Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής

Πρόγραμμα Μεταπτυχιακών Σπουδών
«Προηγμένα Συστήματα Πληροφορικής»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Ανάπτυξη Εφαρμογής MineTweet**
Ονοματεπώνυμο Φοιτητή	**Αθανάσιος Ρουμελιώτης του Νικολάου**
Αριθμός Μητρώου	**ΜΠΣΠ/08013**
Κατεύθυνση	**Συστήματα Υποστήριξης Αποφάσεων**
Επιβλέπων	**Νικόλαος Πελέκης, Λέκτορας**

Ημερομηνία Παράδοσης: **Απρίλιος 2012**

Τριμελής Εξεταστική Επιτροπή

Ν. Πελέκης (επιβλέπων) Λέκτορας	Ι. Σίσκος Καθηγητής	Ι. Θεοδωρίδης Αναπληρωτής Καθηγητής

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΡΟΛΟΓΟΣ – ΕΥΧΑΡΙΣΤΙΕΣ

Τα επιτεύγματα κάθε ανθρώπου είναι προϊόν προσωπικής προσπάθειας, αλλά πάντα υποβοηθούμενα από το περιβάλλον του. Δεν θα μπορούσε να συμβαίνει κάτι διαφορετικό με αυτή την μεταπτυχιακή διατριβή.

Ευχαριστώ τον επιβλέποντα καθηγητή μου Νίκο Πελέκη για την καθοδήγηση του στην παρούσα εργασία, όπως και όλους τους καθηγητές μου στο μεταπτυχιακό πρόγραμμα (αλλά και σε όλους τους προηγούμενους κύκλους εκπαίδευσης) οι οποίοι συνέβαλλαν άλλος πολύ και άλλος λίγο στην εξέλιξη μου. Σίγουρα από όλους κάτι έμαθα.

Στους συναδέλφους μου στην Νέσσος Πληροφορική, για την συμπαράσταση κατά διάρκεια του μεταπτυχιακού προγράμματος, αλλά και για την συνδρομή τους στην βελτίωση μου ως προγραμματιστής οφείλω ακόμα ένα ευχαριστώ. Είναι πολύ όμορφο να είσαι μέλος μιας ομάδας εργασίας στην οποία μπορείς να ανταλλάσσεις καθημερινά ιδέες και γνώσεις.

Τέλος στην οικογένεια μου μαζί με την αγάπη μου οφείλω και ευγνωμοσύνη. Σίγουρα ένα μεγάλο κομμάτι από όσα έχω καταφέρει έως σήμερα, αλλά και από όσα πρόκειται να ακολουθήσουν είναι και δικό τους.

Αφιερωμένο στους Έλληνες που προσπαθούν και αντιστέκονται

ΠΕΡΙΛΗΨΗ

Η εξάπλωση των κοινωνικών δικτύων σήμερα και η καθημερινή χρήση τους από την πλειονότητα των χρηστών του διαδικτύου είναι αδιαμφισβήτητη. Ένα από τα κοινωνικά δίκτυα με μεγάλη απήχηση στους χρήστες του διαδικτύου είναι και το Twitter. Όπως συμβαίνει και στα υπόλοιπα κοινωνικά δίκτυα, ένα προφίλ στο Twitter μπορεί να συνδεθεί με άλλα και έτσι δημιουργείται ο γράφος του. Τα προφίλ τα οποία είναι μέλη μεγάλων γράφων, ή που τα συνδεδεμένα με σημαντικά ενεργά προφίλ, δέχονται ένα πολύ μεγάλο όγκο πληροφοριών, τον οποίο είναι πολλές φορές δύσκολο να παρακολουθήσει ένας χρήστης. Σε αυτή τη μεταπτυχιακή διπλωματική διατριβή επιχειρείται η δημιουργία ενός νέου συστήματος με την ονομασία MineTweet, το οποίο αφού λαμβάνει τα μηνύματα ενός λογαριασμού του Twitter εφαρμόζοντας τις κατάλληλες τεχνικές συσταδοποίησης θα παρουσιάζει τα δεδομένα με έναν πιο πρακτικό τρόπο.

ABSTRACT

The spread of social networks today and their daily usage by many Internet users is indisputable. One of the social networks with great impact to Internet users is Twitter. Like other social networks, a profile on Twitter can be linked to other, thus creating a graph of profiles. Profiles that are members of large graphs, or associated with profiles that are significantly active, receive a large volume of information which is often hard to be tracked by a user. Object of this post-graduate diploma thesis is the development of a new application called MineTweet which targets to deliver and present Twitters data in a more convenient way, after applying the appropriate clustering techniques.

1 ΕΙΣΑΓΩΓΗ

1.1 ΚΙΝΗΤΡΟ

Η ραγδαία εξέλιξη της τεχνολογίας και των υποδομών έχει οδηγήσει στην καθημερινή ενασχόληση όλο και περισσότερων ανθρώπων με τις υπηρεσίες που παρέχει το διαδίκτυο. Τα τελευταία χρόνια ειδικότερα, παρατηρείται μια επίσης ραγδαία αύξηση των ανθρώπων που χρησιμοποιούν τα κοινωνικά δίκτυα. Η βασική ιδέα των σελίδων κοινωνικής δικτύωσης είναι ότι ο κάθε χρήστης της πλατφόρμας μπορεί να δημιουργήσει ένα δίκτυο με άλλα μέλη της πλατφόρμας. Τα μέλη ενός τέτοιου γράφου δύνανται να ανταλλάσουν πληροφορίες μεταξύ τους ενώ πολλά μέλη παραθέτουν πληροφορίες οι οποίες είναι προσβάσιμες σε όλους χωρίς περιορισμούς.

Ο όγκος της πληροφορίας που διακινείται καθημερινά στα κοινωνικά δίκτυα είναι τεράστιος και η γνώση που μπορεί να αποκτήσει κάποιος που μπορεί να εκμεταλλευτεί τις δυνατότητες που παρέχουν οι τεχνικές εξόρυξης γνώσης είναι αντίστοιχα πολύτιμή. Αύτη η δυνατότητα εξόρυξης γνώσης από ένα τόσο μεγάλο όγκο πληροφορίας η οποία προέρχεται από διαφορετικές και ποικιλόμορφες πηγές έχει τραβήξει το ενδιαφέρον της επιστημονικής κοινότητας. Έτσι πολλές ερευνητικές εργασίες έχουν γραφεί και αντίστοιχα εργαλεία έχουν αναπτυχθεί με σκοπό να εξάγουν γνώση από τις σελίδες κοινωνικής δικτύωσης.

Μια πλατφόρμα κοινωνικής δικτύωσης είναι και το δημοφιλές Twitter. Το Twitter έχει μια πολύ ισχυρή δυναμική όσον αφορά την εξάπλωση του και την χρήση του από όλο και περισσότερους χρήστες. Ακόμα παρουσιάζει αρκετές ιδιομορφίες σε σχέση με τις υπόλοιπες σελίδες κοινωνικής δικτύωσης ως προς το περιεχόμενο του, αλλά και ως προς τον τρόπο που παρουσιάζεται αυτή η πληροφορία. Τα δύο αυτά στοιχεία οδήγησαν στην απόφαση ότι θα ήταν εξαιρετικά ενδιαφέρουσα η υλοποίηση ενός εργαλείου το οποίου θα κατηγοριοποιούσε την πληροφορία που υπάρχει στο Twitter και θα την παρουσίαζε στον τελικό χρήστη σε ένα πιο εύχρηστο και φιλικό περιβάλλον.

1.2 ΣΚΟΠΟΣ

Στην παρούσα μεταπτυχιακή διατριβή θα υλοποιηθεί ένα νέο σύστημα με την επωνυμία MineTweet το οποίο θα δίνει την δυνατότητα στους χρήστες του Twitter να παρακολουθούν το ρεύμα μηνυμάτων που καταφθάνει στο λογαριασμό τους με έναν πιο πρακτικό τρόπο. Το νέο σύστημα θα παρέχει σε πρώτη φάση δυο βασικές δυνατότητες στους χρήστες του Twitter. Η πρώτη δυνατότητα θα είναι αυτή της αναζήτησης σε μεγαλύτερο πεδίο χρόνου. Η άλλη και ίσως η πιο σημαντική δυνατότητα θα είναι τρόπος παρουσίασης των μηνυμάτων ενός λογαριασμού. Πιο συγκεκριμένα εκτός από τον καθιερωμένο τρόπο παρουσίασης των μηνυμάτων με αύξουσα χρονικά ταξινόμηση, τα μηνύματα θα παρουσιάζονται και κατάλληλα ομαδοποιημένα με την χρήση των κατάλληλων τεχνικών εξόρυξης γνώσης.

Για την υλοποίηση του συστήματος θα μελετηθούν αντίστοιχες έρευνες και υλοποιήσεις και θα εξεταστεί ποια τεχνική είναι καταλληλότερη για την ομαδοποίηση των tweets. Με βάση αυτή την επιλεγμένη τεχνική θα δημιουργηθεί ένα σύστημα το οποίο θα ομαδοποιεί τα tweets σε κατηγορίες. Ακόμα θα σχεδιαστεί και υλοποιηθεί ένα interface το οποίο θα παρουσιάζει τα αποτελέσματα της κατηγοριοποίησης στον τελικό χρήστη, αλλά θα του δίνει και την βασική λειτουργικότητα της πλατφόρμας. Η γλώσσα στην οποία θα δοθεί βαρύτητα είναι η ελληνική και με βάση αυτή θα σχεδιαστούν οι κατάλληλες τεχνικές που θα επεξεργάζονται το κείμενο, ώστε ο αλγόριθμός που θα εξάγει τα αποτελέσματα να έχει την μέγιστη δυνατή απόδοση.

Τέλος το σύστημα που θα υλοποιηθεί θα αξιολογηθεί πειραματικά ώστε να διαπιστωθεί η ακρίβεια των αποτελεσμάτων που θα παρέχει στον χρήστη.

1.3 ΔΟΜΗ

Η παρούσα εργασία έχει δομηθεί ως ακολούθως: Στο επόμενο κεφάλαιο παρουσιάζονται εργασίες που ερευνούν θέματα σχετικά με την παρούσα διατριβή, καθώς και υλοποιήσεις παρεμφερών συστημάτων. Στη συνέχεια υπάρχει μια αναλυτική παρουσίαση του συστήματος Minetweet. όπου θα δούμε πως έχει σχεδιαστεί, ποιες τεχνολογίες χρησιμοποιήθηκαν για την υλοποίηση, αλλά και πως λειτουργεί το σύστημα στον πυρήνα του. Ειδικότερα θα παρουσιαστεί η επικοινωνία του συστήματος με το Twitter, η διαχείριση των δεδομένων, οι τεχνικές κατηγοριοποίησης των δεδομένων και η web διεπαφή. Συνεχίζοντας θα παρουσιαστεί η διαδικασία αξιολόγησης των αποτελεσμάτων. Τέλος θα γίνει μια επισκόπηση των συμπερασμάτων και των στόχων που επετεύχθησαν και θα προταθούν ιδέες για την μελλοντική επέκταση του συστήματος.

2 TWITTER

2.1 ΠΕΡΙΓΡΑΦΗ

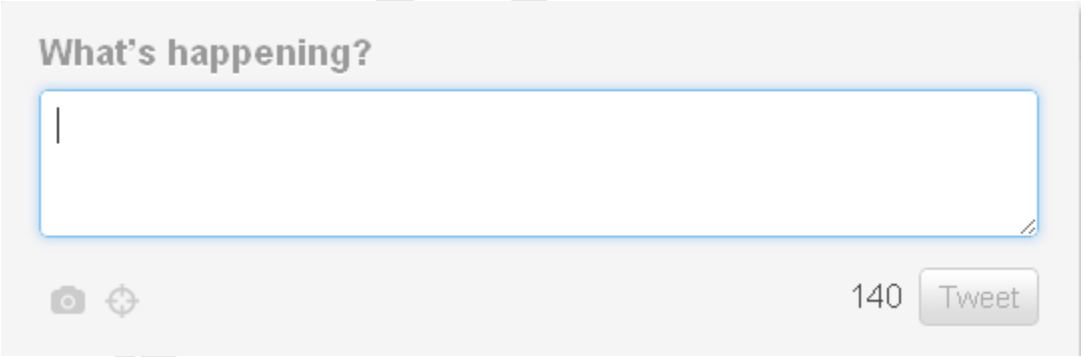
Το Twitter είναι μια υπηρεσία κοινωνικού δικτύου και microblogging η οποία δίνει την δυνατότητα στους χρήστες του να ανταλλάσουν μηνύματα κειμένου μεγέθους έως 140 χαρακτήρων τα οποία και ονομάζει tweets. Αυτά τα μηνύματα είναι εξ ορισμού διαθέσιμα προς όλους, άλλα δίνεται η δυνατότητα να υπάρχουν προσωπικές ανταλλαγές μηνυμάτων. Η υπηρεσία έκανε την εμφάνιση της τον Ιούλιο του 2006 και έγινε άμεσα δημοφιλής με συνεχώς αυξανόμενες τάσεις όσον αφορά την αποδοχή της από τους χρήστες του διαδικτύου.

2.2 ΤΟ TWITTER ΣΕ ΑΡΙΘΜΟΥΣ

Σύμφωνα με τα στοιχεία που έχει δημοσιεύσει το Twitter [1] η χρήση του συνεχίζει να αυξάνεται σε όλο τον κόσμο με πολύ υψηλό ρυθμό. Ενδεικτικά ενώ το 2010 αποστέλλονταν μέσω της υπηρεσίας 65.000.000 tweets την ήμερα, σήμερα αποστέλλονται πάνω από 200 εκατομμύρια tweets ανά ημέρα, ενώ ο συνολικός αριθμός των εγγεγραμμένων χρηστών ξεπερνά τα 100.000.000. Σημαντικό είναι και το ενδιαφέρον που έχει προκαλέσει στην κοινότητα της πληροφορικής αφού έχουν ήδη αναπτυχθεί πάνω από ένα εκατομμύριο εφαρμογές οι οποίες λειτουργούν με βάση το Twitter. Είναι προφανές με βάση τα παραπάνω στοιχεία ότι το Twitter έχει πλέον καθιερωθεί ως ένα από τα δημοφιλέστερα κοινωνικά δίκτυα και ο όγκος της πληροφορίας που διακινείται καθημερινά μέσω αυτού είναι τεράστιος.

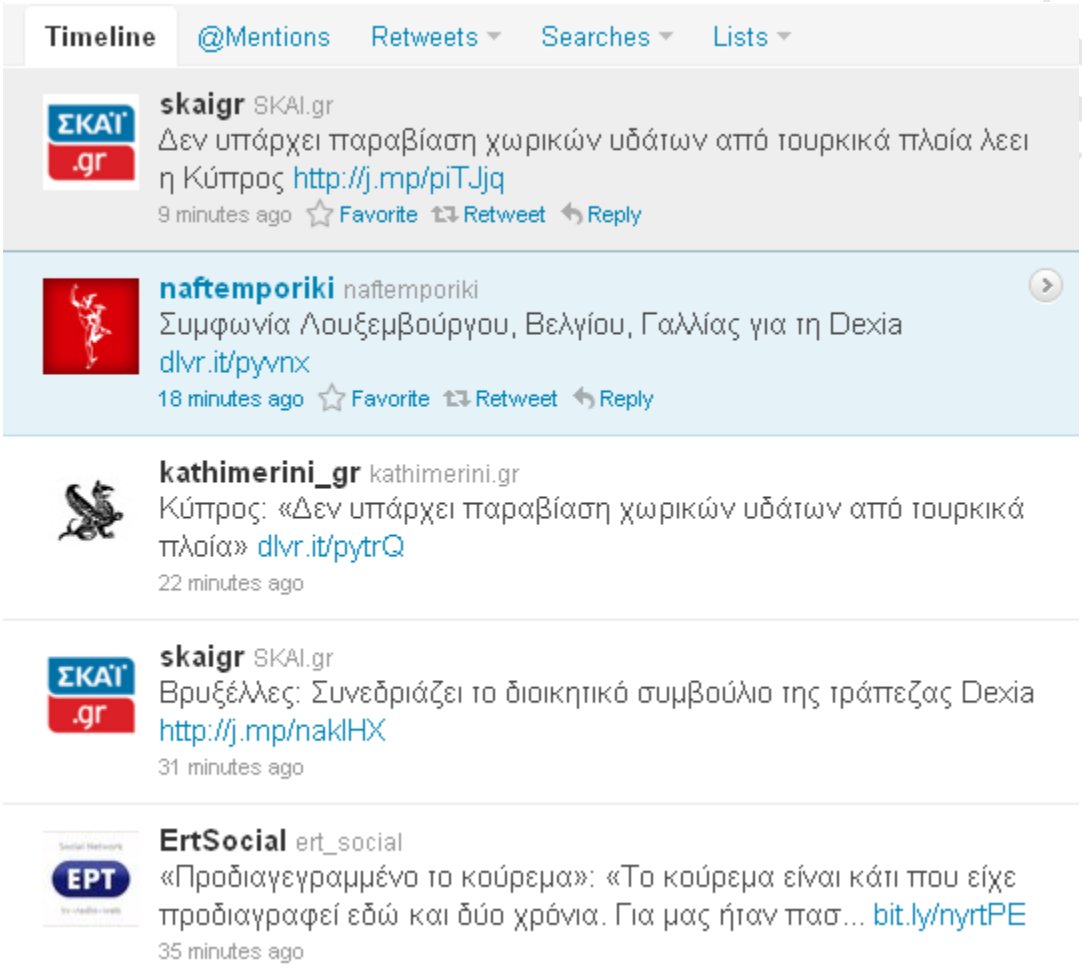
2.3 ΔΙΕΠΑΦΗ

Όσον αφορά την διεπαφή του Twitter, είναι μάλλον μινιμαλιστική, καθώς τα στοιχεία που κυριαρχούν σε αυτή είναι τα άκρως απαραίτητα για την χρήση του. Καταρχήν υπάρχει μια περιοχή όπως φαίνεται και στην Εικόνα 2.3.1 στην οποία ο χρήσης μπορεί να εισάγει ένα νέο μήνυμα.



Εικόνα 2.3.1: Η διεπαφή για την εισαγωγή νέου tweet

Ένα σημαντικό μέρος της βασικής διεπαφής καταλαμβάνει ο χώρος στον οποίο παρουσιάζονται τα tweets τα οποία λαμβάνει ο χρήστης. Τα tweets αυτά εμφανίζονται ταξινομημένα κατά φθίνουσα χρονολογική σειρά. ενώ υπάρχει και η δυνατότητα να παρουσιαστούν tweets τα οποία είναι προωθήσεις άλλων tweets (retweets), ή ακόμα και να παρουσιαστούν μηνύματα απαντήσεις στα tweets ενός χρήστη τα οποία είναι γνωστά και ως replies ή @mentions, όπως φαίνεται και στην Εικόνα 2.3.2.



Εικόνα 2.3.2: Το home timeline ενός χρήστη

Τέλος μια από τις βασικές λειτουργίες που παρέχονται στον χρήστη είναι η δυνατότητα αναζήτησης tweets.Έτσι υπάρχει ένα πεδίο με τη χρήση του οποίου ένας χρήστης μπορεί να αναζητήσει προηγούμενα μηνύματα τα οποία περιέχουν κάποιο από τους όρους της αναζήτησης, όπως φαίνεται στην Εικόνα 2.3.3.

Εικόνα 2.3.3: Το πεδίο αναζήτησης του Twitter

Ένα από τα καλύτερα χαρακτηριστικά του συστήματος αναζήτησης είναι η ικανότητά του να επιστρέφει τα τελευταία μηνύματα για οποιουσδήποτε όρους του τεθούν σε πραγματικό χρόνο, Αυτή η ικανότητα του συστήματος είναι ίσως και ένα από τα μεγάλα του μειονεκτήματα, αφού στην πραγματικότητα το σύστημα αναζήτησης έχει ένα χρονικό περιορισμό. Με άλλα λόγια έχει την δυνατότητα να επιστρέφει μόνο σχετικά πρόσφατα μηνύματα και συγκεκριμένο αριθμό. Σύμφωνα με την τεκμηρίωση του Twitter API [2], το όριο αυτό επί του παρόντος ορίζεται σε 6 με 9 ημέρες, αλλά είναι δυναμικό και με την επιφύλαξη να συρρικνώνεται, καθώς ο αριθμός των tweets ανά ημέρα συνεχίζει να αυξάνεται.

Μια από τις βασικές δυσκολίες στη χρήση του Twitter είναι η παρακολούθηση των εισερχόμενων μηνυμάτων σε λογαριασμούς χρηστών οι οποίοι είτε είναι συνδεδεμένοι με πολλούς χρήστες, είτε είναι οι λογαριασμοί με τους οποίους είναι συνδεδεμένοι είναι εξαιρετικά ενεργοί, δημιουργώντας έτσι μεγάλο όγκο μηνυμάτων στη λίστα με τα εισερχόμενα μηνύματα του χρήστη. Τα δύο παραπάνω στοιχεία είναι αυτά τα οποία θα προσπαθήσει να βελτιώσει η υλοποίηση του συστήματος Minetweet.

3 ΣΧΕΤΙΚΕΣ ΕΡΓΑΣΙΕΣ

3.1 ΕΙΣΑΓΩΓΗ

Το πλεονέκτημα που έχουν οι ιστοσελίδες κοινωνικής δικτύωσης να μην έχουν γεωγραφικό περιορισμό όσον αφορά την δημιουργία επαφών μεταξύ ανθρώπων οδήγησε στην αποδοχή τους από την εμφάνιση τους. Έτσι τα τελευταία χρόνια οι ιστοσελίδες κοινωνικής δικτύωσης έχουν μπει για τα καλά στην ζωή πολλών ανθρώπων και ο ρυθμός με τον οποίο αυξάνουν οι χρήστες που αποκτούν λογαριασμό σε μία η περισσότερες σελίδες κοινωνικής δικτύωσης, παραμένει αμείωτος.

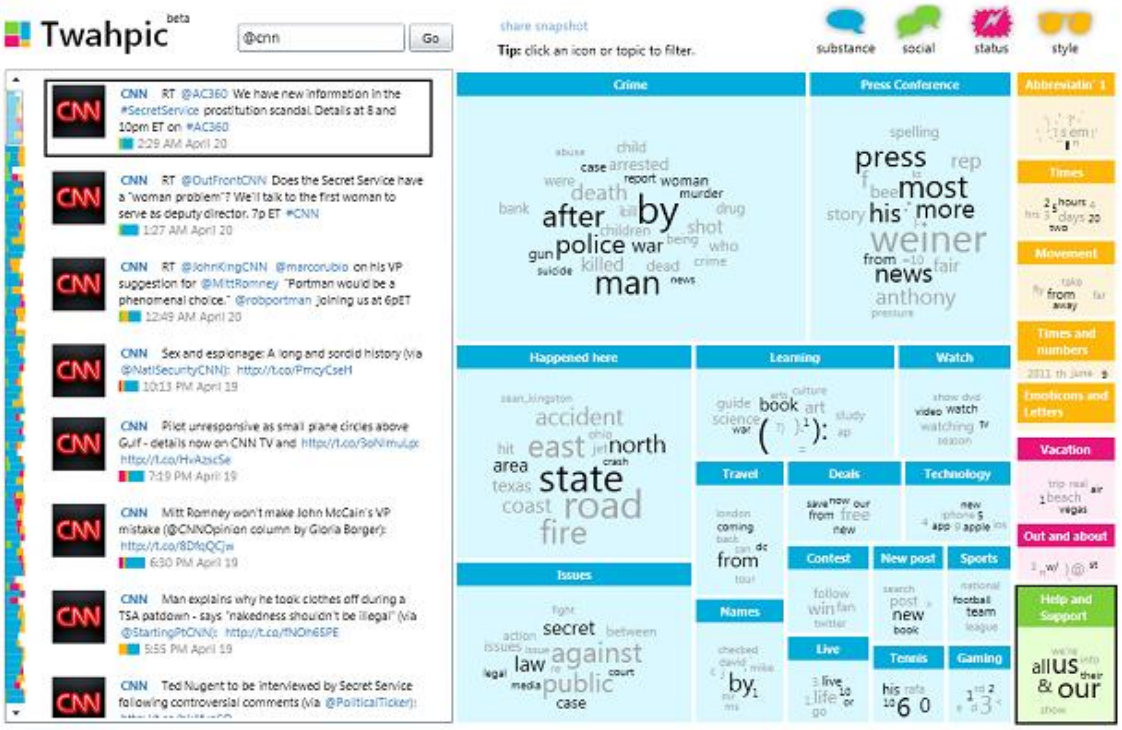
Καθώς λοιπόν τα κοινωνικά δίκτυα είναι τόσο προσιτά σε μια τόσο μεγάλη βάση ανθρώπων η πληροφορία που διακινείτε μεταξύ των μελών των κοινωνικών δικτύων είναι τεράστια Αυτή ακριβώς η ευχέρεια των κοινωνικών δικτύων να είναι τόσο προσιτά και να δέχονται πληροφορίες από ένα τόσο διευρυμένο, όσο και ανομοιογενές σύνολο ανθρώπων, κέντρισε το ενδιαφέρον της ακαδημαϊκής κοινότητας αλλά και της ίδιας της αγοράς που εντόπισαν ένα πολύτιμο μέσο για να εξορύξουν γνώση.

Τα κοινωνικά δίκτυα προσφέρουν πολλά πεδία στα οποία μπορεί να γίνει ανάλυση και να εξορυχτεί γνώση, με την χρήση διάφορων αλγορίθμων. Μερικά από αυτά τα πεδία είναι η ανάλυση γράφων και ο εντοπισμός σχέσεων μεταξύ μελών των κοινωνικών δικτύων, ο εντοπισμός και η ανάλυση κειμένων και ή ανάλυση περιεχομένου όπως φωτογραφιών. Ο τεράστιος όγκος των δεδομένων που είναι προς ανάλυση, αποτελεί ένα ακόμα πρόβλημα που πρέπει να λυθεί καθώς οι ανάγκη για πιο αποτελεσματικούς υπολογισμούς είναι επιτακτική.

Σε αυτή την ενότητα θα παρουσιαστούν μερικές εργασίες που έχουν γίνει στα παραπάνω πεδία έρευνας στα κοινωνικά δίκτυα και ειδικότερα στο Twitter, καθώς έγινε από νωρίς αντιληπτό ότι η ανάλυση αυτής της πληροφορίας θα ήταν δυνατό να αξιοποιηθεί σε πολλά πεδία όπως η κοινωνικές επιστήμες, το marketing, και η πολιτική.

3.2 ΕΡΕΥΝΗΤΙΚΕΣ ΕΡΓΑΣΙΕΣ

Η προσπάθεια για να ανάλυση των κοινωνικών δικτύων έχει οδηγήσει σε σημαντικές εργασίες και υλοποιήσεις. Οι Naaman, Boase και Lai [3] χρησιμοποιώντας τεχνικές κατηγοριοποίησης καταλήγουν στην εξαγωγή ενός συνόλου κατηγοριών. Στη συνέχεια χωρίζοντας τους χρήστες σε informers και meformers (χρήστες που συνήθως αναρτούν θέματα σχετικά με τους ίδιους) αναλύουν περαιτέρω τις κατηγορίες των κειμένων για να κατανοήσουν τις τάσεις των χρηστών όσον αφορά τα κείμενα που αναρτούν. Οι Ramage, Dumais και Liebling [4] χρησιμοποιώντας μια παραλλαγή του LDA την οποία ονομάζουν Labeled LDA και η οποία περιγράφεται αναλυτικά από τους Ramage et al [5], διαχωρίζουν το περιεχόμενο του Twitter σε 4 βασικές διαστάσεις, οι οποίες είναι το περιεχόμενο, το στυλ, το status και το κοινωνικά χαρακτηριστικά της κάθε ανάρτησης. Χρησιμοποιώντας αυτό το μοντέλο κατηγοριοποιούν τους χρήστες και τα tweets και παρουσιάζουν τα αποτελέσματα ως κατηγορίες ανά διάσταση. Η πειραματική υλοποίηση της συγκεκριμένης έρευνας ονομάζεται Twahpic (Εικόνα 3.2.1: Το σύστημα Twahpic , έχει υλοποιηθεί από την Microsoft Research.



Εικόνα 3.2.1: Το σύστημα Twahpic

Ο Christopher Horn [6] προτείνει ένα σύστημα το οποίο χρησιμοποιώντας μάθηση με επίβλεψη και ποιο συγκεκριμένα τον αλγόριθμό Support Vector Machine (SVM) κατηγοριοποιεί ένα σύνολο από tweets τρεις βασικές κατηγορίες.

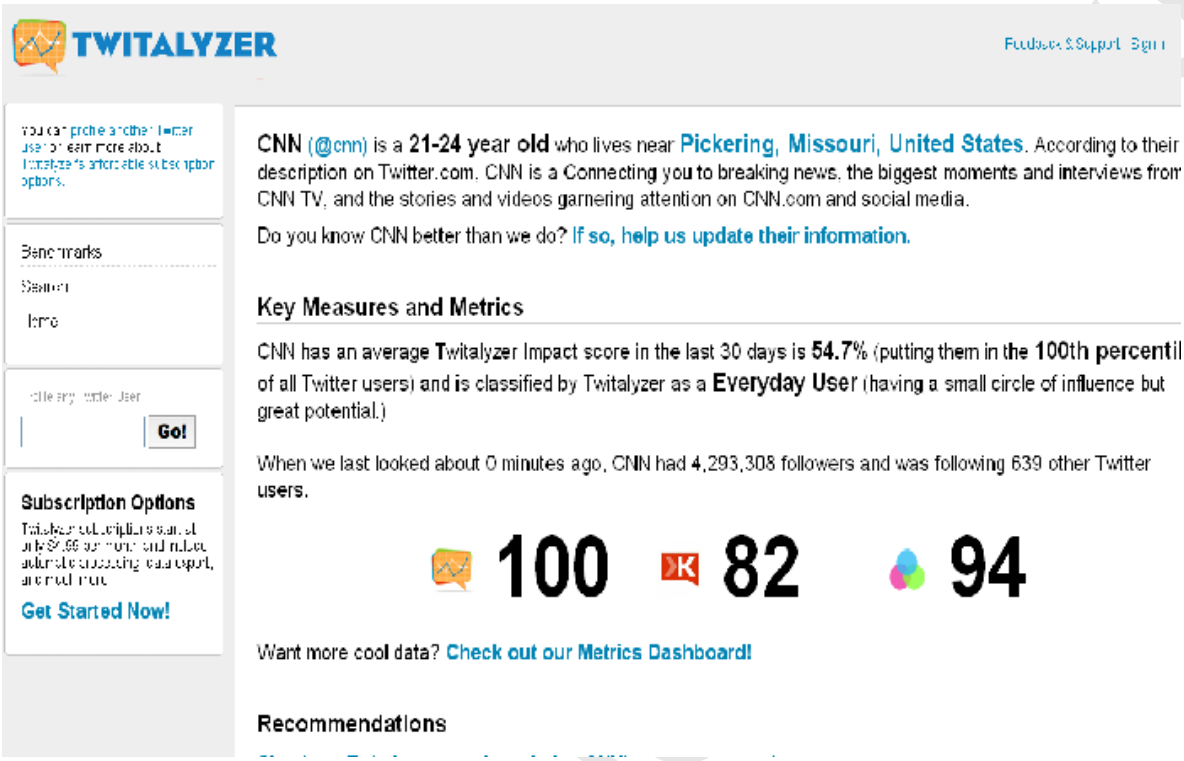
Η επόμενη ενδιαφέρουσα εργασία είναι αυτή των Achrekar et al [7] στην οποία χρησιμοποιούνται δεδομένα του Twitter στα οποία γίνεται classification με βάση το αν έχουν όρους σχετικούς με επιδημία γρίπης. Από αυτή την επεξεργασία προκύπτουν κάποιες τάσεις σχετικά με την πιθανότητα ξεσπάσματος επιδημίας, Στη συνέχεια συγκρίνονται τα δεδομένα αυτά με τα επίσημα reports των κρατικών υπηρεσιών υγείας και το αποτέλεσμα είναι παρεμφερές. Έτσι προκύπτει ως αποτέλεσμα η δυνατότητα πρόβλεψης επιδημίας σε σχεδόν πραγματικό χρόνο και σίγουρα σε καλύτερο χρονικό πλαίσιο από τις προβλέψεις που εκδίδουν οι οργανισμοί υγείας.

Σχετικά με την κατηγοριοποίηση ειδήσεων έχουν γίνει αρκετές εργασίες. Σε μια από αυτές ο Markos Katelanis [8] υλοποιεί ένα σύστημα κατηγοριοποίησης ειδήσεων από ειδησεογραφικά portals με την χρήση του αλγορίθμου συσταδοποίησης k-means. Μια συνδυαστική εργασία είναι αυτή των Phelan et al [9] στην οποία χρησιμοποιώντας τα tweets κάποιου χρήστη στο twitter επιχειρείται να δοθεί μια λίστα από προτεινόμενα RSS feeds προς τον χρήστη.

Ενδιαφέρουσες υλοποιήσεις είναι το Archivist Εικόνα 3.2.2: Το σύστημα Archivist me το οποίο μπορεί κάποιος να δημιουργήσει ένα αρχείο από tweets με βάση ερωτήματα τα οποία περιέχουν συγκεκριμένους όρους ή χρήστες. Από αυτό το αρχείο εξάγονται και παρουσιάζονται σε μορφή γραφημάτων ενδιαφέροντα στοιχεία όπως η συχνότητα εμφάνισης εγγραφών στο αρχείο διαχρονικά ή οι λέξεις με τις περισσότερες εμφανίσεις στο συγκεκριμένο αρχείο.

Εικόνα 3.2.2: Το σύστημα Archivist

Το σύστημα Twitalyzer Εικόνα 3.2.3 δίνει αναλυτικά στοιχεία για ένα χρήστη του Twitter. Ενδεικτικά μερικά από αυτά είναι δημογραφικά όπως ο αριθμός των followers και εκτιμήσεις σχετικά με το φύλο και την ηλικία αυτών. Παρουσιάζονται ακόμα δείκτες όπως ο δείκτης επιρροής ο οποίος προκύπτει από τον αριθμό των followers σε συνδυασμό με τα retweets που γίνονται σε tweets του χρήστη καθώς και τις αναφορές που εμφανίζονται για τον συγκεκριμένο χρήστη. Τέλος ένας ενδιαφέρον δείκτης δίνει την βαρύτητα του συγκεκριμένου λογαριασμού υπολογίζοντας την συχνότητα των tweets του χρήστη σε συνδυασμό με τα στοιχεία του δείκτη επιρροής.



Εικόνα 3.2.3΄Το Σύστημα Twitalyser

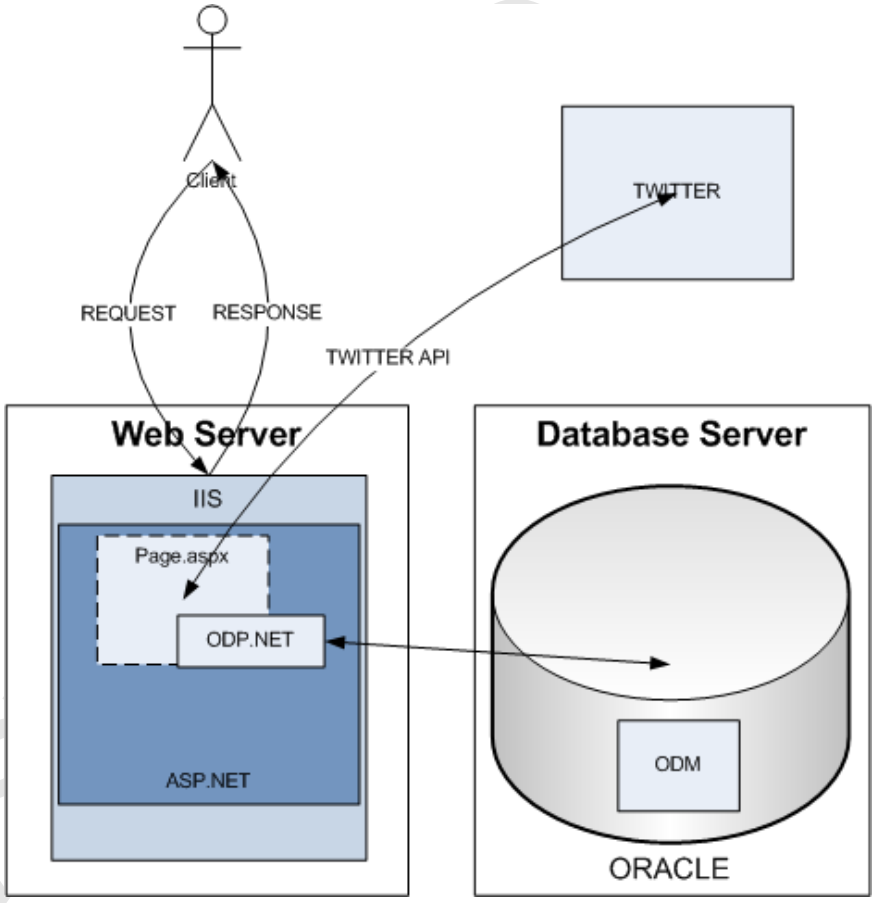
Οι παραπάνω ερευνητικές εργασίες αποτελούν ένα μόνο μικρό μέρος από τις πολλές αξιόλογες εργασίες που έχουν γίνει σχετικά με το Twitter. Κάτι που δεν εντοπίσαμε κατά την έρευνα ήταν ένα εξατομικευμένο σύστημα το οποίο να κατηγοριοποιεί δυναμικά τα tweets που λαμβάνει ο κάθε χρήστης. Την λύση αυτή προσπαθεί να δώσει το σύστημα που παρουσιάζουμε εδώ και του οποίου την υλοποίηση θα δούμε στο επόμενο κεφάλαιο.

4 ΑΝΑΠΤΥΞΗ ΕΡΓΑΛΕΙΟΥ MINETWEET

4.1 ΑΡΧΙΤΕΚΤΟΝΙΚΗ

Το προτεινόμενο σύστημα θα πρέπει να προσφέρει στο χρήστη τις εξής λειτουργικότητες. Καταρχήν θα πρέπει να δίνει τη δυνατότητα να παρακολουθούν όλοι οι χρήστες ακόμα και οι μη πιστοποιημένοι την ροή των public tweets που εισάγονται στο Twitter. Εάν ένας χρήστης πιστοποιηθεί στο σύστημα τότε θα έχει την λειτουργικότητα που του παρέχει το Twitter, και περαιτερω την δυνατότητα να παρακολουθεί τα μηνύματα των λογαριασμών που ακολουθεί καθώς και να δημιουργεί νέα tweets.

Η Εικόνα 4.1.1παρουσιάζει την γενική αρχιτεκτονική της εφαρμογής. Ο χρήστης επικοινωνεί μέσω HTTP πρωτοκόλλου με τον Web Server ο οποίος επεξεργάζεται την κλήση εσωτερικά και με την σειρά του επικοινωνεί με την βάση δεδομένων Oracle η οποία έχει εγκατασταθεί στον Database Server σε περίπτωση που αυτό είναι αναγκαίο, μέσω του driver ODP.NET. Ακόμα η εφαρμογή επικοινωνεί το Twitter μέσω του Twitter API από το οποίο αντλεί δεδομένα τα οποία είτε εμφανίζει στον χρήστη είτε τα αποθηκεύει στην βάση δεδομένων. Όπως θα δούμε και στην συνέχεια το Oracle RDBMS περιέχει και το πακέτο ODM (Oracle Data Mining) με την χρήση του οποίου θα υλοποιηθούν όλες οι λειτουργίες εξόρυξης γνώσης.



Εικόνα 4.1.1 Απεικόνιση της γενικής αρχιτεκτονικής της εφαρμογής Minetweet.

4.2 ΤΕΧΝΟΛΟΓΙΕΣ

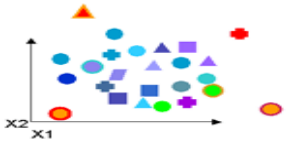
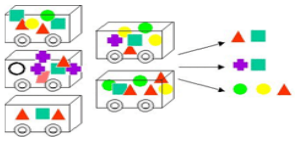
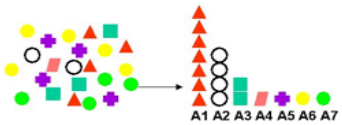
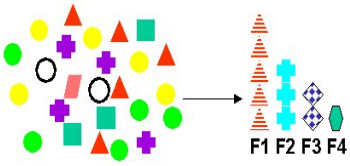
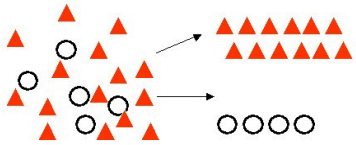
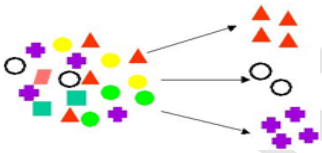
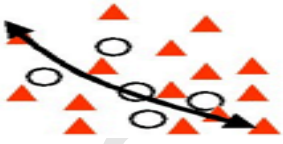
Για την υλοποίηση του συστήματος θα χρησιμοποιηθούν οι τεχνολογίες και οι τεχνικές προγραμματισμού που περιγράφονται στις επόμενες υποενότητες. Πρόκειται για ευρέως διαδομένες τεχνολογίες και τεχνικές προγραμματισμού που δίνουν την δυνατότητα για γρήγορη και αποτελεσματική δημιουργία εφαρμογών (rapid application development).

4.2.1 ORACLE DATABASE

Για την υλοποίηση του συστήματος επιλέχθηκε ως σύστημα βάσης δεδομένων το Oracle ORDBMS. Το σύστημα βάσης δεδομένων της Oracle αποτελεί ένα από τα κορυφαία συστήματα βάσεων δεδομένων προσφέροντας αποδεδειγμένη επεκτασιμότητα ανεξάρτητα από την υπάρχουσα υποδομή και μπορεί να χρησιμοποιηθεί για τη διαχείριση πολύ μεγάλων όγκων πληροφοριών, παρέχοντας παράλληλα πολύ υψηλά επίπεδα ασφάλειας. Ένας από τους βασικούς λόγους που οδήγησαν σε αυτή την επιλογή είναι και το ότι η έκδοση της Oracle 11g R2 Enterprise Edition περιλαμβάνει ενσωματωμένο και το πακέτο Oracle Data Mining.

Το πακέτο Oracle Data Mining (ODM) προσφέρει ισχυρή λειτουργικότητα εξόρυξης δεδομένων ως εγγενείς SQL συναρτήσεις στο εσωτερικό της βάσης δεδομένων. Με το γραφικό περιβάλλον Oracle Data Miner δίνεται η δυνατότητα να ερευνηθούν δεδομένα, να δημιουργηθούν και να αξιολογηθούν μοντέλα εξόρυξης γνώσης, ενώ με την χρήση του SQL API γίνεται εφικτή η ενσωμάτωση της τεχνολογίας σε εφαρμογές οι οποίες δίνουν την δυνατότητα για εκτέλεση αλγορίθμων εξόρυξης γνώσης παρέχοντας πληροφορίες σε πραγματικό χρόνο. Τα πλεονεκτήματα που παρουσιάζει η συγκεκριμένη λύση είναι ότι τα δεδομένα, τα μοντέλα και τα αποτελέσματα παραμένουν στην βάση δεδομένων οπότε η πληροφορία ανακτάται ταχύτατα καθώς δεν χρειάζεται μετακίνηση δεδομένων, ενώ παράλληλα διατηρείται και το υψηλό επίπεδο ασφάλειας των δεδομένων.

Καθώς η εφαρμογή που θα υλοποιηθεί θα εκτελεί λειτουργίες εξόρυξης γνώσης σε κείμενο έχει σημασία να δούμε σε αυτό το σημείο (Πίνακας 4.2.1) ποιες λειτουργίες εξόρυξης γνώσης από κείμενο υποστηρίζει το πακέτο Oracle Data Mining [10].σημειώνοντας ότι σαν εξόρυξη γνώσης από κείμενο εννοούμε την εκτέλεση αλγορίθμων σε λέξεις, ακολουθίες λέξεων και υλικό που μπορεί να εξαχθεί από έγγραφα.

ΛΕΙΤΟΥΡΓΙΑ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ	ORACLE DATA MINING
Ανίχνευση Ακραίων Σημείων (Anomaly Detection) 	Διαδικασία εξόρυξης γνώσης σε δεδομένα κειμένου ή μικτού τύπου δεδομένα, με χρήση του αλγόριθμου One Class Support Vector Machines.
Κανόνες Συσχέτισης (Association Rules) 	Διαδικασία εξόρυξης γνώσης σε δεδομένα κειμένου ή μικτού τύπου δεδομένα, με χρήση του αλγόριθμου Minimum Description Length.
Αξιολόγηση Χαρακτηριστικών (Attribute Importance) 	Διαδικασία εξόρυξης γνώσης σε δεδομένα κειμένου ή μικτού τύπου δεδομένα, με χρήση του αλγόριθμου Apriori.
Εξαγωγή Χαρακτηριστικών (Feature extraction) 	Διαδικασία εξόρυξης γνώσης σε δεδομένα κειμένου ή μικτού τύπου δεδομένα, με χρήση του αλγόριθμου Non-Negative Matrix Factorization.
Κατηγοριοποίηση (Classification) 	Διαδικασία εξόρυξης γνώσης σε δεδομένα κειμένου ή μικτού τύπου δεδομένα, με χρήση των αλγόριθμων Support Vector Machines, Generalized Linear Models, Naive Bayes.
Συσταδοποίηση (Clustering) 	Διαδικασία εξόρυξης γνώσης σε δεδομένα κειμένου ή μικτού τύπου δεδομένα, με χρήση του αλγόριθμου K-Means.
Παλινδρόμηση (Regression) 	Διαδικασία εξόρυξης γνώσης σε δεδομένα κειμένου ή μικτού τύπου δεδομένα, με χρήση των αλγόριθμων Support Vector Machines ή Generalized Linear Models.

Πίνακας 4.2.1: Δυνατότητες Εξόρυξης Γνώσης σε κείμενο με το πακέτο ODM

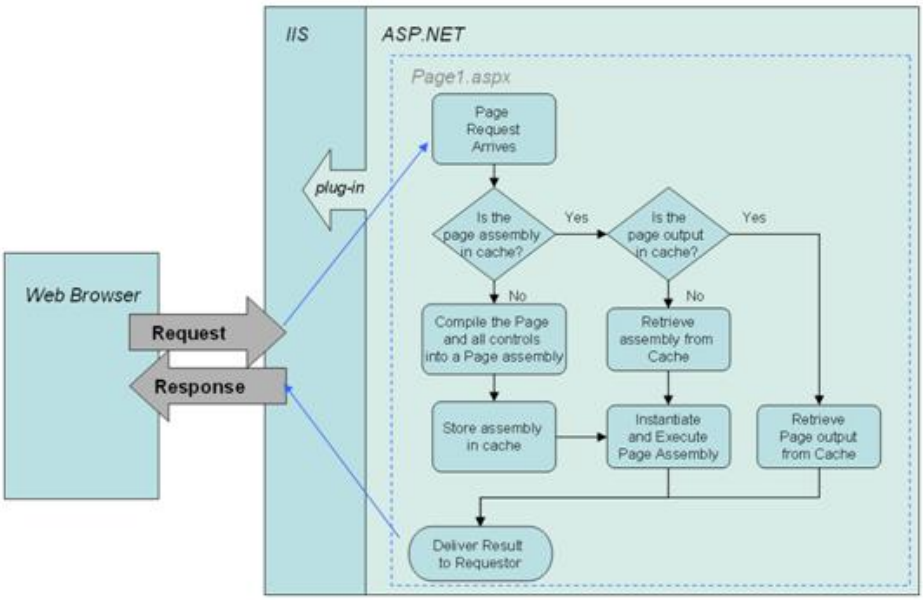
4.2.2 ASP.NET KAI C#

Τον Φεβρουάριο του 2002 η Microsoft παρουσίασε το .NET Framework ως ένα στοιχείο του λειτουργικού Microsoft Windows το οποίο χρησιμοποιείται για την δημιουργία και εκτέλεση

εφαρμογών Windows. Έκτοτε το συγκεκριμένο πακέτο λογισμικού έχει αναβαθμιστεί αρκετές φορές φτάνοντας στην τελευταία σταθερή έκδοση 4.0 τον Απρίλιο του 2010. Το .NET Framework είναι ένα πακέτο λογισμικού στο οποίο μπορούν να βασιστούν οι προγραμματιστές για να δημιουργήσουν οποιαδήποτε μορφή εφαρμογής. Το .NET Framework έχει ως βασικά στοιχεία το Κοινό Περιβάλλον Εκτέλεσης (Common Language Runtime) και μία Κοινή Βιβλιοθήκη (Base Class Library). Το.NET Framework υποστηρίζει την ανάπτυξη χρησιμοποιώντας διαφορετικές γλώσσες. Σε αυτές συμπεριλαμβάνονται οι C#, F#, Visual Basic και Managed C++, ενώ τρίτες εταιρείες παρέχουν επιπλέον γλώσσες για το .NET Framework.

Το ASP.NET είναι μέρος του .NET Framework, και παρέχει ένα τυποποιημένο περιβάλλον εκτέλεσης σε εφαρμογές του .NET Framework. Με άλλα λόγια, το ASP.NET παρέχει το περιβάλλον στο οποίο ενεργοποιούνται και εκτελούνται οι εφαρμογές Web που έχουν γραφτεί για το .NET Framework. Το ASP.NET, όμως, είναι παραπάνω από ένα απλό περιβάλλον: Είναι μια ολοκληρωμένη αρχιτεκτονική για την ανάπτυξη εφαρμογών Web. Οι εφαρμογές web που δημιουργούνται με το ASP.NET χρησιμοποιούν το μοντέλο προγραμματισμού, τα εργαλεία ανάπτυξης και τις πρακτικές εγκατάστασης και ρύθμισης που είναι κοινά για όλες τις εφαρμογές .NET. Οι εφαρμογές ASP.NET εκμεταλλεύονται όλες τις κοινές υπηρεσίες του CLR, περιλαμβανομένης και της υποστήριξης για διαφορετικές γλώσσες προγραμματισμού.

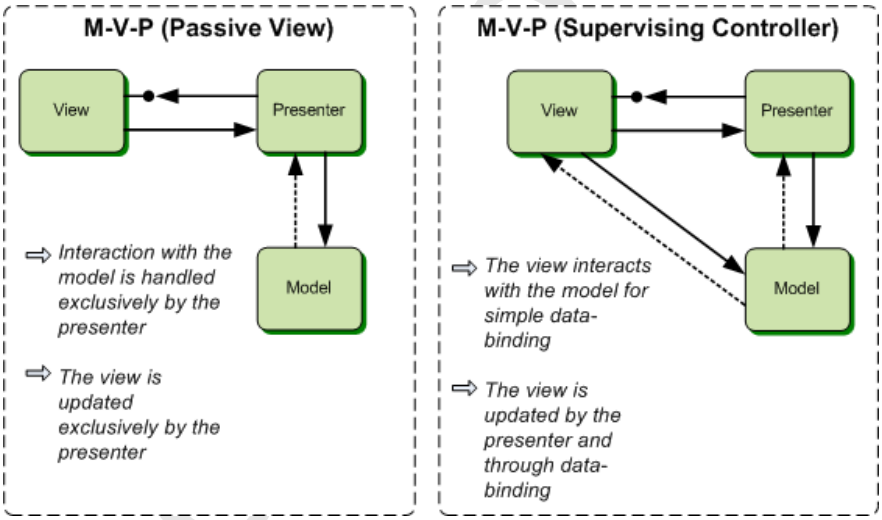
Όσον αφορά την διαδικασία εξυπηρέτησης κλήσης μιας σελίδας από το ASP.NET την πρώτη φορά που ένας browser ή μια συσκευή ζητάει μια σελίδα, το ASP.NET δημιουργεί αυτόματα ένα.NET assembly με τον εκτελέσιμο κώδικα της κλάσης Page που αντιστοιχεί στη σελίδα, συμπεριλαμβανομένων όλων των controls και της λογικής code-behind. Ο εκτελέσιμος κώδικας διατηρείται στη μνήμη για καλύτερη απόδοση. Σε αυτή την απλή περίπτωση η κλάση εκτελείται αμέσως μετά. Το ASP.NET δημιουργεί ένα αντικείμενο αυτής της κλάσης για κάθε κλήση της σελίδας. Το αντικείμενο αυτό ελέγχει τη δημιουργία της HTML για τη συγκεκριμένη κλήση (Εικόνα 4.2.1). Αν έχει ενεργοποιηθεί η λειτουργία cache για τα αποτελέσματα της σελίδας, όσες κλήσεις ακολουθήσουν θα εξυπηρετηθούν από την cache.



Εικόνα 4.2.1: Η διαδικασία εξυπηρέτησης κλήσης μιας σελίδας από το ASP.NET

4.2.3 MVP DESIGN PATTERN

Το MVP Design Pattern εμφανίστηκε την δεκαετία του 1990 από τον M. Potel [11], και είναι ένα design pattern το οποίο σχεδιάστηκε με σκοπό να διευκολύνει τον αυτοματοποιημένο έλεγχο μονάδων του κώδικα, αλλά και να βελτιστοποιήσει τον διαχωρισμό της λογικής από την παρουσίαση. Το μοντέλο περιλαμβάνει τρείς έννοιες και χωρίζει το μοντέλο προγραμματισμού σε τρία μέρη. Το πρώτο είναι το model, το δεύτερο το view και το τρίτο ο presenter. Το κάθε ένα μέρος του μοντέλου προγραμματισμού είναι υπεύθυνο για συγκεκριμένες λειτουργίες. Το model καθορίζει τα δεδομένα τα οποία θα εμφανίζονται και θα αλληλεπιδρούν με την διεπαφή του χρήστη. Το view είναι υπεύθυνο για την παρουσίαση των δεδομένων (model) στον χρήστη καθώς και για την δρομολόγηση εντολών από την διεπαφή του χρήστη προς τον presenter ώστε αυτός με την σειρά του να εκτελέσει οποιαδήποτε εργασία απαιτείται στα δεδομένα. Τέλος ο presenter επιδρά και στο model και στο view. Είναι υπεύθυνος για την ανάκτηση δεδομένων από την βάση δεδομένων, την αποθήκευση τους καθώς και για την τροποποίηση τους, ώστε να παρουσιάζονται ορθά στο view. Εν ολίγοις όλη η λογική βρίσκεται στον presenter. Ο βαθμός κατά τον οποίο επιτρέπεται να υπάρχει λογική στο view εξαρτάται από την υλοποίηση. Από τις δύο επικρατέστερες υλοποιήσεις (Passive View και Supervising Controller) Εικόνα 4.2.2 τις οποίες περιγράφει αναλυτικά ο Martin Fowler [12], καλύτερη στις περιπτώσεις Web εφαρμογών θεωρείται η υλοποίηση Supervising Controller, καθώς πολλές φορές κρίνεται σκόπιμο να διαχειριστεί μια λειτουργία στο View το οποίο εκτελείται στον φυλλομετρητή του χρήστη. Για περισσότερα design patterns μπορεί κανείς να ανατρέξει στα βιβλία [13] και [14].



Εικόνα 4.2.2: Διαφορετικές υλοποιήσεις του design pattern Model-View-Presenter

4.2.4 LINQ2TWITTER

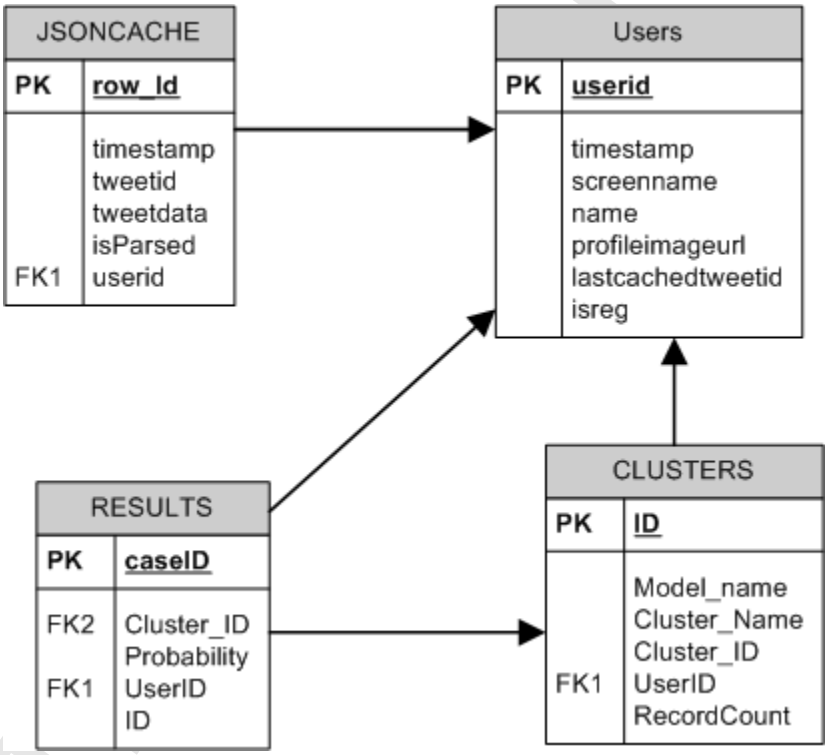
Για την επικοινωνία της εφαρμογής που θα υλοποιήσουμε με το Twitter θα χρησιμοποιηθεί το Twitter API. Το Twitter API είναι βασισμένο στην τεχνολογία REST (Representational State Transfer) την οποία εισήγαγε ο R. Fielding το 2000 [15], και η οποία είναι ένα πρωτόκολλο υλοποίησης Web Services. Η χρήση ενός REST Web Service υλοποιείται με μια HTTP κλήση μέσω ενός URL η οποία επιστρέφει το κείμενο της απάντησης συνήθως σε XML format.

Το Twitter API είναι μια ολοκληρωμένη πλατφόρμα η οποία προσφέρει πολλές εναλλακτικές διεπαφές προγραμματισμού με την οποία έχουμε την δυνατότητα να έχουμε πρόσβαση στα δεδομένα που θέλουμε να συμπεριλάβουμε στην εφαρμογή μας, καθώς και να δώσουμε στους χρήστες της εφαρμογής όλη την λειτουργικότητα που προσφέρει το Twitter.

Καθώς το ενδιαφέρον για την δημιουργία εφαρμογών σχετικών με το Twitter είναι συνεχώς αυξανόμενο και το «οικοσύστημα» του Twitter συνεχώς μεγαλώνει, το API συγχρόνως επεκτείνεται και περιλαμβάνει όλο και περισσότερες εναλλακτικές διεπαφές προγραμματισμού. Παράλληλα δημιουργούνται βιβλιοθήκες σε συγκεκριμένες γλώσσες προγραμματισμού που ενσωματώνουν μέρος της λειτουργικότητας που προσφέρει το API, δίνοντας έτσι την δυνατότητα στους προγραμματιστές να κάνουν κλήσεις προς το API χρησιμοποιώντας βιβλιοθήκες οι οποίες είναι γραμμένες στην γλώσσα προγραμματισμού που χρησιμοποιεί ο εκάστοτε προγραμματιστής. Για την υλοποίηση της εφαρμογής Minetweet θα χρησιμοποιήσουμε την βιβλιοθήκη Linq2Twitter. Το Linq2Twitter είναι μία ανοιχτού κώδικα βιβλιοθήκη η οποία χρησιμοποιεί την τεχνολογία LINQ του .NET Framework για να πραγματοποιήσει κλήσεις προς το twitter API. Χρησιμοποιώντας την συγκεκριμένη βιβλιοθήκη έχουμε το πλεονέκτημα μιας βιβλιοθήκης γραμμένης σε C# για την εκτέλεση των κλήσεων προς το Twitter API. Για περισσότερές πληροφορίες σχετικά με την βιβλιοθήκη μπορούν να βρεθούν στον δικτυακό τόπο codeplex.com [16].

4.3 ΒΑΣΗ ΔΕΔΟΜΕΝΩΝ

Για την υλοποίηση του συστήματος θα χρησιμοποιήσουμε το Oracle RDBMS όπως αναφέραμε και στην προηγούμενη ενότητα. Το σχήμα της βάσης που δημιουργήσαμε φαίνεται στην Εικόνα 4.3.1.



Εικόνα 4.3.1: Σχήμα της βάσης δεδομένων

Όπως φαίνεται στο παραπάνω σχήμα, οι βασικοί πίνακες της βάσης δεδομένων είναι οι εξής τέσσερις:

ΠΙΝΑΚΑΣ USERS		
COLUMN NAME	DATA TYPE	NULLABLE
ID	NUMBER(38,0)	No
USERID	NUMBER(38,0)	Yes
TIMESTAMP	TIMESTAMP(6)	Yes
SCREENNAME	VARCHAR2(100 BYTE)	Yes

Πίνακας 4.3.1: Πίνακας Users

Στον πίνακα Users (Πίνακας 4.2.1) αποθηκεύονται τα δεδομένα για τους χρήστες που έχουν πιστοποιηθεί στο σύστημα.

ΠΙΝΑΚΑΣ JSONCACHE		
COLUMN NAME	DATA TYPE	NULLABLE
ID	NUMBER(38,0)	No
TIMESTAMP	TIMESTAMP(9)	Yes
TWEETID	NUMBER(38,0)	Yes
TWEETDATA	VARCHAR2(500 CHAR)	Yes
ISPARSED	CHAR(1 CHAR)	Yes
USERID	NUMBER(38,0)	Yes
NATIVEDATA	VARCHAR2(500 CHAR)	Yes
CREATEDAT	TIMESTAMP(9)	Yes

Πίνακας 4.3.2: Πίνακας Jsoncache

Ο πίνακας JSONCACHE (Πίνακας 4.3.2) είναι ουσιαστικά ο πίνακας στον οποίο αποθηκεύονται τα tweets που με την βοήθεια του Twitter API λαμβάνονται από το timeline του κάθε πιστοποιημένου χρήστη.

ΠΙΝΑΚΑΣ CLUSTERS		
COLUMN NAME	DATA TYPE	NULLABLE
ID	NUMBER	No
MODEL_NAME	VARCHAR2(40 BYTE)	Yes
CLUSTER_NAME	VARCHAR2(1999 BYTE)	Yes
CLUSTER_ID	NUMBER	Yes
USERID	NUMBER	Yes
RECORD_COUNT	NUMBER	Yes

Πίνακας 4.3.3:Πίνακας Clusters

Ο πίνακας Clusters (Πίνακας 4.3.3) είναι ο πίνακας στον οποίο αποθηκεύονται οι πληροφορίες για το κάθε cluster που προκύπτει μετά την εκτέλεση του αλγορίθμου k-means με το οποίο γίνεται η συσταδοποίηση. Μερικά από τα στοιχεία που αποθηκεύονται είναι το αναγνωριστικό του Cluster, το όνομα του, το οποίο θα δούμε στην συνέχεια πως προκύπτει και ο αριθμός των εγγραφών που αποτελούν το συγκεκριμένο Cluster.

Ο τελευταίος σημαντικός πίνακας του συστήματος είναι ο πίνακας Results(Πίνακας 4.3.4).

ΠΙΝΑΚΑΣ RESULTS		
COLUMN NAME	DATA TYPE	NULLABLE
DMR$CASE_ID	NUMBER	No
ISPARSED	CHAR(4 BYTE)	Yes
TWEETDATA1	VARCHAR2(2000 BYTE)	Yes
USERID	NUMBER	Yes
NATIVEDATA	VARCHAR2(2000 BYTE)	Yes
ID	NUMBER	No
TWEETID	NUMBER	Yes
CLUSTER_ID	NUMBER	Yes
PROBABILITY	NUMBER	Yes

Πίνακας 4.3.4: Πίνακας Results

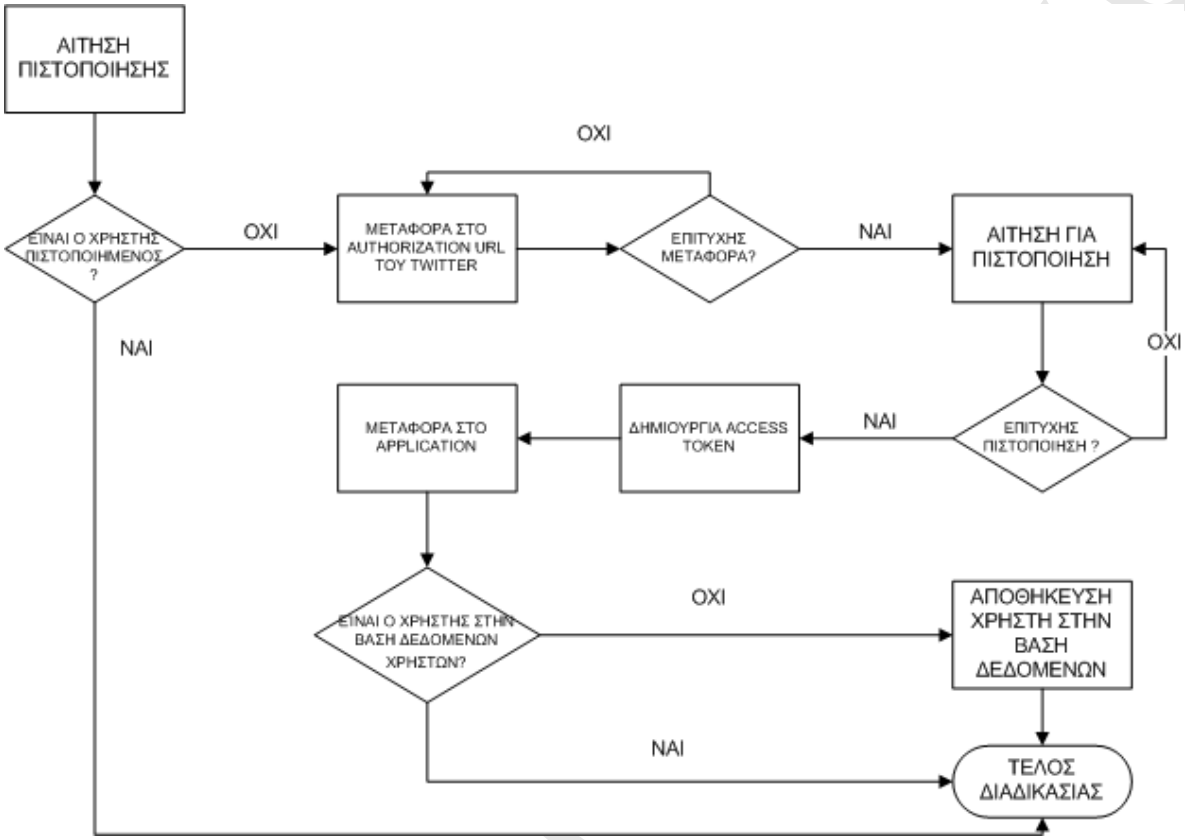
Ο συγκεκριμένος πίνακας είναι ο πίνακας στον οποίο συγκεντρώνονται τα αποτελέσματα της συσταδοποίησης. Μετά την ολοκλήρωση της συσταδοποίησης οι εγγραφές που αποτελούσαν την πηγή στην οποία εκτελέστηκε ο αλγόριθμος εισάγονται σε αυτόν το πίνακα μαζί με τα στοιχεία που περιέχουν το cluster στο οποίο ανήκει η κάθε εγγραφή καθώς και η πιθανότητα με την οποία η εκάστοτε εγγραφή έχει τοποθετηθεί σωστά στο συγκεκριμένο cluster.

4.4 AUTHENTICATION ΧΡΗΣΤΗ

Ξεκινώντας τν υλοποίηση του Minetweet αρχικά δημιουργούμε ένα Twitter Application στην σελίδα Twitter Developers. Η συγκεκριμένη διαδικασία είναι μια εύκολη διαδικασία μέσω ενός οδηγού η οποία όταν ολοκληρωθεί θα δημιουργήσει ένα consumer key και ένα secret key με τα οποία θα υπογράφονται τα requests της εφαρμογής μας προς την υπηρεσία.

Για την πιστοποίηση του χρήστη το Twitter χρησιμοποιείται το πρωτόκολλο OAuth [17], το οποίο δημιουργήθηκε από τον οργανισμό *Internet Engineering Task Force* ο ποίος είναι υπεύθυνος για την δημιουργία standards. Το πρωτόκολλο OAuth είναι ένα ανοιχτό πρωτόκολλο επικοινωνίας μεταξύ εφαρμογών. Στο ένα άκρο του πρωτοκόλλου είναι μια διαδικτυακή υπηρεσία, στην περίπτωση μας το Twitter ενώ στο άλλο άκρο μπορεί να βρίσκεται μια άλλη διαδικτυακή εφαρμογή, mobile εφαρμογή ή desktop εφαρμογή. Το συγκεκριμένο μοντέλο Authentication παρέχει αυξημένο επίπεδο ασφαλείας καθώς η πιστοποίηση του χρήστη γίνεται στην υπηρεσία πάροχο με αποτέλεσμα να μην γνωστοποιούνται τα στοιχεία πιστοποίησης (username και password) σε τρίτες εφαρμογές. Όπως γίνεται κατανοητό το επίπεδο ασφάλειας αυξάνεται και ανά πάσα στιγμή μέσω του Twitter η εξουσιοδότηση μπορεί να αρθεί.

Η διαδικασία πιστοποίησης χρήστη στην εφαρμογή Minetweet φαίνεται στην Εικόνα 4.4.1. Αρχικά ένας χρήστης κάνει μια αίτηση για πιστοποίηση και η εφαρμογή αρχικά εξετάζει αν ο αιτών είναι ήδη πιστοποιημένος. Σε περίπτωση που δεν είναι πιστοποιημένος τότε η εφαρμογή ανακατευθύνει τον browser του στην σελίδα του Twitter όπου γίνεται η πιστοποίηση ενημερώνοντας τον χρήστη για τα δικαιώματα πρόσβασης που θα αποκτήσει η εφαρμογή στο λογαριασμό του. Όταν η πιστοποίηση ολοκληρωθεί γίνεται ανακατεύθυνση στην σελίδα της εφαρμογής όπου τα κοινά στοιχεία του χρήστη αποθηκεύονται στον πίνακα με τους πιστοποιημένους χρήστες.



Εικόνα 4.4.1:Η διαδικασία Authentication του Minetweet.

Μόλις η παραπάνω διαδικασία ολοκληρωθεί η εφαρμογή μπορεί να ανακτήσει δεδομένα από το λογαριασμό του χρήστη καλώντας το Twitter API.

4.5 ΕΠΙΚΟΙΝΩΝΙΑ ΚΑΙ ΑΝΑΚΤΗΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕ ΤΟ TWITTER API

Για να είναι δυνατή η εξόρυξη γνώσης από τα δεδομένα, πρέπει αυτά να αποθηκευθούν στην βάση δεδομένων του συστήματος. Η ανάκτηση και αποθήκευση των δεδομένων επιτυγχάνεται με την χρήση του Twitter API. Για την επικοινωνία αυτή όπως αναφέραμε στην υποενότητα 4.2.4 χρησιμοποιείται η βιβλιοθήκη Linq2Twitter. Παρακάτω θα δούμε ένα παραδείγματα κλήσης προς το API.

Η κλήση που θα περιγράψουμε είναι προς την οντότητα Status και μετά την επιτυχή της εκτέλεση επιστρέφει τα tweets που είναι στην timeline του λογαριασμού, σύμφωνα με τις παραμέτρους που εισάγουμε κατά την κλήση. Οι παράμεροι που μπορούν να εισαχθούν στην ρουτίνα είναι οι εξής:

Όνομα	Σκοπός	Τύπος	Υποχρεωτικό
ContributorDetails	Πρόσθετες πληροφορίες χρήστη	bool	όχι
Count	Αριθμός Tweets που επιστρέφονται. Μέγιστο 200	int	όχι
ExcludeReplies	Αποκλεισμός απαντήσεων	bool	όχι

ID	ID ή ScreenName του χρήστη	string	Μόνο αν δεν εισαχθεί UserID ή ScreenName
MaxID	Επιστροφή tweets έως το ID	ulong	όχι
Page	Σελίδα που θα ανακτηθεί	int	όχι
ScreenName	Το sreenname του χρήστη	string	Μόνο αν δεν εισαχθεί ID ή UserID
SinceID	Επιστροφή tweets μετα από το ID	ulong	όχι
UserID	ID Χρήστη	string	Μόνο αν δεν εισαχθεί ID ή ScreenName

Πίνακας 4.5.1: Οι Παράμετροι για κλησεις στο Home Timeline

Παράδειγμα της κλήσης της παραπάνω μεθόδου στο σύστημα μας αποτελεί το παρακάτω απόσπασμα κώδικα με το οποίο αποθηκεύουμε τα πιο πρόσφατα tweets στην βάση δεδομένων

Απόσπασμα 4.5.1: Αποθήκευση πρόσφατων tweets

```
using (var session = sessionFactory.OpenSession())
{
        // populate the database
        using (var transaction = session.BeginTransaction())
        {
                var jsonCacheList = session.CreateCriteria<JsonCache>().Add(Restrictions.Eq("User.Id",
                TUserId)).List<JsonCache>();
                var maxID = (jsonCacheList.Count == 0) ? 0 : jsonCacheList.Max(id => id.TweetId);

                List<Status> last1000 = new List<Status>();
                for (int i = 1; i < 3; i++)
                {
                        last1000.AddRange((from tweet in twitterCtx.Status where (tweet.Type ==
                        StatusType.Friends) && tweet.Count == 200 && tweet.Page == I select
                        tweet).ToList());
                }
                var user = session.CreateCriteria(typeof(TUser)).Add(Restrictions.Eq("UserId",
                Int32.Parse(currentUser.UserID))).List<TUser>().First();
                foreach (var tweet in last1000.Where(aa => Int64.Parse(aa.StatusID) > maxID))
                {
                        var jsonCacheRow = new JsonCache { TimeStamp = DateTime.Now, CreatedAt =
                        tweet.CreatedAt, NativeData = tweet.Text, TweetData =
                        (tweet.Text.RemoveLink()).ToUpperGreek(), TweetId =
                        Int64.Parse(tweet.StatusID)};
                        jsonCacheRow.User = user;
                        session.SaveOrUpdate(jsonCacheRow);
                }
                transaction.Commit();
        }
}
```

Με την κλήση αντίστοιχων μεθόδων είναι δυνατό να εισάγουμε νέα tweets στο Twitter να διαγράψουμε κάποιο tweet και γενικότερα να κάνουμε οποιαδήποτε ενέργεια μπορεί να κάνει ένας χρήστης απευθείας στο Twitter.

4.6 ΠΡΟΕΤΟΙΜΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Πριν από την εκτέλεση του αλγορίθμου θα επιχειρήσουμε να προετοιμάσουμε τα δεδομένα ώστε αλγόριθμός να αποδώσει τα καλύτερα δυνατόν αποτελέσματα. Η προετοιμασία των δεδομένων χωρίζεται σε δύο διακριτά μέρη, αφενός τον καθαρισμό και μετασχηματισμό των δεδομένων και αφετέρου την δημιουργία εμφωλευμένων πινάκων για τα δεδομένα κειμένου. Ειδικά στο πρώτο στάδιο αφαιρούνται από τα δεδομένα URLs και μία λίστα κοινών λέξεων, στοιχεία τα οποία υπάρχει η πιθανότητα να οδηγήσουν τον αλγόριθμο να αποδώσει φτωχά αποτελέσματα. Παράλληλα γίνεται και ο κατάλληλος μετασχηματισμός τους ώστε ο αλγόριθμός να αποδώσει καλύτερα στην ελληνική γλώσσα. Στο δεύτερο στάδιο τα δεδομένα μετασχηματίζονται σύμφωνα με τις επιταγές του Oracle RDBMS ώστε να είναι εφικτή η χρήση τους ως είσοδος στον αλγόριθμο συσταδοποίησης.

4.6.1 ΚΑΘΑΡΙΣΜΟΣ ΔΕΔΟΜΕΝΩΝ

Κατά την είσοδο των δεδομένων στην βάση μετατρέπουμε το κείμενο σε κεφαλαία, αφενός για να αποφύγουμε πιθανή εμπλοκή του αλγόριθμου από λέξεις που είναι στην ελληνική γλώσσα λόγω του τονικού συστήματος και αφετέρου για να περιορίσουμε τον αριθμό των λέξεων που δεν θα λάβει υπόψη του ο αλγόριθμος. Για να επιτύχουμε την κεφαλαιοποίηση αλλά και την αφαίρεση των τόνων χρησιμοποιούμε τις δύο μεθόδους που ακολουθούν.

Απόσπασμα 4.6.1:Κεφαλαιοποίηση Κειμένου

```
public static string ToUpperGreek(this string inputstring)
{
        return RemoveDiacritics(inputstring.ToUpper());
}

public static string RemoveDiacritics(string stIn)
{
        string stFormD = stIn.Normalize(NormalizationForm.FormD);
        StringBuilder sb = new StringBuilder();

        for (int ich = 0; ich < stFormD.Length; ich++)
        {
                UnicodeCategory uc = CharUnicodeInfo.GetUnicodeCategory(stFormD[ich]);
                if (uc != UnicodeCategory.NonSpacingMark)
                {
                        sb.Append(stFormD[ich]);
                }
        }
        return (sb.ToString().Normalize(NormalizationForm.FormC));
}
```

Η επόμενη εργασία που θα εκτελέσουμε είναι η αφαίρεση όλων των υπερσυνδέσμων που υπάρχουν στο κείμενο. Αυτό επιτυγχάνεται με την χρήση regular expressions. Με την χρήση αυτή της τεχνικής θα αντικαταστήσουμε κάθε γραμματοσειρά που εμπίπτει στην φόρμα του regular expression με μια κενή γραμματοσειρά (Απόσπασμα 4.6.2)

Απόσπασμα 4.6.2: Αφαίρεση υπερσυνδέσμων από το κείμενο

```
public static string RemoveLink(this string txt)
{
        Regex regx = new Regex("http(s)?://([\\w+?\\.\\w+])+([a-zA-Z0-9\\~\\!\\@\\#\\$\\%\\^\\&amp;\\*\\(\\)_\\-
\\=\\+\\\\\\/\\?\\.\\:\\;\\'\\,]*)?", RegexOptions.IgnoreCase);
```

```
        return regx.Replace(txt, String.Empty);
}
```

4.6.2 ΑΦΑΙΡΕΣΗ ΚΟΙΝΩΝ ΛΕΞΕΩΝ

Για την επίτευξη καλύτερων αποτελεσμάτων χρησιμοποιείται μια λίστα από κοινές λέξεις (stoplist) οι οποίες δεν λαμβάνονται υπόψη κατά την εκτέλεση του αλγορίθμου. Η σημασία της λίστας είναι καίριας σημασίας για τα αποτελέσματα του αλγόριθμού, λόγω του περιορισμένου αριθμού λέξεων που περιέχεται σε κάθε tweet. Παράλληλα βελτιώνεται και ο χρόνος εκτέλεσης καθώς τα δεδομένα που εισάγονται στον αλγόριθμο μειώνονται σημαντικά. Για την εισαγωγή της λίστας στην βάση έγινε χρήση της μεθόδου CTX_DDL.CREATE_STOPLIST της Oracle παράδειγμα χρήσης της οποίας παρουσιάζεται στο Απόσπασμα 4.6.3. Η συνολική λίστα κοινών λέξεων είναι διαθέσιμή στο ΠΑΡΑΡΤΗΜΑ Α. ΚΟΙΝΕΣ ΛΕΞΕΙΣ (STOPLIST)

Απόσπασμα 4.6.3: Δημιουργία Stoplist

```
BEGIN
CTX_DDL.DROP_STOPLIST('GREEKSTOPLIST');
        CTX_DDL.CREATE_STOPLIST('GREEKSTOPLIST', 'BASIC_STOPLIST');
        CTX_DDL.ADD_STOPWORD('GREEKSTOPLIST', 'ΚΑΙ');
        CTX_DDL.ADD_STOPWORD('GREEKSTOPLIST', 'ΑΠΟ');
END;
```

4.6.3 ΔΗΜΙΟΥΡΓΙΑ ΕΜΦΩΛΕΥΜΕΝΩΝ ΠΙΝΑΚΩΝ ΜΕ TERMS ΚΕΙΜΕΝΟΥ

Για να εφαρμοστούν οι λειτουργίες εξόρυξης γνώσης της τεχνολογίας Oracle Data Mining σε κείμενο, πρέπει πρώτα αυτό να μετασχηματιστεί σε έναν εμφωλευμένο πίνακα από terms. Επομένως σε αυτό το στάδιο εκτελούνται οι απαραίτητες διαδικασίες για τον μετασχηματισμό των δεδομένων text σε κατάλληλη μορφή, ώστε να είναι δυνατή η επεξεργασία τους από τους αλγόριθμους εξόρυξης γνώσης. Ο εμφωλευμένος πίνακας αποτελείται από δύο χαρακτηριστικά. Στο ένα είναι αποθηκευμένα τα terms των κειμένων και στο άλλο ο βαθμός σημαντικότητας κάθε term σε σχέση με τα υπόλοιπα terms του κειμένου. Ο εμφωλευμένος πίνακας αποτελεί ένα χαρακτηριστικό του πίνακα που εισάγεται στον αλγόριθμο και μπορεί να χρησιμοποιηθεί σαν κάθε άλλο χαρακτηριστικό του πίνακα, κατά τη δημιουργία και εφαρμογή των μοντέλων εξόρυξης γνώσης.

Κατά την εκτέλεση αυτού του σταδίου, δημιουργούνται νέα χαρακτηριστικά τύπου DM_NESTED_NUMERICALS, εξάγοντας terms από στήλες πίνακα που υποστηρίζουν κείμενο. Κάθε σειρά του εμφωλευμένου πίνακα, αποτελείται μια στήλη Attribute η οποία περιέχει το term και μια στήλη Value η οποία περιέχει μία αριθμητική τιμή που αντιπροσωπεύει τη συχνότητα εμφάνισης του term μέσα στο κείμενο. Η δομή του εμφωλευμένου πίνακα παρουσιάζεται στον Πίνακας 4.6.1.

ΧΑΡΑΚΤΗΡΙΣΤΙΚΟ	ΤΥΠΟΣ
Attribute	Varchar2(4000)
Value	Number

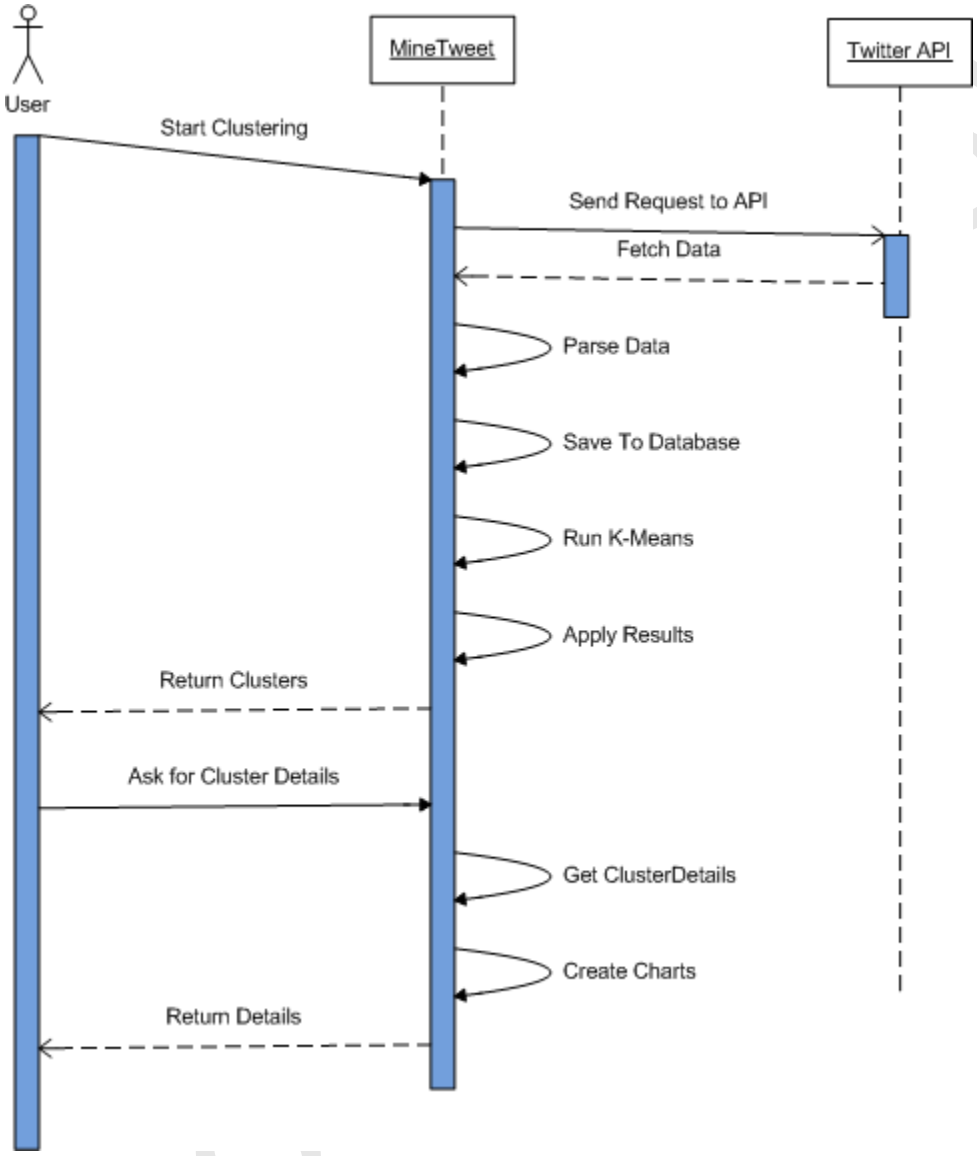
Πίνακας 4.6.1:Η δομή του εμφωλευμένου πίνακα

Μετά την ολοκλήρωση της προετοιμασίας των δεδομένων ακολουθεί η διαδικασία της συσταδοποίησης.

4.7 ΣΥΣΤΑΔΟΠΟΙΗΣΗ

Το πιο σημαντικό μέρος της εφαρμογής είναι η εξαγωγή των κατηγοριών από τα tweets του κάθε χρήστη. Ο σκοπός της υλοποίησης είναι να δημιουργήσουμε κατηγορίες tweets με μία διαδικασία χωρίς επίβλεψη. Ο αλγόριθμος που χρησιμοποιηθεί θα μπορεί να παραμετροποιείται από τον χρήστη της εφαρμογής ώστε να δίνει τα αποτελέσματα σύμφωνα με τις επιθυμίες του. Το Oracle Data Mining υποστηρίζει τον αλγόριθμο k-means για την εφαρμογή συσταδοποίησης σε δεδομένα κειμένου.

Μία αίτηση για συσταδοποίηση και εμφάνιση των αποτελεσμάτων εξυπηρετείται όπως παρουσιάζεται στην Εικόνα 4.7.1. Ο χρήστης κάνει μια αίτηση για συσταδοποίηση των αποτελεσμάτων. Το backend της εφαρμογής στέλνει μία αίτηση στο Twitter API για να λάβει τα τελευταία tweets του λογαριασμού. Μόλις λάβει την απάντηση αναλύει το κείμενο το μετασχηματίζει και το αποθηκεύει στην βάση δεδομένων. Στη συνέχεια τρέχει τον αλγόριθμο αφού πρώτα αφαιρέσει τα κοινές λέξεις από το κείμενο και μετασχηματίσει κατάλληλα το κείμενο σε εμφωλευμένους πίνακες. Στην συνέχεια εφαρμόζει τα αποτελέσματα σε κάθε εγγραφή που συμμετείχε στον αλγόριθμο, ενημερώνοντας την για το cluster στο οποίο ανήκει καθώς και για την πιθανότητα αυτή η τοποθέτηση της να είναι σωστή. Στην συνέχεια επιστρέφονται τα αποτελέσματα (clusters) στον χρήστη, οποίος με μία νέα κλήση μπορεί να δει τις λεπτομέρειες του κάθε cluster.



Εικόνα 4.7.1: Sequence Diagram για την εκτέλεση clustering και παρουσίαση των αποτελεσμάτων

4.7.1 ΑΛΓΟΡΙΘΜΟΣ ENHANCED K-MEANS

Ο αλγόριθμος k-means είναι ένας διαμεριστικός αλγόριθμος συσταδοποίησης ο οποίος ομαδοποιεί n δεδομένα σε έναν δοθέντα αριθμό k clusters. Ο αλγόριθμος εκχωρεί το κάθε στοιχείο σε αυτό το cluster του οποίου ο μέσος (centroid) έχει την μικρότερη απόσταση από το στοιχείο. Για τον υπολογισμό της απόστασης χρησιμοποιείται η ευκλείδεια απόσταση ή η ομοιότητα συνημίτονου.

Η εκδοχή του αλγόριθμου που προσφέρει το πακέτο Oracle Data Mining ονομάζεται enhanced k-means [10] και έχει τα εξής χαρακτηριστικά.

- Ο αλγόριθμος δημιουργεί μοντέλα με ιεραρχικό τρόπο, από πάνω προς τα κάτω, χρησιμοποιώντας δυαδικές διαιρέσεις και στο τέλος βελτιστοποιεί όλους τους κόμβους. Υπό αυτή την έννοια, ο αλγόριθμος είναι παρόμοιος με τον αλγόριθμο Bisecting K-

Means. Το κέντρο βάρους των εσωτερικών κόμβων επαναϋπολογίζεται ώστε να αντικατοπτρίζει τις αλλαγές όπως εξελίσσεται το δένδρο. Στο τέλος της διαδικασίας επιστρέφεται ολόκληρο το δένδρο.

- Ο αλγόριθμος αναπτύσσει το δένδρο κατά ένα κόμβο τη φορά Ανάλογα με τις παραμέτρους που έχουν δηλωθεί, ο κόμβος με τη μεγαλύτερη διακύμανση, διαιρείται για να αυξήσει το μέγεθος του δέντρου, μέχρι να ικανοποιηθεί ο αριθμός των συστάδων που έχει δηλωθεί. Υπάρχει παράμετρος στις ρυθμίσεις του μοντέλου που δηλώνεται ο επιθυμητός αριθμός των συστάδων που θα παραχθούν.
- Ο αλγόριθμος παρέχει πιθανοτική βαθμολόγηση των αποτελεσμάτων και ανάθεση των δεδομένων στις συστάδες..

Οι παράμετροι του αλγόριθμου είναι οι ακόλουθες :

Απόσπασμα 4.7.1: Οι παράμετροι του αλγορίθμου k-means

CLUS_NUM_CLUSTERS *(τιμή>=1): Ο αριθμός των συστάδων που δημιουργεί ο αλγόριθμος.*
KMNS_BLOCK_GROWTH *(1,τιμή<=5): Ο παράγοντας δέσμευσης της μνήμης που θα διατεθεί, για την προσωρινή αποθήκευση των δεδομένων μιας συστάδας.*
KMNS_CONV_TOLERANCE *(0<τιμή<=0.5): Η ανοχή σύγκλισης, η οποία αποτελεί το κριτήριο για να ολοκληρωθεί η διαδικασία εκπαίδευσης του μοντέλου.*

KMNS_DISTANCE *(KMNS_COSINE, KMNS_EUCLIDEAN): Η συνάρτηση απόστασης.*
KMNS_ITERATIONS *(0<τιμή<=20): Ο αριθμός επαναλήψεων του αλγόριθμου.*
KMNS_MIN_PCT_ATTR_SUPPORT *(0<τιμή<=1): Μία τιμή κλάσματος, που δηλώνει το ποσοστό των τιμών ενός χαρακτηριστικού που πρέπει να μην είναι NULL, ώστε το χαρακτηριστικό να συμπεριληφθεί στο κανόνα που περιγράφει τη συστάδα. Η αύξηση της συγκεκριμένης τιμής, είναι ανάλογη της αύξησης της πιθανότητας σχηματισμού πολύ περιορισμένων σε μέγεθος ή ακόμη και κενών κανόνων.*
KMNS_NUM_BINS *(>0): Το πλήθος των «κάδων», που περιέχει το ιστόγραμμα ενός χαρακτηριστικού, το οποίο έχει παραχθεί από τον αλγόριθμο. Τα όρια κάθε «κάδου» για κάθε χαρακτηριστικό, υπολογίζονται ανάλογα με το μέγεθος του συνόλου των δεδομένων εκπαίδευσης. Η μέθοδος binning που εφαρμόζεται είναι η ίσου πλάτους (equi-width). Όλα τα χαρακτηριστικά έχουν τον ίδιο αριθμό κάδων, εκτός από αυτά που έχουν μόνο μία τιμή και επομένως έχουν μόνο ένα κάδο.*
KMNS_SPLIT_CRITERION *(KMNS-SIZE, KMNS-VARIANCE): Το κριτήριο διάσπασης των συστάδων.*

Κάποιες από αυτές τις μεταβλητές όπως ο αριθμός των clusters θα επιλέγονται από το χρήστη στο UI. Για τις υπόλοιπες θα δοθούν οι βέλτιστες τιμές όπως αυτές θα προκύψουν από την πειραματική αξιολόγηση του αλγορίθμου.

4.7.2 ΕΦΑΡΜΟΓΗ ΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΣΤΑ ΔΕΔΟΜΕΝΑ

Έχοντας ολοκληρώσει την συσταδοποίηση θα πρέπει να ενημερωθεί η κάθε εγγραφή που συμμετείχε στον αλγόριθμο με τα στοιχεία του cluster στο οποίο έχει τοποθετηθεί. Κατά την εκτέλεση του ο αλγόριθμος k-means προσδιορίζει πιθανότητες υποθέτοντας ότι τα δεδομένα σε κάθε cluster ακολουθούν μια κανονική κατανομή. Ο μέσος αυτής της κατανομής για κάθε cluster είναι το centroid του. Η διακύμανση της κατανομής είναι η ίδια για όλα τα clusters και υπολογίζεται ως η μέση διακύμανση όλων των clusters.
Με βάση αυτό το bayesian probability που προκύπτει κατά την εκτέλεση του αλγορίθμου εφαρμόζουμε το μοντέλο μας στα δεδομένα και εισάγουμε τα αποτελέσματα στον πίνακα RESULTS. Ένα δείγμα των εγγραφών που προκύπτουν στον πίνακα RESULTS παρουσιάζεται στον Πίνακας 4.7.1: Δείγμα εγγραφών στον πίνακα Results.

ID	TWEETDATA	TWEETID	USERiD	CASEID	CLUSTER_ID	PROBABILITY
225867	Παραιτήθηκε η ολλανδική κυβέρνηση http://t.co/X0HPACJt	194437767407730688	2	255656	36	0.999999999981
225868	Πρωτιά στον αριθμό των spam καταγράφει η Ινδία! http://t.co/vicRu	194437622242877441	2	255657	59	0.603212085864
225869	Πιθανότατα Σεπτέμβριο ή Οκτώβριο οι εκλογές στην Ολλανδία, σύμ	194435207665950720	2	255658	50	0.999999999702
225870	FT: «Οι Έλληνες θα παραμείνουν στην Ευρωζώνη» http://t.co/MOyF	194435012530143234	2	255659	28	0.999989532592
225871	Κυρώσεις κατά προμηθευτών τεχνολογικών μέσων από τον Ομπάμ	194434554226950145	2	255660	34	0.999999623919
225872	Το ΠΑΣΟΚ απαντάει στο «Ζάππειο 3» http://t.co/XJODG0wt	194434381174149120	2	255661	2	0.999442906686
225873	Ντέμης:"Μαζί μπορούμε, μόνος μου όχι": Το Sport24.gr αποκαλύπτ	194433238247276545	2	255662	20	0.999999099672
225874	Οι χρυσοί πυρσοί της Ολυμπιακής Φλόγας: Ολόχρυσοι και πολύ εν	194433235323863040	2	255663	59	0.580990192927
225875	Το τρελό πάρτι της Ντόρτμουντ!: Η κατάκτηση του δεύτερου σερί τίτ	194433230777225216	2	255664	44	0.999999999650

Πίνακας 4.7.1: Δείγμα εγγραφών στον πίνακα Results

Το επόμενο και τελευταίο βήμα πριν την επιστροφή των αποτελεσμάτων στο χρήστη είναι διαδικασία ονοματοδοσίας του κάθε cluster με ένα όνομα το οποίο θα μπορεί να αναγνωριστεί από τον χρήστη και το οποίο θα προκύπτει από τις λέξεις που περιέχονται σε αυτό.

4.7.3 ΟΝΟΜΑΤΟΔΟΣΙΑ ΣΥΣΤΑΔΩΝ

Τελικό βήμα της εκτέλεσης της διαδικασίας συσταδοποίησης είναι η απόδοση ονόματος σε κάθε συστάδα. Με αυτή την διαδικασία θα αποδώσουμε σε κάθε συστάδα ένα όνομα ώστε να είναι η δυνατή η παρουσίαση της στην διεπαφή του χρήστη. Ως όνομα αποφασίστηκε να είναι μια γραμματοσειρά η οποία αποτελείται από τα τρία term που εμφανίζουν την μεγαλύτερη συχνότητα εμφάνισής σε κάθε centroid. Σημειώνεται σε αυτό το σημείο ότι το centroid είναι ένα υβριδικό αντικείμενο και δεν αντιστοιχεί σε φυσική εγγραφή [10].

Για την εξαγωγή αυτών των terms χρησιμοποιείται η συνάρτηση GET_MODEL_DETAILS_KM . Για να επιτύχουμε την εξαγωγή αυτών των terms που θα συνθέσουν το όνομα θα πρέπει πρώτα να ανακτήσουμε τα τελικά clusters, δηλαδή τα clusters που είναι φύλλα στο δέντρο που εξήχθη από τον αλγόριθμο.. Για να λάβουμε την λίστα με τα leaf clusters χρησιμοποιούμε τον κώδικα του Απόσπασμα 4.7.2: Ανάκτηση leaf clustersΑπόσπασμα 4.7.2

Απόσπασμα 4.7.2: Ανάκτηση leaf clusters

```
CURSOR ClusterLeafIds IS
   --Obtain leaf clusters
   SELECT CLUSTER_ID, RECORD_COUNT
   FROM (
    SELECT distinct clus.ID AS CLUSTER_ID,
         clus.RECORD_COUNT RECORD_COUNT,
         clus.DISPERSION DISPERSION,
         clus.PARENT PARENT_CLUSTER_ID,
         clus.TREE_LEVEL TREE_LEVEL,
         CASE WHEN chl.id IS NULL THEN 'YES'
                        ELSE 'NO' END IS_LEAF
     FROM (SELECT *
         FROM TABLE(dbms_data_mining.get_model_details_km('MINETWEET_CLUST',null,null,0,0,0)))
clus,
            table(clus.child) chl
   )
   WHERE is_leaf='YES'
   ORDER BY cluster_id;
```

Αποθηκεύουμε τα αποτελέσματα σε ένα cursor δηλαδή σε ένα μηχανισμό για την αποθήκευση αποτελεσμάτων. Στην συνέχεια (Απόσπασμα 4.5.1) για κάθε εγγραφή του cursor δηλαδή για το κάθε cluster εξάγουμε το centroid και τα terms που το αποτελούν ταξινομημένα σε φθίνουσα σειρά από το πιο σημαντικό στο λιγότερο σημαντικό. Από αυτά τα στοιχεία επιλέγουμε τα τρία πιο σημαντικά terms τα οποία θα χρησιμοποιηθούν ως όνομα του cluster.

Απόσπασμα 4.7.3: Ανάκτηση top 3 text terms από κάθε cluster.

```
SELECT id, term || ' - ' ||
        LEAD(term, 1) OVER (ORDER BY id) || ' - ' ||
        LEAD(term, 2) OVER (ORDER BY id) cluster_name
      FROM
      (
       SELECT id, text term, centroid_mean
        FROM
        (
         SELECT rownum id, a.*
          FROM (
              SELECT cd.attribute_subname  term,
                   cd.mean             centroid_mean
              FROM (
                SELECT *
                 FROM TABLE(dbms_data_mining.get_model_details_km('MINETWEET_CLUST',  c.cluster_id,
null, 1, 0, 0, null)) ) a,
                 TABLE(a.centroid) cd
              order by cd.mean desc
              ) a
         WHERE rownum < 4
        ) x,
      minetweet_map y
      WHERE x.term=y.attribute_id
      ORDER BY centroid_mean
      )
```

Υπόδειγμα των αποτελεσμάτων που εξάγονται από την παραπάνω διαδικασία περιέχει ο Πίνακας 4.7.2.

CLUSTER_ID	CLUSTERNAME
46	ΕΦΑΡΜΟΓΗ-ΝΟΜΟΥ-ΑΕΙ
30	ΑΠΟΦΥΛΑΚΙΣΗΣ-ΑΙΤΗΣΗ-ΑΚΗ
22	ΦΛΟΓΑ-ΟΛΥΜΠΙΑΚΗ-ΚΑΣΤΕΛΟΡΙΖΟ
6	ΕΥΡΩ-ΠΑΡΑΜΕΙΝΕΙ-ΕΛΛΑΔΑ
4	ΕΛΛΕΙΜΜΑ-2011-9,1

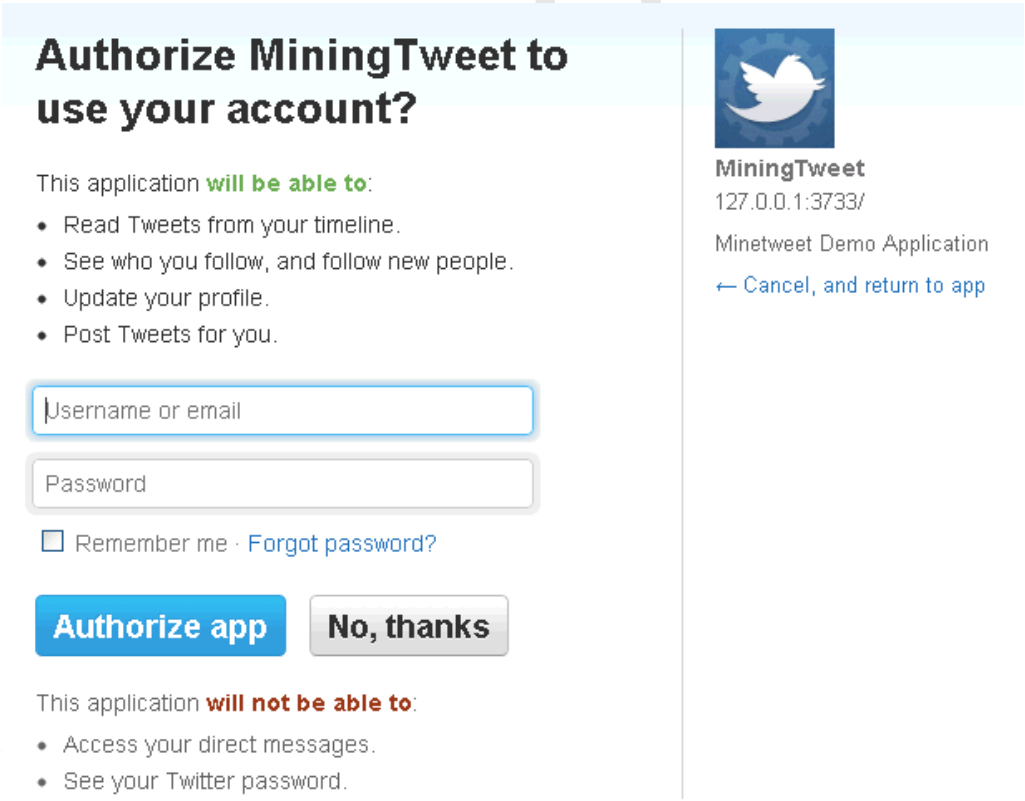
Πίνακας 4.7.2: Υπόδειγμα αποτελεσμάτων διαδικασίας ονοματοδοσίας συστάδας

4.8 ΔΙΕΠΑΦΗ ΧΡΗΣΤΗ

Στην web εφαρμογή που σχεδιάστηκε έγινε προσπάθεια να είναι όσο το δυνατόν πιο απλή ώστε να διευκολύνει στην χρήση της. Στην αρχική σελίδα ο χρήστης μπορεί να δει μια λίστα με τα public tweets. (Εικόνα 4.8.1)

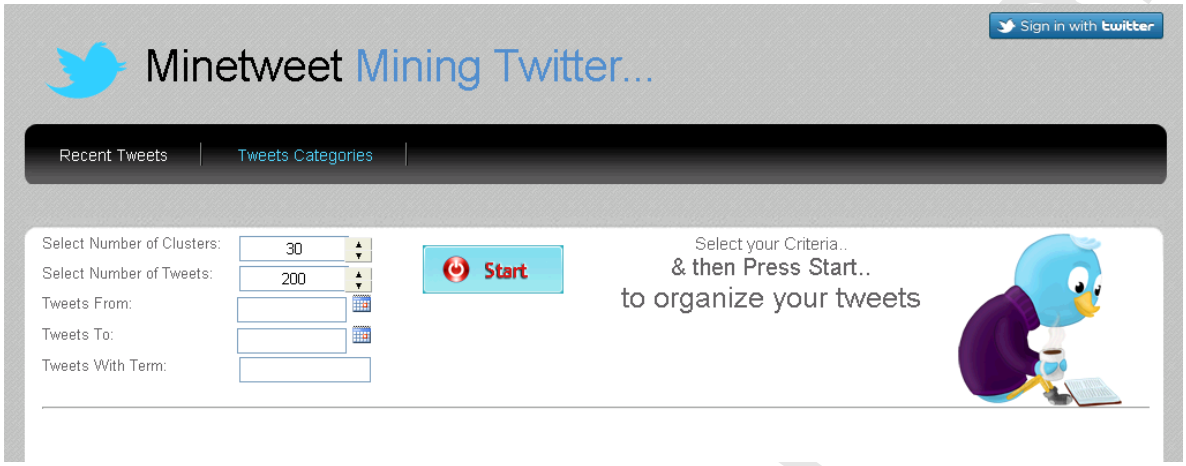
Εικόνα 4.8.1:Η αρχική σελίδα της εφαρμογής Minetweet.

Για να συνεχίσει την περιήγηση ο χρήστης θα πρέπει να πιστοποιηθεί. Αυτή η διαδικασία πραγματοποιείται όπως αναφέραμε σε προηγούμενο κεφάλαιο στην σελίδα του Twitter (Εικόνα 4.8.2)



Εικόνα 4.8.2: Οθόνη Authentication

Με την ολοκλήρωση της πιστοποίησης ο χρήστης επιστρέφει στο Minetweet όπου πλέον μπορεί να έχει πρόσβαση στα στοιχεία του λογαριασμού του, και να δημιουργήσει νέα μηνύματα. Στην σελίδα Categorized Tweets έχει την δυνατότητα να επιλέξει τις παραμέτρους της αρεσκείας του και να εκτελέσει τον αλγόριθμο ώστε να δει τα tweets κατηγοριοποιημένα.(Εικόνα 4.8.3).

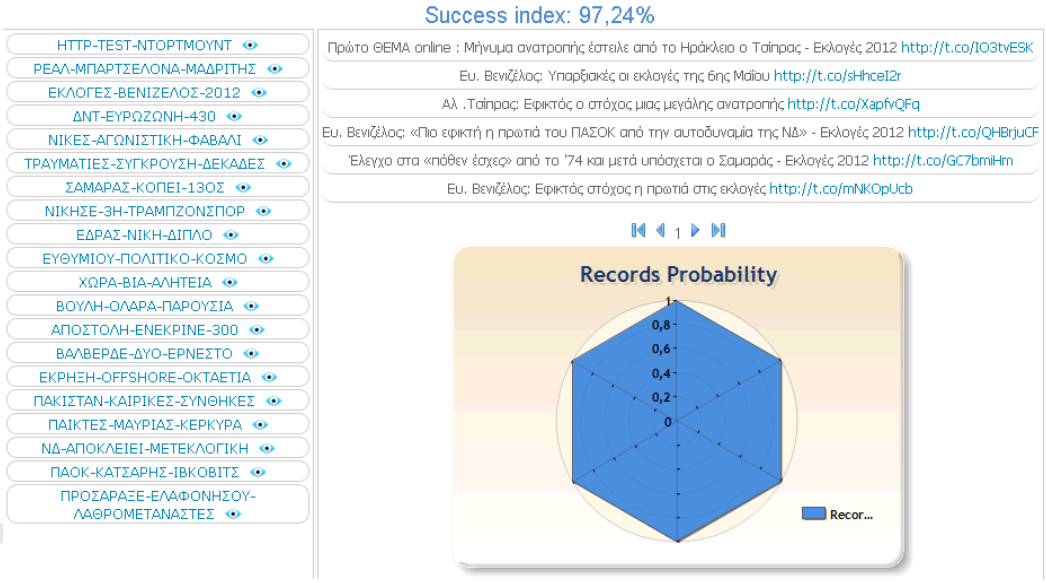


Εικόνα 4.8.3: Επιλογή Παραμέτρων Αλγόριθμου

Οι παράμετροι τις οποίες μπορεί να επιλέξει χρήστης είναι οι παρακάτω:

- Αριθμός Clusters που θα εξάγει ο αλγόριθμος
- Αριθμός tweets που θα εξετάσει ο αλγόριθμος
- Ημερομηνία από την οποία θα ληφθούν tweets
- Ημερομηνία έως την οποία θα ληφθούν tweets
- Term βάση του οποίου θα γίνει η αναζήτηση των tweets

Με τη ολοκλήρωση της εκτέλεσης του αλγορίθμου εμφανίζεται μια λίστα με τα clusters που δημιουργήθηκαν και τα οποία χαρακτηρίζονται από τα πρώτα τρία terms ως προς την βαρύτητα, του centroid. Επιλέγοντας κάθε ένα από τα clusters εμφανίζονται οι εγγραφές που το αποτελούν.



Εικόνα 4.8.4: Κατηγοριοποιημένα tweets

Εκτός από την λίστα με τα tweets που αποτελούν το cluster ο χρήστης μπορεί να δει και δύο διαγράμματα. Στο πρώτο παρουσιάζεται η ακρίβεια με την οποία έχει τοποθετηθεί η κάθε εγγραφή στο συγκεκριμένο cluster,ενώ στο δεύτερο παρουσιάζεται το ποσοστό που αντιπροσωπεύουν οι εγγραφές του συγκεκριμένου cluster σε σχέση με τον συνολικό αριθμό των εγγραφών που αποτέλεσαν την βάση για την εκτέλεση του αλγόριθμου.

5 ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Για την να αξιολογήσουμε την λειτουργία της εφαρμογής δημιουργήθηκε ένα account στο Twitter το οποίο ακολουθεί τα accounts ertsocial, skaigr, tovimagr, kathimerinigr, protothema και naftemporiki. Από τα δεδομένα που συγκεντρώθηκαν στην βάση δεδομένων επιλέχθηκε αρχικά ένα τυχαίο υποσύνολο 500 tweets από το σύνολο των καταχωρημένων tweets στο σύστημα την περίοδο από 25 Μαρτίου έως 5 Απριλίου 2012. Σε αυτά τα δεδομένα εκτελέστηκαν αρκετές πειραματικές εκτελέσεις του αλγόριθμού με διαφορετικές επιλογές των παραμέτρων του.

Με την ολοκλήρωση της 1ης πειραματικής διαδικασίας και αφού επιλέχθηκαν οι καλύτερες δυνατές τιμές για την απόδοση του αλγορίθμου, εφαρμόσαμε τον αλγόριθμο σε ένα ακόμα σύνολο δεδομένων, ώστε να εξετάσουμε τα αποτελέσματα που επιτυγχάνει. Το σύνολο αποτέλεσαν 500 εγγραφές από ένα δεύτερο account που ακολουθούσε accounts σχετικά με τον κινηματογράφο, το θέατρο και την μουσική και πιο συγκεκριμένα τα odeon.gr villagegr, theatrotechnis, ellthea, και texnospito.gr. Στην συνέχεια παρουσιάζονται τα αποτελέσματα αυτής της πειραματικής αξιολόγησης.

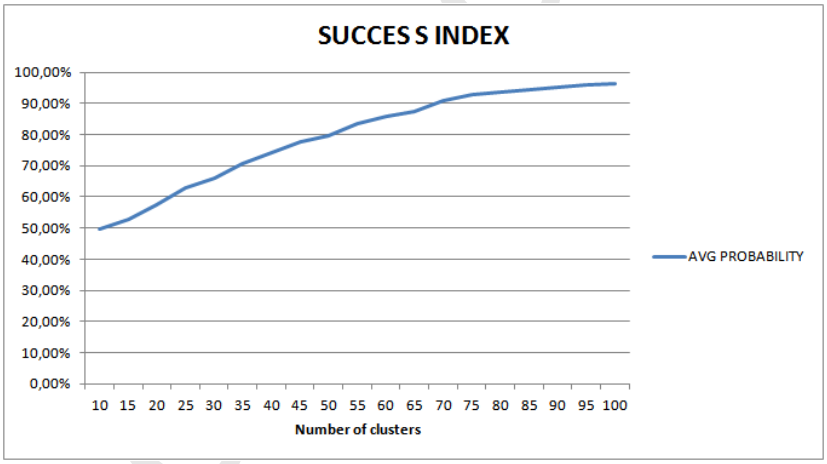
5.1 ΠΡΩΤΗ ΦΑΣΗ ΑΞΙΟΛΟΓΗΣΗΣ

Στο πρώτο μέρος της αξιολόγησης εκτελέστηκε διαδοχικά ο αλγόριθμός για ένα εύρος τιμών σε κάθε παράμετρο εισόδου με σκοπό να εξαχθούν τα καλύτερα δυνατά αποτελέσματα και να ανατεθούν οι τιμές στις παραμέτρους. Η διαδικασία πραγματοποιήθηκε σειριακά ξεκινώντας με τις προτεινόμενες τιμές και ορίζοντας σε κάθε σύνολο εκτελέσεων του αλγόριθμου από μία παράμετρο εισόδου. Για την αξιολόγηση της απόδοσης του αλγορίθμου χρησιμοποιήθηκε ένας δείκτης ο οποίος προκύπτει από την μέση πιθανότητα σωστής απόδοσης της κάθε εγγραφής στο cluster που τοποθετήθηκε.

Η επιλογή του αριθμού των clusters που θα προκύψουν από την εκτέλεση του αλγορίθμου είναι μια από τις βασικές παραμέτρους (CLUS_NUM_CLUSTERS) που επηρεάζουν την αποτελεσματικότητα του αλγορίθμου, και είναι και άμεσα εξαρτημένη από τον δεδομένα εισόδου του αλγόριθμου. Αυτός είναι και ένας λόγος για τον οποίο επιλέξαμε να δώσουμε την δυνατότητα στον κάθε χρήστη να την μεταβάλει δυναμικά πριν από την εκτέλεση του αλγορίθμου. Σημειώνεται εδώ ότι η συγκεκριμένη μεταβλητή επηρεάζει τον χρόνο εκτελέσεως του αλγορίθμου καθώς μεγαλύτερος αριθμός clusters αντιστοιχεί σε μεγαλύτερο χρόνο εκτέλεσης. Η αρχική τιμή που δίνεται στην συγκεκριμένη παράμετρο είναι η τιμή 10. Εξετάστηκε ένα σύνολο τιμών από 10 έως 100 με βήμα 5, με τα αποτελέσματα να παρουσιάζονται στον Πίνακας 5.1.1 και στην Εικόνα 5.1.1.

ΤΙΜΗ ΠΑΡΑΜΕΤΡΟΥ	SUCCESS INDEX
10	49,56%
15	52,89%
20	57,53%
25	62,85%
30	65,90%
35	70,82%
40	74,29%
45	77,54%
50	79,76%
55	83,66%
60	85,77%
65	87,43%
70	90,92%
75	92,98%
80	93,61%
85	94,34%
90	95,08%
95	95,81%
100	96,52%

Πίνακας 5.1.1: Αποτελέσματα πειραματικής αξιολόγησης μεταβλητής CLUS_NUM_CLUSTERS



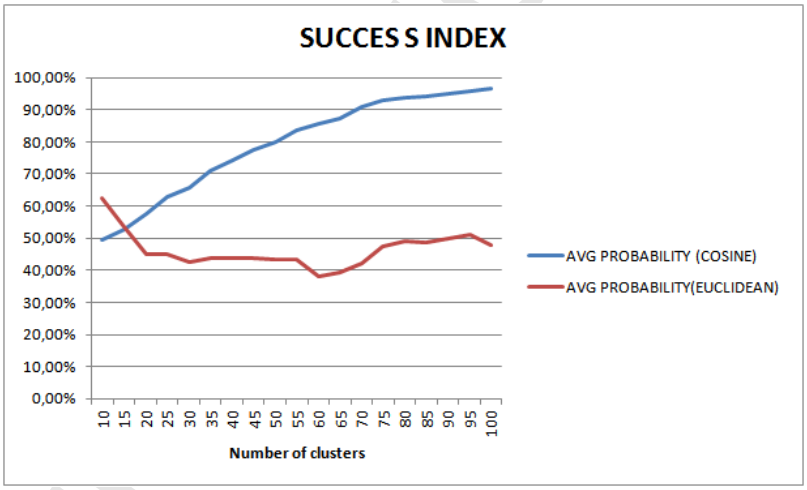
Εικόνα 5.1.1: Δείκτης επιτυχίας για την μεταβλητή CLUS_NUM_CLUSTERS

Παρατηρούμε ότι δεδομένου του σταθερού αριθμού εισόδου η απόδοση του αλγορίθμου αυξάνεται καθώς αυξάνεται ο αριθμός των clusters.

Η επόμενη μεταβλητή που εξετάσαμε είναι η μεταβλητή για την συνάρτηση απόστασης που χρησιμοποιείται (KMNS_DISTANCE). Δεδομένου ότι η αξιολόγηση της μεταβλητής (CLUS_NUM_CLUSTERS) εξετάστηκε με την χρήση της συνάρτησης συνημίτονου για την μεταβλητή KMNS_DISTANCE, θα επαναλάβουμε την ίδια πειραματική διαδικασία με τιμή μεταβλητής την ευκλείδεια απόσταση. Τα αποτελέσματα της διαδικασίας παρουσιάζονται στον Πίνακας 5.1.2 και στην Εικόνα 5.1.2 εμφανίζεται το γράφημα της απόδοσης συγκρινόμενο με το γράφημα των αποτελεσμάτων για την απόσταση συνημίτονου.

ΤΙΜΗ ΠΑΡΑΜΕΤΡΟΥ	SUCCESS INDEX
10	62,46%
15	53,34%
20	44,82%
25	45,04%
30	42,76%
35	43,63%
40	43,63%
45	43,81%
50	43,29%
55	43,28%
60	38,24%
65	39,49%
70	42,09%
75	47,29%
80	49,22%
85	48,75%
90	49,84%
95	51,04%
100	47,79%

Πίνακας 5.1.2:Αποτελέσματα αξιολόγησης μεταβλητής KMNS_DISTANCE με τιμή EUCLIDEAN



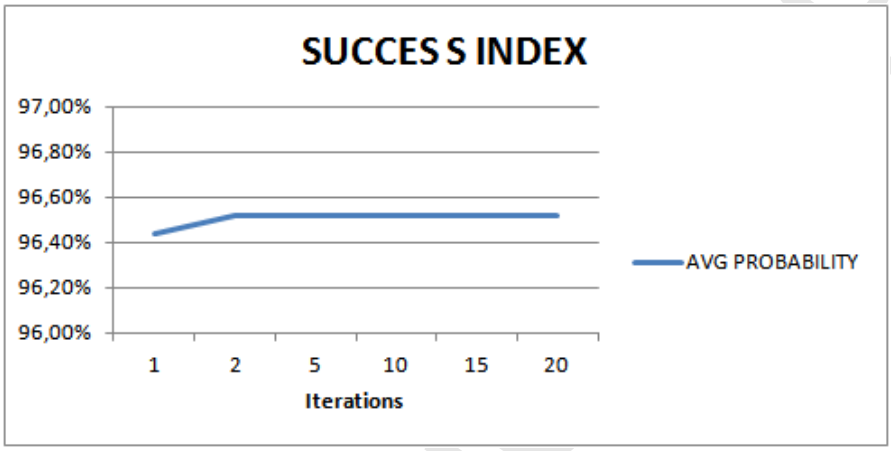
Εικόνα 5.1.2: Συγκριτικά αποτελέσματα ευκλείδειας και απόστασης συνημίτονου

Από τα αποτελέσματα προκύπτει ότι η μεταβλητή με χρήση της απόστασης συνημίτονου δίνει πολύ καλύτερα αποτελέσματα από την ευκλείδεια απόσταση η οποία παρουσιάζει σχετικά φτωχά αποτελέσματα. Επομένως η τιμή που θα χρησιμοποιηθεί είναι αυτή της απόστασης συνημίτονου.

Η επόμενη μεταβλητή που εξετάστηκε είναι η KMNS_ITERATIONS η οποία δίνει των αριθμό των επαναλήψεων της εκτέλεσης του αλγόριθμου. Η εύρος τιμών που δέχεται η μεταβλητή είναι από 1 έως 20. Εκτελέσαμε των αλγόριθμο με τιμές εισόδου 1,2,5,10,15,20 και τα αποτελέσματα παρουσιάζονται στον Πίνακας 5.1.3 και στην Εικόνα 5.1.3.

ΤΙΜΗ ΠΑΡΑΜΕΤΡΟΥ	SUCCESS INDEX
1	96,44%
2	96,52%
5	96,52%
10	96,52%
15	96,52%
20	96,52%

Πίνακας 5.1.3: Αποτελέσματα πειραματικής αξιολόγησης μεταβλητής KMNS_ITERATIONS



Εικόνα 5.1.3: Επίδραση του αριθμού επαναλήψεων στην απόδοση του αλγορίθμου

Παρατηρούμε ότι η εκτός από την τιμή 1 οι υπόλοιπες τιμές δεν επηρεάζουν την απόδοση του αλγορίθμου. Επομένως στην υλοποίηση χρησιμοποιήθηκε η προεπιλεγμένη τιμή 3.

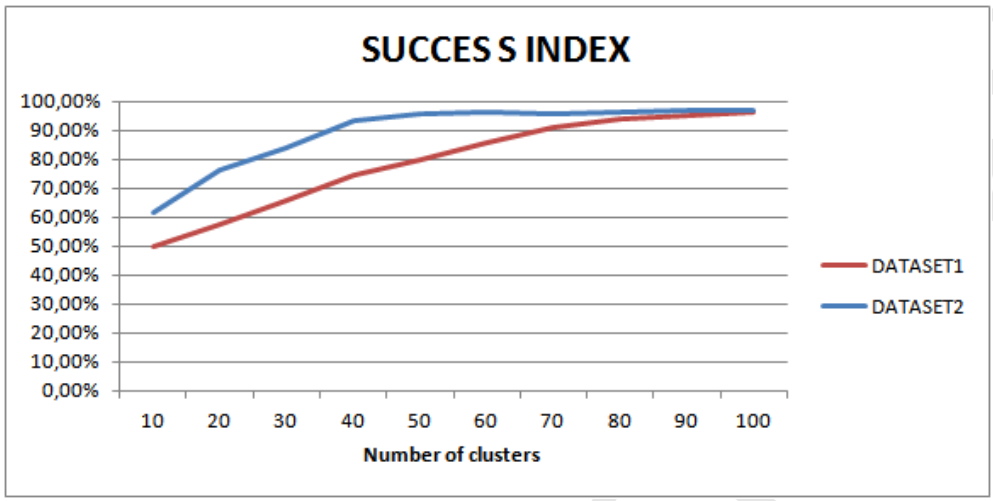
Με αντίστοιχη πειραματική διαδικασία εξετάστηκαν και οι υπόλοιπες παράμετροι του αλγορίθμου k-means, για τις οποίες εξήχθη το συμπέρασμα ότι δεν επηρεάζουν την απόδοση του αλγορίθμου. Επομένως και για αυτές θα χρησιμοποιηθούν οι προεπιλεγμένες τιμές.

5.2 ΔΕΥΤΕΡΗ ΦΑΣΗ ΑΞΙΟΛΟΓΗΣΗΣ

Με την μεταβλητές του αλγορίθμου καθορισμένες, εισάγουμε στον αλγόριθμο 500 εγγραφές από το δεύτερο account που ακολουθεί λογαριασμούς διαφορετικής θεματολογίας. Επαναλαμβάνουμε το πείραμα που εκτελέσαμε και στην πρώτη φάση με τις τελικές τιμές σε όλες τα μεταβλητές και αυξάνοντας τον αριθμό των clusters από 10 σε 100 με βήμα 10. Τα αποτελέσματα της διαδικασίας παρουσιάζονται στον Πίνακας 5.2.1 και στην Εικόνα 5.2.1

ΤΙΜΗ ΠΑΡΑΜΕΤΡΟΥ	SUCCESS INDEX
10	61,84%
20	76,46%
30	83,74%
40	93,47%
50	95,39%
60	96,15%
70	95,90%
80	96,44%
90	96,83%
100	96,88%

Πίνακας 5.2.1: Αποτελέσματα πειραματικής αξιολόγησης μεταβλητής CLUS_NUM_CLUSTERS



Εικόνα 5.2.1: Συγκριτικά αποτελέσματα αλγορίθμου για τα δεδομένα 1ης και 2ης Φάσης Αξιολόγησης

Παρατηρούμε ότι ο αλγόριθμος έχει την ίδια περίπου συμπεριφορά και τα αποτελέσματα διαφέρουν ως προς τις απόλυτες τιμές, με το αποτέλεσμα να κρίνεται φυσιολογικό λόγω της διαφορετικότητας των δεδομένων εισόδου.

Τα αποτελέσματα που εξάγονται από την αξιολόγηση της εφαρμογής οδηγούν στο συμπέρασμα ότι ο αλγόριθμός, και κατά επέκταση η εφαρμογή δίνει ικανοποιητικά αποτελέσματα. Η παράμετρος που βαρύνει περισσότερο στην απόδοση του συστήματος είναι ο αριθμός των clusters στα οποία γίνεται η συσταδοποίηση του συνόλου δεδομένων εισόδου.

6 ΣΥΜΠΕΡΑΣΜΑΤΑ

Σε αυτή την μεταπτυχιακή διατριβή προτάθηκε ένα σύστημα για την εξατομικευμένη ομαδοποίηση και παρουσίαση της πληροφορίας που δέχεται ένας χρήστης του Twitter στο λογαριασμό του. Η υλοποίηση απέδωσε μια web εφαρμογή η οποία μπορεί να χρησιμοποιηθεί με ασφάλεια από οποιοδήποτε χρήστη του Twitter, με ένα φιλικό στην χρήση interface. Η υλοποίηση του συστήματος έγινε εξετάζοντας διάφορες τεχνικές προγραμματισμού και υιοθετώντας αυτές οι οποίες επιτρέπουν την εύκολη μετατροπή και επέκταση του.

Βασικό μέρος της εργασίας αποτέλεσε η παραμετροποίηση του αλγορίθμου k-means για την συσταδοποίηση των δεδομένων ώστε να αποδίδει καλά αποτελέσματα για την ελληνική γλώσσα. Για τον σκοπό αυτό δημιουργήθηκαν οι κατάλληλες μέθοδοι για των καθαρισμό και τον μετασχηματισμό των δεδομένων. Παράλληλα κατά την υλοποίηση του συστήματος δημιουργήθηκε μια λίστα κοινών λέξεων η οποία εμπλουτίστηκε με βάση τα αποτελέσματα που προέκυπταν σε όλη την διάρκειας της υλοποίησης και η οποία είναι κριτικής σημασίας λόγω του μικρού μεγέθους των δεδομένων που περιέχει το κάθε tweet.

Το αποτέλεσμα που προέκυψε από την αξιολόγηση των πειραματικών εκτελέσεων του αλγόριθμου είναι άκρως ικανοποιητικό φτάνοντας σε ποσοστό επιτυχίας 96,88%, το οποίο δύναται να επιτευχθεί για κάθε σύνολο δεδομένων εισόδου, αφού αντί μιας άκαμπτης υλοποίησης επιλέχθηκε η δυναμική παραμετροποίηση του αλγορίθμου από τον εκάστοτε χρήστη.

Τέλος επιτυγχάνεται με την συγκεκριμένη εφαρμογή η αποθήκευση των δεδομένων του κάθε χρήστη διαχρονικά, και παρέχεται η δυνατότητα πρόσβασης και αναζήτησης σε αυτά χωρίς χρονικό περιορισμό, λειτουργία την οποία αυτή την στιγμή δεν διαθέτει το Twitter.

7 ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ

Το σύστημα που υλοποιήθηκε στην παρούσα μεταπτυχιακή διατριβή είναι πρωτότυπο, και είχε ως σκοπό να προσφέρει συγκεκριμένη λειτουργικότητα στον χρήστη. Ήδη από την περίοδο υλοποίησης του έχουν διαμορφωθεί κάποιες ιδέες οι οποίες μπορούν να υλοποιηθούν και να επεκτείνουν το σύστημα.

Μια πρώτη επέκταση θα μπορούσε να είναι η προσαρμογή του συστήματος ώστε να παρουσιάζει καλά αποτελέσματα και στην αγγλική γλώσσα καθώς και σε υβριδικές γλώσσες όπως είναι τα greeklish. Αυτό ο στόχος θα μπορούσε να επιτευχθεί με την ανάπτυξη μιας λίστας κοινών λέξεων για την αγγλική γλώσσα και με την δημιουργία ενός υποπρογράμματος το οποίο θα μεταφράζει τα μεταφράζει τα greeklish σε ελληνικά.

Μία ακόμα ενδιαφέρουσα επέκταση του συστήματος δεδομένης της μόνιμης αποθήκευσης των δεδομένων στη βάση θα ήταν η χρονική ανάλυση τους ώστε να εξάγεται πληροφορία σχετική με την εξέλιξη στον χρόνο.

Για την ολοκλήρωση του συστήματος και τον εμπλουτισμό του, προτείνεται η υλοποίηση υποσυστημάτων από τα οποία εξάγεται πληροφορία σχετική με τον γράφο στον οποίο ανήκει ο χρήστης και ποιο συγκεκριμένα κατανομές δημογραφικών και γεωγραφικών δεδομένων, η υλοποίηση των οποίων έχει αποτελέσει αντικείμενο εκτεταμένης έρευνας.

8 ΑΝΑΦΟΡΕΣ

1. Twitter. Your world, more connected. *Twitter Blog.* [Online] 2011. http://blog.twitter.com/2011/08/your-world-more-connected.html.
2. —. Using the Twitter Search API. *Twitter Developers.* [Online] 2011. https://dev.twitter.com/docs/using-search.
3. *Is it Really About Me? Message Content in Social Awareness Streams.* Naaman, Mor, Boase, Jeffrey and Lai, Chih-Hui. Savannah, Georgia, USA : s.n. ACM conference on Computer supported cooperative work. pp. 189-192.
4. *Characterizing Microblogs with Topic Models.* Ramage, Daniel; Dumais, Susan; Liebling, Dan.
5. *Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora.* Ramage, Daniel, et al. 2009.
6. Horn, Christopher. *Analysis and Classification of Twitter Messages.* 2010.
7. *Predicting Flu Trends using Twitter Data.* Achrekar, Harshavardhan, et al. 2011. IEEE Infocom, 2011 workshop on on Cyber-Physical.
8. *Βελτιστοποίηση και επέκταση του συστήματος συλλογής και ταξινόμησης ειδησεογραφικών πηγών NewsMiner.* Katelanis, Markos. 2010.
9. *Using Twitter to Recommend Real-Time Topical News.* Phelan, Owen, McCarthy, Kevin and Smyth, Barry. 2010. Proceedings of the third ACM conference on Recommender systems.
10. Oracle. Oracle Documents. [Online] http://www.oracle.com/technetwork/indexes/documentation/index.html.
11. Potel, Mike. MVP: Model-View-Presenter The Taligent Programming Model for C++ and Java. Technical Report : Taligent Inc, 1996.
12. Fowler, Martin. GUI Architectures. [Online] 2006. http://martinfowler.com/eaaDev/uiArchs.html.
13. Gamma, Erich, et al. *Design Patterns Elements of Reusable Object-Oriented Software.* s.l. : Addison Wesley, 1994.
14. Fowler, Martin. *Patterns of Enterprise Application Architecture.* s.l. : Addison-Wesley, 2002.
15. Fielding, Roy Thomas. Architectural Styles and the Design of Network-based Software Architectures. Irvine, USA : University Of California, 2000.
16. Linq2Twitter. *Codeplex.* [Online] http://linqtotwitter.codeplex.com/.
17. Hammer-Lahav, Eran. The OAuth 1.0 Protocol. *Internet Engineering Task Force.* [Online] 2010. http://tools.ietf.org/html/rfc5849.

ΠΑΡΑΡΤΗΜΑΤΑ

Α. ΚΟΙΝΕΣ ΛΕΞΕΙΣ (STOPLIST)

Α	ΕΞΗΣ	ΜΕ	ΠΟΛΥ
ΑΔΙΑΚΟΠΑ	ΕΞΙΣΟΥ	ΜΕΘΑΥΡΙΟ	ΠΟΣΕΣ
ΑΙ	ΕΞΩ	ΜΕΙΟΝ	ΠΟΣΗ
ΑΚΟΜΑ	ΕΠΑΝΩ	ΜΕΛΕΙ	ΠΟΣΗΝ
ΑΚΟΜΗ	ΕΠΕΙΔΗ	ΜΕΛΛΕΤΑΙ	ΠΟΣΗΣ
ΑΚΡΙΒΩΣ	ΕΠΕΙΤΑ	ΜΕΜΙΑΣ	ΠΟΣΟΙ
ΑΛΗΘΕΙΑ	ΕΠΙ	ΜΕΝ	ΠΟΣΟΣ
ΑΛΗΘΙΝΑ	ΕΠΙΣΗΣ	ΜΕΡΙΚΑ	ΠΟΣΟΥΣ
ΑΛΛΑ	ΕΠΟΜΕΝΩΣ	ΜΕΡΙΚΕΣ	ΠΟΤΕ
ΑΛΛΑΧΟΥ	ΕΣΑΣ	ΜΕΡΙΚΟΙ	ΠΟΥ
ΑΛΛΕΣ	ΕΣΕΙΣ	ΜΕΡΙΚΟΥΣ	ΠΟΥΘΕ
ΑΛΛΗ	ΕΣΕΝΑ	ΜΕΡΙΚΩΝ	ΠΟΥΘΕΝΑ
ΑΛΛΗΝ	ΕΣΤΩ	ΜΕΣΑ	ΠΡΕΠΕΙ
ΑΛΛΗΣ	ΕΣΥ	ΜΕΤ	ΠΡΙΝ
ΑΛΛΙΩΣ	ΕΤΕΡΑ	ΜΕΤΑ	ΠΡΟ
ΑΛΛΙΩΤΙΚΑ	ΕΤΕΡΑΙ	ΜΕΤΑΞΥ	ΠΡΟΚΕΙΜΕΝΟΥ
ΑΛΛΟ	ΕΤΕΡΑΣ	ΜΕΧΡΙ	ΠΡΟΚΕΙΤΑΙ
ΑΛΛΟΙ	ΕΤΕΡΕΣ	ΜΗ	ΠΡΟΠΕΡΣΙ
ΑΛΛΟΙΩΣ	ΕΤΕΡΗ	ΜΗΔΕ	ΠΡΟΣ
ΑΛΛΟΙΩΤΙΚΑ	ΕΤΕΡΗΣ	ΜΗΝ	ΠΡΟΤΟΥ
ΑΛΛΟΝ	ΕΤΕΡΟ	ΜΗΠΩΣ	ΠΡΟΧΘΕΣ
ΑΛΛΟΣ	ΕΤΕΡΟΙ	ΜΗΤΕ	ΠΡΟΧΤΕΣ
ΑΛΛΟΤΕ	ΕΤΕΡΟΝ	ΜΙΑ	ΠΡΩΤΥΤΕΡΑ
ΑΛΛΟΥ	ΕΤΕΡΟΣ	ΜΙΑΝ	ΠΩΣ
ΑΛΛΟΥΣ	ΕΤΕΡΟΥ	ΜΙΑΣ	Ρ
ΑΛΛΩΝ	ΕΤΕΡΟΥΣ	ΜΟΛΙΣ	Σ
ΑΜΑ	ΕΤΕΡΩΝ	ΜΟΛΟΝΟΤΙ	ΣΑΝ
ΑΜΕΣΑ	ΕΤΟΥΤΑ	ΜΟΝΑΧΑ	ΣΑΣ
ΑΜΕΣΩΣ	ΕΤΟΥΤΕΣ	ΜΟΝΕΣ	ΣΕ
ΑΝ	ΕΤΟΥΤΗ	ΜΟΝΗ	ΣΕΙΣ
ΑΝΑ	ΕΤΟΥΤΗΝ	ΜΟΝΗΝ	ΣΗΜΕΡΑ
ΑΝΑΜΕΣΑ	ΕΤΟΥΤΗΣ	ΜΟΝΗΣ	ΣΙΓΑ
ΑΝΑΜΕΤΑΞΥ	ΕΤΟΥΤΟ	ΜΟΝΟ	ΣΟΥ
ΑΝΕΥ	ΕΤΟΥΤΟΙ	ΜΟΝΟΙ	ΣΤΑ
ΑΝΤΙ	ΕΤΟΥΤΟΝ	ΜΟΝΟΜΙΑΣ	ΣΤΗ

ΑΝΤΙΠΕΡΑ	ΕΤΟΥΤΟΣ	ΜΟΝΟΣ	ΣΤΗΝ
ΑΝΤΙΣ	ΕΤΟΥΤΟΥ	ΜΟΝΟΥ	ΣΤΗΣ
ΑΝΩ	ΕΤΟΥΤΟΥΣ	ΜΟΝΟΥΣ	ΣΤΙΣ
ΑΝΩΤΕΡΩ	ΕΤΟΥΤΩΝ	ΜΟΝΩΝ	ΣΤΟ
ΑΞΑΦΝΑ	ΕΤΣΙ	ΜΟΥ	ΣΤΟΝ
ΑΠ	ΕΥ	ΜΠΟΡΕΙ	ΣΤΟΥ
ΑΠΕΝΑΝΤΙ	ΕΥΓΕ	ΜΠΟΡΟΥΝ	ΣΤΟΥΣ
ΑΠΟ	ΕΥΘΥΣ	ΜΠΡΑΒΟ	ΣΤΩΝ
ΑΠΟΨΕ	ΕΥΤΥΧΩΣ	ΜΠΡΟΣ	ΣΥΓΧΡΟΝΩΣ
ΑΡΑ	ΕΦΕΞΗΣ	Ν	ΣΥΝ
ΑΡΑΓΕ	ΕΧΕΙ	ΝΑ	ΣΥΝΑΜΑ
ΑΡΓΑ	ΕΧΕΙΣ	ΝΑΙ	ΣΥΝΕΠΩΣ
ΑΡΓΟΤΕΡΟ	ΕΧΕΤΕ	ΝΩΡΙΣ	ΣΥΝΗΘΩΣ
ΑΡΙΣΤΕΡΑ	ΕΧΘΕΣ	Ξ	ΣΥΧΝΑ
ΑΡΚΕΤΑ	ΕΧΟΜΕ	ΞΑΝΑ	ΣΥΧΝΑΣ
ΑΡΧΙΚΑ	ΕΧΟΥΜΕ	ΞΑΦΝΙΚΑ	ΣΥΧΝΕΣ
ΑΣ	ΕΧΟΥΝ	Ο	ΣΥΧΝΗ
ΑΥΡΙΟ	ΕΧΤΕΣ	ΟΙ	ΣΥΧΝΗΝ
ΑΥΤΟΝ	ΕΧΩ	ΟΛΑ	ΣΥΧΝΗΣ
ΑΥΤΑ	ΕΩΣ	ΟΛΕΣ	ΣΥΧΝΟ
ΑΥΤΕΣ	Ζ	ΟΛΗ	ΣΥΧΝΟΙ
ΑΥΤΗ	Η	ΟΛΗΝ	ΣΥΧΝΟΝ
ΑΥΤΗΝ	ΗΔΗ	ΟΛΗΣ	ΣΥΧΝΟΣ
ΑΥΤΗΣ	ΗΜΑΣΤΑΝ	ΟΛΟ	ΣΥΧΝΟΥ
ΑΥΤΟ	ΗΜΑΣΤΕ	ΟΛΟΓΥΡΑ	ΣΥΧΝΟΥΣ
ΑΥΤΟΙ	ΗΜΟΥΝ	ΟΛΟΙ	ΣΥΧΝΩΝ
ΑΥΤΟΝ	ΗΣΑΣΤΑΝ	ΟΛΟΝ	ΣΥΧΝΩΣ
ΑΥΤΟΣ	ΗΣΑΣΤΕ	ΟΛΟΝΕΝ	ΣΧΕΔΟΝ
ΑΥΤΟΥ	ΗΣΟΥΝ	ΟΛΟΣ	ΣΩΣΤΑ
ΑΥΤΟΥΣ	ΗΤΑΝ	ΟΛΟΤΕΛΑ	Τ
ΑΦΟΤΟΥ	ΗΤΑΝΕ	ΟΛΟΥ	ΤΑ
ΑΦΟΥ	ΗΤΟΙ	ΟΛΟΥΣ	ΤΑΔΕ
Β	ΗΤΤΟΝ	ΟΛΩΝ	ΤΑΥΤΩΝ
ΒΕΒΑΙΑ	Θ	ΟΛΩΣ	ΤΑΥΤΑ
ΒΕΒΑΙΟΤΑΤΑ	ΘΑ	ΟΛΩΣΔΙΟΛΟΥ	ΤΑΥΤΕΣ
Γ	Ι	ΟΜΩΣ	ΤΑΥΤΗ
ΓΙ	ΙΔΙΑ	ΟΠΟΙΑ	ΤΑΥΤΗΝ
ΓΙΑ	ΙΔΙΑΝ	ΟΠΟΙΑΔΗΠΟΤΕ	ΤΑΥΤΗΣ

ΓΡΗΓΟΡΑ	ΙΔΙΑΣ	ΟΠΟΙΑΝ	ΤΑΥΤΟ
ΓΥΡΩ	ΙΔΙΕΣ	ΟΠΟΙΑΝΔΗΠΟΤΕ	ΤΑΥΤΟΝ
Δ	ΙΔΙΟ	ΟΠΟΙΑΣ	ΤΑΥΤΟΣ
ΔΑ	ΙΔΙΟΙ	ΟΠΟΙΑΣΔΗΠΟΤΕ	ΤΑΥΤΟΥ
ΔΕ	ΙΔΙΟΝ	ΟΠΟΙΔΗΠΟΤΕ	ΤΑΧΑ
ΔΕΙΝΑ	ΙΔΙΟΣ	ΟΠΟΙΕΣ	ΤΑΧΑΤΕ
ΔΕΝ	ΙΔΙΟΥ	ΟΠΟΙΕΣΔΗΠΟΤΕ	ΤΕΛΙΚΑ
ΔΕΞΙΑ	ΙΔΙΟΥΣ	ΟΠΟΙΟ	ΤΕΛΙΚΩΣ
ΔΗΘΕΝ	ΙΔΙΩΝ	ΟΠΟΙΟΔΗΠΟΤΕ	ΤΕΣ
ΔΗΛΑΔΗ	ΙΔΙΩΣ	ΟΠΟΙΟΙ	ΤΕΤΟΙΑ
ΔΙ	II	ΟΠΟΙΟΝ	ΤΕΤΟΙΑΝ
ΔΙΑ	III	ΟΠΟΙΟΝΔΗΠΟΤΕ	ΤΕΤΟΙΑΣ
ΔΙΑΡΚΩΣ	ΙΣΑΜΕ	ΟΠΟΙΟΣ	ΤΕΤΟΙΕΣ
ΔΙΚΑ	ΙΣΙΑ	ΟΠΟΙΟΣΔΗΠΟΤΕ	ΤΕΤΟΙΟ
ΔΙΚΟ	ΙΣΩΣ	ΟΠΟΙΟΥ	ΤΕΤΟΙΟΙ
ΔΙΚΟΙ	Κ	ΟΠΟΙΟΥΔΗΠΟΤΕ	ΤΕΤΟΙΟΝ
ΔΙΚΟΣ	ΚΑΘΕ	ΟΠΟΙΟΥΣ	ΤΕΤΟΙΟΣ
ΔΙΚΟΥ	ΚΑΘΕΜΙΑ	ΟΠΟΙΟΥΣΔΗΠΟΤΕ	ΤΕΤΟΙΟΥ
ΔΙΚΟΥΣ	ΚΑΘΕΜΙΑΣ	ΟΠΟΙΩΝ	ΤΕΤΟΙΟΥΣ
ΔΙΟΛΟΥ	ΚΑΘΕΝΑ	ΟΠΟΙΩΝΔΗΠΟΤΕ	ΤΕΤΟΙΩΝ
ΔΙΠΛΑ	ΚΑΘΕΝΑΣ	ΟΠΟΤΕ	ΤΗ
ΔΙΧΩΣ	ΚΑΘΕΝΟΣ	ΟΠΟΤΕΔΗΠΟΤΕ	ΤΗΝ
Ε	ΚΑΘΕΤΙ	ΟΠΟΥ	ΤΗΣ
ΕΑΝ	ΚΑΘΟΛΟΥ	ΟΠΟΥΔΗΠΟΤΕ	ΤΙ
ΕΑΥΤ?Ν	ΚΑΘΩΣ	ΟΠΩΣ	ΤΙΠΟΤΑ
ΕΑΥΤΟ	ΚΑΙ	ΟΡΙΣΜΕΝΑ	ΤΙΠΟΤΕ
ΕΑΥΤΟΝ	ΚΑΚΑ	ΟΡΙΣΜΕΝΕΣ	ΤΙΣ
ΕΑΥΤΟΥ	ΚΑΚΩΣ	ΟΡΙΣΜΕΝΩΝ	ΤΟ
ΕΑΥΤΟΥΣ	ΚΑΛΑ	ΟΡΙΣΜΕΝΩΣ	ΤΟΙ
ΕΓΚΑΙΡΑ	ΚΑΛΩΣ	ΟΣΑ	ΤΟΝ
ΕΓΚΑΙΡΩΣ	ΚΑΜΙΑ	ΟΣΑΔΗΠΟΤΕ	ΤΟΣ
ΕΓΩ	ΚΑΜΙΑΝ	ΟΣΕΣ	ΤΟΣ?Ν
ΕΔΩ	ΚΑΜΙΑΣ	ΟΣΕΣΔΗΠΟΤΕ	ΤΟΣΑ
ΕΙΔΕΜΗ	ΚΑΜΠΟΣΑ	ΟΣΗ	ΤΟΣΕΣ
ΕΙΘΕ	ΚΑΜΠΟΣΕΣ	ΟΣΗΔΗΠΟΤΕ	ΤΟΣΗ
ΕΙΜΑΙ	ΚΑΜΠΟΣΗ	ΟΣΗΝ	ΤΟΣΗΝ
ΕΙΜΑΣΤΕ	ΚΑΜΠΟΣΗΝ	ΟΣΗΝΔΗΠΟΤΕ	ΤΟΣΗΣ
ΕΙΝΑΙ	ΚΑΜΠΟΣΗΣ	ΟΣΗΣ	ΤΟΣΟ

ΕΙΣ	ΚΑΜΠΟΣΟ	ΟΣΗΣΔΗΠΟΤΕ	ΤΟΣΟΙ
ΕΙΣΑΙ	ΚΑΜΠΟΣΟΙ	ΟΣΟ	ΤΟΣΟΝ
ΕΙΣΑΣΤΕ	ΚΑΜΠΟΣΟΝ	ΟΣΟΔΗΠΟΤΕ	ΤΟΣΟΣ
ΕΙΣΤΕ	ΚΑΜΠΟΣΟΣ	ΟΣΟΙ	ΤΟΣΟΥ
ΕΙΤΕ	ΚΑΜΠΟΣΟΥ	ΟΣΟΙΔΗΠΟΤΕ	ΤΟΣΟΥΣ
ΕΙΧΑ	ΚΑΜΠΟΣΟΥΣ	ΟΣΟΝ	ΤΟΤΕ
ΕΙΧΑΜΕ	ΚΑΜΠΟΣΩΝ	ΟΣΟΝΔΗΠΟΤΕ	ΤΟΥ
ΕΙΧΑΝ	ΚΑΝΕΙΣ	ΟΣΟΣ	ΤΟΥΛΑΧΙΣΤΟ
ΕΙΧΑΤΕ	ΚΑΝΕΝ	ΟΣΟΣΔΗΠΟΤΕ	ΤΟΥΛΑΧΙΣΤΟΝ
ΕΙΧΕ	ΚΑΝΕΝΑ	ΟΣΟΥ	ΤΟΥΣ
ΕΙΧΕΣ	ΚΑΝΕΝΑΝ	ΟΣΟΥΔΗΠΟΤΕ	ΤΟΥΤ?Ν
ΕΚ	ΚΑΝΕΝΑΣ	ΟΣΟΥΣ	ΤΟΥΤΑ
ΕΚΑΣΤΑ	ΚΑΝΕΝΟΣ	ΟΣΟΥΣΔΗΠΟΤΕ	ΤΟΥΤΕΣ
ΕΚΑΣΤΕΣ	ΚΑΠΟΙΑ	ΟΣΩΝ	ΤΟΥΤΗ
ΕΚΑΣΤΗ	ΚΑΠΟΙΑΝ	ΟΣΩΝΔΗΠΟΤΕ	ΤΟΥΤΗΝ
ΕΚΑΣΤΗΝ	ΚΑΠΟΙΑΣ	ΟΤΑΝ	ΤΟΥΤΗΣ
ΕΚΑΣΤΗΣ	ΚΑΠΟΙΕΣ	ΟΤΙ	ΤΟΥΤΟ
ΕΚΑΣΤΟ	ΚΑΠΟΙΟ	ΟΤΙΔΗΠΟΤΕ	ΤΟΥΤΟΙ
ΕΚΑΣΤΟΙ	ΚΑΠΟΙΟΙ	ΟΤΟΥ	ΤΟΥΤΟΙΣ
ΕΚΑΣΤΟΝ	ΚΑΠΟΙΟΝ	ΟΥ	ΤΟΥΤΟΝ
ΕΚΑΣΤΟΣ	ΚΑΠΟΙΟΣ	ΟΥΔΕ	ΤΟΥΤΟΣ
ΕΚΑΣΤΟΥ	ΚΑΠΟΙΟΥ	ΟΥΤΕ	ΤΟΥΤΟΥ
ΕΚΑΣΤΟΥΣ	ΚΑΠΟΙΟΥΣ	ΟΧΙ	ΤΟΥΤΟΥΣ
ΕΚΑΣΤΩΝ	ΚΑΠΟΙΩΝ	Π	ΤΥΧΟΝ
ΕΚΕΙ	ΚΑΠΟΤΕ	ΠΑΛΙ	ΤΩΝ
ΕΚΕΙΝΑ	ΚΑΠΟΥ	ΠΑΝΤΟΤΕ	ΤΩΡΑ
ΕΚΕΙΝΕΣ	ΚΑΠΩΣ	ΠΑΝΤΟΥ	Υ
ΕΚΕΙΝΗ	ΚΑΤ	ΠΑΝΤΩΣ	ΥΠ
ΕΚΕΙΝΗΝ	ΚΑΤΑ	ΠΑΡΑ	ΥΠΕΡ
ΕΚΕΙΝΗΣ	ΚΑΤΙ	ΠΕΡΑ	ΥΠΟ
ΕΚΕΙΝΟ	ΚΑΤΙΤΙ	ΠΕΡΙ	ΥΠΟΨΗ
ΕΚΕΙΝΟΙ	ΚΑΤΟΠΙΝ	ΠΕΡΙΠΟΥ	ΥΠΟΨΙΝ
ΕΚΕΙΝΟΝ	ΚΑΤΩ	ΠΕΡΙΣΣΟΤΕΡΟ	ΥΣΤΕΡΑ
ΕΚΕΙΝΟΣ	ΚΙΟΛΑΣ	ΠΕΡΣΙ	Φ
ΕΚΕΙΝΟΥ	ΚΛΠ	ΠΕΡΥΣΙ	ΦΕΤΟΣ
ΕΚΕΙΝΟΥΣ	ΚΟΝΤΑ	ΠΙΑ	Χ
ΕΚΕΙΝΩΝ	ΚΤΛ	ΠΙΘΑΝΟΝ	ΧΑΜΗΛΑ
ΕΚΤΟΣ	ΚΥΡΙΩΣ	ΠΙΟ	ΧΘΕΣ

ΕΜΑΣ	Λ	ΠΙΣΩ	ΧΤΕΣ
ΕΜΕΙΣ	ΛΙΓΑΚΙ	ΠΛΑΙ	ΧΩΡΙΣ
ΕΜΕΝΑ	ΛΙΓΟ	ΠΛΕΟΝ	ΧΩΡΙΣΤΑ
ΕΜΠΡΟΣ	ΛΙΓΩΤΕΡΟ	ΠΛΗΝ	Ψ
ΕΝ	ΛΟΓΩ	ΠΟΙΑ	ΨΗΛΑ
ΕΝΑ	ΛΟΙΠΑ	ΠΟΙΑΝ	Ω
ΕΝΑΝ	ΛΟΙΠΟΝ	ΠΟΙΑΣ	ΩΡΑΙΑ
ΕΝΑΣ	Μ	ΠΟΙΕΣ	ΩΣ
ΕΝΟΣ	ΜΑ	ΠΟΙΟ	ΩΣΑΝ
ΕΝΤΕΛΩΣ	ΜΑΖΙ	ΠΟΙΟΙ	ΩΣΟΤΟΥ
ΕΝΤΟΣ	ΜΑΚΑΡΙ	ΠΟΙΟΝ	ΩΣΠΟΥ
ΕΝΤΩΜΕΤΑΞΥ	ΜΑΚΡΥΑ	ΠΟΙΟΣ	ΩΣΤΕ
ΕΝΩ	ΜΑΛΙΣΤΑ	ΠΟΙΟΥ	ΩΣΤΟΣΟ
ΕΞ	ΜΑΛΛΟΝ	ΠΟΙΟΥΣ	ΩΧ
ΕΞΑΦΝΑ	ΜΑΣ	ΠΟΙΩΝ	