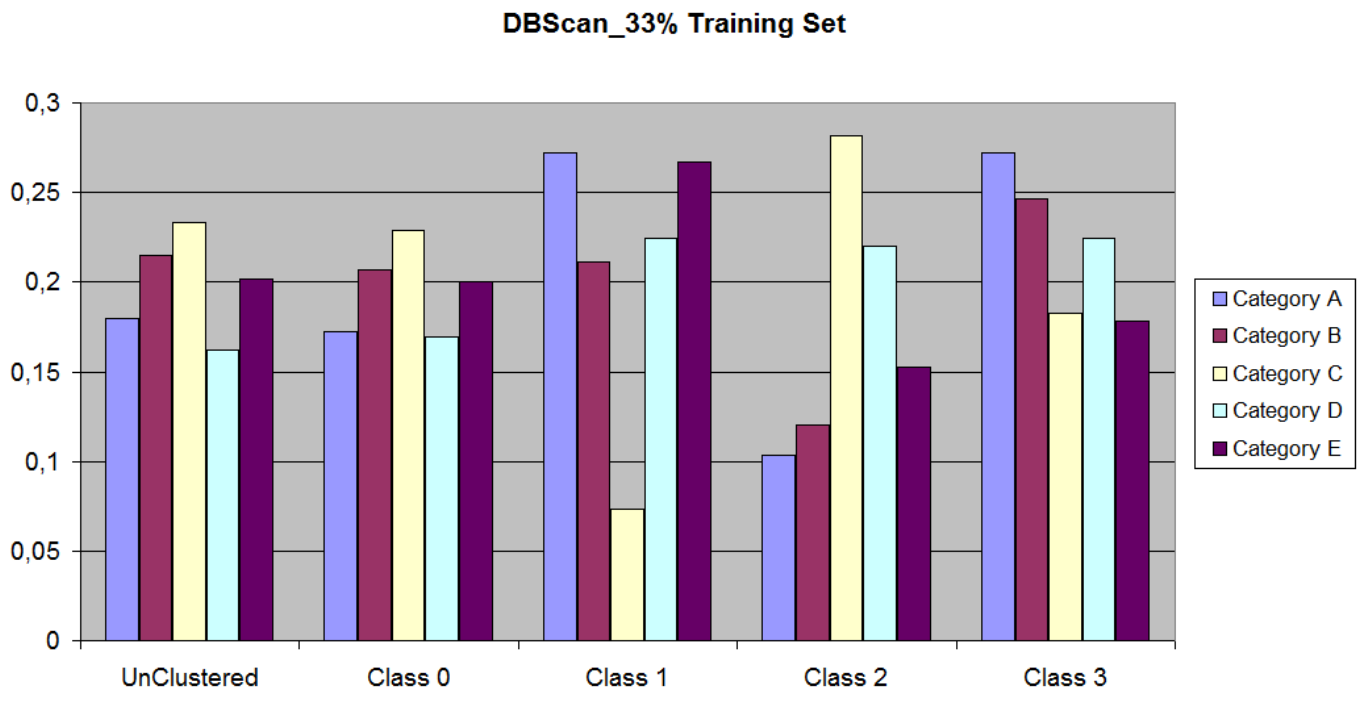
EM_75% Training Set						
	Category A	Category B	Category C	Category D	Category E	Μέσο Ποσοστό Επιτυχίας Ταύτισης παραγόμενων κλάσεων με αρχικές κατηγορίες hashtag
Class 0	0,398238195	0,230869502	0,081805121	0,330093101	0,105342821	0,229269748
Class 1	0,199674288	0,225864594	0,470951129	0,216091304	0,273757139	0,277267691
Class 2	0,233453008	0,305127751	0,182426017	0,272329305	0,200418188	0,238750854
Class 3	0,168634509	0,238138153	0,264817733	0,18148629	0,420481853	0,254711707



Γράφημα 4 ΕΜ αλγόριθμος με 75% training set

DBScan Algorithm 33% Training Set

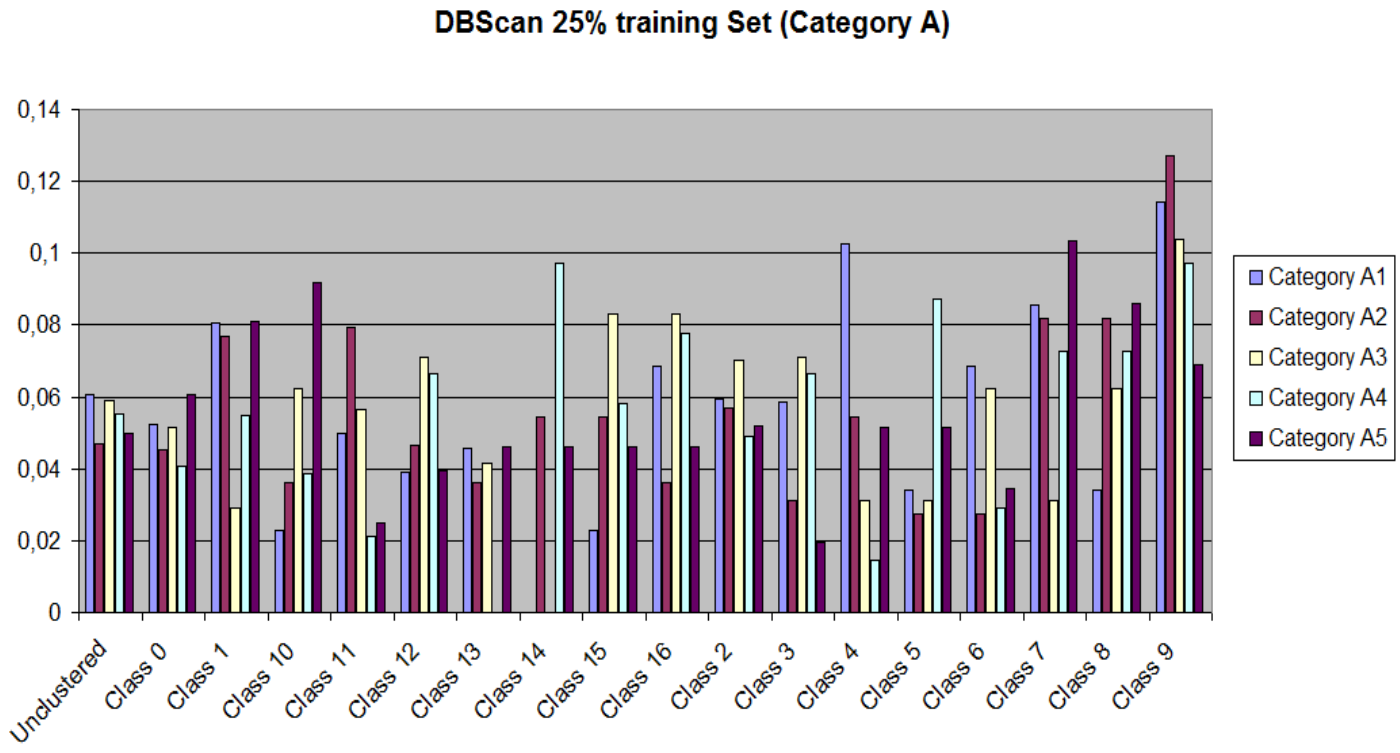
DBScan_33% Training Set						
	Category A	Category B	Category C	Category D	Category E	Μέσο Ποσοστό Επιτυχίας Ταύτισης παραγόμενων κλάσεων με αρχικές κατηγορίες hashtag
UnClustered	0,179669923	0,215101314	0,233159099	0,162232928	0,201709639	0,198374581
Class 0	0,172328829	0,207210553	0,229142366	0,169499671	0,19992728	0,19562174
Class 1	0,272161177	0,210982927	0,073086132	0,224280919	0,267353945	0,20957302
Class 2	0,103678893	0,120559864	0,281899266	0,219705562	0,152773173	0,175723352
Class 3	0,272161177	0,246145342	0,182713137	0,224280919	0,178235963	0,220707308



Γράφημα 5 DBScan αλγόριθμος με 33% training set

DBScan Algorithm 25% Training Set (Category A)

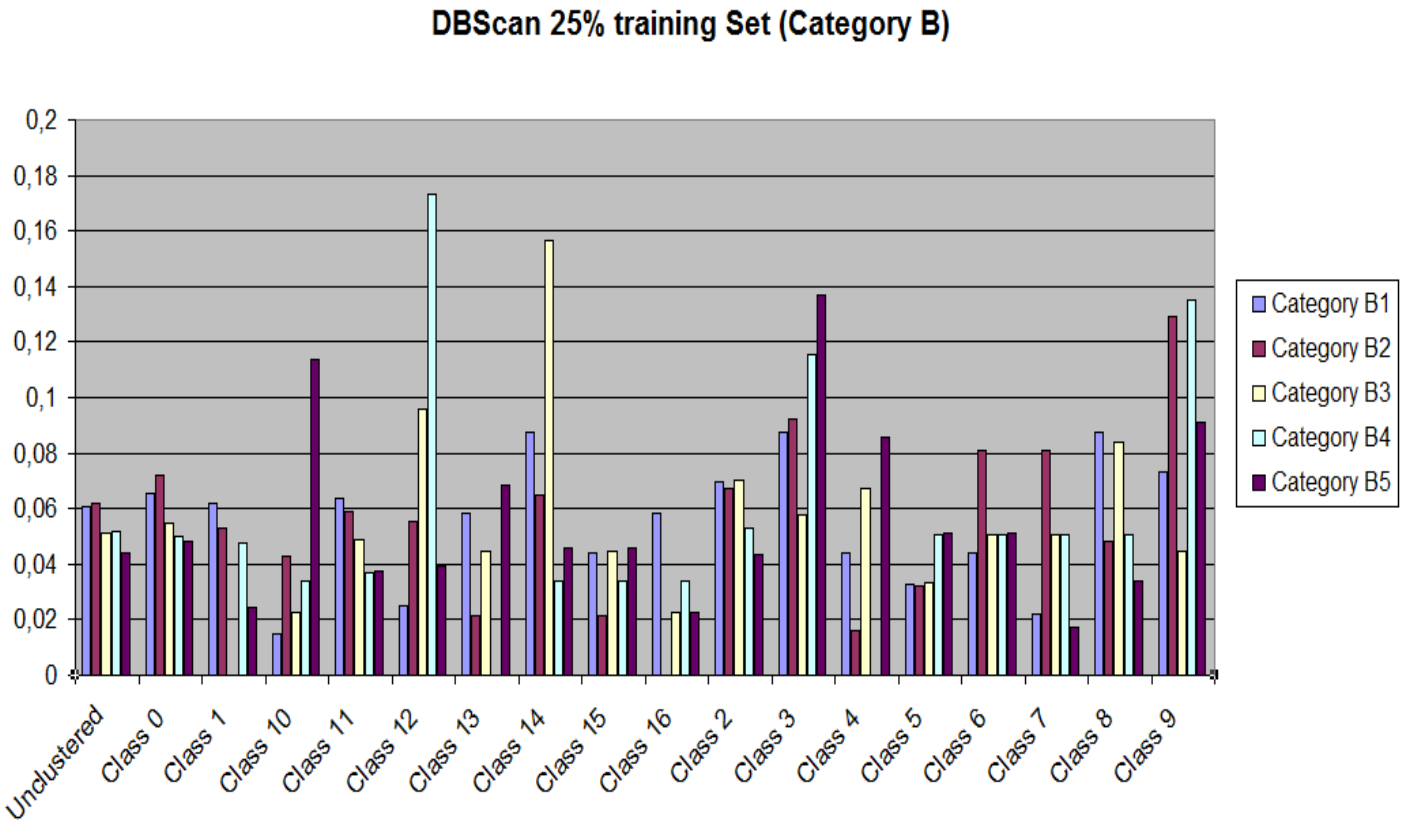
DBScan_25% Training Set (Category A)						
	Category A1	Category A2	Category A3	Category A4	Category A5	Μέσο Ποσοστό Επιτυχίας Ταύτισης παραγόμενων κλάσεων με αρχικές κατηγορίες hashtag
Unclustered	0,060692374	0,046789988	0,058962674	0,055215042	0,04983893	0,054299802
Class 0	0,052447747	0,045193977	0,051625057	0,040906292	0,060522988	0,050139212
Class 1	0,080556998	0,07689111	0,029277559	0,054833389	0,081128854	0,064537582
Class 10	0,022824483	0,036309691	0,062214812	0,038840317	0,091946034	0,050427067
Class 11	0,049798871	0,079221143	0,05655892	0,021185628	0,025076191	0,046368151
Class12	0,039127685	0,046683888	0,071102643	0,066583401	0,039405443	0,052580612
Class 13	0,045648965	0,036309691	0,041476542	0	0,045973017	0,033881643
Class 14	0	0,054464536	0	0,097100793	0,045973017	0,039507669
Class 15	0,022824483	0,054464536	0,082953083	0,058260476	0,045973017	0,052895119
Class 16	0,068473448	0,036309691	0,082953083	0,077680634	0,045973017	0,062277975
Class 2	0,059542129	0,056832559	0,070329788	0,048972574	0,051969498	0,057529309
Class 3	0,058691527	0,031122592	0,071102643	0,066583401	0,019702722	0,049440577
Class 4	0,102710172	0,054464536	0,031107406	0,014565119	0,051719644	0,050913375
Class 5	0,034236724	0,027232268	0,031107406	0,087390714	0,051719644	0,046337351
Class 6	0,068473448	0,027232268	0,062214812	0,029130238	0,034479763	0,044306106
Class 7	0,08559181	0,081696804	0,031107406	0,072825595	0,103439288	0,074932181
Class 8	0,034236724	0,081696804	0,062214812	0,072825595	0,086199407	0,067434668
Class 9	0,114122413	0,127083918	0,103691354	0,097100793	0,068959526	0,102191601



Γράφημα 6 DBScan αλγόριθμος για την κατηγορία hashtag A, με 25% training set

DBScan Algorithm 25% Training Set (Category B)

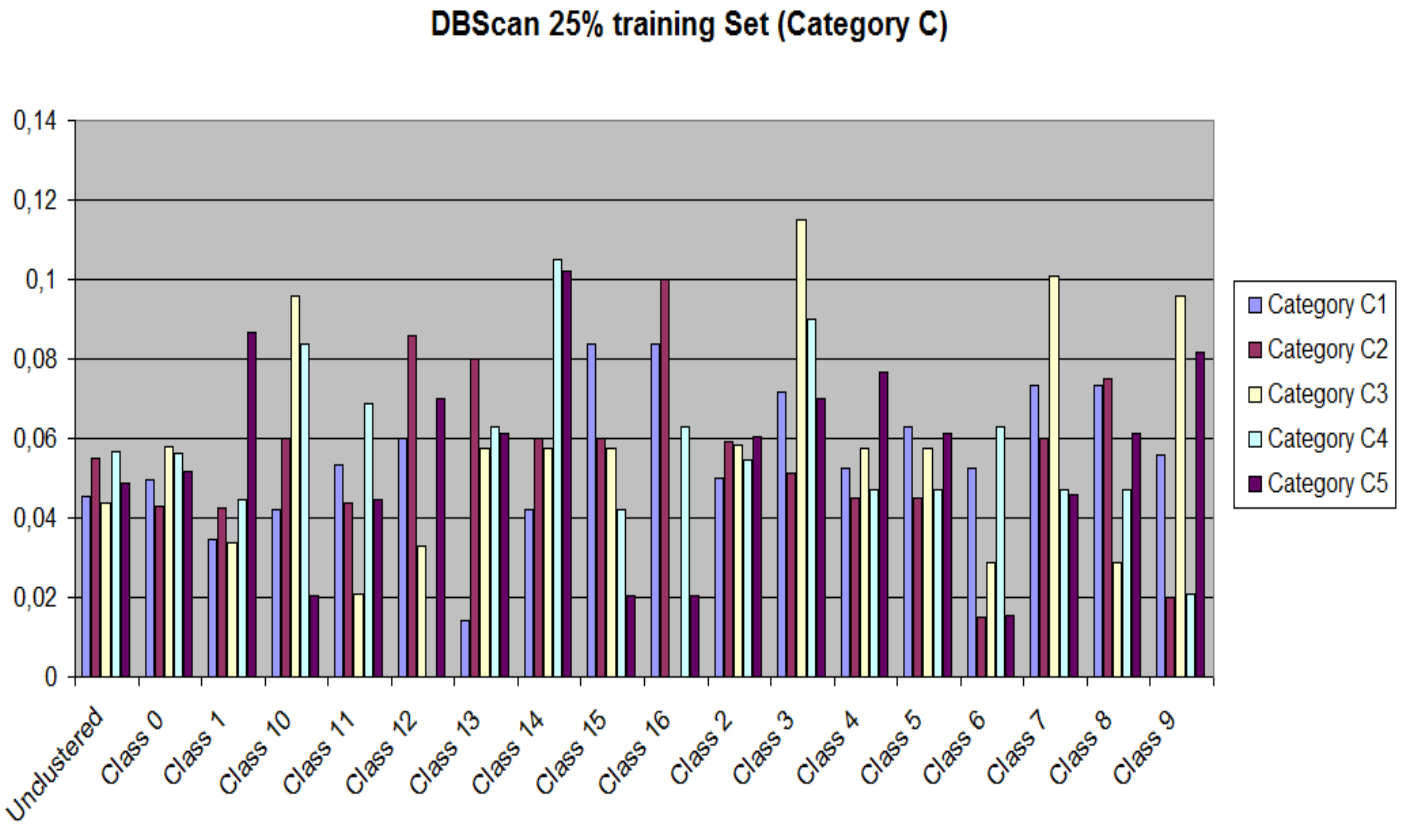
DBScan_25% Training Set (Category B)						
	Category B1	Category B2	Category B3	Category B4	Category B5	Μέσο Ποσοστό Επιτυχίας Ταύτισης παραγόμενων κλάσεων με αρχικές κατηγορίες hashtag
Unclustered	0,060618804	0,061716116	0,051306093	0,051726994	0,044291991	0,053932
Class 0	0,065695255	0,072220987	0,054678416	0,050039447	0,048011923	0,058129206
Class 1	0,061830828	0,053245277	0	0,047602358	0,024134336	0,03736256
Class 10	0,014598946	0,043103319	0,022395517	0,033718337	0,113967696	0,045556763
Class 11	0,06370449	0,058777254	0,048862945	0,03678364	0,037298519	0,04908537
Class12	0,025026764	0,055418553	0,095980785	0,17340859	0,039074639	0,077781866
Class 13	0,058395782	0,02155166	0,044791033	0	0,068380617	0,038623819
Class 14	0,087593674	0,064654979	0,156768616	0,033718337	0,045587078	0,077664537
Class 15	0,043796837	0,02155166	0,044791033	0,033718337	0,045587078	0,037888989
Class 16	0,058395782	0	0,022395517	0,033718337	0,022793539	0,027460635
Class 2	0,069821044	0,067466065	0,070107704	0,052776527	0,043605031	0,060755274
Class 3	0,087593674	0,092364256	0,057588471	0,115605727	0,136761235	0,097982672
Class 4	0,043796837	0,016163745	0,06718655	0	0,085475772	0,042524581
Class 5	0,032847628	0,03232749	0,033593275	0,050577505	0,051285463	0,040126272
Class 6	0,043796837	0,080818724	0,050389912	0,050577505	0,051285463	0,055373688
Class 7	0,021898418	0,080818724	0,050389912	0,050577505	0,017095154	0,044155943
Class 8	0,087593674	0,048491234	0,083983187	0,050577505	0,034190309	0,060967182
Class 9	0,072994728	0,129309958	0,044791033	0,134873348	0,091174157	0,094628645



Γράφημα 7 DBScan αλγόριθμος για την κατηγορία hashtag B, με 25% training set

DBScan Algorithm 25% Training Set (Category C)

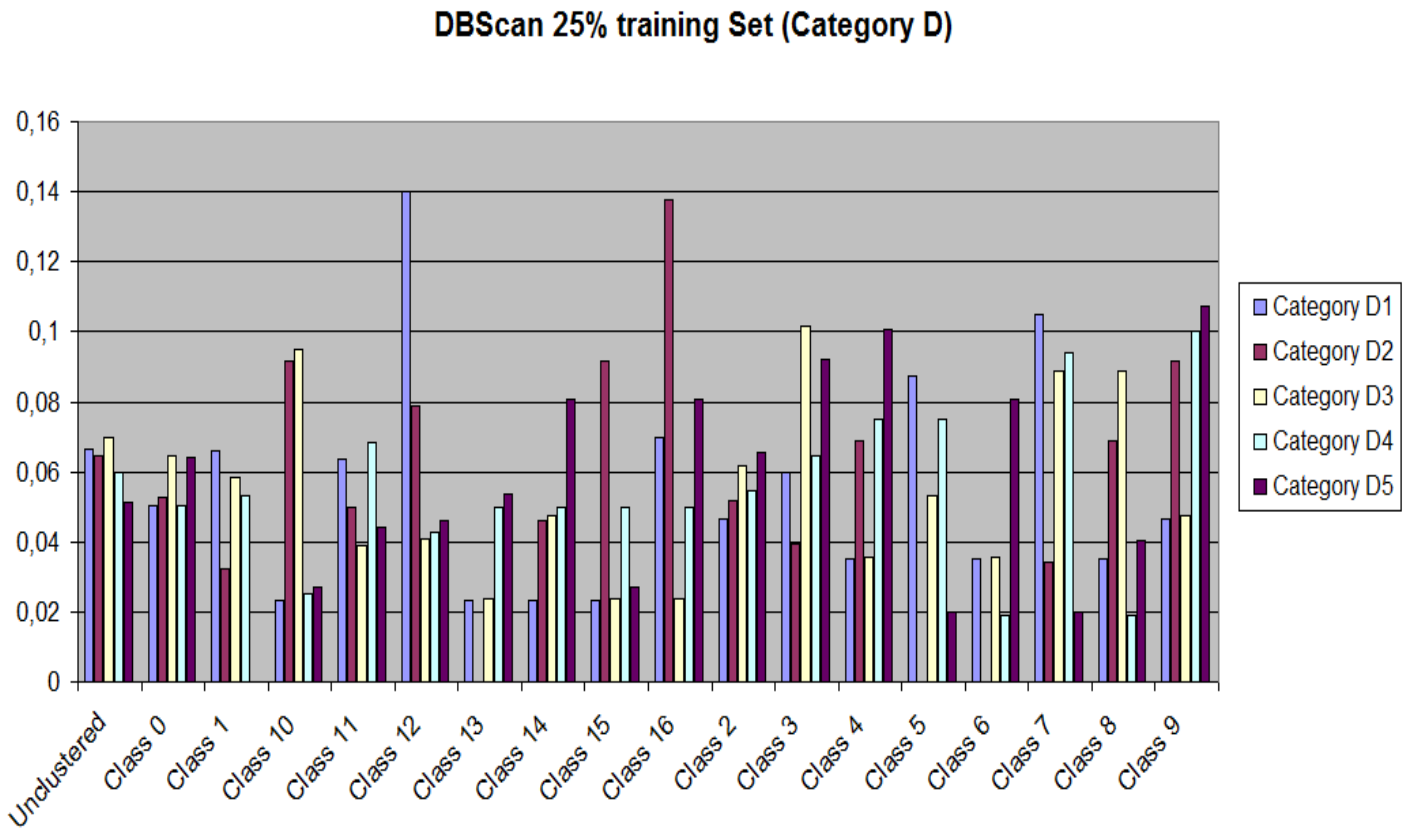
DBScan_25% Training Set (Category C)						
	Category C1	Category C2	Category C3	Category C4	Category C5	Μέσο Ποσοστό Επιτυχίας Ταύτισης παραγόμενων κλάσεων με αρχικές κατηγορίες hashtag
Unclustered	0,045437155	0,055195537	0,043912178	0,056644582	0,048929124	0,050023715
Class 0	0,049492632	0,043060348	0,057809915	0,056242848	0,051874303	0,051696009
Class 1	0,034516324	0,042328603	0,033825907	0,044427123	0,086591512	0,048337894
Class 10	0,04191268	0,059965521	0,09584007	0,0839179	0,020445218	0,060416278
Class 11	0,05334341	0,043611288	0,020910561	0,0686601	0,044607748	0,046226622
Class12	0,059875257	0,08566503	0,032859453	0	0,07009789	0,049699526
Class 13	0,013970893	0,079954028	0,057504042	0,062938425	0,061335654	0,055140609
Class 14	0,04191268	0,059965521	0,057504042	0,104897374	0,10222609	0,073301142
Class 15	0,083825359	0,059965521	0,057504042	0,04195895	0,020445218	0,052739818
Class 16	0,083825359	0,099942535	0	0,062938425	0,020445218	0,053430308
Class 2	0,049809271	0,059096456	0,058337434	0,054729065	0,060446732	0,056483792
Class 3	0,071850308	0,051399018	0,115008084	0,089912035	0,07009789	0,079653467
Class 4	0,05239085	0,044974141	0,057504042	0,047203819	0,076669568	0,055748484
Class 5	0,062869019	0,044974141	0,057504042	0,047203819	0,061335654	0,054777335
Class 6	0,05239085	0,01499138	0,028752021	0,062938425	0,015333914	0,034881318
Class 7	0,073347189	0,059965521	0,100632074	0,047203819	0,046001741	0,065430069
Class 8	0,073347189	0,074956902	0,028752021	0,047203819	0,061335654	0,057119117
Class 9	0,055883573	0,019988507	0,09584007	0,020979475	0,081780872	0,054894499



Γράφημα 8 DBScan αλγόριθμος για την κατηγορία hashtag C, με 25% training set

DBScan Algorithm 25% Training Set (Category D)

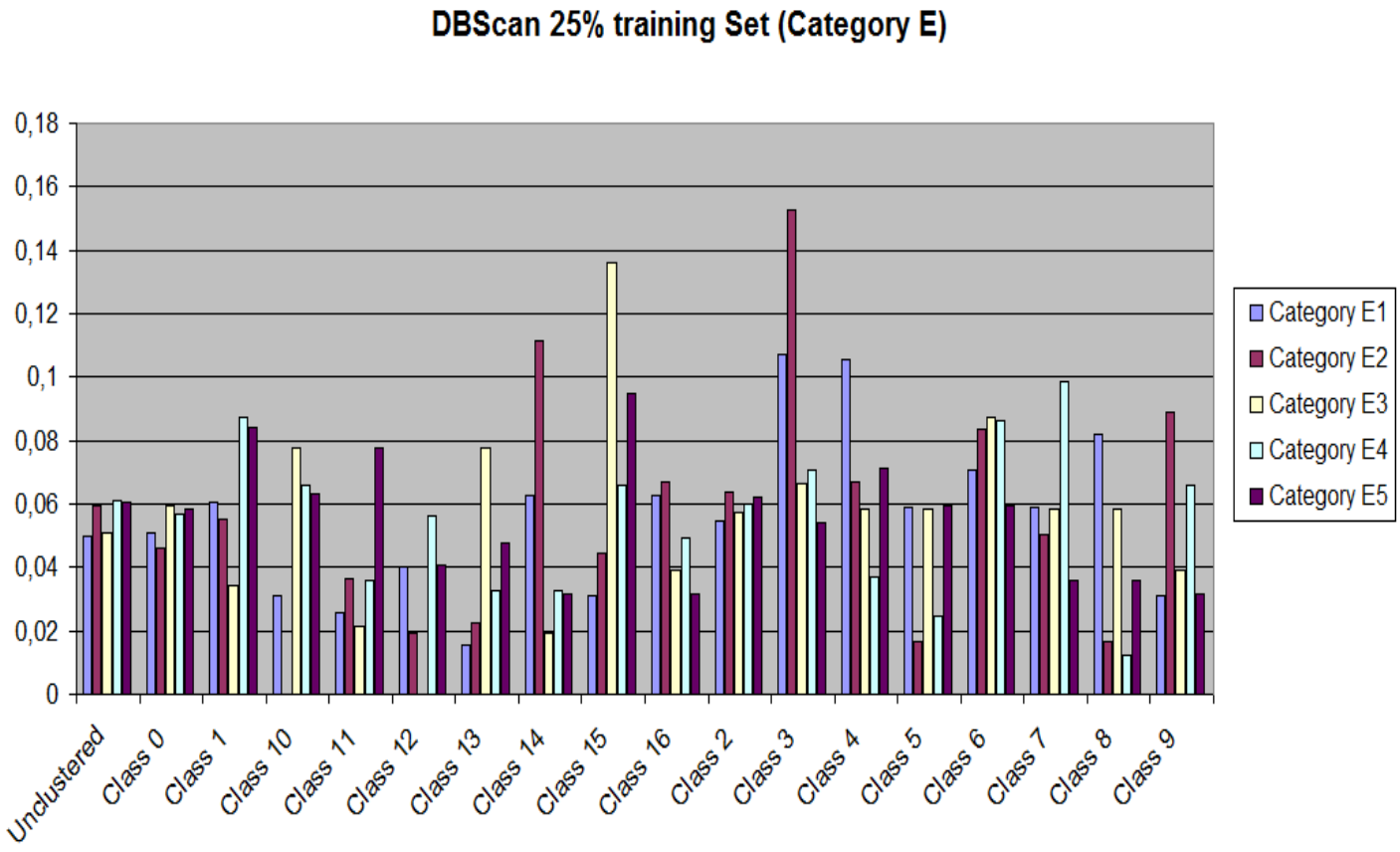
DBScan_25% Training Set (Category D)						
	Category D1	Category D2	Category D3	Category D4	Category D5	Μέσο Ποσοστό Επιτυχίας Ταύτισης παραγόμενων κλάσεων με αρχικές κατηγορίες hashtag
Unclustered	0,066321527	0,064714869	0,06989717	0,059891882	0,051054165	0,062375922
Class 0	0,050251272	0,052681854	0,064754027	0,05031833	0,064241113	0,056449319
Class 1	0,065863104	0,032366629	0,058628345	0,052996336	0	0,041970883
Class 10	0,023326516	0,09170545	0,094922083	0,025026048	0,02683851	0,052363721
Class 11	0,063617771	0,050021155	0,038831761	0,068252857	0,043917561	0,052928221
Class12	0,139959097	0,078604672	0,040680893	0,042901796	0,046008874	0,069631066
Class 13	0,023326516	0	0,023730521	0,050052095	0,053677019	0,03015723
Class 14	0,023326516	0,045852725	0,047461041	0,050052095	0,080515529	0,049441581
Class 15	0,023326516	0,09170545	0,023730521	0,050052095	0,02683851	0,043130618
Class 16	0,069979548	0,137558175	0,023730521	0,050052095	0,080515529	0,072367174
Class 2	0,046653032	0,051833515	0,061905706	0,054404451	0,065345936	0,056028528
Class 3	0,05998247	0,039302336	0,101702232	0,064352694	0,092017747	0,071471496
Class 4	0,034989774	0,068779088	0,035595781	0,075078143	0,100644411	0,063017439
Class 5	0,087474436	0	0,053393672	0,075078143	0,020128882	0,047215026
Class 6	0,034989774	0	0,035595781	0,018769536	0,080515529	0,033974124
Class 7	0,104969323	0,034389544	0,088989453	0,093847678	0,020128882	0,068464976
Class 8	0,034989774	0,068779088	0,088989453	0,018769536	0,040257764	0,050357123
Class 9	0,046653032	0,09170545	0,047461041	0,10010419	0,107354038	0,078655551



Γράφημα 9 DBScan αλγόριθμος για την κατηγορία hashtag D, με 25% training set

DBScan Algorithm 25% Training Set (Category E)

DBScan_25% Training Set (Category E)						
	Category E1	Category E2	Category E3	Category E4	Category E5	Μέσο Ποσοστό Επιτυχίας Ταύτισης παραγόμενων κλάσεων με αρχικές κατηγορίες hashtag
Unclustered	0,049958045	0,059211432	0,050914884	0,061267428	0,060591734	0,056388705
Class 0	0,050966472	0,045837982	0,059581247	0,056726448	0,058392535	0,054300937
Class 1	0,060783543	0,055013939	0,03431763	0,087128857	0,083863992	0,064221593
Class 10	0,031312735	0	0,077786629	0,065830692	0,063363905	0,047658792
Class 11	0,02561951	0,036437804	0,021214535	0,03590765	0,077764793	0,039388858
Class12	0,04025923	0,019086469	0	0,056426308	0,040733939	0,031301189
Class 13	0,015656367	0,022267547	0,077786629	0,032915346	0,047522929	0,039229764
Class 14	0,062625469	0,111337734	0,019446657	0,032915346	0,031681953	0,051601432
Class 15	0,031312735	0,044535094	0,1361266	0,065830692	0,095045858	0,074570196
Class 16	0,062625469	0,066802641	0,038893314	0,049373019	0,031681953	0,049875279
Class 2	0,05445693	0,063898178	0,057494465	0,060106284	0,061986429	0,059588457
Class 3	0,107357947	0,15269175	0,066674253	0,070532885	0,054311919	0,090313751
Class 4	0,105680479	0,066802641	0,058339971	0,037029764	0,071284393	0,06782745
Class 5	0,058711377	0,01670066	0,058339971	0,02468651	0,059403661	0,043568436
Class 6	0,070453653	0,083503301	0,087509957	0,086402784	0,059403661	0,077454671
Class 7	0,058711377	0,050101981	0,058339971	0,098746038	0,035642197	0,060308313
Class 8	0,082195928	0,01670066	0,058339971	0,012343255	0,035642197	0,041044402
Class 9	0,031312735	0,089070188	0,038893314	0,065830692	0,031681953	0,051357776



Γράφημα 10 DBScan αλγόριθμος για την κατηγορία hashtag E, με 25% training set

6. Επίλογος

Σε αυτό το σημείο θα συνοψίσουμε το περιεχόμενο της πτυχιακής εργασίας και θα εκθέσουμε τα συμπεράσματα μας σχετικά με το αντικείμενο της. Τέλος θα προτείνουμε μελλοντικές επεκτάσεις που επιδέχεται, προκειμένου να βελτιωθούν τα αποτελέσματα στα οποία οδηγούμαστε.

6.1. Σύνοψη και συμπεράσματα

Το Twitter από τη στιγμή που πρωτοεμφανίστηκε σαν υπηρεσία, έχει γνωρίσει τεράστια ανάπτυξη και είναι μια από τις πρωταρχικές υπηρεσίες social network. Η επίδραση που έχει στην κοινωνία είναι ενδεικτική του πόσο δημοφιλής υπηρεσίας κοινωνικής δικτύωσης είναι. Λόγω αυτής της μαζικής χρήσης του, τα δεδομένα που διακινούνται είναι πάρα πολλά και η επεξεργασία τους είναι προφανώς μεγάλης σημασίας.

Κινούμενοι προς αυτή την κατεύθυνση, στην παρούσα πτυχιακή εργασία, προσπαθήσαμε να συλλέξουμε ένα επαρκές σύνολο δεδομένων από το Twitter (tweets) που ικανοποιούσαν συγκεκριμένα κριτήρια. Τα κριτήρια αυτά ήταν η αναφορά του περιεχομένου, των tweets αυτών, σε συγκεκριμένες κατηγορίες hashtag που ορίστηκαν εξαρχής. Ο επιλεκτικός διαχωρισμός και η ειδικότερη επεξεργασία τους οδήγησε στο τελικό στάδιο της κατηγοριοποίησης-συσταδοποίησης του συνόλου των tweets από το πρόγραμμα WEKA.

Σκοπός της εργασίας μας αποτέλεσε η σύγκριση και η πιθανότητα ταύτισης των παραγόμενων κλάσεων από το WEKA με τις κατηγορίες hashtag εκ των οποίων αντλήθηκαν τα δεδομένα. Αξιολογώντας λοιπόν τα αποτελέσματα των αλγορίθμων κατηγοριοποίησης, αντιλαμβανόμαστε ότι η σύγκριση αυτή δεν στέφθηκε με την απόλυτη επιτυχία που αναμέναμε. Το γεγονός αυτό οφείλεται στον τρόπο που συντάσσουν τα tweets οι χρήστες του Twitter., ειδικότερα όταν παρατηρείται το φαινόμενο της σύνταξης tweets με αυθαίρετες λέξεις, άνευ σημασίας, σε άσχετα με

τη θεματική ενότητα hashtag.

Η αποτελεσματικότητα της έρευνας που πραγματοποιήθηκε στην πτυχιακή εργασία κρίνεται, εκ μέρους μας, επαρκής για την εξαγωγή συμπεράσματος, καθώς αντιλαμβανόμαστε ποια παραγόμενη κλάση αντιστοιχεί σε μια κατηγορία hashtag. Με τη χρήση των γραφημάτων, που δημιουργήσαμε από τα αποτελέσματα των αλγορίθμων συσταδοποίησης, η κατάληξη σε κάποιο συμπέρασμα αποτελεί εύκολη υπόθεση ακόμα και για κάποιον μη ειδικό στον τομέα της εξόρυξης δεδομένων, αρκεί να εντοπίσει σε κάθε γράφημα τη μέγιστη πιθανότητα ταύτισης για κάθε παραγόμενη κλάση.

6.2. Μελλοντικές Επεκτάσεις

Η έρευνα που πραγματοποιήσαμε επιδέχεται αρκετές αλλαγές για την βελτίωση των αποτελεσμάτων. Τα περιθώρια βελτίωσης εντοπίζονται κυρίως στο στάδιο της άντλησης των δεδομένων από το Twitter καθώς επίσης και στην προεπεξεργασία των δεδομένων προτού ακολουθήσει η συσταδοποίηση τους.

Αρχικά στο στάδιο της συλλογής δεδομένων το πρόγραμμα άντλησης που κατασκευάστηκε θα μπορούσε να πραγματοποιεί επικοινωνία με το Streaming API του Twitter και να αναζητά μεγαλύτερο όγκο πιο επίκαιρων δεδομένων. Αν και το σενάριο αυτό αποτελούσε αρχική σκέψη για την έρευνα που πραγματοποιήσαμε, εγκαταλείφτηκε εξαιτίας της επικύρωσης και αυθεντικοποίησης του χρήστη που απαιτούνταν. Η αυθεντικοποίηση αυτή, περιόριζε το σύνολο των request που μπορούσαμε να πραγματοποιήσουμε στην επικοινωνία μας με το Streaming API του Server του Twitter. Παρόλα αυτά, θεωρούμε πως η χρήση κατάλληλων περιορισμών στον αλγόριθμο του προγράμματος και η απεριόριστη δυνατότητα λειτουργίας του, χωρίς την πίεση του χρόνου, ενδεχομένως να οδηγήσει στην άντληση περισσότερων και καταλληλότερων δεδομένων.

Επίσης, να επισημάνουμε ότι το σύνολο των δεδομένων που αντλήθηκαν ανά κατηγορία hashtag ανέρχονταν στα 200 tweets για κάθε μια, με αποτέλεσμα να

συλλέξουμε ένα πλήθος 5000 tweets. Αν το σύνολο αυτό ήταν μεγαλύτερο, ενδεχομένως η εκπαίδευση που θα πραγματοποιούνταν στους αλγόριθμους συσταδοποίησης να ήταν και καλύτερη, αφού θα τους δίδονταν μεγαλύτερος αριθμός tweets ως training set και θα είχαν την δυνατότητα να εξάγουν καλύτερα αποτελέσματα.

Τα προβλήματα που εντοπίζονται γενικότερα στην ανάλυση των δεδομένων από το Twitter και ειδικότερα στο στάδιο της προ-επεξεργασίας των δεδομένων αυτών, οφείλονται, όπως αναφέρουμε και παραπάνω, στον τρόπο σύνταξης των tweet από τους χρήστες. Τα αποτελέσματα της πτυχιακής μας εργασίας προκύπτουν βάση της κατηγοριοποίησης στην οποία υπόκεινται τα παραπάνω tweets. Επομένως η χρήση tweets που εμφανίζουν προβλήματα στην σύνταξη τους από τους χρήστες, έχει ως αποτέλεσμα την αλλοίωση των αποτελεσμάτων και επομένως την έλλειψη εγκυρότητας. Η χρήση εργαλείων επεξεργασίας φυσικής γλώσσας (Natural Language Processing - NLP), σε συνδυασμό με τις μεθόδους επεξεργασίας που ήδη πραγματοποιούνται στην πτυχιακή μας εργασίας, ενδεχομένως να οδηγούσε σε καλύτερα και προφανέστερα αποτελέσματα.

Η άποψη αυτή στηρίζεται στη δυνατότητα εξαγωγής, που προσφέρουν τα NLP εργαλεία, πιο σύνθετων πληροφοριών από ένα tweet. Εφαρμόζοντας, συντακτική ανάλυση στο κειμενικό περιεχόμενο ενός tweet προκύπτουν πληροφορίες, όπως ποιο είναι το ρήμα, το υποκείμενο, το αντικείμενο ή και ορισμένοι προσδιορισμοί καθώς και πως αυτά συσχετίζονται μεταξύ τους. Γνωρίζοντας επομένως επιπλέον πληροφορίες για τη σημασιολογία ενός tweet, το στάδιο της δεικτοδότησης, που αποτελεί και από τα κυριότερα στάδια για την εξαγωγή έγκυρων αποτελεσμάτων, πραγματοποιείται βάση αυτών των σημασιολογικών πληροφοριών.

Παράρτημα Ι

Στο παρακάτω παράρτημα ακολουθεί η παράθεση του κώδικα του προγράμματος που αποτέλεσε τον αυτοματοποιημένο τρόπο συλλογής των tweets από το Twitter. Όπως αναφέραμε και στην ενότητα 5.1.1.1 η γλώσσα προγραμματισμού που χρησιμοποιήθηκε ήταν η Python και ειδικότερα η χρήση της βιβλιοθήκης που παρέχεται για Data Mining στο API του κοινωνικού δικτύου Twitter, που ονομάζεται Twython.

```
1  from twython import Twython
2
3  twitter = Twython()
4
5  search_results = twitter.search(q="#hashtag_name", rpp="200", until="Date_that_we_want(month/day/year)")
6
7  f = open ('hashtag_name.txt', 'w', 'r')
8
9      for tweet in search_results["results"]:
10         print "Tweet from @%s Date: %s" % (tweet['from_user'].encode('utf-8'),tweet['created_at'])
11         print tweet['text'].encode('utf-8'),"\n"
12         f.write(tweet['text'].encode('utf-8'))
13
14 f.close()
```

Στην πρώτη γραμμή του κώδικα με τη χρήση της εντολής ***«import»*** γίνεται η εισαγωγή της βιβλιοθήκης Twython στο πρόγραμμα, προκειμένου να γίνει η δυνατή η χρήση του συνόλου των εργαλείων που προσφέρει.

Στην πέμπτη γραμμή του κώδικα η χρήση της εντολής ***«twitter.search»*** ξεκινά τη διαδικασία της αναζήτησης στο Search API του Twitter, με τα εξής κριτήρια:

- Η αναζήτηση γινόταν για tweets που αναφέρονταν σε συγκεκριμένες κατηγορίες hashtag, που εμφανίζονται στην Εικόνα 9, στο κεφάλαιο 5.1. Το όνομα της κάθε κατηγορίας hashtag που αναζητούσαμε, γραφόταν κάθε φορά πριν την εκκίνηση του προγράμματος στη θέση

«"#hashtag_name"» που εντοπίζεται πάνω στον κώδικα.

- Σε κάθε μια κατηγορία hashtag, ο επιθυμητός αριθμός συλλογής tweets ήταν 200. Επομένως με τη χρήση της εντολής ***«rpp:200»*** σε κάθε σελίδα αποτελεσμάτων, που εμφάνιζε το πρόγραμμα μας στην επικοινωνία του με το Search API, εμφανίζονταν συνολικά 200 tweets.
- Ο τελευταίος περιορισμός επιτυγχάνεται και με τη χρήση της εντολής ***«until»*** και αφορά την ημερομηνία δημοσίευσης των tweets. Βασική προϋπόθεση στην πτυχιακή μας εργασία αποτέλεσε η επικαιρότητα του συνόλου των δεδομένων που θα αντλούσαμε. Επομένως θεωρήθηκε ως ορθό η συλλογή tweets που είχαν δημοσιευτεί 2 μέρες πριν την εκκίνηση του προγράμματος. Η ημερομηνία των 2 ημερών πρίν, όπως βλέπουμε και στον κώδικα μας, εισάγονταν με την μορφή ***" month/day/year".***

Στην έβδομη γραμμή του κώδικα πραγματοποιείται η δημιουργία ενός αρχείου απλού κειμένου κατάληξης .txt, το οποίο είχε το όνομα κάθε κατηγορίας hashtag που έτρεχε το πρόγραμμα.

Η εμφάνιση των δεδομένων (tweets) που συλλέγαμε εμφανίζονταν στην οθόνη του υπολογιστή μας με τις εντολές που εκτελούνται στις γραμμές δέκα και έντεκα του κώδικά μας. Η εμφάνιση των αποτελεσμάτων, για κάθε ένα tweet ξεχωριστά, γινόταν με την εξής μορφή:

- Στο επάνω μέρος εμφάνιζε τον τίτλο ***"Tweet from user Date: month/day/year"***, όπου στο πεδίο user εμφάνιζε το χρήστη που έκανε την δημοσίευση και στο πεδίο date την ημερομηνία δημοσίευσης.
- Ακριβώς κάτω από τον τίτλο που προαναφέρουμε εμφανίζονταν το κειμενικό περιεχόμενο του κάθε tweet, που είχε αντλήσει το πρόγραμμα μας.

Τέλος στην γραμμή 12 του κώδικα μας πραγματοποιούνταν η αποθήκευση των περιεχομένων του κάθε tweet, που είχε αντληθεί, στο αρχείο που είχαμε δημιουργήσει στην γραμμή 7 του κώδικα μας.

Βιβλιογραφία

[1]http://news.cnet.com/8301-1023_3-57541566-93/report-twitter-hits-half-a-billion-tweets-a-day/

[2] Wikipedia http://en.wikipedia.org/wiki/Social_media

[3] Μαλαματής Αλέξης, Κασφίκης Νικόλαος, Κατάνου Βασηλιάνα: «ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ»

[4] Βικιπαιδεια http://el.wikipedia.org/wiki/Twitter

[5] ^ "Twitter.com – Traffic Details from Alexa". Alexa Internet. August 26, 2010. Retrieved August 26, 2010.

[6] The Smartest People Prefer Twitter To LinkedIn And Facebook, Research Shows [STUDY] by Shea Bennett http://www.mediabistro.com/alltwitter/smart-twitter-users_b30041

[7] ^ a b Miller, Claire Cain (August 25, 2009). "Who's Driving Twitter's Popularity? Not Teens". *The New York Times*. Retrieved September 18, 2009.

[8] ^ Cheng, Alex; Evans, Mark (June 2009). "Inside Twitter – An In-Depth Look Inside the Twitter World".Sysomos. Retrieved February 23, 2011.

[9] http://www.zephoria.org/thoughts/archives/2009/08/16/twitter_pointle.html

[10] WordNet http://wordnetweb.princeton.edu/perl/webwn?s=data%20mining

[Baeza-Yates,99] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, Addison Wesley Longman Inc., 1999.

[Μακρής] Μακρής Χ., Σημειώσεις του μαθήματος "Ανάκτηση Πληροφορίας", Πανεπιστήμιο Πατρών, 2002.

[Salton,71] G. Salton. The SMART Retrieval System – Experiments in Automatic Document Processing. Prentice Hall Inc., 1971.

[Salton,68] G. Salton, M. E. Lesk. Computer evaluation of indexing and text processing,

Journal of the ACM, 15(1): 8-36, January 1968.

[Salton,88] G. Salton, C. Buckley. Term-weighting approaches in automatic retrieval, Information Processing & Management, 24(5): 513-523, 1988

[11] Βικιπαιδεια http://en.wikipedia.org/wiki/Cluster_analysis

[12] Wikibooks http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/Hybrid_Hierarchical_Clustering

[13] Wikipedia http://en.wikipedia.org/wiki/Hierarchical_clustering

[14] Standford University http://nlp.stanford.edu/IR-book/html/htmledition/hierarchical-agglomerative-clustering-1.html

[15] Xu and Wunsch || - 2005 http://axon.cs.byu.edu/Dan/678/papers/Cluster/Xu.pdf

[16] Βικιπαιδεια http://el.wikipedia.org/wiki/Συσταδοποίηση

[17] Οικονομάκου – Βεζιργιάννης A Review of Web Document Clustering Approaches

[18] Wikipedia http://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm

[19] Martin Ester, Hans-Peter Kriegel, J.Sander and Xiaowei Xu A destiny-base algorithm for discovering clusters in large spatial databases with noise.

[20] Wikipedia http://en.wikipedia.org/wiki/DBSCAN

[21] http://en.wikipedia.org/wiki/Open_Directory_Project

[22] https://dev.twitter.com/docs/api/1

[23] https://dev.twitter.com/docs/using-search

[24] https://dev.twitter.com/docs/streaming-apis

[25] http://en.wikipedia.org/wiki/Hypertext_Transfer_Protocol