Ατομική Διπλωματική Εργασία

TWITTERNOW: A SERVICE FOR IDENTIFYING EVENTS USING TWITTER

Κασιανή Πάρη

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ



ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάιος 2015

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

TwitterNow: a service that brings into the users attention events happening in close proximity of their location

Κασιανή Πάρη

Επιβλέπων Καθηγητής
Γιώργος Πάλλης

Η Ατομική Διπλωματική Εργασία υποβλήθηκε προς μερική εκπλήρωση των απαιτήσεων απόκτησης του πτυχίου Πληροφορικής του Τμήματος Πληροφορικής του Πανεπιστημίου Κύπρου

Μάιος 2015

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέπων καθηγητή μου κύριο Γιώργο Πάλλη για την εμπιστοσύνη που έδειξε προς το πρόσωπό μου και μου ανάθεσε την παρούσα διπλωματική εργασία καθώς επίσης και για την ανεκτίμητη βοήθεια του στη μελέτη και την ολοκλήρωση της εργασίας αυτής.

Θα ήθελα να εκφράσω ένα μεγάλο ευχαριστώ στους Δημήτρη Αντωνιάδη και Χαρίτων Ευσταθιάδη για την πολύτιμη υποστήριξη που μου προσέφεραν, καθώς επίσης και για τις πολύτιμες συμβουλές που μου έδιναν όλο αυτό το διάστημα.

Τέλος, θα ήθελα να εκφράσω ιδιαίτερες ευχαριστίες στην οικογένεια μου που με στήριξε ηθικά και ψυχολογικά σε όλη την πορεία κατά τη διάρκεια φοίτησης μου στο Πανεπιστήμιο Κύπρου, αλλά και κατά τη δύσκολη αυτή περίοδο της ολοκλήρωσης της διπλωματικής μου εργασίας.

Περίληψη

Τα κοινωνικά δίκτυα σήμερα έχουν κατακλύσει σε μεγάλο βαθμό τη καθημερινότητα μας αντικαθιστώντας ακόμα και τα μέσα μαζικής ενημέρωσης, καθιστώντας τα ίσως και τη πρώτη πηγή ενημέρωσης. Έρευνες έχουν δείξει πως ένα μεγάλο ποσοστό του ανθρώπινου χρόνου περνιέται σε αυτά αναζητώντας πληροφορίες που θεωρούνται ενδιαφέρον. Παραδείγματα τέτοιων πληροφοριών είναι διάφορες εκδηλώσεις οι οποίες διοργανώνονται ανά τον κόσμο και προσελκύουν το ευρύ κοινό.

Σκοπός της παρούσας διπλωματικής εργασίας είναι η υλοποίηση ενός συστήματος το οποίο θα χρησιμοποιεί το κοινωνικό μέσο Twitter για να εντοπίσει διάφορες πολιτιστικές εκδηλώσεις ή πολιτιστικά γεγονότα με καθορισμένη ημερομηνία και ώρα οι οποίες διοργανώνονται στην περιοχή την οποία ενδιαφέρεται ο χρήστης του συστήματος.

Συγκεκριμένα, το σύστημα θα παρουσιάζεται μέσω μιας ιστοσελίδας ένα ημερολόγιο εκδηλώσεων για τη περιοχή ενδιαφέροντος του χρήστη. Ο χρήστης θα είναι σε θέση να επιλέξει το τον τίτλο της πολιτιστικής εκδήλωσης ή πολιτιστικού γεγονότος και να ενημερωθεί με περισσότερες πληροφορίες για την εκδήλωση την οποία επέλεξε.

Περιεχόμενα

Ευχαριστίες iii
Περίληψη iv
Κεφάλαιο 1 Εισαγωγή 1
1.1 Πρόλογος 1
1.2 Σκοπός Ατομικής Διπλωματικής Εργασίας 2
1.3 Δομή Ατομικής Διπλωματικής Εργασίας 3
Κεφάλαιο 2 Σχετική Βιβλιογραφία 5
2.1 EDCoW αλγόριθμος (Event Detection with Clustering of Wavelet-based Signals) 5
2.1.1 Ανάλυση wavelet 6
2.2 Πρώτη προσέγγιση Naive Bayes αλγόριθμου 8
2.3 Δεύτερη προσέγγιση Naive Bayes αλγόριθμου 9
2.4 Emerging Topic Detection προσέγγιση 11
Κεφάλαιο 3 Εισαγωγή στο Twitter 13
3.1 Τι είναι το Twitter 14
3.2 Τι είναι το Tweet 15
3.3 Hashtag 16
3.4 Σύνδεσμοι 17
3.4.1 Ορισμός και Είδη Υπερσυνδέσμων 17
3.4.1.1 “Επιφανειακοί” σύνδεσμοι (surface links) 18
3.4.1.2 “Βαθιοί” σύνδεσμοι (deep links) 18
3.4.1.3 Σύνδεσμοι “πλαισίου” (framing links) 18
3.4.1.4 “Αυτόματοι ή δυναμικοί υπερσύνδεσμοι” (automatic ή inline/embedded links) 20
3.4.1.5“Περιτυλιγμένοι υπερσυνδέσμοι”(wrap hyperlinks) 20

3.5 Περιτύλιγμα συνδέσμου στο Twitter....20
3.6 Κλειδολέξεις....21
3.6.1 Κλειδολέξεις και Twitter....22
3.6.1.1 Αναζήτηση....22
3.6.1.1 Χρονοδιάγραμμα....22
Κεφάλαιο 4 Τεχνικές Συλλογής και Ανάλυσης Δεδομένων....24
4.1 Τι είναι η Διεπαφή Προγραμματισμού Εφαρμογών (API)....24
4.1.1 Περιγραφή Διεπαφής Προγραμματισμού Εφαρμογών....25
4.2 Twitter API....25
4.2.1 Τελικά σημεία (Endpoints)....26
4.2.2 Είδη API....26
4.2.2.1 REST APIs....26
4.2.2.2 Streaming APIs....27
4.2.2.2.1 Public streams....28
4.2.2.2.2 User streams....29
4.2.2.2.3 Site streams....30
4.2.2.2.4 Διαφορές μεταξύ REST API και Streaming API....30
4.3 Βιβλιοθήκες του Twitter API....31
4.4 Μέθοδοι του Twitter API....35
Κεφάλαιο 5 Στρατηγική Ερευνητικής Μεθοδολογίας....38
5.1 Εισαγωγή....38
5.2 Έρευνα και μέθοδοι συλλογής δεδομένων:....38
5.2.1. Συλλογή δεδομένων με τη χρήση διεπαφών του Twitter....38
5.2.2 Συλλογή δεδομένων με ιχνηλάτιση ιστοσελίδων(web crawling)....39
Κεφάλαιο 6 Μεθοδολογία....41
6.1 Εισαγωγή....41
6.2 Ανάλυση μεθοδολογίας....42
6.3 Παρουσίαση ψευδοκώδικα δεύτερης προσέγγισης....50
6.4.1 Γραφικές αναπαραστάσεις....51

6.4.1 Βήματα αλγορίθμου πρώτη προσέγγισης σε γραφική αναπαράσταση.....51
Επεξήγηση αλγορίθμου πρώτης προσέγγισης: 52
6.4.2 Βήματα αλγορίθμου δεύτερης προσέγγισης σε γραφική αναπαράσταση.....54
Επεξήγηση αλγορίθμου δεύτερης προσέγγισης:.....55
6.5 Αξιολόγηση και η εγκυρότητα.....57
Κεφάλαιο 7 Διαδικτυακή εφαρμογή.....59
7.1 Παρουσίαση διαδικτυακής εφαρμογής.....67
7.2 Τεχνολογίες διαδικτύου.....72
Κεφάλαιο 8 Επίλογος.....73
Βιβλιογραφία.....75
Παράρτημα Α.....77
Παράρτημα Β.....78

Κεφάλαιο 1 Εισαγωγή

1.1 Πρόλογος
1.2 Σκοπός Ατομικής Διπλωματικής Εργασίας
1.3 Δομή Ατομικής Διπλωματικής Εργασίας

1.1 Πρόλογος

Τα τελευταία χρόνια παρατηρείται ραγδαία εξέλιξη των μέσων κοινωνικής δικτύωσης καθιστώντας τα ένα αναπόσπαστο κομμάτι της καθημερινότητας μας. Τα μέσα αυτά φτάνουν πολλές φορές σε σημείο να αντικαταστήσουν ακόμα και τα μέσα μαζικής ενημέρωσης που μέχρι πρότινος αποτελούσαν τη πρώτη πηγή ενημέρωσης για τον άνθρωπο. Όλα τα μέσα κοινωνικής δικτύωσης και πιο συγκεκριμένα το Twitter , δίνουν τη δυνατότητα σε όλους τους χρήστες του να ενημερώνονται άμεσα, έγκαιρα για όλα τα γεγονότα και εκδηλώσεις οι οποίες πραγματοποιούνται σε πραγματικό χρόνο . Με αυτό τον τρόπο γεφυρώνονται όλες οι γεωγραφικές αποστάσεις δίνοντας στους χρήστες τη ψευδαίσθηση πως είναι παρόντες σε αυτά.

Για αυτούς του λόγους όλα τα μέσα κοινωνικής δικτύωσης αποτελούν μια τεράστια πηγή πληροφόρησης και κίνηση πληροφορίας μεταξύ των χρηστών τους στο δίκτυο. Αναλύοντας και επεξεργάζοντας κανείς όλο αυτό τον όγκο πληροφοριών τον οποίο αντλεί από τα μέσα κοινωνικής δικτύωσης, μπορεί να παράξει πιο αναλυτικές και πιο ενδιαφέρον πληροφορίες που θα τον οδηγήσουν σε έγκυρα συμπεράσματα. Συμπεράσματα είτε για τους ίδιους τους χρήστες είτε για συγκεκριμένα θέματα που τους αφορούν.

Μέσα από το μεγάλο όγκο πληροφοριών που υπάρχει στα μέσα κοινωνικής δικτύωσης, στην παρούσα διπλωματική εργασία καταβάλλεται η προσπάθεια να ανιχνευθούν πολιτιστικά γεγονότα με συγκεκριμένη ημερομηνία. Για κάποιον ο οποίος ενδιαφέρεται για πολιτιστικά γεγονότα τα οποία συμβαίνουν σε συγκεκριμένη

γεωγραφική περιοχή είναι δύσκολο να φτάσει σε πληροφορίες οι οποίες έχουν σχέση με αυτά. Αυτή η δυσκολία υπάρχει διότι μπορεί ο ενδιαφερόμενος να μην είναι συνδεδεμένος με τα κατάλληλα άτομα (να μην είναι follower σε κάποιον). Ένα άλλο αίτιο αυτής της δυσκολίας είναι ότι οι πληροφορίες που σχετίζονται με αυτό το πολιτιστικό γεγονός χάνονται μέσα στον μεγάλο όγκο πληροφοριών άλλου θέματος ή γεγονότος.

Για παράδειγμα, σε μεγάλες και πολυπληθής χώρες υπάρχει τεράστιος όγκος πληροφοριών ο οποίος αποτελείται από δεδομένα και πληροφορίες που σχετίζονται με εργασιακά θέματα. Αν κάποιος ο οποίος ενδιαφέρεται για πολιτιστικά θέματα και γεγονότα είναι δύσκολο να πληροφορηθεί για αυτά διότι χάνονται μέσα στο σύνολο των πληροφοριών που σχετίζονται με τα εργασιακά θέματα

1.2 Σκοπός Ατομικής Διπλωματικής Εργασίας

Σκοπός της παρούσας ατομικής διπλωματικής εργασίας είναι η δημιουργία ενός συστήματος που θα δίνει τη δυνατότητα στο χρήστη να αναζητήσει εκδηλώσεις που πραγματοποιούνται σε προγραμματισμένη μέρα και ώρα σε κοντινή γεωγραφική περιοχή στην οποία ενδιαφέρεται π.χ μουσικής, θεάτρου, κινηματογράφου, αθλητισμού κλπ. Αυτό το σύστημα θα υλοποιηθεί μέσω δεδομένων του μέσου κοινωνικής δικτύωσης Twitter. Πιο συγκεκριμένα ο χρήστης θα έχει τη δυνατότητα μπαίνοντας στην ιστοσελίδα να του παρουσιάζεται ένα ημερολόγιο όπου θα εμφανίζονται όλες οι εκδηλώσεις με το τίτλο τους οι οποίες είναι προγραμματισμένες για τη χρονική περίοδο και περιοχή που επέλεξε ο χρήστης. Από εκεί ο χρήστης θα μπορεί επιλέγοντας κάποια από τις εκδηλώσεις να μεταβαίνει αυτόματα στη σελίδα της κάθε εκδήλωσης όπου θα μπορεί να ενημερωθεί καλύτερα και σε μεγαλύτερο βαθμό για τη συγκεκριμένη εκδήλωση.

Μια ερώτηση η οποία μπορεί να γεννηθεί από τον σκοπό της Διπλωματικής Εργασίας είναι το πώς θα ανιχνευτούν τα πολιτιστικά γεγονότα από το Twitter. Η απάντηση σε αυτή την ερώτηση είναι ότι τα γεγονότα θα ανιχνευτούν μέσω των ίδιων των χρηστών του Twitter. Αυτό οφείλεται στο γεγονός ότι αν ένα πολιτιστικό γεγονός πρόκειται να διοργανωθεί , οι χρήστες του Twitter τείνουν να αναρτούν συνέχεια tweets (κείμενο) στο οποίο αναφέρουν το γεγονός και λεπτομέρειες για αυτό. Επίσης, οι

χρήστες του τείνουν να κάνουν retweet (να ξανά-αναρτούν) ένα tweet κάποιου άλλου χρήστη ο οποίος αναφέρει το γεγονός και δίνει πληροφορίες σχετικά με αυτό.

Μια συναφής ερώτηση η οποία ενδέχεται να γεννηθεί είναι το πώς ένα γεγονός μπορεί να διαφοροποιηθεί από την πληθώρα αναρτήσεων tweets τα οποία υπάρχουν στο Twitter. Η συμπεριφορά των χρηστών του Twitter αλλάζει όταν πρόκειται να πραγματοποιηθεί ένα πολιτιστικό γεγονός. Δηλαδή, οι αναρτήσεις των χρηστών αναφέρουν τους συμμετέχοντες του γεγονότος, την ημερομηνία και την τοποθεσία του γεγονότος κτλπ. Επίσης, οι χρήστες μέσα από τις αναρτήσεις τους εκφράζουν τα συναισθήματά τους για το πολιτιστικό γεγονός και έχουν υψηλή επικοινωνία μεταξύ τους. Δηλαδή, ή αναρτούν ξανά ένα ήδη υπάρχον tweet από άλλο χρήστη, ή σχολιάζουν το tweet ενός άλλου χρήστη. Επιπλέον, οι διοργανωτές ενός πολιτιστικού γεγονότος αναρτούν tweets και δίνουν πληροφορίες και λεπτομέρειες για αυτό.

1.3 Δομή Ατομικής Διπλωματικής Εργασίας

Στο κεφάλαιο 2, παρουσιάζεται το γνωστικό υπόβαθρο της παρούσας διπλωματικής εργασίας. Επεξηγούνται περιγραμματικά διάφοροι αλγόριθμοι, τεχνικές , προσεγγίσεις και μέθοδοι τα οποία χρησιμοποιήθηκαν σε παρόμοιες έρευνες.

Στο κεφάλαιο 3, γίνεται μια εισαγωγή για το Twitter και γίνεται επεξήγηση των χαρακτηριστικών του και πως αυτά θα βοηθήσουν στη ολοκλήρωσης της παρούσας διπλωματικής εργασίας.

Στο κεφάλαιο 4, γίνεται αναφορά στις τεχνικές συλλογής και ανάλυσης δεδομένων από το Twitter. Δηλαδή, περιγράφεται η διεπαφή προγραμματισμού εφαρμογών του Twitter(API) και πώς αυτή η διεπαφή χρησιμοποιήθηκε για την επίτευξη του στόχου της παρούσας διπλωματικής εργασίας

Το πέμπτο κεφάλαιο αφορά την στρατηγική της ερευνητικής μεθοδολογίας. Παρουσιάζονται οι τρόποι με τους οποίους συλλέχθηκαν τα δεδομένα για χρήση και επεξεργασία και εξαγωγή πληροφοριών

Στο κεφάλαιο 6, παρουσιάζεται η μεθοδολογία η οποία ακολουθήθηκε για την ολοκλήρωση της παρούσας διπλωματικής εργασίας. Επίσης, παρουσιάζονται και οι δύο προσεγγίσεις της.

Στο κεφάλαιο 7, παρουσιάζεται η διαδικτυακή εφαρμογή η οποία εμφανίζει τα αποτελέσματα της μεθοδολογίας που χρησιμοποιήθηκε (δηλαδή τα πολιτιστικά γεγονότα τα οποία ανιχνεύθηκαν). Επιπλέον, επεξηγούνται ποιες τεχνολογίες διαδικτύου χρησιμοποιήθηκαν έτσι ώστε να υλοποιηθεί η εφαρμογή.

Στο κεφάλαιο 8, γίνεται επίλογος και αναφορά σε μελλοντικές εργασίες

Κεφάλαιο 2 Σχετική Βιβλιογραφία

2.1 EDCoW αλγόριθμος (Event Detection with Clustering of Wavelet-based Signals)
5.1.1 Ανάλυση wavelet
2.2 Πρώτη προσέγγιση Naive Bayes αλγόριθμου
2.3 Δεύτερη προσέγγιση Naive Bayes αλγόριθμου
2.4 Emerging Topic Detection προσέγγιση

Σε αυτό το κεφάλαιο θα παρουσιαστεί συνοπτικά η βιβλιογραφία που ερευνήθηκε στα πλαίσια της παρούσας διπλωματικής εργασίας. Τα άρθρα τα οποία μελετήθηκαν αναφέρουν μεθόδους, τεχνικές, προσεγγίσεις και αλγόριθμους τα οποία είναι σχετικά με την λεξικογραφική ανάλυση, την εύρεση συχνότητας σημάνσεων, την ταξινόμηση και ομαδοποίησή τους.

2.1 EDCoW αλγόριθμος (Event Detection with Clustering of Wavelet-based Signals)

Ο αλγόριθμος EDCoW μπορεί να χρησιμοποιηθεί για ανίχνευση γεγονότων από το Twitter και να βοηθήσει στην ολοκλήρωση της παρούσας διπλωματικής εργασίας. Ο συγκεκριμένος αλγόριθμος μπορεί να δημιουργήσει σημάνσεις για μεμονωμένες λέξεις εφαρμόζοντας κάποια ανάλυση. Αυτή η ανάλυση ονομάζεται wavelet και ασχολείται με τον υπολογισμό της συχνότητας των σημάνσεων των μοναδικών λέξεων. Στη συνέχεια φιλτράρει τις τετριμμένες λέξεις λαμβάνοντας υπόψη τις αντίστοιχες σημάνσεις των συσχετίσεων μεταξύ τους. Μετά το φιλτράρισμα , ο αλγόριθμος EDCoW μετρά την συσχέτιση των υπολοίπων σημάνσεων των λέξεων οι οποίες δεν έτυχαν φιλτραρίσματος , και δημιουργεί ομάδες οι οποίες δομούνται σε μορφή γράφου. Για να διαφοροποιηθούν τα σημαντικά γεγονότα από τα ασήμαντα, ο αλγόριθμος EDCoW προσπαθεί να μετρήσει την σημαντικότητα των γεγονότων τα οποία εμφανίζονται. Η σημαντικότητα εξαρτάται από δύο παράγοντες: τον αριθμό των μοναδικών λέξεων και τη συσχέτιση μεταξύ τους.

2.1.1 Ανάλυση wavelet

Η ανάλυση wavelet εφαρμόζεται στον αλγόριθμο EDCoW έτσι ώστε να δημιουργηθούν σημάνσεις για τις μοναδικές λέξεις. Ο πυρήνας της ανάλυσης wavelet είναι η μετατροπή wavelet. Η μετατροπή wavelet μετατρέπει τις σημάνσεις από το time domain στο time-scale domain (κλίμακα - αντίθετο της συχνότητας-). Χωρίζει μια σήμανση σε παραμέτρους wavelet και ένα σύνολο από γραμμικά ανεξάρτητες συναρτήσεις, την λεγόμενη οικογένεια wavelet. Μια οικογένεια wavelet δημιουργείται από την κλιμάκωση και την μετάφραση μιας «μητέρας» της οικογένειας. Η κλιμάκωση της «μητέρας» αντιστοιχεί στο πόσο εκτείνεται και πόσο συρρικνώνεται σε βάθος χρόνου.

Η μετατροπή wavelet χωρίζεται σε δύο άλλες μετατροπές οι οποίες ονομάζονται συνεχόμενη μετατροπή wavelet (continuous wavelet transformation (CWT)) και διακριτή μετατροπή wavelet (discrete wavelet transformation (DWT)). Οι δύο αυτές μετατροπές έχουν μια κύρια διαφορά. Η συνεχόμενη μετατροπή wavelet (continuous wavelet transformation (CWT)) προσφέρει περιττές αναπαραστάσεις των σημάνσεων όταν προβεί σε ανάλυση. Επίσης είναι πολύ χρονοβόρο για να υπολογισθούν οι σημάνσεις των λέξεων κατευθείαν. Σε αντίθεση, η διακριτή μετατροπή wavelet (discrete wavelet transformation (DWT)) δεν προσφέρει περιττή αναπαράσταση των σημάνσεων των λέξεων αλλά προσφέρει μια αποδοτική αναπαράσταση των σημάνσεων των λέξεων.

Γενικά, ο αλγόριθμος EDCoW αποτελείται από τρία κύρια στοιχεία τα οποία είναι σημαντικά για την υλοποίησή του. Τα τρία αυτά στοιχεία είναι η κατασκευή σημάνσεων, ο υπολογισμός της συσχέτισης και ο σχεδιασμός γράφου.

Για την κατασκευή των σημάνσεων πρέπει να δημιουργηθούν n-grams. Τα n-grams είναι μια ακολουθία από τερματικά σύμβολα (λέξεις ή χαρακτήρες) τα οποία στην περίπτωσή μας εμφανίζονται στο κείμενο του tweet. Η δημιουργία των n-grams γίνεται σε δύο στάδια.

Στο πρώτο στάδιο η δημιουργία των n-grams μπορεί να υπολογισθεί από τον πολλαπλασιασμό της διαίρεσης του αριθμού των tweets που αναφέρουν μια συγκεκριμένη λέξη με όλο τον αριθμό των tweets του συνόλου δεδομένων με τον

λογάριθμο της διαίρεσης του αθροίσματος του αριθμού των tweets που αναφέρουν μια συγκεκριμένη λέξη διά όλο τον αριθμό των tweets του συνόλου δεδομένων σε βάθος χρόνου.

Στο δεύτερο στάδιο η σήμανση της λέξης δημιουργείται με τη βοήθεια ενός sliding windo. Η εξίσωση για την δημιουργία των n-grams είναι η ίδια όπως εξηγήθηκε πιο πάνω όμως προστίθεται και το sliding window. Κάθε σημείο σε αυτό το στάδιο ελέγχεται αν υπάρχει αλλαγή στο sliding window.

Για να μπορεί κάποιος να μιλήσει τον υπολογισμό της συσχέτισης μεταξύ των λέξεων αξίζει να επαναληφθεί ότι ο αλγόριθμος EDCoW ανιχνεύει γεγονότα με το να ομαδοποιεί τις λέξεις που έχουν παρόμοια σημασία. Για την επίτευξη αυτού, οι ομοιότητες μεταξύ των λέξεων πρέπει να υπολογιστούν. Για αυτό το λόγο εφαρμόζεται η συσχέτιση πάνω στις ίδιες της σημάνσεις.

Για τη συσχέτιση μεταξύ δύο σημάνσεων σε βάθος χρόνου υπολογίζεται το γινόμενο των δύο σημάνσεων. Με αυτό τον τρόπο μπορεί να μετρηθεί η ομοιότητα μεταξύ των δύο σημάνσεων ως συνάρτηση του χρόνου καθυστέρησης που εφαρμόζεται σε αυτά.

Τρίτο και τελευταίο στοιχείο του αλγορίθμου EDCoW είναι η κατασκευή κατευθυνόμενου γράφου χωρίς να προσδιοριστεί από την αρχή το μέγεθος του γράφου. Κάθε κορυφή αποτελείται από τις σημάνσεις που δημιουργήθηκαν από τη συσχέτιση των λέξεων. Στον γράφο, υπάρχει μια ακμή ανάμεσα σε δύο κορυφές εάν οι δύο σημάνσεις δεν έτυχαν φιλτράρισμα. Άρα έχουν συσχέτιση μεταξύ τους. Για την ανίχνευση των γεγονότων μπορεί να ειπωθεί ότι είναι ένας διαχωρισμός του γράφου σε υπό-γράφους θα οδηγήσει σε γεγονότα. Δηλαδή, οι υπό-γράφοι είναι γεγονότα. Για να μετρηθεί η ποιότητα του διαχωρισμού του γράφου σε υπό-γράφους πρέπει να υπολογιστεί το άθροισμα των βαρών των ακμών που συνδέουν τους υπό-γράφους μεταξύ τους (modularity)

Για να βρεθεί πόσο σημαντικό είναι το γεγονός το ποίο ανιχνεύθηκε, πρέπει να ελεγχθεί ο γράφος. Ο γράφος δεν μπορεί να αποτελείται το λιγότερο από δύο κορυφές (δηλαδή, ένα γεγονός να περιγράφεται το λιγότερο από δύο λέξεις).

2.2 Πρώτη προσέγγιση Naive Bayes αλγόριθμου

Ο αλγόριθμος Naïve Bayes είναι ένας αλγόριθμος ταξινόμησης που βασίζεται σε θεωρήματα του Bayes. Ο αλγόριθμος χρησιμοποιεί Bayesian τεχνικές (πιθανολογικές και στατιστικές τεχνικές) και δεν λαμβάνει υπόψη εξαρτήσεις οι οποίες μπορεί να υπάρξουν. Επίσης, ο Naïve Bayes ενεργεί τους λιγότερους υπολογισμούς σε σύγκριση με παρόμοιους αλγορίθμους και είναι γρήγορος στη δημιουργία μοντέλων εξόρυξης για να την ανακάλυψη σχέσεων μεταξύ δεδομένων εισόδου και δεδομένων πρόβλεψης. Χρησιμοποιείται για να διευρυνθούν τα δεδομένα και στη συνέχεια να εφαρμοστούν αποτελέσματα για τη δημιουργία μοντέλων εξόρυξης σε συνδυασμό με άλλους αλγόριθμους που παρέχουν πιο ακριβή αποτελέσματα.

Σε αυτή την προσέγγιση χρησιμοποιούνται τα μοντέλα CAN και TAN. Τα μοντέλα CAN και TAN ασχολούνται με την επεξεργασία και την ταξινόμηση με βάση το κείμενο και τα χαρακτηριστικά του κειμένου. Σε συνδυασμό με πιθανολογικά μοντέλα αποφασίζουν κατά πόσο ένα tweet ανήκει σε ένα πολιτιστικό γεγονός ή όχι.

Στην περίπτωση αυτή της προσέγγισης, εντάχθηκαν τα n-grams στον γνωστό αλγόριθμο Naïve Bayes έτσι ώστε να αντιμετωπισθούν οι ελλείψεις οι οποίες υπάρχουν στον πρωτότυπο Naïve Bayes. Το αποτέλεσμα αυτής της ένταξης είναι η επιστροφή ενός μοντέλου CAN (Chain Augmented Naive Bayes).

Το μοντέλο CAN προσφέρει πολλά πλεονεκτήματα. Το πρώτο κύριο πλεονέκτημα του είναι ότι μειώνει τις μη-εξαρτημένες υποθέσεις του πρωτότυπου Naïve Bayes εξακολουθώντας να παράγει αποτελεσματική εξαγωγή συμπερασμάτων. Το δεύτερο κύριο πλεονέκτημα του είναι ότι το μοντέλο αυτό επιτρέπει τη σωστή εφαρμογή εξελιγμένων τεχνικών για μοντελοποίηση με χρήση στατιστικής γλώσσας. Αυτό επιτρέπει να εξάγονται καλύτερες εκτιμήσεις από ότι πριν. Στην περίπτωση αυτή, θα χρησιμοποιηθεί αυτή η τεχνική σε διάφορα προβλήματα ταξινόμησης κειμένου. Πρέπει να αποδειχθεί ότι μια γλώσσα είναι ανεξάρτητη έτσι ώστε να μπορούν να δημιουργηθούν ανεξάρτητοι ταξινομητές.

Η ταξινόμηση κειμένου αναφέρεται στο πρόβλημα της αυτόματης ανάθεσης ενός δοθέντος κειμένου σε καθορισμένες κατηγορίες. Λόγω της ραγδαίας ανάπτυξης του κειμένου σε ψηφιακή μορφή, η κατηγοριοποίηση κειμένου έγινε σημαντικό πεδίο της έρευνας μεγάλου όγκου δεδομένων. Οι Naive Bayes ταξινομητές επεκτάθηκαν με πολλούς τρόπους που βοηθούν σε καλύτερα και σημαντικότερα αποτελέσματα. Μια

πολύ ενδιαφέρον επέκταση είναι ο ταξινομητής Tree Augmented Naive Bayes classifier (TAN) ο οποίος επιτρέπει να δομηθεί ως ένα δένδρο το οποίο λαμβάνει υπόψη την εξάρτηση μεταξύ των μεταβλητών. Ο εν λόγω ταξινομητής μπορεί να μοντελοποιηθεί στο μοντέλο CAN το οποίο αναφέρθηκε πιο πάνω. Δηλαδή, οι μεταβλητές που χρησιμοποιούνται στο δένδρο το οποίο εξάχθηκε από το μοντέλο TAN σχηματίζουν το CAN. Επίσης οι μεταβλητές που εξάχθηκαν χρησιμοποιούν τεχνικές για να εξάγουν στατιστικές από τα n-grams. Το αποτέλεσμα αυτού του συνδυασμού είναι ένας απλός και εξαιρετικός ταξινομητής.

Εκτός από τη βελτίωση του πρωτότυπου Naïve Bayes, τα μοντέλα CAN έχουν και άλλα πλεονεκτήματα σε σχέση με άλλες ταξινομητές κειμένου. Ένα κοινό σημείο που έχουν όλοι οι ταξινομητές κειμένου είναι ότι μεταχειρίζονται την ταξινόμηση κειμένου ως τυπικό πρόβλημα ταξινόμησης και μειώνουν την διαδικασία μάθησης σε δύο απλά βήματα. Αυτά τα δύο απλά βήματα είναι το feature engineering και το discriminant learning σε όλα τα χαρακτηριστικά. Το feature engineering προσφέρει καλή απόδοση αλλά με κάποια μειονεκτήματα. Το πρώτο μειονέκτημα είναι ότι η κατασκευή των ταξινομητών εξαρτάται από την γλώσσα στην οποία είναι γραμμένες οι λέξεις. Ένα δεύτερο μειονέκτημα είναι ότι υπάρχει πιθανότητα να υπάρχουν πολλά χαρακτηριστικά τα οποία πρέπει να ληφθούν υπόψη για την ταξινόμηση κειμένου και οι τυπικές προσεγγίσεις δεν επιστρέφουν αποδοτικά αποτελέσματα. Συνήθως, δεν υπάρχει ακρίβεια στην ταξινόμηση άρα οι πληροφορίες που θα εξαχθούν είναι λίγες και ανακριβείς.

2.3 Δεύτερη προσέγγιση Naive Bayes αλγόριθμου

Στόχος αυτής της προσέγγισης είναι να κατηγοριοποιηθούν τα tweets σε δύο κατηγορίες (classification model). Αυτές οι κατηγορίες είναι η κατηγορία στην οποία τα κείμενα σχετίζονται με γεγονότα και η δεύτερη κατηγορία είναι τα κείμενα τα οποία δεν σχετίζονται με γεγονότα. Για να υλοποιηθούν αυτές οι δύο κατηγορίες πρέπει να εξαχθούν από τα κείμενα των tweets χαρακτηριστικά κειμένου και n-grams.

Το μοντέλο ταξινόμησης είναι βασισμένο σε κατηγοριοποίηση κειμένου (text classification). Η κατηγοριοποίηση θα γίνει με τον αλγόριθμο Naive Bayes ο οποίος θα χρησιμοποιήσει τα χαρακτηριστικά του κειμένου και n-grams που εξάχθηκαν. Στη συνλεχεια θα ποχωρήσει σε κατηγοριοποίηση

Για να δοκιμαστεί το μοντέλο ταξινόμησης, μέσα από τα tweets θα ψάχνονται για τα τρία κύρια χαρακτηριστικά από τα οποία αποτελούνται τα γεγονότα. Τα τρία αυτά χαρακτηριστικά είναι η ώρα, τα πρόσωπα που εμπλέκονται στα γεγονότα και οι τοποθεσίες στις οποίες πραγματοποιούνται τα γεγονότα. Όσα tweets έχουν αυτά τα χαρακτηριστικά ή λείπουν κάποια και τα υπόλοιπα μπορούν να συμπεραθούν, εντάσσονται στην κατηγορία των tweets που σχετίζονται με γεγονότα αλλιώς εντάσσονται στην κατηγορία των tweets που δεν σχετίζονται.

Όπως αναφέρθηκε και πιο πάνω ο τρόπος με τον οποίο θα υλοποιηθεί ο Naïve Bayes αλγόριθμος θα είναι με την χρήση των n-grams διότι δεν πρέπει να υπάρχουν ορθογραφικά λάθη και δεν μπορούν να προκύψουν stemming procedures.

Για να βρεθεί η πιθανότητα αν ένα tweet ανήκει στην κατηγορία στην οποία τα tweets σχετίζονται με γεγονότα ή όχι πρέπει να χρησιμοποιηθούν πιθανολογικές εξισώσεις. Σε αυτή την εξίσωση λαμβάνεται υπόψη ο αριθμός όλων των tweets και ο αριθμός των tweets που πιστεύεται ότι αναφέρεται σε ένα γεγονός.

Για να βρεθεί αν ένα n-gram ανήκει σε μια από τις δύο κατηγορίες διαιρείται ο αριθμός εμφάνισης του n-gram στην κατηγορία με το συνολικό αριθμό n-grams της κατηγορίας. Για να βρεθεί η πιθανότητα της κάθε κατηγορίας διαιρείται ο αριθμός των n-grams της συγκεκριμένης κατηγορίας με τον συνολικό αριθμό των n-grams και των δύο κατηγοριών.

Αφού κατηγοριοποιηθούν τα tweets πρέπει να τα "ονομάσουμε" . Αυτό μπορεί να γίνει με δύο τρόπους. Ο πρώτος τρόπος είναι χειροκίνητα (manually). Δηλαδή, να χρησιμοποιηθούν τα URLs που μας προσφέρουν υπηρεσίες όπως οι bit.ly και oil.ly και να ληφθούν υπόψη μόνο τα tweets που συμπεριλαμβάνουν τις λέξεις “/event/”, “/artists/” και “/venue/”. Ο δεύτερος τρόπος είναι ο ημι-αυτόματος. Για αυτόν τον τρόπο πρέπει να συμπεριληφθούν τα tweets ενός γεγονότος στην κατηγορία που σχετίζονται με γεγονότα όταν αναφέρουν την ώρα , τα ονόματα και λέξεις με κεφαλαία γράμματα καθώς επίσης να προσδιορίζονται η ώρα και τα πρόσωπα. Για να υπάρξει καλύτερη και αποδοτικότερη κατηγοριοποίηση πρέπει να αφαιρεθούν οι τερματικές λέξεις, τα σημεία στίξης, τα hashtags και τα "μικρά" URLs

2.4 Emerging Topic Detection προσέγγιση

Ένα tweet θεωρείτε αναδυόμενο όταν εμφανίζεται πολλές φορές σε σύντομο χρονικό διάστημα ενώ στο παρελθόν εμφανιζόταν σπάνια. Επίσης για να κατανοηθεί η σημαντικότητα ενός tweet βασιζόμαστε στις πηγές που το έχουν αναφέρει χρησιμοποιώντας τον αλγόριθμο Page Rank. Δημιουργείται ένας γράφος ο οποίος ενώνει τους σημαντικούς όρους με σχετικές λέξεις κλειδιά έτσι ώστε να ανιχνευθούν τα σημαντικά γεγονότα.

Για να ονομαστεί ένα tweet ως σημαντικό και να παραπέμπει σε ένα σημαντικό γεγονός, ακολουθείται μια διαδικασία πέντε βημάτων: Πρώτο βήμα είναι η συλλογή των tweets και η αντιμετώπισή του ως vectors που αποτελούνται από όρους με την συχνότητα που εμφανίζονται. Το δεύτερο βήμα είναι να οριστεί ένας κατευθυνόμενος γράφος των ενεργών χρηστών βασισμένος στις κοινωνικές τους σχέσεις και να μετρηθεί η αξιοπιστία λαμβάνοντας πληροφορίες από τον αλγόριθμο Page Rank. Το τρίτο βήμα είναι να μοντελοποιηθεί για κάθε όρο ο κύκλος ζωής του σύμφωνα με μια θεωρία η οποία εξετάζει πόσο έμπιστες είναι οι πηγές. Τέταρτο βήμα είναι να επιλεχθεί ένα σύνολο σημαντικών tweets κατατάσσοντας τις λέξεις κλειδιά ανάλογα με το πόσες φορές χρησιμοποιήθηκαν. Το πέμπτο και τελευταίο βήμα είναι η δημιουργία ενός γράφου μέσα στον οποίο μπορεί κάποιος να πλοηγηθεί εύκολα. Αυτός ο γράφος ενώνει τους όρους που ορίστηκαν ως σημαντικοί με τα θέματα που ήδη είναι σίγουρο ότι είναι σημαντικά.

Γίνεται χρήση τεχνικών ομαδοποίησης (clustering) για να ομαδοποιηθούν τα tweets που περιέχουν παρόμοιο περιεχόμενο και στη συνέχεια να φιλτραριστούν.
Για να μπορούν να προτείνουν θέματα στους χρήστες, γίνεται χρήση δύο τεχνικών αλλά με διαφορετικό αλγόριθμο για το κάθε ένα. Ο πρώτος αλγόριθμος αναφέρει και λαμβάνει υπόψη το τι επέλεξαν οι υπόλοιποι χρήστες που πραγματοποίησαν αναζήτηση για παρόμοια θέματα. Ο δεύτερος αλγόριθμος αναλύει τη σημασιολογία του περιεχομένου χωρίς να ληφθεί υπόψη η προέλευση τους

Ακόμα ένας τρόπος για να μπορούμε να ανταπεξέλθουμε με τις πολλές πληροφορίες που υπάρχουν στο διαδίκτυο είναι η ανάλυση των λέξεων κλειδιών ανάλογα με τη διάρκεια τους μέσα στο χρόνο, δηλαδή ανάλογα με τη διαχρονικότητα τους .

Επίσης, η αντιμετώπιση του περιεχομένου του διαδικτύου το οποίο είναι πραγματικού χρόνου και αυτός μπορεί να λειτουργήσει ως ένα αισθητήρας ο οποίος μπορεί να καταλάβει το τι συμβαίνει στον κόσμο χωρίς καθυστέρηση. Πρέπει όμως να αποφευχθεί η σύγκρουση των σημαντικών όρων μέσω σχετικών αλγορίθμων . Με αυτόν τον τρόπο μειώνεται η συχνότητα εμφάνισης της λέξης αφού ήδη έχει επισημανθεί ως σημαντική.

Κεφάλαιο 3 Εισαγωγή στο Twitter

3.1 Τι είναι το Twitter

3.2 Τι είναι το Tweet

3. 3 Hashtag

3.4 Σύνδεσμοι (Links)

3.4.1 Ορισμός & Είδη Υπερσυνδέσμων (Hyperlinks)

3.4.1.1 "Επιφανειακοί" σύνδεσμοι (Surface Links)

3.4.1.2 "Βαθιοί" σύνδεσμοι (Deep Links)

3.4.1.3 Σύνδεσμοι "πλαισίου" (Framing Links)

3.4.1.4 "Αυτόματοι ή δυναμικοί υπερσύνδεσμοι" (Automaic ή Inline/Embedded Link)

3.5 Περιτύλιγμα συνδέσμου στο Twitter

3.6 Κλειδολέξεις στο Twitter (keywords)

3.6.1 Αναζήτηση

3.6.2 Χρονοδιάγραμμα

3.7 Τερματικές λέξεις (Stopwords)

Σε αυτό το κεφάλαιο θα γίνει μια εισαγωγή στο Twitter και στα χαρακτηριστικά του. Γίνεται μια εισαγωγή για τη σημασία του Twitter και στη συνέχεια επεξηγείται αναλυτικά και ξεχωριστά κάθε χαρακτηριστικό του. Με τον όρο χαρακτηριστικά εννοούμε τον τρόπο με τον οποίο ένας άνθρωπος μπορεί να χρησιμοποιήσει το Twitter αλλά και να περιηγηθεί σε αυτό. Στις επόμενες σελίδες αυτού του κεφαλαίου εξηγούνται χαρακτηριστικά όπως το tweet,το hashtag, οι κλειδολέξεις (keywords), οι τερματικές λέξεις (stopwords), οι συνδέσμοι (URLs) και ο τρόπος περιτυλίγματος τους που χρησιμοποιεί το Twitter

3.1 Τι είναι το Twitter

Το Twitter είναι μια υπηρεσία κοινωνικής δικτύωσης η οποία επιτρέπει στους χρήστες να μοιράζονται σύντομα μηνύματα κειμένου μεγέθους 140 χαρακτήρων τα οποία ονομάζονται tweets.

Όσοι χρήστες έχουν λογαριασμό στο Twitter μπορούν να διαβάσουν και να δημοσιεύσουν tweets, αλλά όσοι χρήστες δεν έχουν λογαριασμό μπορούν να τα διαβάσουν μόνο. Οι χρήστες έχουν πρόσβαση στο Twitter μέσω της ιστοσελίδας του Twitter και μέσω του κινητού τους τηλεφώνου χρησιμοποιώντας εφαρμογή Android ή iOS.

Η υπηρεσία αυτή μέσα σε πολύ λίγα χρόνια κέρδισε εκατομμύρια ενεργούς χρήστες ανά τον κόσμο οι οποίοι δημοσιεύουν πάνω από τριακόσια εκατομμύρια tweets (μηνύματα κειμένου) κάθε μέρα. Αυτά τα μηνύματα, είναι ένας τρόπος για να ενώνει τους χρήστες οι οποίοι εκφράζουν τα συναισθήματά τους ή τις σκέψεις τους ή τις απόψεις τους ή το τι συμβαίνει ή θα συμβεί κάποια στιγμή, καθώς επίσης να ενημερώνονται από αυτό.

Ένα tweet αποσκοπεί σε διάφορους σκοπούς: στην ανταλλαγή, στην οργάνωση και στην εύρεση περιεχομένου από τους χρήστες και την έμμεση επαφή μεταξύ τους.

Το Twitter μπορεί να χρησιμοποιηθεί από τις επιχειρήσεις για την ενημέρωση του κοινού για τις υπηρεσίες τους, για την συγκέντρωση πληροφοριών για την αγορά σε πραγματικό χρόνο, καθώς επίσης και για τη δημιουργία επαγγελματικών σχέσεων με τυχόν πελάτες και συνεργάτες.

Άλλη χρήση του Twitter είναι να χρησιμοποιηθεί ως μέσο ενημέρωσης. Το Twitter αντικαθιστά την τηλεόραση, το ραδιόφωνο, την ψυχαγωγία και τη διάδοση των ειδήσεων. Το Twitter έχει την ικανότητα να διαδίδει πληροφορίες και ειδήσεις σε πραγματικό χρόνο. Τα μέσα μαζικής ενημέρωσης χρειάζονται πολύ περισσότερο χρόνο από το Twitter. Μέσα από το Twitter μπορεί κάποιος χρήστης να ενημερώνει τους υπόλοιπους χρήστες για μεγάλα , ενδιαφέρον και σημαντικά γεγονότα που συμβαίνουν ανά τον κόσμο τη δεδομένη στιγμή.

3.2 Τι είναι το Tweet

Το tweet είναι ένα μικρό μήνυμα που περιορίζεται στους 140 χαρακτήρες και δημοσιεύεται στην ιστοσελίδα του Twitter. Σε αυτό το μήνυμα μπορεί να γραφτεί κείμενο, σύνδεσμοι ιστοσελίδων, σύνδεσμοι εικόνων ή βίντεο μια ιστοσελίδας, αναφορές συγκεκριμένων χρηστών (mention) και η δίεση(hashtag).

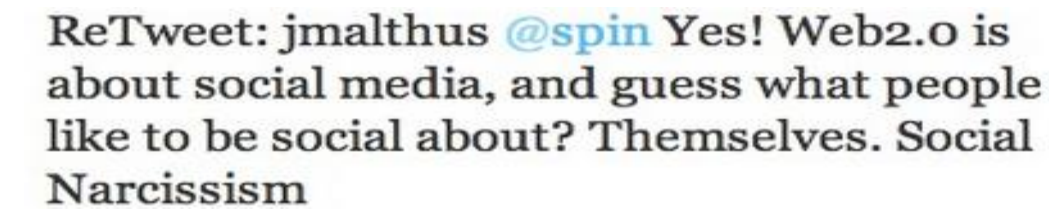


Εικόνα 1: Παράδειγμα tweet στο οποίο χρησιμοποιούνται το κείμενο, σύνδεσμος βίντεο και δίεση

Ένας χρήστης έχει την δυνατότητα να κάνει retweet. Με τον όρο αυτό εννοούμε ότι ένας χρήστης δημοσιεύει ξανά ένα tweet ενός άλλου χρήστη το οποίο βρίσκει ενδιαφέρον και θέλει να το δημοσιεύσει και στο δικό του προφίλ. Όπως φαίνεται στην εικόνα 1, ένα retweet ξεκινά από RT (συντομογραφία του Retweet) ή ReTweet(όπως στην εικόνα 2): @το όνομα του χρήστη από τον οποίο έγινε το retweet και στη συνέχεια το αυθεντικό μήνυμα μαζί με προαιρετικά σχόλια αυτού που κάνει το retweet χωρίς να υπερβαίνει το όριο των 140 χαρακτήρων. Οι χρήστες συνήθως χρησιμοποιούν τα retweets για να μοιραστούν με τους υπόλοιπους χρήστες ένα ευχάριστο ή ένα πολύ σημαντικό γεγονός το οποίο θεωρούν ότι πρέπει να πάρει περισσότερες διαστάσεις.



Εικόνα 2: Παράδειγμα retweet χρησιμοποιώντας τα αρχικά "RT"



Εικόνα 3: Παράδειγμα retweet χρησιμοποιώντας τη λέξη ReTweet

3.3 Hashtag

Το hashtag (δίεση) έγινε ένα από τα πιο αναγνωρίσιμα σύμβολα της τελευταίας δεκαετίας και έγινε ένα από τα πιο δυνατά εργαλεία. Είναι ένα είδος ετικέτας το οποίο χρησιμοποιούν όλα τα μέσα κοινωνικής δικτύωσης. Οι χρήστες χρησιμοποιούν το hashtag για αναζήτηση ενός συγκεκριμένου θέματος ή προσωπικού περιεχομένου ή/και επαγγελματικού επιπέδου. Η χρήση ή/και η δημιουργία του hashtag γίνεται με την τοποθέτηση μιας δίεσης στην αρχή ή στο τέλος μιας λέξης ή αριθμού ή κειμένου χωρίς κενά (spaces). Τότε το tweet κατηγοριοποιείται με αποτέλεσμα όταν γίνει αναζήτηση ενός συγκεκριμένου hashtag ή πατώντας πάνω σε ένα συγκεκριμένο hashtag θα παρουσιαστούν όσα μηνύματα κειμένου (tweets) έχουν επισημανθεί με αυτό.



Εικόνα 4 : Παράδειγμα αναζήτησης με το hashtag χρησιμοποιώντας την λέξη "ajax"

3.4 Σύνδεσμοι

Οι υπερσύνδεσμοι είναι γραμμές κώδικα λογισμικού, που εμπεριέχουν συγκεκριμένες εντολές προς τον φυλλομετρητή του διαδικτυακού χρήστη για τη μετάβαση μεταξύ διάφορων ιστοσελίδων. Η δυνατότητα δημιουργίας υπερσυνδέσμων βοήθησε στη διευκόλυνση της ελεύθερης διασύνδεσης μεταξύ κόμβων (ιστοσελίδων του Παγκόσμιου Ιστού) στον Παγκόσμιο Ιστό αλλά και στην ευκολότερη περιήγηση ενός χρήστη σε αυτόν.

Σήμερα το διαδίκτυο έχει αναδειχθεί ως το πιο εξελιγμένο μέσο ανθρώπινης επικοινωνίας, συσσωρεύοντας στους διακομιστές (servers) του το σύνολο της ανθρώπινης γνώσης. Η λειτουργία των υπερσυνδέσμων αποτελεί τη βάση, πάνω στην οποία οικοδομείται ο παγκόσμιος ιστός, η υπερκείμενη τεχνολογία του διαδικτύου, που κοινωνικοποίησε και παγκοσμιοποίησε τη χρήση του, αναδεικνύοντας τις απεριόριστες δυνατότητες επικοινωνίας, πληροφόρησης, επιμόρφωσης, ψυχαγωγίας και επιχειρηματικότητας.

3.4.1 Ορισμός και Είδη Υπερσυνδέσμων

Ως σύνδεσμος ή υπερσύνδεσμος (link ή hyperlink) ορίζεται το στοιχείο κώδικα γλώσσας προγραμματισμού HTML, που εμπεριέχει εντολή προς τον φυλλομετρητή (browser) του χρήστ. Ο σύνδεσμός είναι υπεύθυνος για την καταφόρτωση (download) των δεδομένων, που βρίσκονται σε ηλεκτρονική διεύθυνση URL η οποία είναι αυστηρά ορισμένη. Η διεύθυνση URL παρέχει τη δυνατότητα στο ψηφιακό περιβάλλον να μεταβαίνει απεθείας σε διαφορετικό τμήμα της ίδιας ιστοσελίδας ή σε άλλη ιστοσελίδα.

Ανάλογα με την τεχνική τους φύση, οι υπερσύνδεσμοι μπορούν ενδεικτικά να διακριθούν σε τέσσερις κατηγορίες :

3.4.1.1 "Επιφανειακοί" σύνδεσμοι (surface links)

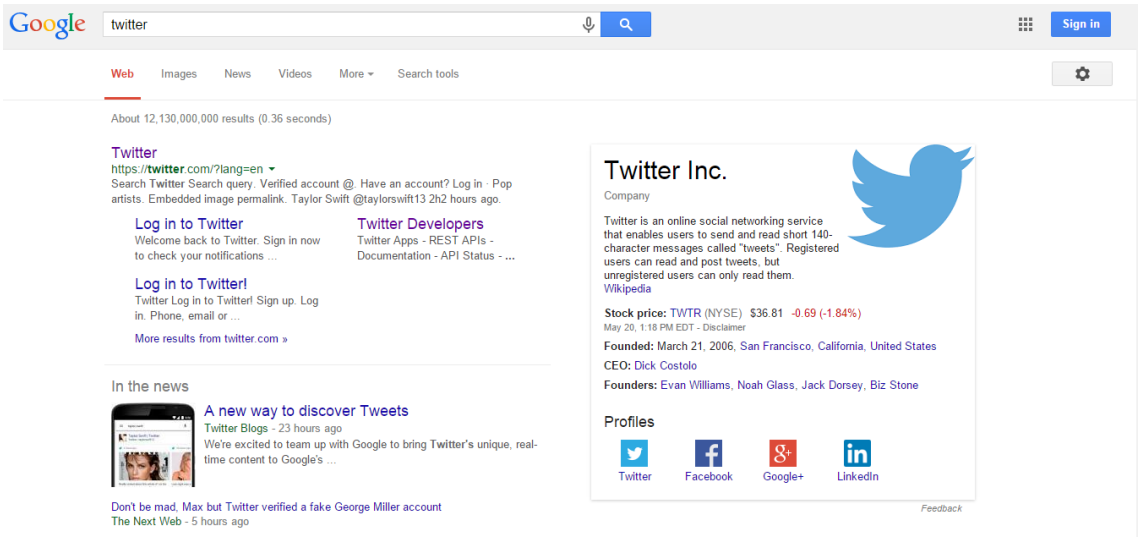
Με την ενεργοποίησή τους δίνεται εντολή στον φυλλομετρητή (browser) του χρήστη να μεταβεί από τη σελίδα ενός ιστότοπου στην αρχική σελίδα (homepage) ενός άλλου ιστότοπου.

3.4.1.2 "Βαθιοί" σύνδεσμοι (deep links)

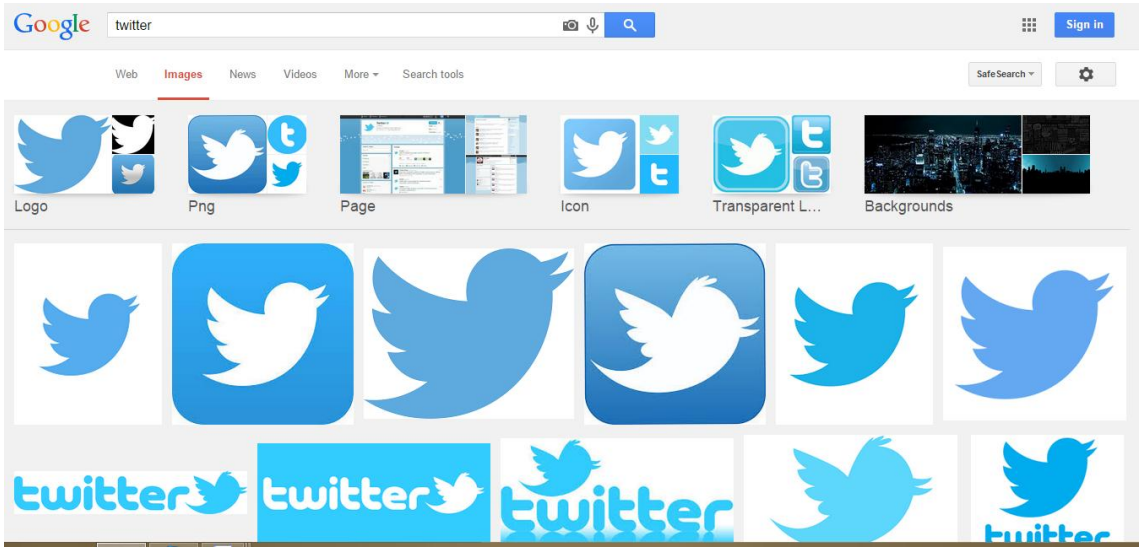
Με την ενεργοποίησή τους δίνεται εντολή στον φυλλομετρητή (browser) του χρήστη να μεταβεί από τη σελίδα ενός ιστότοπου αλλά όχι στην αρχική αλλά σε (βαθύτερη) υποσελίδα ενός άλλου. Για παράδειγμα: http://example.com/path/page

3.4.1.3 Σύνδεσμοι "πλαισίου" (framing links)

Με την ενεργοποίησή τους δίνεται εντολή στον φυλλομετρητή (browser) του χρήστη για την καταφόρτωση (download) της σελίδας ενός ιστότοπου, η οποία όμως "πλαισιώνεται" από την αρχική ιστοσελίδα
Για παράδειγμα, ένας χρήσης ο οποίος πλοηγείται στον παγκόσμιο ιστό και επισκέπτεται το **www.google.com** το οποίο έχει ένα συγκεκριμένο πλαίσιο, και θέλει να συνεχίσει την πλοήγησή του για την εμφάνιση εικόνων τότε το πλαίσιο παραμένει το ίδιο αλλά το περιεχόμενο της σελίδας αλλάζει όπως φαίνεται στην εικόνα 5 και 6.



Εικόνα 5: Η ιστοσελίδα της www.google.com για γενική αναζήτηση της λέξης Twitter



Εικόνα 6: Η ιστοσελίδα της www.google.com για αναζήτηση της λέξης Twitter για εμφάνιση εικόνων

3.4.1.4 “Αυτόματοι ή δυναμικοί υπερσύνδεσμοι” (automatic ή inline/embedded links)

Μπορεί να είναι οποιεσδήποτε από τις παραπάνω κατηγορίες υπερσυνδέσμων με την διαφορά ότι ενεργοποιούνται αυτόματα χωρίς την παρέμβαση του χρήστη. Χρησιμοποιούνται σε κώδικα HTML στον οποίο γράφονται οι αυτόματοι υπερσυνδέσμοι έτσι όπως φαίνεται στη εικόνα 7 και στον χρήστη εμφανίζεται όπως φαίνεται στην εικόνα 8

```
<p><a href="http://www.w3schools.com/html/">Visit our HTML tutorial</a></p>
```

Εικόνα 7 : Κώδικας HTML στον οποίο γράφονται οι αυτόματοι υπερσυνδέσμοι

Visit our HTML tutorial

Εικόνα 8 : Ο τρόπος με τον οποίο εμφανίζεται ο υπερσύνδεσμος στον χρήστη

3.4.1.5“Περιτυλιγμένοι υπερσυνδέσμοι”(wrap hyperlinks)

Μπορεί να είναι οποιεσδήποτε από τις παραπάνω κατηγορίες υπερσυνδέσμων -βκυρίως από τα μέσα κοινωνικής δικτύωσης- με την διαφορά ότι οι πρωτότυποι υπερσυνδέσμοι περικλείονται σε μικρότερους υπερσυνδέσμους. Αυτή η διαδικασία γίνεται με σκοπό οι χρήστες να προστατευθούν από κακόβουλο περιεχόμενο. Αν ο χρήστης επιλέξει έναν περιτυλιγμένο σύνδεσμο, θα οδηγηθεί στον υπερσύνδεσμο στον οποίο περίκλυσε. Όπως για παράδειγμα η υπηρεσία “bitly” παρέχει τον περιτυλιγμένο σύνδεσμο: http://bit.ly/x, όπου x η επέκταση του περιτυλιγμένου συνδέσμου .

3.5 Περιτύλιγμα συνδέσμου στο Twitter

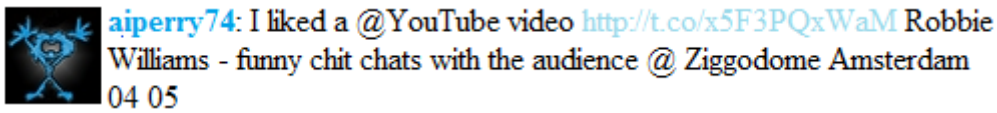
Όπως αναφέρθηκε και πιο πάνω όλα τα μέσα κοινωνικής δικτύωσης περικλείουν τους υπερσυνδέσμους σε μικρότερους. Όταν ένας χρήστης του Twitter γράψει ένα tweet το οποίο περιέχει υπερσύνδεσμο, τότε ο μικρότερος υπερσύνδεσμος ο οποίος παράγεται αυτόματα, εμφανίζεται με την συντομογραφία ’ t.co’.

Τρόπος με τον οποίο το Twitter περικλείει έναν υπερσύνδεσμο σε μικρότερο:

- Twitter περικλείει όλους τους υπερσυνδέσμους οι οποίοι δεν περιέχουν έγκυρα πρωτόκολλα όπως URL, HTTP, HTTPS. Δεν περικλείει τους υπερσυνδέσμους που περιέχουν τα έγκυρα πρωτόκολλα .

- Οι τελείες και τα κόμματα, εκτός από ένα κωδικοποιημένο URL, αγνοούνται και δεν θεωρούνται μέρος ενός URL το οποίο πρέπει να περιτυλιχθεί. Δύο διευθύνσεις URL που συνδέονται με μια χρονική περίοδο ή κόμμα δεν θα πρέπει να προβούν σε ανάλυση.

- Η διεύθυνση URL δεν πρέπει να περιλαμβάνει πληροφορίες σύνδεσης του χρήστη διότι δε θα μπορέσει να γίνει μέρος του περιτυλίγματος.



Εικόνα 9: Παράδειγμα tweet στο οποίο υπάρχει περιτυλιγμένος σύνδεσμος

3.6 Κλειδολέξεις

Μια κλειδολέξη ή μια λέξη κλειδί είναι η λέξη που χρησιμεύει ως κλειδί, ως προς την έννοια μιας άλλης λέξης ή μια φράσης. Επίσης ονομάζεται σύνθημα (catchword). Η κλειδολέξη ίναι μια σημαντική λέξη ή όρος σε ένα τίτλο, μια περίληψη, ή το κείμενο ενός εγγράφου ή άλλου στοιχείου που αναπροσαρμόζεται, και χρησιμοποιείται ως καταχώρηση ενός ευρετηρίου ή μιας μηχανής αναζήτησης. Μπορεί να χαρακτηριστεί ως μια λέξη που χρησιμοποιείται για να οργανωθεί το ψηφιακό περιεχόμενο του διαδικτύου, ή να διευκολύνει μια ηλεκτρονική έρευνα για πληροφορίες που ενδιαφέρουν τους χρήστες που την αναζητούν.

3.6.1 Κλειδολέξεις και Twitter

Τα τελευταία χρόνια όλο και περισσότεροι χρήστες των μέσων κοινωνικών δικτύων χρησιμοποιούν τις λέξεις κλειδιά. Μια στοχευόμενη λέξη κλειδί (keyword targeting) προσφέρει μια πολύ σημαντική δυνατότητα στους χρήστες να συνδέονται μεταξύ τους βασισμένοι σε λέξεις και φράσεις που έχουν ενσωματώσει στην ανάρτησή (tweet) τους πρόσφατα. Αυτό σημαίνει ότι ένας χρήστης θα μπορεί προσεγγίσει τους φίλους του και τα ενδιαφέροντά του πιο εύκολα.

Υπάρχουν 2 είδη στοχευόμενων κλειδολέξεων:

3.6.1.1 Αναζήτηση

Οι όροι, οι φράσεις και τα hashtags που αναζητούν οι χρήστες στο Twitter μπορούν να δείξουν το τι έχουν μέσα στο μυαλό τους την συγκεκριμένη στιγμή που πραγματοποιούν την αναζήτηση. Η στόχευση κλειδολέξεων (keyword targeting) επιτρέπει την στόχευση των tweets, έτσι ώστε να εμφανίζονται στα αποτελέσματα αναζήτησης όροοι και φράσεις σχετικές με την σκέψη και το ενδιαφέρον τους. Αυτός ο τύπος στόχευσης είναι ιδιαίτερα αποτελεσματικός κατά τη διάρκεια εκδηλώσεων που είναι ακριβώς ότι τους ενδιαφέρει, όπως κατά τη διάρκεια πολιτιστικών γεγονότων ή γεγονότων που δημιουργούν αιχμές συζήτησης στο Twitter.

3.6.1.1 Χρονοδιάγραμμα

Τα tweets όπως αναφέραμε στο κεφάλαιο 3, παρουσιάζουν σημαντικά μηνύματα σχετικά με το τι οι χρήστες αντιμετωπίζουν ή αισθάνονται μια συγκεκριμένη στιγμή: τις σκέψεις, τα συναισθήματα, τις επιθυμίες και τις ανάγκες τους. Η στόχευση των κλειδολέξεων (keyword targeting) στο χρονοδιάγραμμα στοχεύει τους χρήστες που βασίζονται ή ασχολούνται με αυτά που έχουν γράψει πρόσφατα και μπορούν οι υπόλοιποι χρήστες να έρθουν επαφή μαζί τους.

3.7 Τερματικές λέξεις (Stopwords)

Στην πληροφορική οι τερματικές λέξεις είναι οι λέξεις οι οποίες φιλτράρονται μετά από μερική επεξεργασία κειμένου φυσικής γλώσσας. Οι τερματικές λέξεις είναι λέξεις φυσικής γλώσσας μικρής σημασίας, παραδείγματος χάρη “και”, ”το” και παρόμοιες λέξεις. Εφόσον οι τερματικές λέξεις έχουν μικρή σημασία, οι μηχανές αναζήτησης αλλά και οι προγραμματιστές οι οποίοι επεξεργάζονται κείμενα από τα μέσα κοινωνικής δικτύωσης τις αγνοούν έτσι ώστε να εξάγουν σημαντικές λέξεις οι οποίες θα τους δώσουν χρήσιμες πληροφορίες για αυτούς.

Γενικά όμως, δεν υπάρχουν συγκεκριμένες λέξεις οι οποίες μπορεί να θεωρηθούν ως τερματικές και να μπορούν να συμπεριληφθούν σε μια λίστα. Οπότε ούτε ένα παρεμφερές εργαλείο μπορεί να χρησιμοποιήσει μια τέτοια λίστα και να τις αφαιρέσει από το κείμενο της φυσικής γλώσσας το οποίο πρόκειται να προβεί σε επεξεργασία.

Ένας προγραμματιστής που πρόκειται να επεξεργαστεί ένα κείμενο φυσικής γλώσσας , μπορεί να επιλέξει μια οποιαδήποτε ομάδα λέξεων και να τις θεωρήσει ως τερματικές ανάλογα με το ποιος είναι ο στόχος του.

Για την ολοκλήρωση της παρούσας διπλωματικής εργασίας, επιλέχθηκαν τα πιο πάνω χαρακτηριστικά με σκοπό να παρέχουν ικανοποιητικά αποτελέσματα. Κατά την διάρκεια υλοποίησης της διπλωματικής εργασίας, επεξεργάζονται και αναλύονται τα προαναφερθέντα χαρακτηριστικά. Αυτά τα χαρακτηριστικά φάνηκαν σωστή επιλογή γιατί βοήθησαν στην ανίχνευση πολιτιστικών γεγονότων.

Κεφάλαιο 4 Τεχνικές Συλλογής και Ανάλυσης Δεδομένων

4.2 Τι είναι η Διεπαφή Προγραμματισμού Εφαρμογών (API)
4.1.1 Περιγραφή Διεπαφής Προγραμματισμού Εφαρμογών
4.2 Twitter API
4.2.1 Τελικά σημεία (Endpoints)
4.2.2 Είδη API
4.2.2.1 REST APIs
4.2.2.2 Streaming APIs
4.2.2.2.1 Public streams
4.2.2.2.2 User streams
4.2.2.2.3 Site streams
4.2.2.3 Διαφορές μεταξύ REST API και Streaming API
4.2.2.4 Ads APIs
4.2.2.4.1 Πλατφόρμα Ads APIs
4.3 Βιβλιοθήκες API
4.3.1 Βιβλιοθήκες που χρησιμοποιήθηκαν
4.4 Μέθοδοι API
4.4.1 Μέθοδοι που χρησιμοποιήθηκαν

4.1 Τι είναι η Διεπαφή Προγραμματισμού Εφαρμογών (API)

Η Διεπαφή Προγραμματισμού Εφαρμογών μπορεί να ονομαστεί και ως Διασύνδεση Προγραμματισμού Εφαρμογών (σε συντομία διεπαφή ή διασύνδεση). Μπορεί να ονομαστεί ως η διεπαφή των προγραμματιστικών διαδικασιών που ένα λειτουργικό σύστημα ή βιβλιοθήκη ή εφαρμογή παρέχει προκειμένου να επιτρέψει να γίνονται αιτήσεις (queries) προς αυτό από άλλα προγράμματα αλλά επίσης χρησιμοποιείται και για ανταλλαγή δεδομένων.

4.1.1 Περιγραφή Διεπαφής Προγραμματισμού Εφαρμογών

Ένας από τους βασικούς σκοπούς μίας διεπαφής είναι να ορίζει και να διατυπώνει το σύνολο των λειτουργιών και υπηρεσιών που μπορεί να παρέχει μια βιβλιοθήκη ή ένα λειτουργικό σύστημα ή μια εφαρμογή σε άλλα προγράμματα, χωρίς να επιτρέπει πρόσβαση στον κώδικα που υλοποιεί αυτές τις υπηρεσίες. Η διεπαφή, μεταξύ καλούντος και καλούμενου, διαχωρίζει την προγραμματιστική υλοποίηση κάποιων υπηρεσιών από τη χρήση τους. Για παράδειγμα, το λειτουργικό σύστημα Windows έχει τη δική του διεπαφή (κλήσεις συστήματος), η οποία παρέχεται από την εταιρεία Microsoft. Η Microsoft περιγράφει τους τρόπους αξιοποίησης του συνόλου των υπηρεσιών που παρέχει το λειτουργικό το οποίο προσφέρει. Το τμήμα του λειτουργικού συστήματος το οποίο υλοποιεί τις υπηρεσίες που περιγράφονται στη διεπαφή, ονομάζεται “υλοποίηση της διεπαφής”.

Πρόκειται για ένα σύνολο από ρουτίνες (routines), δομές δεδομένων (data structures), κλάσεις αντικειμένων (object classes) και/ή πρωτόκολλα (protocols) που υποστηρίζουν την ανάπτυξη λογισμικού και υπηρεσιών βασισμένων σε μια πρωτογενή πηγή (υπολογιστικό σύστημα, βιβλιοθήκη, εφαρμογή).

4.2 Twitter API

Το Twitter API μπορεί να φαίνεται λίγο περίπλοκο, αλλά στην πραγματικότητα είναι αρκετά απλό. Υπάρχει ένας αριθμός από βιβλιοθήκες για τις διάφορες γλώσσες προγραμματισμού που κάνουν τη διαδικασία αρκετά εύκολη. Με την πάροδο του χρόνου το Twitter API documentation (τεκμηρίωση του API) έχει βελτιωθεί σημαντικά και βοηθά στην εύκολη ανάπτυξη μιας διαδικτυακής εφαρμογής η οποία έχει ως βάση της την επικοινωνία με το Twitter.

4.2.1 Τελικά σημεία (Endpoints)

Τελικά σημεία (endpoints) ονομάζονται οι διαδικτυακές υπηρεσίες οι οποίες ορίζονται από ένα WSDL αρχείο και η κύρια λειτουργία του είναι να ορίζει την διεύθυνση ή το σημείο σύνδεσης μιας διαδικτυακής υπηρεσίας. Τυπικά, αναπαριστάται από μια συμβολοσειρά HTTP URL. Στα REST APIs, τα τελικά σημεία παρέχουν διάφορες μεθόδους όπως οι GET, POST, DELETE και PUT. Η κάθε μια από αυτές τις μεθόδους, έχει παρόμοια λειτουργία όπως και το όνομά της, και χρησιμοποιούνται από τον χρήστη και τον προγραμματιστή μιας υπηρεσίας διαδικτύου. Δηλαδή ένας προγραμματιστής ο οποίος χρησιμοποιεί το Twitter REST API χρησιμοποιεί τις πιο πάνω μεθόδους για να αλληλεπιδράσει ανάλογα με το Twitter.

4.2.2 Είδη API

Για να καταστεί δυνατή η ανάπτυξη μια εφαρμογής από προγραμματιστές , η οποία έχει ως βάση της το Twitter, μπορούν να χρησιμοποιηθούν τρία είδη APIs τα οποία προσφέρει το Twitter. Τα REST APIs , τα Streaming APIs και τα Ads APIs. Κάθε ένα από αυτά τα APIs προσφέρει διαφορετικές δυνατότητες στον προγραμματιστή.

4.2.2.1 REST APIs

Οι προγραμματιστές που θέλουν να αναπτύξουν μια εφαρμογή κάνοντας χρήση των REST APIs μπορούν να γράψουν αλλά και να διαβάσουν με προγραμματιστικό τρόπο δεδομένα από το Twitter. Για παράδειγμα, ο προγραμματιστής μπορεί να αναρτήσει ένα καινούργιο tweet , να διαβάσει το προφίλ ενός χρήστη αλλά και των «ακόλουθων» του. Τα REST APIs ταυτοποιούν τις εφαρμογές που αναπτύσσονται μέσω του Twitter με το oAuth (ταυτοποίηση) το οποίο ανταποκρίνεται και απαντά πίσω στον προγραμματιστή εμφανίζοντας του δεδομένα σε μορφή Json.

Είναι σημαντικό να σημειώσουμε ότι στα REST APIs υπάρχουν rate limits. Δηλαδή, όταν ένας προγραμματιστής χρησιμοποιεί τα REST APIs για την ανάπτυξη μια εφαρμογής, το rate limit παραθύρου για μια εφαρμογή είναι δεκαπέντε αιτήσεις. Στις

τελευταίες εκδόσεις του Twitter API, τα rate limits χωρίζονται σε διαστήματα των δεκαπέντε λεπτών. Επίσης, τα τελικά σημεία (endpoints) χρειάζονται ταυτοποίηση οπότε δεν υπάρχει περίπτωση για μη-ταυτοποιημένες ή μη-εξουσιοδοτημένες αιτήσεις χρηστών και rate limits. Για το rate limit παίζουν ρόλο και οι κεφαλίδες HTTP. Οι συνδέσεις HTTP πρέπει να επαληθευτούν για να μπορέσουν να παρέχουν τα κατάλληλα δεδομένα τα οποία είναι σχετικά με τις αιτήσεις του προγραμματιστή. Επίσης, μπορούν να ελέγξουν το rate limit έτσι ώστε ένας προγραμματιστής να μην το υπερβεί. Αν ένα τελικό σημείο/ ένας προγραμματιστής υπερβεί τον επιτρεπόμενο αριθμό από αιτήσεις μέσα στον παράθυρο των δεκαπέντε λεπτών ή μέσα στο παράθυρο των δεκαπέντε αιτήσεων τις οποίες δικαιούται, τότε η επικεφαλίδα HTTP που θα του επιστραφεί είναι HTTP 429: "Too many requests".

GET and POST Request Limits: Όπως αναφέρθηκε και στην αρχή αυτού του κεφαλαίου, τα REST APIs χρησιμοποιούνται έτσι ώστε να γραφτούν και να διαβαστούν δεδομένα από το Twitter . Όμως το διάβασμα και το γράψιμο των δεδομένων δεν μπορεί να γίνει απεριόριστες φορές. Το Twitter αποφάσισε να βάλει ένα όριο στον αριθμό των φορών που ένας προγραμματιστής μπορεί να διαβάσει και να γράψει. Αυτός ο περιορισμός γίνεται σε επίπεδο χρήστη και όχι εφαρμογής.

4.2.2.2 Streaming APIs

Για να αναπτυχθεί μια εφαρμογή χρησιμοποιώντας τα Streaming APIs είναι απαραίτητο να υπάρχει συνεχής σύνδεση στο διαδίκτυο έτσι ώστε ένας προγραμματιστής με τη χρήση του HTTP πρωτοκόλλου να μπορεί να λαμβάνει τα δεδομένα που χρειάζεται σε πραγματικό χρόνο.

Με τη χρήση των Streaming APIs ένας προγραμματιστής είναι σε θέση να εντάξει στο πρόγραμμά του μια συγκεκριμένη λέξη. Με τη χρήση αυτής της λέξης θα μπορεί να λαμβάνει ενημερώσεις με τα τελευταία tweets που την αναφέρουν. Επίσης, ενημερώνονται και ανανεώνονται προφίλ των χρηστών και γίνεται αναζήτηση μέσα από αυτά.

Τα Streaming APIs δίνουν στον προγραμματιστή low latency πρόσβαση στα δεδομένα του Twitter που μπορεί οποιοσδήποτε να κοιτάξει απλά με μια περιήγηση στο

δίκτυο. Αυτό γίνεται είτε είναι χρήστης και έχει λογαριασμό στο Twitter είτε δεν έχει. Αυτά τα δεδομένα λέγονται «δημόσια δεδομένα» (public data) .

Τα Streaming APIs παρέχουν τα ακόλουθα «endpoints» τα οποία χρησιμοποιούνται για διαφορετικούς σκοπούς το κάθε ένα. Για παράδειγμα, αν η απάντηση του Twitter στον προγραμματιστή θα είναι σε μορφή json, τότε το endpoint θα είναι : " endpoint1 = https://api.twitter.com/1.1/search/tweets.json"

4.2.2.2.1 Public streams

Ένας χρήστης ο οποίος προγραμματίζει χρησιμοποιώντας τα public streams μπορεί να συλλέξει δεδομένα του Twitter τα οποία είναι δημόσια. Με αυτόν τον τρόπο μπορεί να αναζητήσει για ένα συγκεκριμένο θέμα, συγκεκριμένο χρήστη και συγκεκριμένα δεδομένα. Μόλις γίνει η χρήση του public streams και η δημιουργία σύνδεσης στο Twitter με τη χρήση του HTTP πρωτόκολλου , ο προγραμματιστής θα λάβει πληθώρα tweets για το θέμα το οποίο αναζήτησε (query) χωρίς να ανησυχήσει για τον περιορισμένο αριθμό αιτήσεων με τον οποίο έγινε σύνδεση με το Twitter.

Τα τελικά σημεία των Public streams είναι οι μέθοδοι POST statuses / filter, GET statuses / sample και GET statuses . Για να χρησιμοποιηθούν τα Public streams πρέπει ο προγραμματιστής να συνδεθεί μέσω των προαναφερθέντων τελικών σημείων χρησιμοποιώντας καλές πρακτικές και παραμέτρους έτσι ώστε να του επιστραφούν όσο πιο διευκρινιστικές , κατανοητές και επιθυμητές απαντήσεις από το Twitter public stream.

Ο κάθε προγραμματιστής ο οποίος θέλει να χρησιμοποιήσει αυτά τα streams πρέπει να δημιουργήσει έναν λογαριασμό από τον οποίο μπορεί να πραγματοποιήσει μόνο μια σύνδεση στα τελικά σημεία των public streams. Εάν επιθυμεί να συνδεθεί σε ένα Public stream περισσότερο από μια φορά πρέπει να τερματίσει την προηγούμενή του σύνδεση και να πραγματοποιήσει μια καινούργια.

4.2.2.2.2 User streams

Με τη χρήση των User streams ένας προγραμματιστής μπορεί να λάβει πληροφορίες και δεδομένα που αφορούν ένα συγκεκριμένο χρήστη. Δηλαδή τα User streams προσφέρουν δεδομένα και γεγονότα σε πραγματικό χρόνο για έναν χρήστη ο οποίος έχει πιστοποιημένο λογαριασμό στο Twitter.

Ένα πρόβλημα το οποίο θα αντιμετωπίσει ένας προγραμματιστής όταν χρησιμοποιήσει τα User streams είναι ο περιορισμός των αριθμών των συνδέσεων που μπορεί να πραγματοποιήσει και να αιτηθεί από το OAuth με τον λογαριασμό του συγκεκριμένου χρήστη από τον οποίο επιθυμεί να συλλέξει δεδομένα. Αφού ο προγραμματιστής υπερβεί το όριο του αριθμού των συνδέσεων με την χρήση του τελικού σημείου GET (δεν προσφέρονται οι υπόλοιπες μέθοδοι στα user streams) . Η σύνδεση τερματίζεται αυτόματα και επιστρέφεται το μήνυμα λάθους «HTTP 420» το οποίο υποδεικνύει ότι ο συγκεκριμένος λογαριασμός από τον οποίο ο προγραμματιστής συλλέγει δεδομένα ανταποκρίθηκε σε πολλές συνδέσεις. Όταν γίνει αυτό, το Twitter αυτόματα «κλειδώνει» τον λογαριασμό του χρήστη και δεν επιτρέπει περισσότερες προσβάσεις σε αυτόν .

Με την χρήση των User streams , τα tweets μπορούν να φιλτραριστούν έτσι ώστε να εμφανίζονται πιο κατανοητά στον χρήστη. Τα User streams διατηρούν ένα σύνολο δεδομένων από το χρονοδιάγραμμα του χρήστη για να μπορέσουν να το καθορίσουν. Στο χρονοδιάγραμμα του χρήστη εφαρμόζονται οι λεγόμενες «κοινωνικές εκδηλώσεις» («social events»), οι οποίες περιέχουν σημαντικά γεγονότα του χρήστη συγκεκριμένες ημερομηνίες. Αυτό γίνεται έτσι ώστε τα δεδομένα του χρονοδιαγράμματος ενός χρήστη να είναι συγχρονισμένα με τον διακομιστή (server) του Twitter. Στο σύνολο δεδομένων είναι αναμενόμενο ότι εκτός από έναν μεγάλο αριθμό tweets , θα περιέχονται τα retweets και οι επισημάνσεις των tweets. Σημαντικό είναι να τονισθεί δεν μπορεί να πραγματοποιηθεί φιλτράρισμα σε μπλοκαρισμένους λογαριασμούς του Twitter.

4.2.2.2.3 Site streams

Τα site streams είναι η έκδοση των User streams για πολλούς χρήστες και προορίζονται για διακομιστές "servers" οι οποίοι συνδέονται στο Twitter στο χρονοδιάγραμμα πολλών χρηστών. Το μόνο τελικό σημείο το οποίο προσφέρεται από τα site streams είναι το GET site.

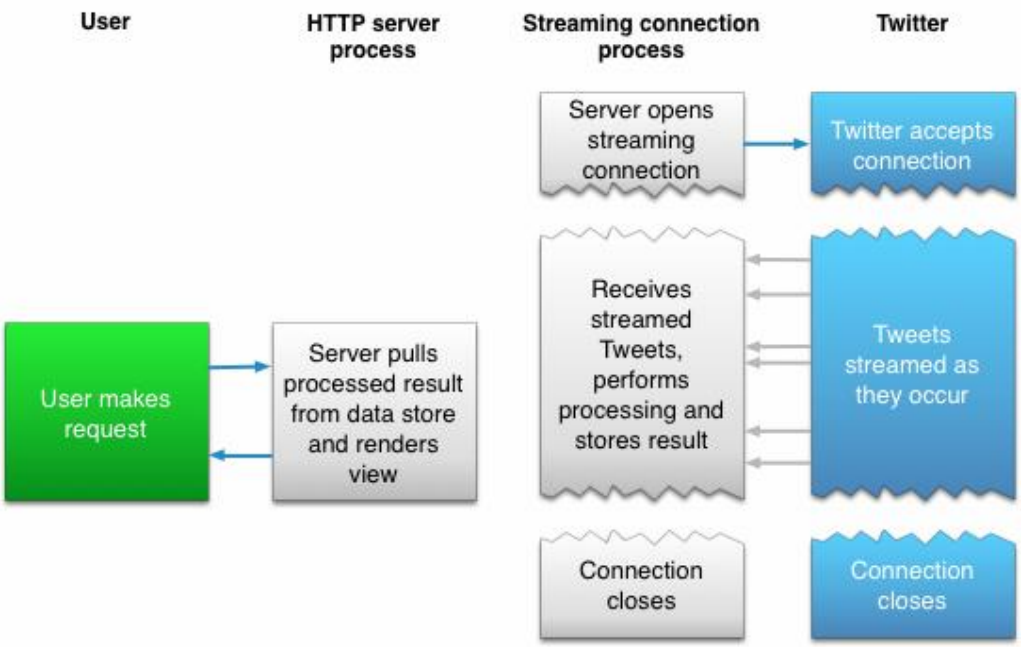
Τα δεδομένα τα οποία επιστρέφονται από τα site streams είναι όπως αυτά των user streams με την μόνη διαφορά ότι τα δεδομένα αυτά περικλείονται με ένα επιπλέον περιτύλιγμα αναφέροντας τον χρήστη στον οποίο απευθύνεται το μήνυμα. Τα site streams μπορούν να επιστρέψουν ιδιωτικά και δημόσια δεδομένα με μόνο ένα stream. Αυτά τα streams αποδέχονται μόνο oAuth αιτήσεις (για εξουσιοδότηση) από τον ιδιοκτήτη της εφαρμογής που πρόκειται να δημιουργηθεί ή να συντηρηθεί. Όταν ο προγραμματιστής καταφέρει να συνδεθεί με το site streams πρέπει να καταγράψει τα ονόματα των χρηστών του Twitter οι οποίοι δέχτηκαν να λάβει ο προγραμματιστής δεδομένα από αυτούς. Αξίζει να σημειωθεί ότι υπάρχουν όρια τόσο στον αριθμό των χρηστών από τους οποίους ένας προγραμματιστής μπορεί να μαζέψει δεδομένα, όσο και στον αριθμό των συνδέσεων ανά δευτερόλεπτο.

4.2.2.2.4 Διαφορές μεταξύ REST API και Streaming API

Όπως αναφέρθηκε και πιο πάνω η σύνδεση με το Streaming API απαιτεί διατήρηση της HTTP σύνδεσης συνέχεια. Σε πολλές περιπτώσεις η ανάπτυξη της εφαρμογής η οποία αναπτύσσεται με το REST API είναι διαφορετική. Για παράδειγμα, εάν υπάρχει μια διαδικτυακή εφαρμογή η οποία δέχεται αιτήσεις από διάφορους χρήστες οι οποίοι χρησιμοποιούν το REST API, το Twitter API ανταποκρίνεται και τους απαντά πίσω.

Μια εφαρμογή η οποία αναπτύσσεται με το Streaming API δε θα είναι σε θέση να εγκαθιδρύσει σύνδεση όταν το Twitter απαντήσει σε ένα αίτημα κάποιου χρήστη. Ο κώδικας όμως για τη διατήρηση της σύνδεσης Streaming τρέχει σε μια διαδικασία ξεχωριστή από αυτή που γίνεται με τα HTTP αιτήματα. Αυτό φαίνεται στο πιο κάτω σχήμα:

Εικόνα 10: Διατήρηση της σύνδεσης HTTP ξεχωριστά απ'ότι της σύνδεσης Streaming



Η διαδικασία με το Streaming API παίρνει είσοδο από το Twitter και αναλύει , φιλτράρει και ομαδοποιεί τα δεδομένα σε μια βάση δεδομένων. Η διαδικασία HTTP "ρωτά" αυτήν την βάση έτσι ώστε να πιάσει το αποτέλεσμα του Streaming.

4.3 Βιβλιοθήκες του Twitter API

Για την χρήση των τριών ειδών APIs τα οποία προσφέρει το Twitter υπάρχει ένας αριθμός από βιβλιοθήκες σε πολλά είδη προγραμματιστικών γλωσσών. Αυτές οι βιβλιοθήκες χρησιμοποιούνται από προγραμματιστές έτσι ώστε να χρησιμοποιήσουν τα προνόμια που τους προσφέρει το Twitter API , να αποκτήσουν πρόσβαση σε αυτά και να αναπτύξουν χρήσιμες εφαρμογές.

Ο προγραμματιστής έχει την δυνατότητα πρώτα από όλα να διαλέξει με ποια ή με ποιες γλώσσες θέλει να ασχοληθεί για να ανάπτυξη μια εφαρμογή χρησιμοποιώντας το Twitter API. Θα διαλέξει την/τις γλώσσα/ες με την οποία/ες είναι εξοικειωμένος περισσότερο και έτσι η ανάπτυξη της εφαρμογής του θα είναι πιο εύκολη και πιο

γρήγορη. Μετά την επιλογή της γλώσσας προγραμματισμού την οποία θα χρησιμοποιήσει για να αναπτύξει την εφαρμογή του, θα πρέπει να επιλέξει μέσα από έναν μεγάλο αριθμό βιβλιοθηκών τις οποίες προσφέρει η κάθε γλώσσα, ποια βιβλιοθήκη προτιμά να χρησιμοποιήσει. Κάθε βιβλιοθήκη ξεχωριστά προσφέρει διαφορετικές μεθόδους με διαφορετικές δυνατότητες στον προγραμματιστή σε σύγκριση με κάποια άλλη.

Για τη παρούσα διπλωματική εργασία έπρεπε να γίνει συλλογή δεδομένων (tweets) από το Twitter έτσι ώστε να επεξεργαστούν και να επιτευχθεί ο στόχος της. Για τη συλλογή δεδομένων επιλέχθηκε η προγραμματιστική γλώσσα Python. Το Twitter API προσφέρει πολλές βιβλιοθήκες οι οποίες περιέχουν μεθόδους και μπορούν να χρησιμοποιηθούν για άντληση και επεξεργασία των δεδομένων του Twitter.

Χρησιμοποιώντας τα δύο είδη APIs του Twitter (Rest και Streaming) , παρατηρείται ότι προσφέρουν βιβλιοθήκες οι οποίες με τη σειρά τους αποτελούνται από μεθόδους οι οποίες βοηθούν στην συλλογή δεδομένων. Οι βιβλιοθήκες που χρησιμοποιήθηκαν είναι οι εξής:

Βιβλιοθήκη Oauth2

Η βιβλιοθήκη Oauth2 χειρίζεται τα αιτήματα των προγραμματιστών όταν επιθυμούν να εγγραφούν στο Twitter ή να συνδεθούν ή να αλληλεπιδράσουν με αυτό. Αυτό σημαίνει ότι ο προγραμματιστής μιας εφαρμογής του Twitter, πρέπει να καταχωρήσει στον κώδικά του τα τέσσερα διακριτικά (tokens) του, τα οποία είναι ξεχωριστές συμβολοσειρές για τον κάθε χρήστη που χρησιμοποιεί το Twitter API. Δημιουργούνται όταν ο προγραμματιστής δημιουργήσει μια εφαρμογή στο Twitter developers. Τα διακριτικά τα οποία δημιουργούνται είναι τα εξής και φαίνονται στη εικόνα 11:
Consumer Key (API Key)(κλειδί χρήστη), Consumer Secret (API Secret) (μυστικό χρήστη), Access Token (διακριτικό πρόσβασης), Access Token Secret (μυστικό διακριτικό πρόσβασης).

Consumer Key (API Key)	3I306T4BdMnBFWOZBggFgz4ya
Consumer Secret (API Secret)	eErhCeeeCzoIN7QF8iGHnEKxFrZZEfixU6jVYOomwrX8N87d4U
Access Token	823701890-5n9AZXRfT6fryPvLaNmhwHtF74qBTbVRirgQ7nsM
Access Token Secret	5dpFve1CYnA2Kf7E1SxmDTT4weJw523m1B06Fvyhno13F

Εικόνα 11, παραδείγματα των τεσσάρων διακριτικών όταν ένας χρήστης δημιουργεί μια εφαρμογή στο Twitter developers

Σε επίπεδο κώδικα, με τη βιβλιοθήκη Oauth2 μπορούμε να δημιουργήσουμε ένα αντικείμενο consumer και σε αυτό το αντικείμενο να αποθηκεύσουμε τα στα χαρακτηριστικά αυτού του αντικειμένου εισάγοντας τα τέσσερα διακριτικά μέσα στον κώδικα . Έτσι επιτυγχάνεται η σύνδεση του προγραμματιστή με το Twitter API.

```
consumer = oauth2.Consumer(key="INSERT API KEY", secret="INSERT API SECRET")
token = oauth2.Token(key="INSERT ACCESS TOKEN", secret="INSERT ACCESS TOKEN
   SECRET")

params["oauth_consumer_key"] = consumer.key
params["oauth_token"] = token.key
```

Εικόνα 12, παράδειγμα χρήσης των χαρακτηριστικών του αντικειμένου consumer όταν τους ανατεθεί η τιμή των διακριτικών του χρήστη.

Επίσης, η βιβλιοθήκη Oauth2 χρησιμοποιείται έτσι ώστε να δηλωθεί το τελικό σημείο (endpoint) το οποίο αναφέρεται σε πιο πάνω υποκεφάλαιο.

```
# Set the API endpoint

url1 = "https://api.twitter.com/1.1/search/tweets.json"
```

Εικόνα 13, παράδειγμα εισαγωγής endpoint στον κώδικα του προγραμματιστή για να επιστραφούν δεδομένα σε μορφή json

<u>Βιβλιοθήκη urllib2</u>

Η βιβλιοθήκη urllib2 χρησιμοποιείται για τον ορισμό των λειτουργιών και των κλάσεων τα οποία βοηθούν στο άνοιγμα των διευθύνσεων URL , για την αυθεντικοποίηση και ανακατευθύνσεις (redirections) κλπ.

Βιβλιοθήκη time

Η βιβλιοθήκη time παρέχει λειτουργίες οι οποίες έχουν σχέση με την ώρα, τον χειρισμό, την ανάλυση, την μορφοποίηση και την μετατροπή της ώρας

Βιβλιοθήκη json

Όπως φαίνεται και στη εικόνα 14, η βιβλιοθήκη json μπορεί να αναλύσει δεδομένα αρχείων τα οποία είναι σε μορφή json και να τα μετατρέψει σε μια λίστα ή ένα λεξικό (dictionary) της Python, Επίσης, μπορεί να μετατρέψει λεξικό της Python ή μια λίστα σε json μορφή.


```
  "in_reply_to_user_id_str": null,
  "contributors": null,
  "text": "Aggressive Ponytail #freebandnames",
  "metadata": {
   "iso_language_code": "en",
   "result_type": "recent"
  },
  "retweet_count": 0,
  "in_reply_to_status_id_str": null,
  "id": 250075927172759552,
  "geo": null,
  "retweeted": false,
  "in_reply_to_user_id": null,
  "place": null,
  "user": {
   "profile_sidebar_fill_color": "DDEEF6",
   "profile_sidebar_border_color": "C0DEED",
   "profile_background_tile": false,
   "name": "Sean Cummings",
   "profile_image_url":
"http://a0.twimg.com/profile_images/2359746665/1v6zfgqo8g0d3mk7ii5s
_normal.jpeg",
   "created_at": "Mon Apr 26 06:01:55 +0000 2010",
   "location": "LA, CA",
   "follow_request_sent": null,
   "profile_link_color": "0084B4",
   "is_translator": false,
   "id_str": "137238150",
   "entities": {
```


Εικόνα 14, παράδειγμα απάντησης του Twitter σε αίτημα του προγραμματιστή σε μορφή json. Οι απαντήσεις είναι σε σχέση με το προφίλ του χρήστη, εάν το tweet που επιστράφηκε έχει γίνει retweet, πότε δημιουργήθηκε κλπ.

4.4 Μέθοδοι του Twitter API

Με βιβλιοθήκες που αναφέρθηκαν στο προηγούμενο κεφάλαιο και οι οποίες παρέχονται από το Twitter API, είναι δυνατό να χρησιμοποιηθούν μέθοδοι οι οποίες βοηθούν στην αλληλεπίδραση του χρήστη με το Twitter API και στην συλλογή δεδομένων από τις απαντήσεις του .

Οι μέθοδοι οι οποίες χρησιμοποιήθηκαν για τη συλλογή δεδομένων από το Twitter είναι οι εξής :

- Oauth2.generate_nonce():

  Αυτή η μέθοδος ανήκει στην βιβλιοθήκη Oauth2 και χρησιμοποιείται για να δημιουργήσει τη συμβολοσειρά ‘’nonce’’. Αυτή η συμβολοσειρά, είναι μια συμβολοσειρά με τυχαία γράμματα και χρησιμοποιείται από τον διακομιστή (server) για να επαληθεύσει ότι η αίτηση του προγραμματιστή δεν έγινε ξανά στο παρελθόν. Με αυτόν τον τρόπο βοηθά στην αποτροπή επιθέσεων από τρίτους, όταν ο προγραμματιστής συνδέεται στον διακομιστή του Twitter από ένα μη-ασφαλές κανάλι. Η συμβολοσειρά ‘’nonce’’ πρέπει να είναι η μοναδική σε όλες τις αιτήσεις με την ίδια χρονική σήμανση και τα διακριτικά του προγραμματιστή .

- Oauth2.Request(method="GET",url=url,parameters=params):

  Αυτή η μέθοδος ανήκει στην βιβλιοθήκη Oauth2 και χρησιμοποιείται για να αιτηθεί δεδομένα από το Twitter. Χρησιμοποιεί τη μέθοδο ‘GET’ από το Streaming API , και του επιστρέφονται τα tweets σε μορφή json. Αυτά τα tweets, περιέχουν τη λέξη την οποία αποθήκευσε στην παράμετρο params με την οποία αιτείται από το Twitter (query).

- Oauth2.SignatureMethod_HMAC_SHA1():

  Αυτή η μέθοδος ανήκει στην βιβλιοθήκη Oauth2 και χρησιμοποιείται έτσι ώστε να δημιουργηθεί μια «υπογραφή» του προγραμματιστή. Επίσης, χρησιμοποιείται για να συνδεθεί ο προγραμματιστής στο Twitter API.

- Oauth2.sign_request(signature_method, consumer, token):

  Αυτή η μέθοδος ανήκει στην βιβλιοθήκη Oauth2 και χρησιμοποιείται από τον προγραμματιστή έτσι ώστε να αιτηθεί σύνδεση με το Twitter API. Οι παράμετροι του είναι το αποτέλεσμα της μεθόδου (signature_method, consumer, token), το αντικείμενο concumer και τα μοναδικά διακριτικά του προγραμματιστή.

- req.to_header():

  Αυτή η μέθοδος ανήκει στη βιβλιοθήκη Oauth2 και ορίζει κλάσεις οι οποίες υλοποιούνται με πρωτόκολλα HTTP και HTTPs . Μπορεί να ανοίξει σύνδεση με τον διακομιστή (server) του Twitter.

- req.to_url():

  Αυτή η μέθοδος ανήκει στη βιβλιοθήκη Oauth2 και ορίζει κλάσεις οι οποίες επιτρέπουν την αίτηση σύνδεσης με το URL το οποίο δηλώθηκε ως endpoint.

- urllib2.Request(url):

  Αυτή η μέθοδος ανήκει στη βιβλιοθήκη Oauth2 και ορίζει κλάσεις οι οποίες επιτρέπουν την σύνδεση με το URL το οποίο δηλώθηκε ως τελικό σημείο (endpoint).

- json.load(urllib2.urlopen(response)):

  Αυτή η μέθοδος ανήκει στη βιβλιοθήκη json και επιτρέπει την φόρτωση των δεδομένων που επιστρέφονται από το Twitter API. Παίρνει ως παράμετρο το αποτέλεσμα της urllib2.Request(url) μεθόδου.

- json.dump(data):

  Αυτή η μέθοδος ανήκει στη βιβλιοθήκη json και είναι υπεύθυνη στο γράψιμο των δεδομένων σε αρχείο σε μορφή json (παράδειγμα εικόνας 14).

- int(time.time()):

  Αυτή η μέθοδος ανήκει στην βιβλιοθήκη time και επιστρέφει την ώρα ως αριθμό κινητής υποδιαστολής εκφράζεται σε δευτερόλεπτα από την εποχή, σε UTC. Στα πλαίσια της παρούσας διπλωματικής εργασίας, αυτή η μέθοδος χρησιμοποιήθηκε έτσι ώστε, όταν επιστρέφεται η τρέχουσα ώρα, το API του Twitter να επιστρέφει δεδομένα τα οποία βρίσκονται στα πλαίσια του χρονικού πλαισίου του Rest API και του Streaming API.

- time.sleep(5):

  Αυτή η μέθοδος ανήκει στη βιβλιοθήκη time και χρησιμοποιείται ως «παύση» μεταξύ των αιτήσεων. Σε αυτή την περίπτωση κάθε 5 δευτερόλεπτα γινόταν αίτηση προς το Twitter API.

Κεφάλαιο 5 Στρατηγική Ερευνητικής Μεθοδολογίας

5.1 Εισαγωγή
5.2 Έρευνα και μέθοδοι συλλογής δεδομένων:
5.2.1. Συλλογή δεδομένων με τη χρήση διεπαφών του Twitter
5.2.2 Συλλογή δεδομένων με web crawling

5.1 Εισαγωγή

Για να πραγματοποιηθεί μια έρευνα ή μελέτη είναι απαραίτητο να ακολουθείται μια στρατηγική μεθοδολογία . Αυτή η έρευνα/μελέτη ολοκληρώνεται με διάφορες μεθοδολογικές προσεγγίσεις ανάλογα με τον τύπο των δεδομένων των μέσων κοινωνικών δικτύων που πρόκειται να επεξεργαστούν και να αναλυθούν.

5.2 Έρευνα και μέθοδοι συλλογής δεδομένων:

Η μελέτη της σχετικής βιβλιογραφίας θεωρείται μέρος της έρευνας και γενικά της μεθοδολογίας. Μέσα από αυτήν, παρέχονται κατευθυντήριες γραμμές για την διεξαγωγή παρόμοιων αναλύσεων και μελετών για τον μεγάλο όγκο δεδομένων που αντλείται από τα μέσα κοινωνικής δικτύωσης. Στη παρούσα διπλωματική εργασία εφαρμόστηκαν δύο μέθοδοι για συλλογή δεδομένων.

5.2.1. Συλλογή δεδομένων με τη χρήση διεπαφών του Twitter

Στην παρούσα έρευνα ήταν δυνατό να εξασφαλιστούν ακατέργαστα (raw) δεδομένα από το κοινωνικό δίκτυο Twitter. Το Twitter προσφέρει διεπαφές οι οποίες θα εξηγηθούν σε μετέπειτα κεφάλαιο (APIs). Αυτές οι διεπαφές χρησιμοποιούνται για τη συλλογή ακατέργαστων δεδομένων από το Twitter. Στην παρούσα διπλωματική εργασία συλλέχθηκαν δεδομένα χρησιμοποιώντας αυτές τις διεπαφές. Συγκεκριμένα συλλέχθηκαν δεδομένα από τους χρήστες οι οποίοι κατοικούν ή βρίσκονται στην Ολλανδία. Η απόφαση για τη συλλογή δεδομένων από την Ολλανδία πάρθηκε διότι σε αυτή τη χώρα διοργανώνονται πολλά πολιτιστικά γεγονότα με συγκεκριμένη

ημερομηνία και ώρα. Αυτό βοήθησε πολύ στην συλλογή μεγάλου και ποιοτικού όγκου δεδομένων τα οποία επεξεργάστηκαν και οδήγησαν στην ανίχνευση πολιτιστικών γεγονότων στην Ολλανδία.
Όπως αναφέρθηκε στο κεφάλαιο 4, η συλλογή δεδομένων έγινε με τη χρήση του REST και του Streaming API. Επιλέχθηκε η προγραμματιστική γλώσσα Python και βιβλιοθήκες και μέθοδοι που προσφέρει. Το σύνολο δεδομένων αποτελείτο από εννιά εκατομμύρια tweets για κάθε μήνα Μάιο μέχρι Ιούλιο του 2014 ξεχωριστά, τα οποία έτυχαν επεξεργασίας.

5.2.2 Συλλογή δεδομένων με ιχνηλάτιση ιστοσελίδων(web crawling)

Ένας web crawler είναι ένα πρόγραμμα το οποίο γράφεται από ένα προγραμματιστή ο οποίος θέλει να περιηγηθεί στο διαδίκτυο και να πραγματοποιήσει αναζήτηση με αυτόματο τρόπο. Το web crawling μπορεί να ασχοληθεί με υπερσυνδέσμους και με τον κώδικα της ιστοσελίδας ο οποίος είναι γραμμένος σε γλώσσα HTML. Ένας προγραμματιστής ο οποίος ασχολείται και προγραμματίζει ένα web crawler έχει σκοπό να αντλήσει , να «παρακολουθήσει» ιστοσελίδες διαδικτύου και να αντλήσει δεδομένα από αυτές. Για αυτό τον σκοπό , στην παρούσα διπλωματική εργασία, δημιουργήθηκε ένας web crawler στην προγραμματιστική γλώσσα JAVA (με τη χρήση της βιβλιοθήκης Jsoup) ο οποίος «παρακολούθησε» επίσημες ιστοσελίδες της Ολλανδίας. Στις ιστοσελίδες αυτές, αναφέρονταν πολιτισμικά γεγονότα.. Αντλήθηκαν πληροφορίες από αυτές για λίγο περισσότερο από εξακόσια πολιτιστικά γεγονότα για τους μήνες Μάιο μέχρι Αύγουστο του 2014. Οι επίσημες ιστοσελίδες της Ολλανδίας οι οποίες χρησιμοποιήθηκαν είναι :

1. http://www.iamsterdam.com
2. http://www.amsterdam.info
3. http://www.holland.com

Αυτά τα δεδομένα χρησιμοποιήθηκαν ως Ground Truth για τα αποτελέσματα που εξάχθηκαν από την ολοκλήρωση της διπλωματικής εργασίας. Με τον όρο Ground Truth εννοούμε ένα σύνολο δεδομένων το οποίο είναι πιο ακριβής και πιο αξιόπιστο από την έρευνα που διεξάγεται και μπορεί να συγκριθεί με τα αποτελέσματα που

διεξάχθηκαν από την έρευνα και να εξετάσει κατά πόσο τα αποτελέσματα αυτά είναι αξιόπιστα και ακριβείς.

Κεφάλαιο 6 Μεθοδολογία

6.1 Εισαγωγή

Στην παρούσα ατομική διπλωματική εργασία σκοπός είναι η δημιουργία ενός συστήματος το οποίο θα ανιχνεύει πολιτιστικές εκδηλώσεις οι οποίες πραγματοποιούνται σε προγραμματισμένη μέρα και ώρα σε κοντινή γεωγραφική περιοχή στην οποία ενδιαφέρεται π.χ μουσικής, θεάτρου, κινηματογράφου, αθλητισμού κλπ. Πιο συγκεκριμένα ο χρήστης ο οποίος ενδιαφέρεται για πολιτιστικά γεγονότα, θα έχει τη δυνατότητα μπαίνοντας στην ιστοσελίδα να του παρουσιάζεται ένα ημερολόγιο. Στο ημερολόγιο αυτό, θα εμφανίζονται όλες οι πολιτιστικές εκδηλώσεις με το τίτλο τους οι οποίες είναι προγραμματισμένες για τη χρονική περίοδο και περιοχή που επέλεξε ο χρήστης. Από εκεί ο χρήστης θα μπορεί επιλέγοντας κάποια από τις εκδηλώσεις να μεταβαίνει αυτόματα στη σελίδα της κάθε εκδήλωσης . Από αυτή την ιστοσελίδα θα μπορεί να ενημερωθεί καλύτερα και σε μεγαλύτερο βαθμό για τη συγκεκριμένη εκδήλωση την οποία επέλεξε. Επίσης, θα είναι σε θέση να παρατηρήσει πόσο έγκυρο είναι το πολιτιστικό γεγονός το οποίο του παρουσιάζεται.

Σε σχετική βιβλιογραφία η οποία μελετήθηκε και επεξηγήθηκε σε προηγούμενο κεφάλαιο, γίνεται προσπάθεια λύσης του πιο πάνω προβλήματος. Στη σχετική βιβλιογραφία, χρησιμοποιούνται διάφοροι αλγόριθμοι όπως ο EDCoW αλγόριθμος με τη χρήση της wavelet ανάλυσης και Naïve Bayes αλγόριθμοι με τη χρήση πιθανοτήτων και TAN και CAN μοντέλων ταξινόμησης. Επιπλέον χρησιμοποιούνται διάφορες προσεγγίσεις όπως η Emerging Topic Detection προσέγγιση η οποία προσπαθεί να ανιχνεύσει πολιτιστικά γεγονότα με βάση τη σημαντικότητα του tweet. Η σημαντικότητα ενός tweet βασίζεται στην ανάλυσή του tweet και στα URLs που υπάρχουν στο κείμενο του. Σε αυτή τη προσέγγιση χρησιμοποιείται ο Page Rank αλγόριθμος έτσι ώστε να κατανοηθεί κατά πόσο η πηγή η οποία αναφέρεται στο tweet είναι έγκυρη

Στο παρόν κεφάλαιο θα παρουσιαστεί η μεθοδολογία η οποία ακολουθήθηκε για την επίτευξη του στόχου της παρούσας διπλωματικής εργασίας. Συγκεκριμένα, η μεθοδολογία αυτή κατάφερε να εξάγει πληροφορίες για πολιτιστικά γεγονότα της Ολλανδίας, μέσα από τον τεράστιο όγκο πληροφοριών ο οποίος είναι αποθηκευμένος στο Twitter. Όπως αναφέραμε και σε προηγούμενο κεφάλαιο η επιλογή της Ολλανδίας για η συλλογή δεδομένων έγινε επειδή σε αυτή πραγματοποιούνται πολλά πολιτιστικά γεγονότα κάθε μέρα. Στη μεθοδολογία αυτή γίνεται επεξεργασία των tweets των κατοίκων της Ολλανδίας έτσι ώστε να μειωθεί ο θόρυβος των δεδομένων και να εξαχθούν επιθυμητά αποτελέσματα. Στη συνέχεια, για να βρεθεί η ημερομηνία του πολιτιστικού γεγονότος, γίνεται ομαδοποίηση για κάθε ημερομηνία ξεχωριστά. Επόμενο βήμα είναι η επεξεργασία και το φιλτράρισμα της ομαδοποίησης για να εξαχθούν τα πιο δημοφιλή hashtags, λέξεις κλειδιά και URLs για κάθε πολιτιστικό γεγονός . Στη συνέχεια, γίνεται χρήση των δεδομένων που συλλέχθηκαν από επίσημες ιστοσελίδες της Ολλανδίας έτσι ώστε να αξιολογηθεί η μεθοδολογία

6.2 Ανάλυση μεθοδολογίας

Για την επίτευξη του στόχου της παρούσας διπλωματικής εργασίας έπρεπε να αποφασιστούν συγκεκριμένα βήματα και τεχνικές τα οποία θα δημιουργούσαν μια συγκεκριμένη μεθοδολογία. Η μεθοδολογία/ο αλγόριθμος που δημιουργήθηκε και στο τέλος χρησιμοποιήθηκε, ήταν πιο απλή/ός από αυτήν/όν που μελετήθηκε στην σχετική μεθοδολογία. Ο λόγος για τον οποίο χρησιμοποιήθηκε μια πιο απλή μεθοδολογία/αλγόριθμος ήταν επειδή έφερε τα επιθυμητά αποτελέσματα που ζητούσε ο στόχος της παρούσας διπλωματικής εργασίας. Δεν χρειάστηκε να χρησιμοποιηθούν σε μεγάλο βαθμό οι δύσκολες και οι περίπλοκες μέθοδοι και τεχνικές της σχετικής βιβλιογραφίας για την επίτευξη του στόχου της.

Στα αρχικά βήματα της παρούσας διπλωματικής εργασίας αποφασίστηκε η συλλογή δεδομένων να γίνεται μόνο με tweets τα οποία θα επεξεργάζονται. Όπως αναφέρθηκε και σε προηγούμενο κεφάλαιο τα tweets αποτελούνται από κείμενο το οποίο περιέχει λέξεις κλειδιά, hahstags και URLs. Έτσι η επεξεργασία των δεδομένων βασιζόταν σε αυτά τα χαρακτηριστικά από τα οποία θα έπρεπε να αντληθούν πληροφορίες και να εξαχθούν αποτελέσματα. Από τα hashtags και τις λέξεις κλειδιά θα μπορούσε να βρεθεί το όνομα και η τοποθεσία της πολιτιστικής εκδήλωσης ή του

γεγονότος, ενώ με τα URLs μπορούσε να βρεθεί κάποια ιστοσελίδα οποία περιέχει λεπτομέρειες σχετικές με πολιτιστικά γεγονότα και πολιτιστικές εκδηλώσεις.

Κατά την διάρκεια ολοκλήρωσης της μεθοδολογίας χρησιμοποιήθηκαν δύο προσεγγίσεις. Η πρώτη προσέγγιση αναφέρεται στον εντοπισμό των tweets τα οποία αναφέρουν συγκεκριμένες τοποθεσίες της Ολλανδίας στις οποίες διεξάγονται συχνά πολιτιστικές εκδηλώσεις και γεγονότα και των tweets που ανέφεραν ημερομηνίες (με χρήση dictionary). Όμως με αυτή την προσέγγιση τα αποτελέσματα δεν ήταν ικανοποιητικά. Δηλαδή, ο αριθμός των tweets τα οποία ανέφεραν τοποθεσίες στις οποίες διεξάγονται συχνά πολιτιστικές εκδηλώσεις ήταν πάρα πολύ μικρός. Έτσι αποφασίστηκε να ακολουθηθεί μια δεύτερη προσέγγιση. Η δεύτερη προσέγγιση αναφέρεται στην ομαδοποίηση των tweets για κάθε ημερομηνία ξεχωριστά (με χρήση dictionary) από την αρχή ενός μήνα μέχρι το τέλος του (Μάιο μέχρι Αύγουστο του 2014). Άρα το συνολικό σύνολο δεδομένων για αυτούς τους μήνες τριάντα έξι εκατομμύρια tweets τα οποία έπρεπε να επεξεργαστούν. Η δεύτερη προσέγγιση έφερε τα επιθυμητά αποτελέσματα οπότε η μεθοδολογία ολοκληρώνεται με αυτήν.

<u>Βήματα μεθοδολογίας/αλγορίθμου πρώτης προσέγγισης:</u>

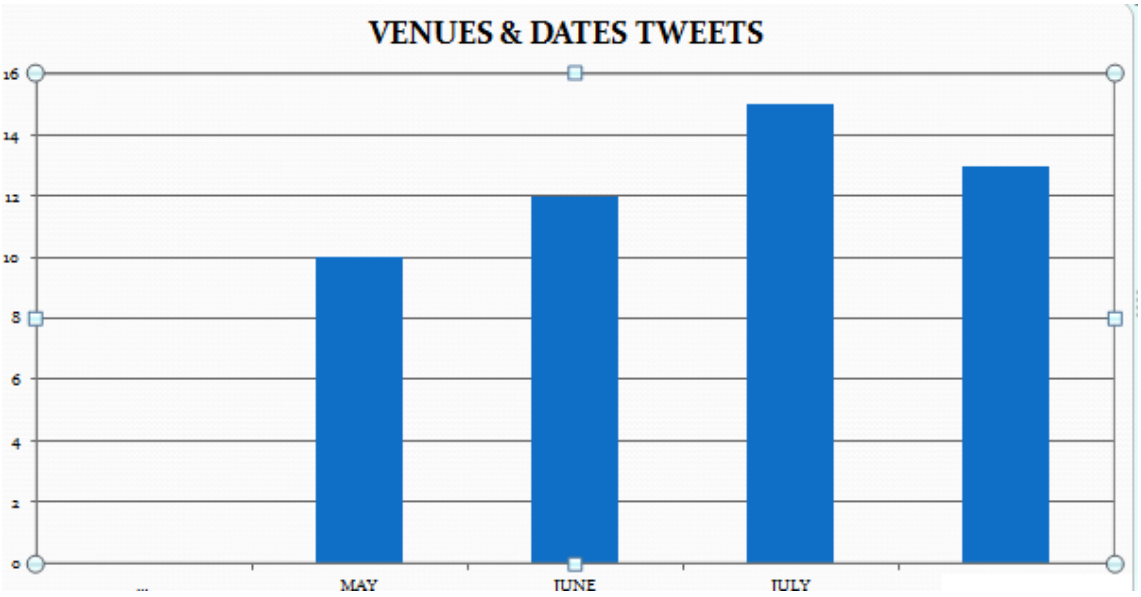
1. Το πρώτο βήμα της μεθοδολογίας/αλγοριθμου ήταν η συλλογή δεδομένων από το κοινωνικό δίκτυο Twitter με τον τρόπο τον οποίο αναφέρθηκε σε προηγούμενο κεφάλαιο των Τεχνικών Συλλογής και Ανάλυσης Δεδομένων (κεφάλαιο 4). Τα δεδομένα αυτά είναι tweets από κατοίκους της Ολλανδίας.
2. Το δεύτερο βήμα ήταν η συλλογή δεδομένων από ιστοσελίδες της Ολλανδίας οι οποίες περιέχουν λεπτομέρειες σχετικά με πολιτιστικές εκδηλώσεις και γεγονότα. Η συλλογή δεδομένων αφορούσε μόνο την ημέρα και την ώρα της πολιτιστικής εκδήλωσης/γεγονότος. Η συγκεκριμένη συλλογή δεδομένων έγινε με σκοπό να χρησιμοποιηθεί σε μετέπειτα στάδιο ως Ground Truth.
3. Στο τρίτο βήμα έγινε εντοπισμός των tweets τα οποία αναφέρουν συγκεκριμένες τοποθεσίες της Ολλανδίας στις οποίες διεξάγονται συχνά πολιτιστικές εκδηλώσεις και γεγονότα (δημιουργία αρχείου με τέτοιες

τοποθεσίες). Μετά τον εντοπισμών των συγκεκριμένων tweets, ακολουθεί η επεξεργασία τους. Συγκεκριμένα, στην παρούσα διπλωματική εργασία μελετήθηκαν τα πιο δημοφιλή hashtags, λέξεις κλειδιά , URLs και ημερομηνίες.

4. Για περαιτέρω ανάλυση των δεδομένων του Twitter, εντοπίστηκαν τα tweets τα οποία αναφέρουν ημερομηνία σε οποιαδήποτε μορφή της. Συγκεκριμένα, δημιουργήθηκε ένα λεξικό (dictionary) στην προγραμματιστική γλώσσα JAVA το οποίο αναγνωρίζει τις ημερομηνίες στις διάφορες μορφές τους. Για παράδειγμα το λεξικό αναγνωρίζει τις ημερομηνίες που έχουν μορφή:
    a. "yyyy.MM.dd"
    b. "dd-MM-yyyy"
    c. "MM:dd:yyyy"
    d. MM/dd/yyyy
    e. Μέρες εβδομάδας στα αγγλικά και στα ολλανδικά
    f. Μήνες και συντομεύσεις τους στα αγγλικά και στα ολλανδικά
    g. Κεφαλαία και μικρά

    - Όπου ‘y’ γίνεται αναφορά στην χρονολογία (year), ‘M’ στον μήνα (Month) και όπου ‘d’ στην μέρα (day)
5. Τα πιο πάνω σύνολα δεδομένων συνδυάστηκαν σε ένα. Δηλαδή το νέο σύνολο δεδομένων περιείχε tweets τα οποία ανέφεραν τοποθεσίες της Ολλανδίας και ημερομηνίες μαζί. Αυτό μπορεί να θεωρηθεί ως ένα επιπλέον φιλτράρισμα
6. Στη συνέχεια έγινε φιλτράρισμα αυτών των tweets και βρέθηκαν τα πιο δημοφιλή hashtags, λέξεις κλειδιά , URLs και ημερομηνίες έγιναν με φιλτράρισμα του κειμένου των tweets. Για το φιλτράρισμα των hashtags και των λέξεων κλειδιών, δημιουργήθηκε ένα αρχείο το οποίο περιείχε τερματικές λέξεις στα Ολλανδικά και στα Αγγλικά. Οπότε, οι λέξεις οι οποίες βρίσκονταν στο αρχείο αφαιρούνταν από το κείμενο του κάθε tweet το οποίο τις περιείχε. Για το φιλτράρισμα των URLs πραγματοποιήθηκε επέκταση (expand) για κάθε περικλειόμενο URL το οποίο περιέχεται στο κείμενο του tweet. Η επέκταση των περικλειόμενων URLs, έγινε με τη δημιουργία σύνδεσης με HTTP πρωτοκόλλου έτσι ώστε να επιστραφεί το πρωτότυπο URL.

7. Αυτός ο τρόπος δεν έφερε ικανοποιητικά αποτελέσματα αφού εντοπίστηκε μόνος ένα μικρός αριθμός από πολιτιστικά γεγονότα.
8. Η ένωση των tweets τα οποία αναφέρουν συγκεκριμένες τοποθεσίες της Ολλανδίας με αυτά που αναφέρουν ημερομηνίες ήταν πολύ μικρός οπότε τα αποτελέσματα δεν ήταν τα επιθυμητά.



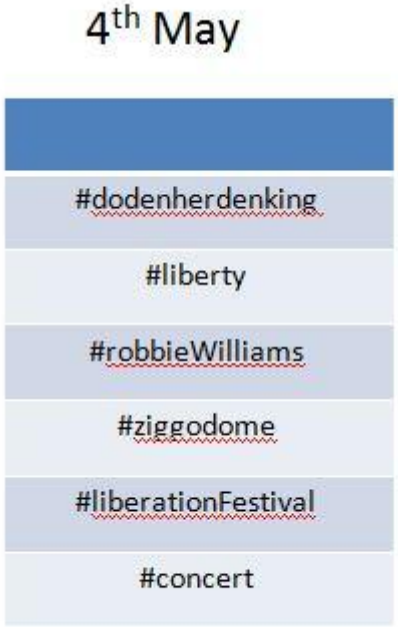
Εικόνα 15, ο αριθμός των tweets τα οποία αναφέρουν στο κείμενο τους τοποθεσίες και ημερομηνίες είναι πολύ μικρός.

Βήματα μεθοδολογίας/αλγορίθμου δεύτερης προσέγγισης:

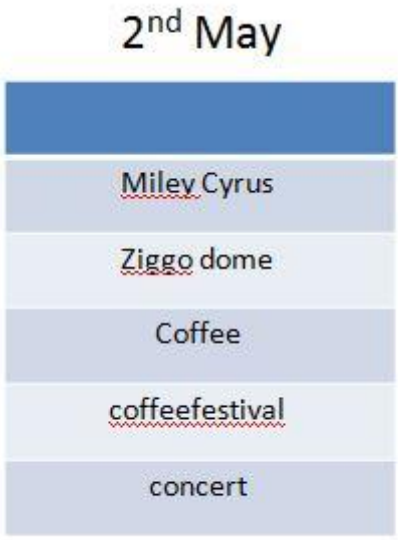
1. Το πρώτο βήμα της μεθοδολογίας/αλγοριθμου ήταν η συλλογή δεδομένων από το κοινωνικό δίκτυο Twitter με τον τρόπο τον οποίο αναφέρθηκε σε προηγούμενο κεφάλαιο των Τεχνικών Συλλογής και Ανάλυσης Δεδομένων (κεφάλαιο 3). Τα δεδομένα αυτά είναι tweets από κατοίκους της Ολλανδίας.
2. Το δεύτερο βήμα ήταν η συλλογή δεδομένων από ιστοσελίδες της Ολλανδίας οι οποίες περιέχουν λεπτομέρειες σχετικά με πολιτιστικές εκδηλώσεις και γεγονότα. Η συλλογή δεδομένων αφορούσε μόνο την ημέρα και την ώρα της πολιτιστικής εκδήλωσης/γεγονότος. Η

συγκεκριμένη συλλογή δεδομένων έγινε με σκοπό να χρησιμοποιηθεί σε μετέπειτα στάδιο ως Ground Truth.

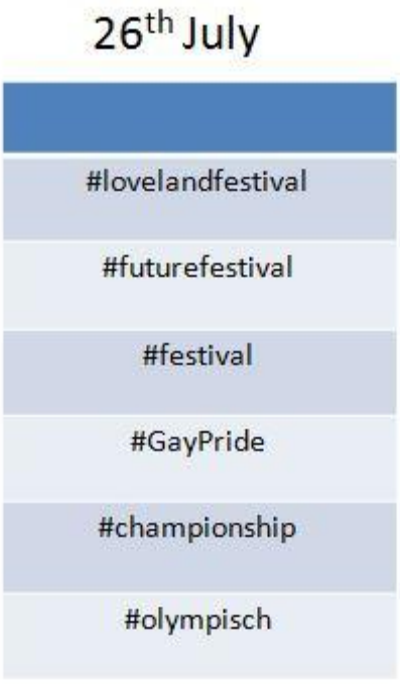
3. Η δεύτερη προσέγγιση αναφέρεται στην ομαδοποίηση των tweets για κάθε ημερομηνία ξεχωριστά από την αρχή ενός μήνα μέχρι το τέλος του ($1^{η}$ μέρα μέχρι την $31^{η}$ μέρα για τους μήνες Μάιο μέχρι Ιούλιο ξεχωριστά). Η ομαδοποίηση των tweets έγινε για όσα tweets αναφέρονται στο σήμερα (ημερομηνία δημιουργίας του tweet), σε ημερομηνία σε οποιαδήποτε μορφή (χρησιμοποιήθηκε το λεξικό της πρώτης προσέγγισης) και στην λέξη "tomorrow"
4. Επαναλαμβάνονται τα βήματα 3 και 4 από την πρώτη προσέγγιση
5. Εύρεση γειτονικών λέξεων για άντληση περισσοτέρων πληροφοριών (π.χ συγκεκριμένη τοποθεσία ή πρόσωπα)
6. Χρήση του Ground Truth το οποίο είχε συλλεχθεί



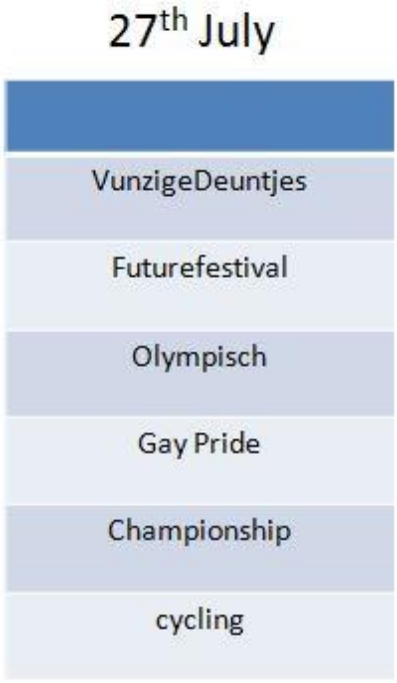
Εικόνα 16, παράδειγμα δημοφιλέστερων hashtags για την ημερομηνία 04/05/201



Εικόνα 17, παράδειγμα δημοφιλέστερων κλειδολέξεων για την ημερομηνία 02/05/2014



Εικόνα 18, παράδειγμα δημοφιλέστερων hashtags για την ημερομηνία 26/07/2014



Εικόνα 19, παράδειγμα δημοφιλέστερων κλειδολέξεων για την ημερομηνία 27/07/2014

7th July	16th July
www.newslocker.com	www.nu.nl
www.nu.nl	www.newslocker.com
www.fonearena.com	www.nos.nl
www.news.google.com	www.ad.nl
www.ad.nl	www.feedproxy.google.com
www.at5.nl	www.rtlnieuws.nl
www.melkweg.nl	www.at5.nl
www.010mycity.nl	www.rtvutrecht.nl
www.arenapoort.nl	www.fifa.com
www.sanzarrugby.com	www.mysterious-event.com

Εικόνα 20, παραδείγματα δημοφιλέστερων, URLs για τις ημερομηνίες 07/07/2014 και 16/07/2014. Τα παραπάνω URL είναι URLs ιστοσελίδων οι οποίες είναι σχετικές με πολιτιστικά γεγονότα, Για παράδειγμα, είναι ιστοσελίδες τοποθεσιών στις οποίες γίνονται συχνά events, ιστοσελίδες οι οποίες προσφέρουν εισιτήρια για εκδηλώσεις κλπ

17th August	28th June
ziggo.heeft-vacatures.nl	www.anko.nl
www.locatieamersfoort.nl	www.avbrothers.com
amsterdammetropole.com	amsterdammetropole.com
www.schouwburgpleinrotterdam.nl	www.amersfoort-productions.nl
www.ajax.nl	www.bilderberg.nl
foursquare.com	www.rtlnieuws.nl
www.melkweg.nl	www.dutchsportsinfrastructure.nl
www.concertgebouw.nl	www.oram.nl
www.arenapoort.nl	www.fifa.com
m.newslocker.com	www.nu.nl

Εικόνα 21, παραδείγματα δημοφιλέστερων, URLs για τις ημερομηνίες 17/08/2014 και 28/06/2014. Τα παραπάνω URL είναι URLs ιστοσελίδων οι οποίες είναι σχετικές με πολιτιστικά γεγονότα, Για παράδειγμα, είναι ιστοσελίδες τοποθεσιών στις οποίες γίνονται συχνά events, ιστοσελίδες οι οποίες προσφέρουν εισιτήρια για εκδηλώσεις κλπ

6.3 Παρουσίαση ψευδοκώδικα δεύτερης προσέγγισης

Ανίχνευση πολιτιστικών γεγονότων:

Συλλογή δεδομένων από Twitter από συγκεκριμένη περιοχή

Συλλογή δεδομένων από ιστοσελίδες συγκεκριμένης περιοχής

Ομαδοποίηση tweets με βάση την ημερομηνία πριν την ημερομηνία διεξαγωγής γεγονότος

Επεξεργασία των ομαδοποιήσεων και φιλτράρισμα

Βρίσκω πιο δημοφιλή κλειδολέξεις, hashtaghs, URLs

Χρήση Ground Truth

Detect entertainment events:

Collect Twitter data from specific region

Collect data from websites related to entertainment events

Group tweets by date (before start event)

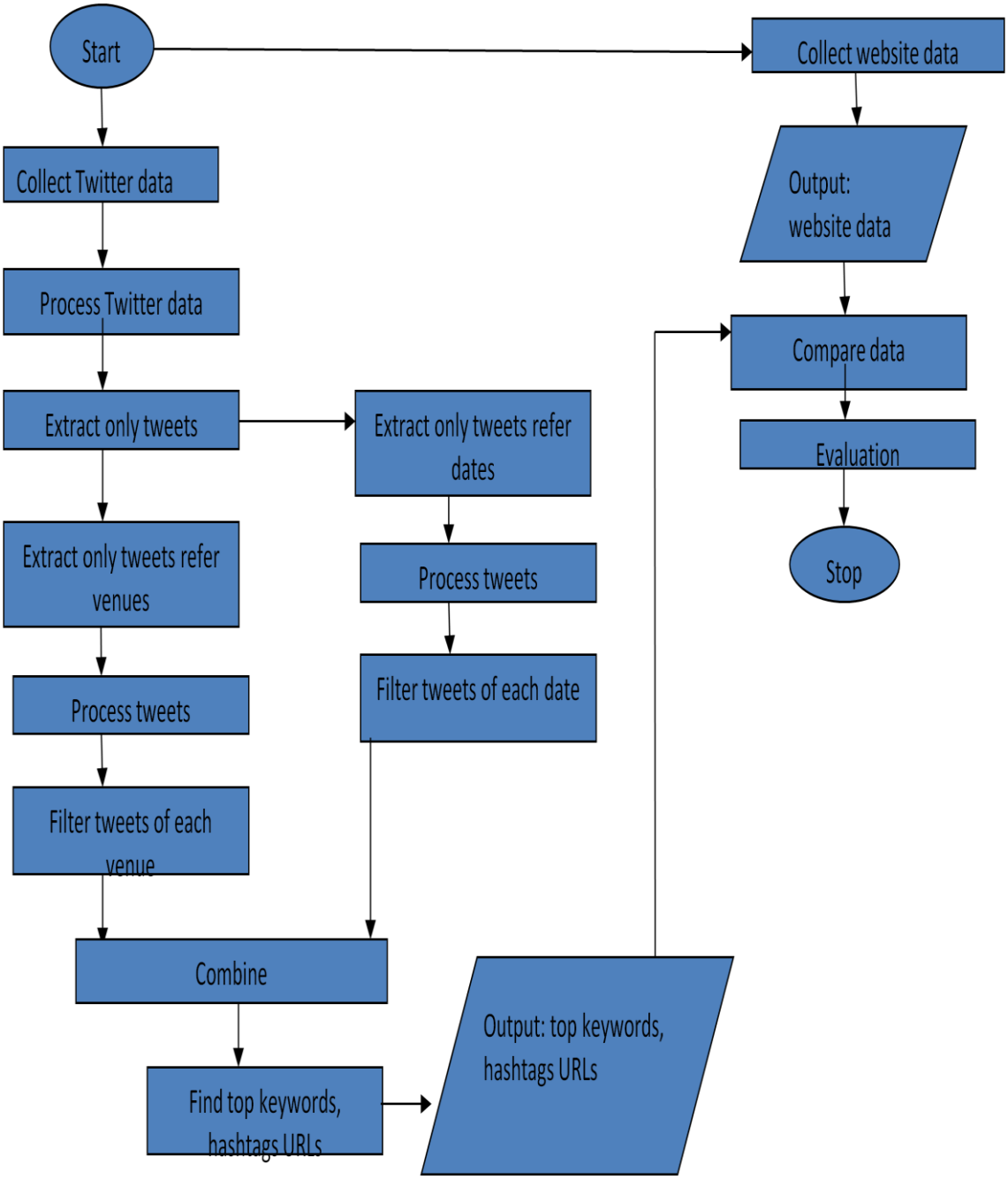
Process groups

Find top keywords,hashtags, URLs,

Use Groung Truth

6.4.1 Γραφικές αναπαραστάσεις

6.4.1 Βήματα αλγορίθμου πρώτη προσέγγισης σε γραφική αναπαράσταση



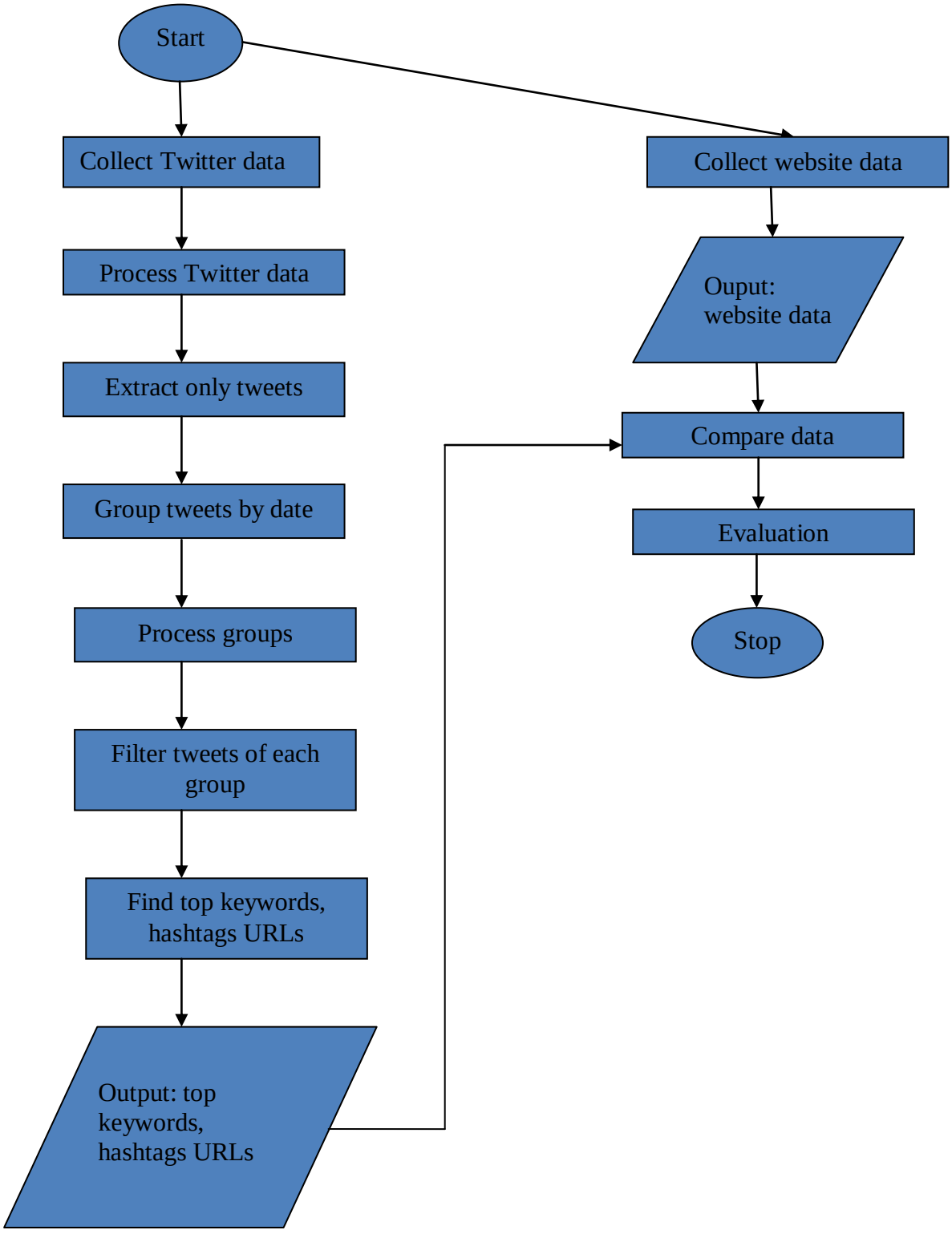
Επεξήγηση αλγορίθμου πρώτης προσέγγισης:

1. Πρώτο βήμα του αλγορίθμου είναι η συλλογή δεδομένων από το Twitter με τη χρήση των Rest και Streaming APIs του Twitter. Παράλληλα με αυτό γίνεται συλλογή δεδομένων από επίσημες και έγκυρες ιστοσελίδες της Ολλανδίας. Αυτές οι ιστοσελίδες αναφέρουν λεπτομέρειες σχετικά με 630 πολιτιστικές εκδηλώσεις(τίτλος γεγονότων και ημερομηνία διεξαγωγής)..
2. Δεύτερο βήμα είναι η επεξεργασία των δεδομένων που συλλέχθηκαν από το Twitter έτσι ώστε να εξαχθούν μόνο τα tweets και όχι οι μετά-πληροφορίες τους (meta-info).
3. Τρίτο βήμα είναι η εξαγωγή των tweets τα οποία αναφέρουν στο κείμενό τους τοποθεσίες της Ολλανδίας στις οποίες διοργανώνονται συχνά πολιτιστικά γεγονότα. Παράλληλα με αυτό εξάγονται τα tweets τα οποία αναφέρονται σε ημερομηνίες σε κάθε μορφή , κάνοντας χρήση του λεξικού των ημερομηνιών το ποιο περιγράφτηκε σε προηγούμενο υποκεφάλαιο.
4. Επόμενο βήμα είναι η επεξεργασία και το φιλτράρισμα των δύο συνόλων τα οποία προκύπτει από το προηγούμενο βήμα (τοποθεσίες και ημερομηνίες). Στο φιλτράρισμα, αφαιρούνται οι τερματικές Ολλανδικές και Αγγλικές λέξεις όπως περιγράφτηκε σε προηγούμενο υποκεφάλαιο.
5. Μετά την εξαγωγή των tweets τα οποία αναφέρουν τοποθεσίες στις οποίες διοργανώνονται συχνά πολιτιστικά γεγονότα και ημερομηνίες, γίνεται ένας συνδυασμός των δύο το οποίο παραπέμπει σε ένα επιπλέον φιλτράρισμα. Δηλαδή δημιουργείται ένα σύνολο δεδομένων στο οποίο υπάρχουν tweets τα οποία αναφέρουν στο κείμενό τους συγκεκριμένες τοποθεσίες και ημερομηνίες μαζί.
6. Στη συνέχεια γίνεται και άλλο φιλτράρισμα αυτών των tweets βρίσκονται τα δημοφιλέστερα hashtags , λέξεις κλειδιά και URLs για κάθε γεγονός.
7. Επόμενο βήμα και τελευταίο είναι η σύγκριση των επεξεργασμένων δεδομένων από τα συνδυασμένο σύνολο το οποίο προκύπτει, με το σύνολο δεδομένων το οποίο συλλέχθηκε από τις επίσημες ιστοσελίδες της Ολλανδίας. Με αυτή τη σύγκριση θα πραγματοποιηθεί η αξιολόγηση και η εγκυρότητα της μεθοδολογίας η οποία ακολουθήθηκε.

Το κατώφλι (threshold) των tweets για την ανίχνευση ενός γεγονότος είναι 5. Δηλαδή, αν 5 tweets αναφέρονταν σε ένα γεγονός τότε αυτό το γεγονός λαμβανόταν υπόψη. Όμως ο αριθμός των tweets τα οποία επιστράφηκαν από αυτόν το συνδυασμό ήταν μικρότερος από πέντε.

Αυτή η προσέγγιση δεν έφερε τα επιθυμητά αποτελέσματα λόγω του ότι ο αριθμός των tweets ο οποίος εξάχθηκε από τον συνδυασμό των tweets τα οποία αναφέρουν τοποθεσίες με αυτά που αναφέρουν ημερομηνίες ήταν μικρότερος από το κατώφλι.

6.4.2 Βήματα αλγορίθμου δεύτερης προσέγγισης σε γραφική αναπαράσταση



Επεξήγηση αλγορίθμου δεύτερης προσέγγισης:

1. Πρώτο βήμα του αλγορίθμου είναι η συλλογή δεδομένων από το Twitter με τη χρήση των Rest και Streaming APIs του Twitter. Παράλληλα με αυτό γίνεται συλλογή δεδομένων από επίσημες και έγκυρες ιστοσελίδες της Ολλανδίας. Αυτές οι ιστοσελίδες αναφέρουν λεπτομέρειες σχετικά με 630 πολιτιστικές εκδηλώσεις (τίτλος γεγονότων και ημερομηνία διεξαγωγής).
2. Δεύτερο βήμα είναι η επεξεργασία των δεδομένων που συλλέχθηκαν από το Twitter έτσι ώστε να εξαχθούν μόνο τα tweets και όχι οι μετά-πληροφορίες τους (meta-info).
3. Τέταρτο βήμα είναι η εξαγωγή των tweets τα οποία αναφέρονται σε ημερομηνίες σε κάθε μορφή, κάνοντας χρήση του λεξικού των ημερομηνιών το ποιο περιγράφτηκε σε προηγούμενο υποκεφάλαιο.
4. Επόμενο βήμα είναι η ομαδοποίηση αυτών των tweets. Η ομαδοποίηση γίνεται για κάθε μέρα του κάθε μήνα ξεχωριστά. Δηλαδή από την $1^{η}$ μέχρι την $31^{η}$ μέρα του μήνα για τους μήνες Μάιο μέχρι Ιούνιο του 2014 ξεχωριστά. Στην ομαδοποίηση χρησιμοποιείται το λεξικό το οποίο εξηγήθηκε σε προηγούμενο υποκεφάλαιο, η ημέρα ανάρτησης του tweet και η λέξη “tomorrow”. Δηλαδή, αν ένα tweet σε αυτή την ομαδοποίηση αναφέρει τη λέξη “tomorrow”, τότε αυτό σημαίνει ότι την επόμενη μέρα ανάρτησης του συγκεκριμένου tweet, θα πραγματοποιηθεί ένα πολιτιστικό γεγονός.
5. Στη συνέχεια γίνεται φιλτράρισμα αυτών των tweets. Δηλαδή, βρίσκονται τα δημοφιλέστερα hashtags , λέξεις κλειδιά και URLs για κάθε γεγονός.
6. Επόμενο βήμα και τελευταίο είναι η σύγκριση των γεγονότων τα οποία ανιχνεύτηκαν από τον αλγόριθμο, με αυτών που συλλέχθηκαν από τις επίσημες ιστοσελίδες της Ολλανδίας για αξιολόγηση και εγκυρότητα.
   Η αξιολόγηση και η εγκυρότητα έγινε με βάση τον αριθμό των πολιτιστικών γεγονότων τα οποία ανιχνεύτηκαν και μετρήθηκαν ως έγκυρα , με αυτά τα οποία συλλέχθηκαν από την ιχνηλάτιση των ιστοσελίδων.

   Το κατώφλι (threshold) των tweets για την ανίχνευση ενός γεγονότος είναι 5. Δηλαδή, αν 5 tweets αναφέρονταν σε ένα γεγονός τότε λαμβανόταν υπόψη.

Αν ένα γεγονός το οποίο ανιχνεύτηκε και υπάρχει μέσα σε μια από τις τρεις ιστοσελίδες οι οποίες ιχνηλατίστηκαν τότε λαμβανόταν υπόψη και μετριόταν ως έγκυρο. Ενώ αν ένα γεγονός το οποίο δεν υπάρχει μέσα στις ιστοσελίδες τότε δεν μετριόταν ως έγκυρο. Ανιχνεύτηκαν περίπου 615 γεγονότα από τα οποία τα 600 μετρήθηκαν ως έγκυρα ενώ τα υπόλοιπα 15 είναι μη-έγκυρα και αποτελούνταν από περίπου 5 tweets το κάθε ένα. Δηλαδή, ο αριθμός των tweets των 15 αυτών γεγονότων τα οποία ανιχνεύτηκαν είναι ίσος με το κατώφλι αλλά δεν μπορούσαν να ληφθούν ως έγκυρα επειδή δεν αναφέρονταν σε καμία από τις τρεις ιστοσελίδες οι οποίες ανιχνεύτηκαν.

6.5 Αξιολόγηση και η εγκυρότητα

Μετά την ανάπτυξη και το τέλος μιας μεθοδολογίας / αλγορίθμου, πρέπει να ακολουθήσει η αξιολόγησή της/του έτσι ώστε να κριθεί εάν είναι σωστή/ός και αν φέρει σωστά , ικανοποιητικά και έγκυρα αποτελέσματα.

Αποφασίστηκε ότι ο πιο σωστός τρόπος να γίνει η αξιολόγηση και η εγκυρότητα της μεθοδολογίας/αλγορίθμου η/ο οποία/ος δημιουργήθηκε στα πλαίσια της παρούσας διπλωματικής εργασίας, είναι να χρησιμοποιηθούν τα δεδομένα τα οποία συλλέχθηκαν από την ιχνηλάτιση (crawling) των τριών ιστοσελίδων της Ολλανδίας.

Όπως αναφέρθηκε και σε προηγούμενο κεφάλαιο, έγινε συλλογή δεδομένων από τρεις επίσημες ιστοσελίδες τις Ολλανδίας:

- http://www.iamsterdam.com
- http://www.amsterdam.info
- http://www.holland.com

Λόγω του ότι αυτές οι ιστοσελίδες είναι επίσημες ιστοσελίδες της χώρας, οι πληροφορίες των γεγονότων οι οποίες συλλέχθηκαν κατά την διαδικασία της ιχνηλάτισης (crawling) είναι έγκυρες και αξιόπιστες και δεν μπορούν να διαψευθούν. Αυτές οι ιστοσελίδες, αναφέρουν και περιέχουν πληροφορίες σχετικά με πολιτιστικά γεγονότα τα οποία διοργανώνονται σε ολόκληρη τη χώρα. Οι πληροφορίες οι οποίες συλλέχθηκαν από αυτές τις ιστοσελίδες (τίτλος γεγονότων και ημερομηνία διεξαγωγής) θα χρησιμοποιηθούν ως Ground Truth για τον αλγόριθμο ο οποίος αναπτύχθηκε στα πλαίσια της παρούσας διπλωματικής εργασίας.

Για την αξιολόγηση του αλγορίθμου ο οποίος αναπτύχθηκε πρέπει να ληφθούν υπόψη δύο παράμετροι:

1. Ο αριθμός των πολιτιστικών γεγονότων ή πολιτιστικών εκδηλώσεων τα οποία εξήχθηκαν από την μεθοδολογία που ακολουθήθηκε
2. Ο αριθμός των πολιτιστικών γεγονότων ή πολιτιστικών εκδηλώσεων τα οποία συλλέχθηκαν κατά την διαδικασία της ιχνηλάτισης (crawling) των τριών επίσημων ιστοσελίδων της Ολλανδίας

Ο αριθμός των πολιτιστικών γεγονότων ή πολιτιστικών εκδηλώσεων τα οποία εξήχθηκαν από την μεθοδολογία που ακολουθήθηκε είναι περίπου 600.
Ο αριθμός των γεγονότων ο οποίος επιστράφηκε από το Ground Truth είναι περίπου 630.

Για την εγκυρότητα των αποτελεσμάτων της μεθοδολογίας / του αλγορίθμου που αναπτύχθηκε, χρειάστηκε το σύνολο των πληροφοριών που συλλέχθηκε από τις τρεις επίσημες ιστοσελίδες (τίτλος γεγονότων και ημερομηνία διεξαγωγής), να διαχωριστεί σε τρία υποσύνολα. Το κάθε υποσύνολο περιέχει πληροφορίες από μία μόνο επίσημη ιστοσελίδα. Επόμενο βήμα είναι η σύγκριση των αποτελεσμάτων του αλγόριθμου που αναπτύχθηκε, με τις πληροφορίες του κάθε υποσυνόλου (δηλαδή της κάθε ιστοσελίδας). Κατά την διάρκεια της σύγκρισης, θα διαπιστωθεί πόσο έγκυρο είναι το πολιτιστικό γεγονός το οποίο ανιχνεύτηκε από τον αλγόριθμο. Για παράδειγμα, εάν ένα πολιτιστικό γεγονός αναφέρεται και στα τρία υποσύνολα, τότε λαμβανόταν υπόψη και μετριόταν ως έγκυρο. Ενώ αν ένα γεγονός το οποίο δεν υπάρχει μέσα στα τρία υποσύνολα τότε δεν μετριόταν ως έγκυρο. Αν ένα γεγονός υπήρχε και στα τρία υποσύνολα τότε είναι πιο έγκυρο από ένα γεγονός το οποίο υπήρχε σε δύο υποσύνολα και ακόμα πιο έγκυρο από ένα γεγονός το οποίο υπήρχε μόνο σε ένα υποσύνολο.

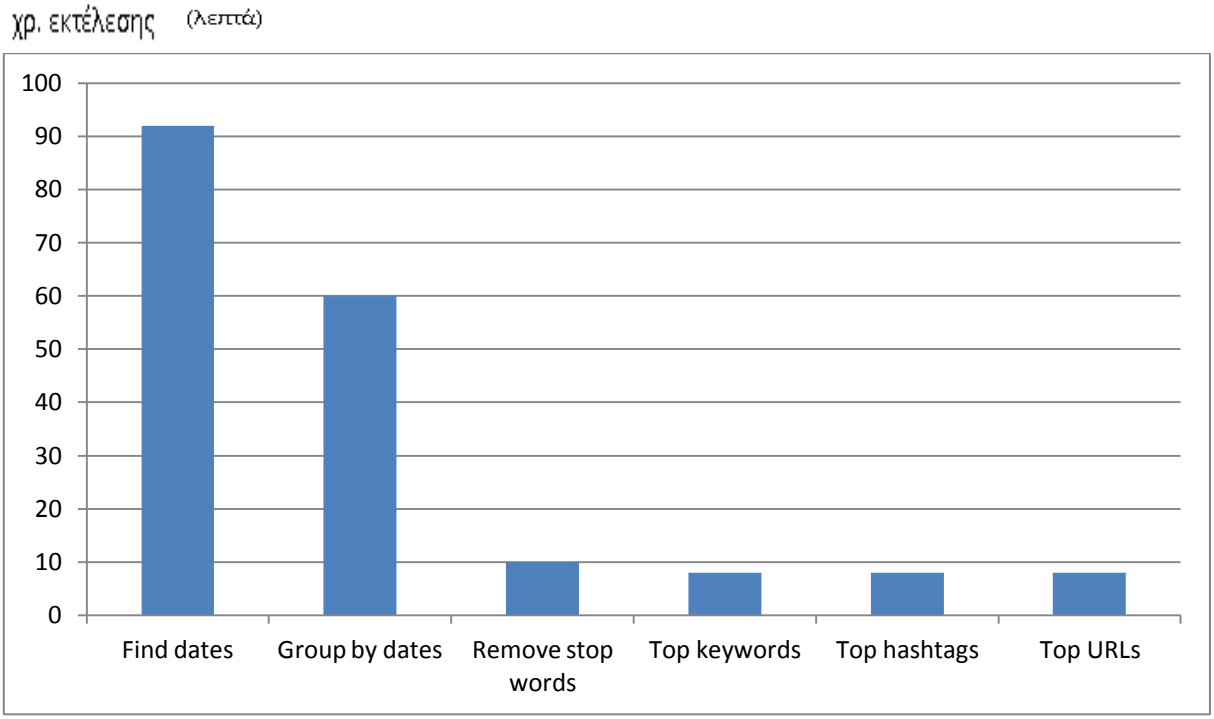
Ανιχνεύτηκαν περίπου 615 γεγονότα από τα οποία τα 600 μετρήθηκαν ως έγκυρα ενώ τα υπόλοιπα 15 είναι μη-έγκυρα αφού αποτελούνταν από περίπου 5 tweets το κάθε ένα. Άρα έστω και ξεπερνούσαν το κατώφλι των πέντε tweets το οποίο αναφέρθηκε, δεν μπορούσε να γίνει έγκυρο γιατί οι ιστοσελίδες οι οποίες ιχνηλατίστηκαν δεν το ανέφερα

Εάν ένα πολιτιστικό γεγονός δεν αναφέρεται σε κανένα υποσύνολο τότε δεν μπορεί να επιβεβαιωθεί άρα δεν μπορεί να εξεταστεί ούτε η εγκυρότητά του. Μερικοί λόγοι οι οποίοι μπορεί να συμβεί αυτό είναι τα tweets τα οποία αναφέρουν το συγκεκριμένο πολιτιστικό γεγονός να είναι ψευδής, ή οι χρήστες οι οποίοι έγραψαν τα tweets αυτά ανήκουν σε μια μικρή ομάδα στην οποία μπορεί αυθόρμητα κάποιος να άρχισε να τραγουδά ή να παίζει μουσική και οι φίλοι του να ανάρτησαν tweets για αυτό.

Με βάση τους πιο πάνω αριθμούς πολιτιστικών γεγονότων, μπορεί να συμπεραθεί ότι το ποσοστό επιτυχίας της μεθοδολογίας/αλγορίθμου ο οποίος αναπτύχθηκε είναι περίπου 95% και το ποσοστό λάθους είναι περίπου 5%

Ο χρόνος που χρειάζεται για την ανάλυση όλου του συνόλου δεδομένων για τους μήνες Μάιο μέχρι Αύγουστο του 2014 αποτελείται από 36 εκατομμύρια tweets (9 εκατομμύρια tweets για κάθε μήνα). Η ανάλυση αυτή γίνεται με την εύρεση ημερομηνιών με τη χρήση του λεξικού ημερομηνιών όπως εξηγήθηκε σε προηγούμενο υποκεφάλαιο , η ομαδοποίηση των ημερομηνιών όπως εξηγήθηκε , η επεξεργασία της κάθε ομάδας η οποία προκύπτει , το φιλτράρισμα της κάθε ομάδας με την αφαίρεση των κλειδολέξεων, και η εύρεση των πιο δημοφιλών κλειδολέξεων, hashtags και URLs. Από την ανάλυση των δεδομένων του συνόλου αυτού έγινε ανίχνευση πολιτιστικών γεγονότων. Ο αριθμός των πολιτιστικών γεγονότων τα οποία ανιχνεύτηκαν είναι 600. Αυτή η ανάλυση και η ανίχνευση γεγονότων χρειάστηκε 188 λεπτά για να ολοκληρωθεί. Δηλαδή 3 ώρες και 13 λεπτά. Στην πιο κάτω γραφική παράσταση φαίνεται καθαρά ο χρόνος που χρειάζεται κάθε βήμα της ανάλυσης των δεδομένων.

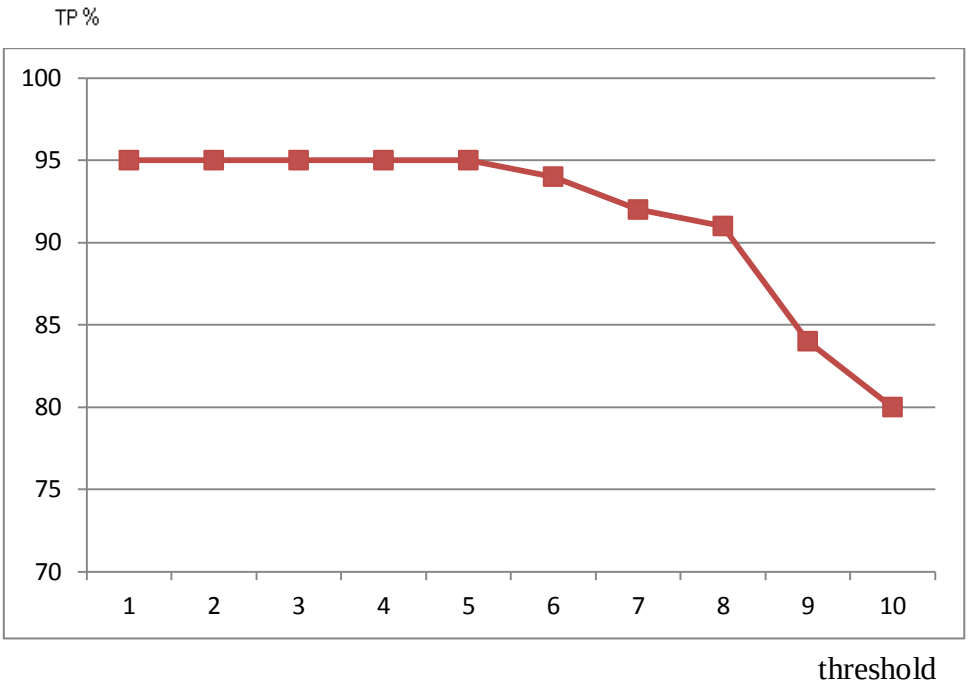
Γραφική παράσταση για την εύρεση και των 600 πολιτιστικών γεγονότων:



Παρατηρείται ότι τον περισσότερο χρόνο της ανάλυσης χρειάζεται το βήμα το οποίο αναγνωρίζει τις ημερομηνίες μέσα από τον μεγάλο όγκο δεδομένων ο οποίος συλλέχθηκε. Αυτό γιατί μέσα στο κείμενο του tweet μπορεί μια ημερομηνία να είναι γραμμένη σε οποιαδήποτε μορφή, οπότε αν βρεθούν ημερομηνίες πρέπει να ελεγχθούν

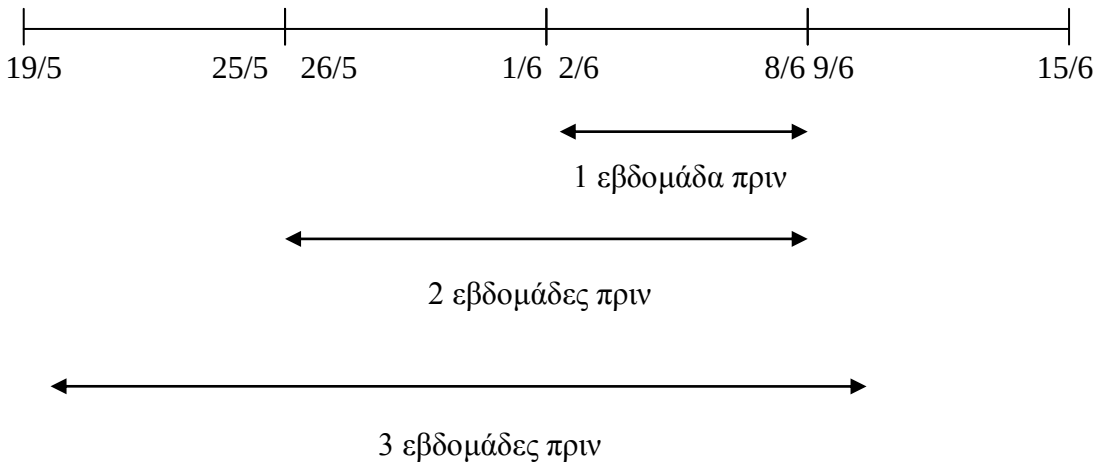
αν είναι έγκυρες με βάση το λεξικό το οποίο δημιουργήθηκε κατά την διαδικασία της μεθοδολογίας που ακολουθήθηκε.

Γραφική παράσταση για True positives για διαφορετικά threshold

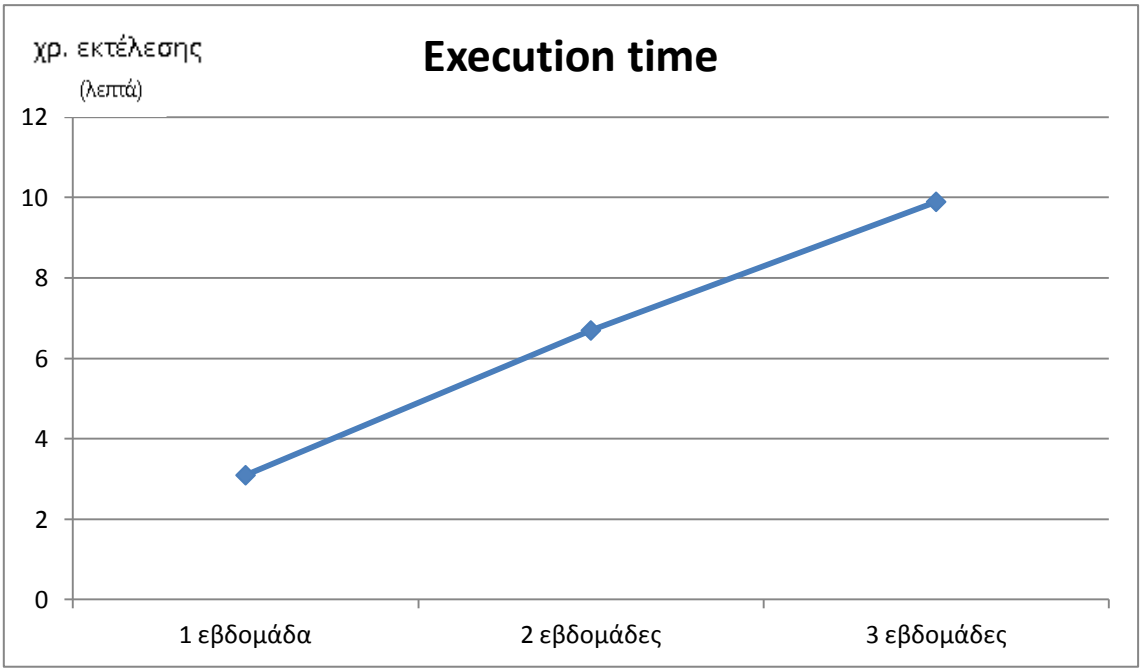


Παρατηρείται ότι με κατώφλι (threshold) που ισούται με 5 υπάρχει το μεγαλύτερο ποσοστό επιτυχίας στην ανίχνευση έγκυρων πολιτιστικών γεγονότων. Σε δοκιμές που έγιναν με κατώφλι μεγαλύτερο του 5, ο αριθμός των έγκυρων πολιτιστικών γεγονότων τα οποία ανιχνεύονταν μειωνόταν άρα μειωνόταν και το ποσοστό των True Positives (με βάση τα 630 πολιτιστικά γεγονότα τα οποία αντλήθηκαν κατά την διάρκεια της ιχνηλάτισης των έγκυρων ιστοσελίδων της Ολλανδίας) όπως φαίνεται και στην πιο πάνω γραφική παράσταση.

Μελέτη χρόνου εκτέλεσης και True Positives για μια συγκεκριμένη εβδομάδα όταν μελετηθούν και αναλυθούν δεδομένα για τις 3 προηγούμενες τις εβδομάδες

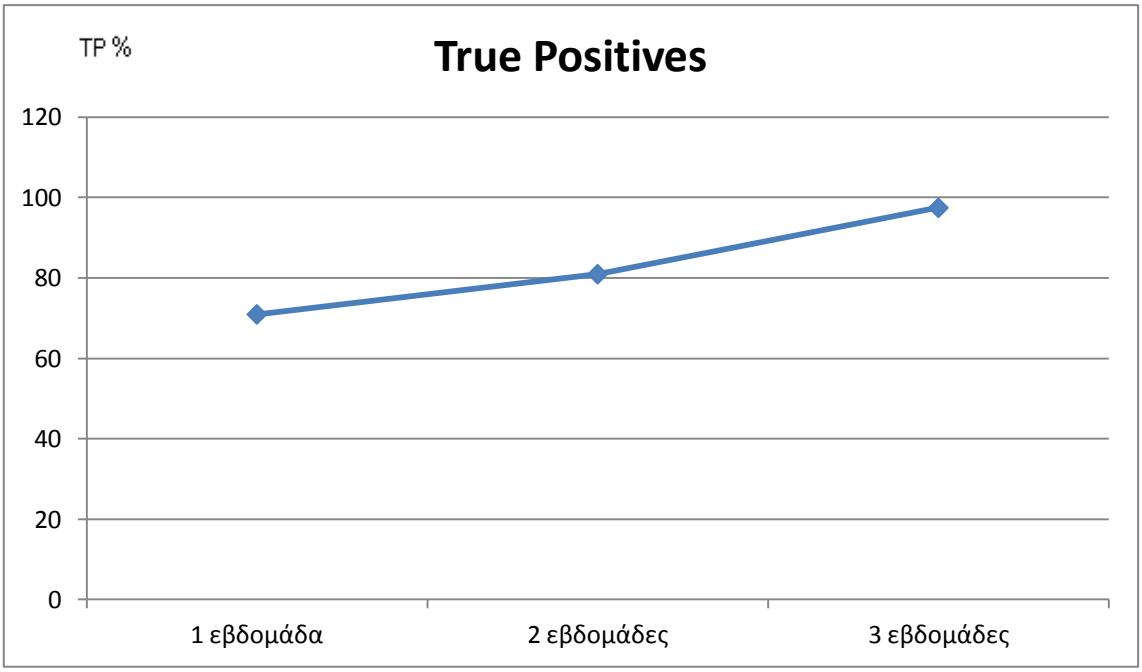


Στο πιο πάνω σχεδιάγραμμα φαίνεται η επιλογή της εβδομάδας 9 Ιουνίου μέχρι 15 Ιουνίου 2014 για την ανάλυση του χρόνου εκτέλεσης και του ποσοστού True Positives όταν μελετηθεί η χρονική περίοδος της μιας εβδομάδας πιο πριν, μέχρι την τρίτη εβδομάδα πιο πριν. Δηλαδή πόσος χρόνος χρειάζεται για να ανιχνευθούν τα πολιτιστικά γεγονότα της συγκεκριμένης εβδομάδας όταν εφαρμοστεί η μεθοδολογία ανίχνευσης στις μία, δύο και τρείς προηγούμενες εβδομάδες. Επίσης, θα βρεθούν τα True Positives για τις μία, δύο και τρείς προηγούμενες εβδομάδες. Επιλέχθηκε η συγκεκριμένη εβδομάδα διότι είναι η εβδομάδα κατά την οποία διοργανώθηκαν τα περισσότερα πολιτιστικά γεγονότα.



εβδομάδες

Στην πιο πάνω γραφική παράσταση φαίνεται ο χρόνος εκτέλεσης για την ανίχνευση των γεγονότων τα οποία διοργανώθηκαν στις 9 Ιουνίου μέχρι 15 Ιουνίου 2014 για τις μέχρι 3 προηγούμενες εβδομάδες της. Αναλύοντας δεδομένα μόνο από της πρώτης προηγούμενης εβδομάδας της, χρειάστηκαν 3,1 λεπτά για την ανίχνευση των πολιτιστικών γεγονότων που διοργανώνονται, για τις 2 προηγούμενες εβδομάδες χρειάστηκαν 6,7 λεπτά ενώ για τις 3 προηγούμενες εβδομάδες χρειάστηκαν 9,9 λεπτά.

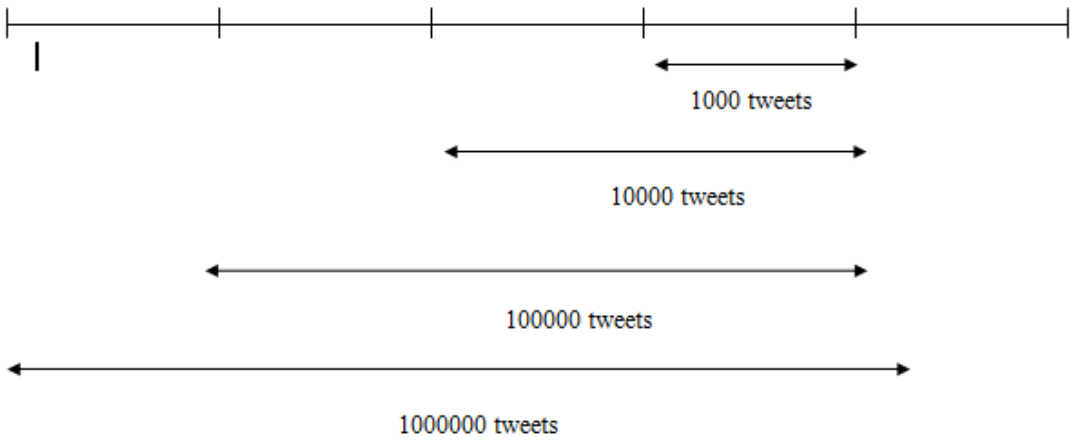


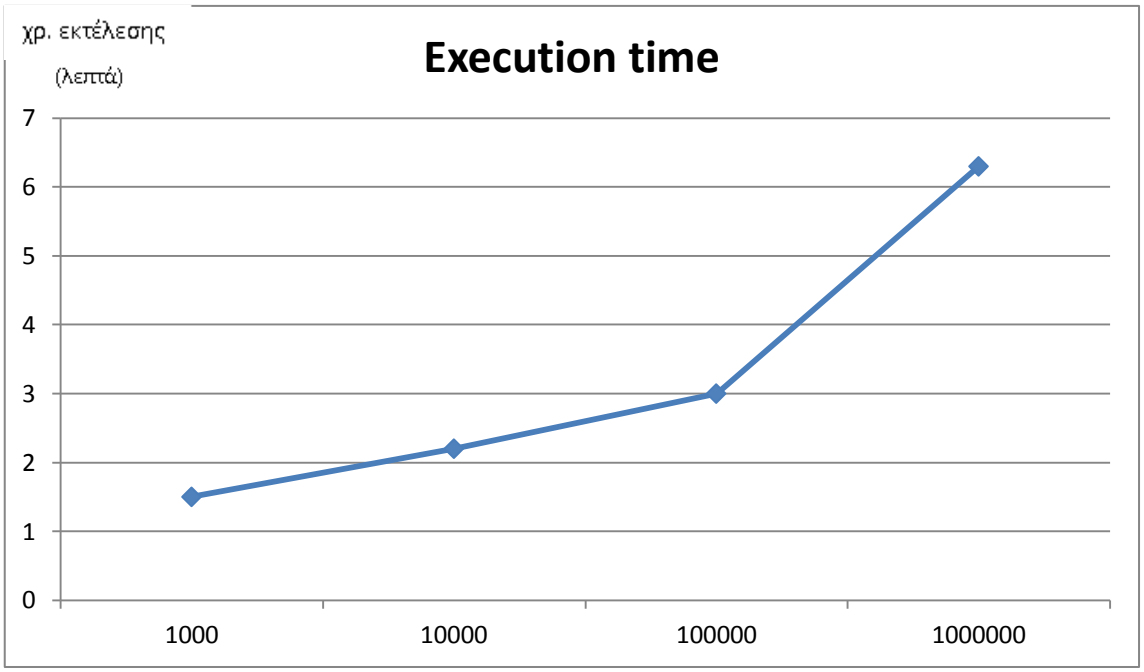
εβδομάδες

Στην πιο πάνω γραφική παράσταση φαίνεται το ποσοστό True Positives για την ανίχνευση πολιτιστικών γεγονότων για την εβδομάδα με ημερομηνίες 9 μέχρι 15 Ιουνίου 2014. Κατά τη διάρκεια αυτής της εβδομάδας διοργανώθηκαν συνολικά 80 πολιτιστικά γεγονότα. Αναλύοντας δεδομένα μόνο της προηγούμενης εβδομάδας της, ανιχνεύτηκαν 57 από τα 80 πολιτιστικά γεγονότα, για τις 2 προηγούμενες εβδομάδες ανιχνεύτηκαν 65 από τα 80 πολιτιστικά γεγονότα ενώ για τις 3 προηγούμενες εβδομάδες ανιχνεύτηκαν 78 από τα 80 πολιτιστικά γεγονότα. Στην πιο πάνω γραφική παράσταση αναπαριστάται το ποσοστό των πολιτιστικών γεγονότων τα οποία ανιχνεύτηκαν σε σχέση με τα 80 που διοργανώθηκαν τη συγκεκριμένη εβδομάδα.

Από τις πιο πάνω γραφικές παραστάσεις μπορεί να συμπεραθεί ότι όταν αναλύονται δεδομένα μία εβδομάδα πιο πριν, το ποσοστό των True Positives είναι πιο μικρό από την δεύτερη εβδομάδα και την τρίτη εβδομάδα έστω και αν χρειάζεται λιγότερος χρόνος να εκτελεστεί. Το ίδιο συμβαίνει και για την δεύτερη εβδομάδα. Η τρίτη εβδομάδα πιο πριν, έχει τον μεγαλύτερο χρόνο εκτέλεσης αλλά έχει μεγαλύτερο ποσοστό True Positives από τις άλλες 2 εβδομάδες (97,5% έναντι 81% της δεύτερης εβδομάδας και 71% της πρώτης εβδομάδας πιο πριν)

Μελέτη χρόνου εκτέλεσης και True Positives για μια συγκεκριμένη εβδομάδα όταν μελετηθούν και αναλυθούν δεδομένα για ορισμένο αριθμό tweets

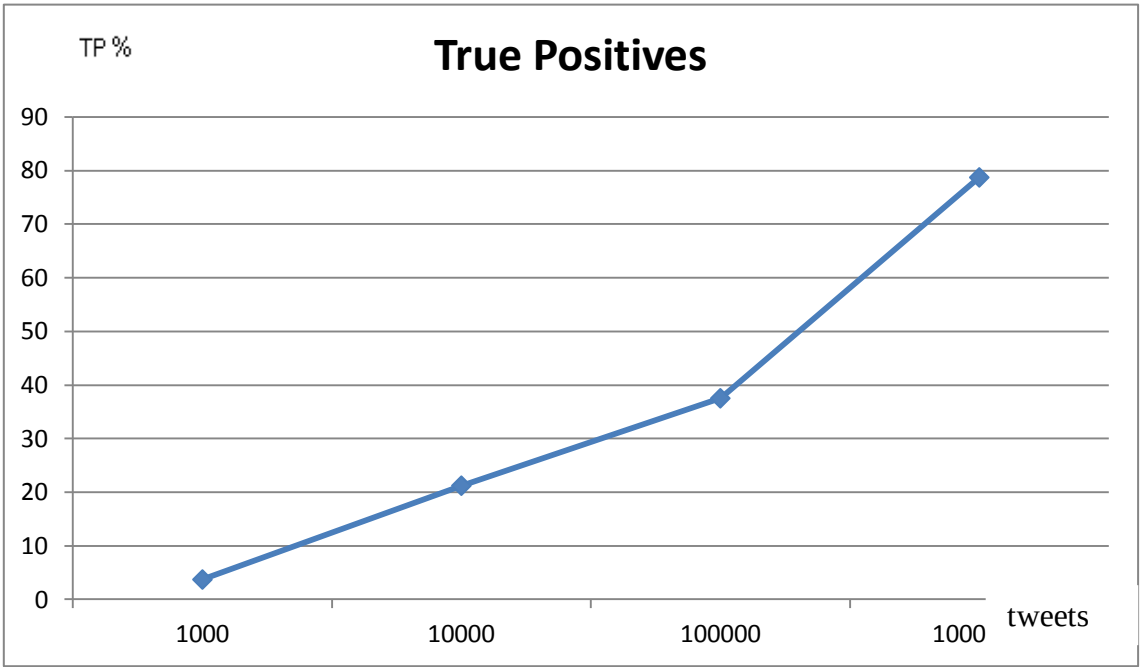




tweets

Στην πιο πάνω γραφική παράσταση φαίνεται ο χρόνος εκτέλεσης των tweets ο οποίος χρειάζεται για να ανιχνευθούν τα πολιτιστικά γεγονότα τα οποία διοργανώνονται για την εβδομάδα με ημερομηνίες 9 μέχρι 15 Ιουνίου 2014. Για τα

1000 tweets χρειάζονται 1,5 λεπτά, για τα 10000 tweets χρειάζονται 2,2 λεπτά για τα 1000000 tweets χρειάζονται 3 λεπτά και για τα 1000000 tweets χρειάζονται 6,3 λεπτά.



Στην πιο πάνω γραφική παράσταση φαίνεται το ποσοστό True Positives για την ανίχνευση πολιτιστικών γεγονότων για την εβδομάδα με ημερομηνίες 9 μέχρι 15 Ιουνίου 2014. Κατά τη διάρκεια αυτής της εβδομάδας διοργανώθηκαν συνολικά 80 πολιτιστικά γεγονότα. Αναλύοντας δεδομένα 1000 tweets, ανιχνεύτηκαν 3 από τα 80 πολιτιστικά γεγονότα, για 10000 tweets ανιχνεύτηκαν 17 από τα 80 πολιτιστικά γεγονότα , για 100000 tweets ανιχνεύτηκαν 30 από τα 80 πολιτιστικά γεγονότα και για 1000000 tweets ανιχνεύτηκαν 63 από τα 80 πολιτιστικά γεγονότα.. Στην πιο πάνω γραφική παράσταση αναπαριστάται το ποσοστό των γεγονότων τα οποία ανιχνεύτηκαν σε σχέση με τα 80 που διοργανώθηκαν τη συγκεκριμένη εβδομάδα.

Από τις πιο πάνω γραφικές παραστάσεις μπορεί να συμπεραθεί ότι όταν αναλύονται δεδομένα για μικρό αριθμό tweets τότε το ποσοστό των True Positives είναι μικρό (3% - 37%) έστω και αν χρειάζεται πιο λίγος χρόνος να εκτελεστεί (1,5 – 3 λεπτά). Ενώ για μεγάλο αριθμό tweets το ποσοστό των True Positives αυξάνεται σε ικανοποιητικά επίπεδα (78,7%) με περισσότερο όμως χρόνο (6,3 λεπτά).

Από τις πιο πάνω γραφικές παραστάσεις μπορεί να εξαχθεί το συμπέρασμα ότι ο αλγόριθμος μπορεί να επιστρέψει πολύ ακριβή αποτελέσματα σε μικρό χρονικό διάστημα. Αυτό δείχνει ότι ο αλγόριθμος μπορεί να τρέξει σε περιοδικά διαστήματα μέσα στη μέρα,έτσι ώστε να αυξηθεί η ακρίβεια των αποτελεσμάτων (περισσότερα έγκυρα πολιτιστικά γεγονότα).

Κεφάλαιο 7 Διαδικτυακή εφαρμογή

7.1 Παρουσίαση διαδικτυακής εφαρμογής

Για την εμφάνιση των αποτελεσμάτων της ανίχνευσης των πολιτιστικών γεγονότων και πολιτιστικών εκδηλώσεων δημιουργήθηκε μια διαδικτυακή εφαρμογή. Στο παρόν κεφάλαιο θα παρουσιαστεί και θα επεξηγηθεί η διαδικτυακή εφαρμογή στην οποία θα περιέχονται τα αποτελέσματα



Εικόνα 22: Στην πιο πάνω εικόνα φαίνεται η αρχική σελίδα της διαδικτυακής εφαρμογής στην οποία εμφανίζονται εικόνες οι οποίες έχουν σχέση με πολιτιστικά γεγονότα και εκδηλώσεις. Καθώς επίσης παρουσιάζονται τα τρία πιο δημοφιλή γεγονότα για τους μήνες Μάιο, Ιούνιο, Ιούλιο και Αύγουστο του 2014 ξεχωριστά.

Ένας χρήστης του Διαδικτύου ο οποίος επισκέφτηκε την συγκεκριμένη ιστοσελίδα, εκτός από τα τρία δημοφιλέστερα γεγονότα για τους πιο πάνω μήνες, είναι σε θέση να επιλέξει τον μήνα που ενδιαφέρεται..Όπως φαίνεται και στην πιο κάτω εικόνα, θα εμφανιστεί μια άλλη σελίδα η οποία θα περιέχει ένα ημερολόγιο το οποίο αναγράφει για κάθε μέρα του μήνα ξεχωριστά τα πολιτιστικά γεγονότα τα οποία ανιχνεύθηκαν για τον μήνα Μάιο.



Εικόνα 23: εμφάνιση ημερολογίου στο οποίο εμφανίζονται πολιτιστικά γεγονότα

Κάθε τίτλος του πολιτιστικού γεγονότος παρουσιάζεται ως ένας σύνδεσμος. Όταν αυτός ο σύνδεσμος επιλεχθεί, εμφανίζεται μια νέα ιστοσελίδα στην οποία αναφέρεται ο τίτλος του πολιτιστικού γεγονότος . Επίσης, αναφέρονται και τα πιο δημοφιλή hashtags, κλειδολέξεις και URLs τα οποία σχετίζονται με το συγκεκριμένο πολιτιστικό γεγονός. Όπως φαίνεται και στη εικόνα , δίπλα από τον τίτλο του πολιτιστικού γεγονότος αναφέρεται πόσο ακριβής και αξιόπιστο είναι. Αν αναφέρεται και στις τρεις ιστοσελίδες οι οποίες ιχνηλατίστηκαν, τότε υπάρχουν τρεις αστέρες δίπλα από το τίτλο, αν αναφέρεται σε δύο ιστοσελίδες τότε θα υπάρχουν δύο αστέρες κ.ό.κ



Εικόνα 24:Η εικόνα αναφέρεται στις πιο δημοφιλής λέξεις κλειδιά και hashtags μαζί με τα πιο δημοφιλή URLs του συγκεκριμένου πολιτιστικού γεγονότος. Καθώς επίσης, το συγκεκριμένο πολιτιστικό γεγονός είναι έγκυρο αφού κατέχει τρεις αστέρες άρα ανιχνεύθηκε και από τις τρεις ιστοσελίδες οι οποίες ιχνηλατίστηκαν



Εικόνα 25: Η εικόνα αναφέρεται στις πιο δημοφιλής λέξεις κλειδιά και hashtags μαζί με τα πιο δημοφιλή URLs του συγκεκριμένου πολιτιστικού γεγονότος. Καθώς επίσης, το συγκεκριμένο πολιτιστικό γεγονός δεν είναι πολύ αξιόπιστο αφού κατέχει μόνο έναν αστέρα άρα ανιχνεύθηκε μόνο από μία ιστοσελίδα από τις τρεις οι οποίες ιχνηλατίστηκαν



Εικόνα 26: Η εικόνα αναφέρεται στις πιο δημοφιλής λέξεις κλειδιά και hashtags μαζί με τα πιο δημοφιλή URLs του συγκεκριμένου πολιτιστικού γεγονότος. Καθώς επίσης, το συγκεκριμένο πολιτιστικό γεγονός δεν είναι αξιόπιστο αφού δεν κατέχει κανένα αστέρα. Αυτό σημαίνει ότι το συγκεκριμένο πολιτιστικό γεγονός δεν αναφέρεται σε καμιά από τις ιστοσελίδες οι οποίες ιχνηλατίστηκαν. Οπότε δεν μπορεί να γίνει εγκυρότητα για αυτό έτσι μπορούμε να υποθέσουμε ότι τα tweets τα οποία αναφέρουν το συγκεκριμένο πολιτιστικό γεγονός να είναι ψευδής, ή οι χρήστες οι οποίοι έγραψαν τα tweets αυτά ανήκουν σε μια μικρή ομάδα στην οποία μπορεί αυθόρμητα κάποιος να άρχισε να τραγουδά ή να παίζει μουσική και οι φίλοι του να ανάρτησαν tweets για αυτό

7.2 Τεχνολογίες διαδικτύου

Χρησιμοποιήθηκαν τεχνολογίες διαδικτύου οι οποίες βοήθησαν στην παρουσίαση των αποτελεσμάτων της μεθοδολογίας. Δηλαδή στην παρουσίαση των πολιτιστικών γεγονότων τα οποία ανίχνευσε ο αλγόριθμος.

Για την δημιουργία της διαδικτυακής εφαρμογής χρησιμοποιήθηκε η προγραμματιστική γλώσσα HTML και css. Η γλώσσα HTML και css χρησιμοποιήθηκαν με τέτοιο τρόπο έτσι ώστε να εμφανίζονται τα αποτελέσματα του αλγόριθμου / μεθοδολογίας. Επίσης, βοήθησε στη δημιουργία της αρχικής σελίδας αλλά και των υπολοίπων σελίδων, έτσι ώστε να εμφανίζονται τα αποτελέσματα του κάθε μήνα ξεχωριστά με ωραίο και όμορφο τρόπο όπως φαίνονται στις πιο πάνω εικόνες. Επιπλέον, κάνοντας χρήση της γλώσσας HTML και css προστέθηκαν εικόνες και τα δεδομένα σε μορφή στηλών (columns).

Επίσης, χρησιμοποιήθηκε η προγραμματιστική γλώσσα JavaScript. Με τη βοήθεια της προγραμματιστικής αυτής γλώσσας δημιουργήθηκε ένα sliding window στο οποίο αλλάζουν κάθε λίγα δευτερόλεπτα εικόνες σχετικά με πολιτιστικές εκδηλώσεις .

Επιπλέον, με τη χρήση της γλώσσας JavaScript ήταν δυνατόν κάθε φορά που ένας χρήστης της διαδικτυακής εφαρμογής επιλέξει ένα τίτλο ενός πολιτιστικού γεγονότος, να ανοίγει δυναμικά καινούργια ιστοσελίδα και να αναγράφονται οι πιο δημοφιλείς λέξεις και hashtags και URLs (όπως ειπώθηκε και πιο πάνω).

Αυτή η σελίδα δημιουργείται δυναμικά. Δηλαδή, όταν κάποιος επιλέξει ένα τίτλο πολιτιστικού γεγονότος, τότε με την χρήση μεθόδων της JavaScript, ανοίγεται αρχείο ή επιλέγεται ο συγκεκριμένος πίνακας από τη βάση δεδομένων, στα οποία είναι αποθηκευμένα οι δημοφιλείς λέξεις , hashtags , URLs κλπ . Στη συνέχεια τα εμφανίζει στην καινούργια ιστοσελίδα και ο χρήστης είναι σε θέση να τα δεί.

Κεφάλαιο 8 Επίλογος

Στην παρούσα διπλωματική εργασία, παρουσιάστηκε μια προσέγγιση κατά την οποία επεξεργάζονται τα δεδομένα και μέσα από αυτά ανιχνεύονται πολιτισμικά γεγονότα.

Στην αρχή παρουσιάζεται η σχετική βιβλιογραφία και σχετικές έρευνες με την παρούσα διπλωματική εργασία. Στη συνέχεια, γίνεται μια εισαγωγή στο Twitter και επεξηγούνται τα χαρακτηριστικά του (tweet, κλειδολέξεις, hashtags, URLs κλπ).

Επεξηγείται το Twitter API, ο τρόπος συλλογής δεδομένων από το Twitter με τη χρήση μεθόδων του REST και του Streaming API. Από τη συγκεκριμένη συλλογή αντλήθηκαν εννιά εκατομμύρια tweets για τους μήνες Μάιο μέχρι Αύγουστο του 2014

Αναφέρεται ο τρόπος δημιουργίας ενός ιχνηλάτη για τρεις διαφορετικές ιστοσελίδες της Ολλανδίας. Αυτές οι ιστοσελίδες, αναφέρουν πληροφορίες για πολιτισμικά γεγονότα και θα χρησιμοποιηθούν μετέπειτα ως Ground Truth. Από τη συγκεκριμένη συλλογή αντλήθηκαν πληροφορίες για περίπου 600 πολιτισμικά γεγονότα για τους μήνες Μάιο μέχρι Αύγουστο του 2014

Στη συνέχεια, αναλύονται οι δύο προσεγγίσεις, - tweets τα οποία αναφέρουν συγκεκριμένες τοποθεσίες της Ολλανδίας στις οποίες γίνονται συχνά πολιτιστικά γεγονότα μαζί με ημερομηνίες, και η ομαδοποίηση tweets με βάση την κάθε ημερομηνία ξεχωριστά. Η πρώτη προσέγγιση έτυχε επεξεργασίας, φιλτραρίσματος κειμένου (λέξεις, hashtags, URLS, ημερομηνίες με χρήση λεξικού). Στη συνέχεια , αξιολογήθηκε αλλά δεν έφερε τα επιθυμητά αποτελέσματα επειδή ο αριθμός των tweets ο οποίος επιστράφηκε ήταν πολύ μικρός.

Έτσι αποφασίστηκε να ακολουθηθεί η δεύτερη προσέγγιση, κατά την οποία έγινε ομαδοποίηση της κάθε ημερομηνίας ξεχωριστά για κάθε μήνα και επαναλήφθηκαν τα βήματα από την πρώτη προσέγγιση. Η αξιολόγηση του αλγορίθμου έγινε με τη σύγκριση του αριθμού των πολιτιστικών γεγονότων τα οποία ανιχνεύθηκαν από τη μεθοδολογία που αναπτύχθηκε, με τον αριθμό των πολιτιστικών γεγονότων που επιστράφηκαν από την ιχνηλάτηση.

Στη συνέχεια παρουσιάζεται η διαδικτυακή εφαρμογή που αναπτύχθηκε με εικόνες της αρχικής σελίδας όπου εμφανίζονται τα τρία δημοφιλέστερα γεγονότα του μήνα.

Επίσης, φαίνεται με τη χρήση εικόνων το ημερολόγιο στο οποίο αναγράφονται για κάθε μέρα του μήνα ξεχωριστά τα γεγονότα που διεξάγονται. Στη συνέχεια, γίνεται παρουσίαση εάν ένας χρήστης επιλέξει ένα γεγονός, να μπορεί να δει τις δημοφιλέστερες λέξεις και hashtags και URLs.

Σε μελλοντική εργασία θα μπορούσαν να ανιχνεύονται γεγονότα τα οποία δεν έχουν προκαθορισμένη ημερομηνία και ώρα. Για παράδειγμα, ανίχνευση ενός δυστυχήματος ή το κλείσιμο ενός δρόμου.

Βιβλιογραφία

[1] Twitter Developers [https://dev.twitter.com]

[2] Alexandre Passant, DERI, NUI Galway, Ireland; Fabien Gandon, INRIA Sophia Antipolis, France; Harith Alani, KMI, The Open University, UK Aligning Tweets with Events: Automation via Semantics

[3] Yuheng Huy Fei Wangx Subbarao Kambhampatiy Listening to the Crowd: Automated Analysis of Events via Aggregated Twitter Sentiment

[4] Elena Ilina1, Claudia Hauff1, Ilknur Celik2, Fabian Abel1, Geert-Jan Houben1 Social Event Detection on Twitter

[5] Fuchun Peng University of Massachusetts – Amherst Augmenting Naive Bayes Classifiers with Statistical Language Models

[6] JianshuWeng Services Platform Lab, HP Labs Singapore Bu-Sung Lee Event Detection in Twitter

[7] Hila Becker, Mor Naaman Luis Gravano Beyond Trending Topics: Real-World Event Identification on Twitter

[8] Richard McCreadie_, Craig Macdonald_ and Iadh Ounis_ Miles Osborne† and Sasa Petrovic† Scalable Distributed Event Detection for Twitter

[9] Rui LI #1, Kin Hou Lei _2, Ravi Khadiwala #3, Kevin Chen-Chuan Chang #4, TEDAS: a Twitter Based Event Detection and Analysis System

[10] Manish Gupta, Rui Li Kevin Chen-Chuan Towards a Social Media Analytics Platform: Event Detection and User Profiling for Twitter

[11] Mario Cataldi, Luigi Di Caro, Claudio Schifanella Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation.

[12] Streaming Trend Detection in Twitter James Benhardus

[13] Chenliang Li, Aixin Sun, Anwitaman Datta Twevent: Segment-based Event Detection from Tweets

[14] Hamed Abdelhaq, Christian Sengstock, and Michael Gertz, EvenTweet: Online Localized Event Detection from Twitter

[15] David A. Smith Perseus Project, Tufts University Medford, MA 02155, U.S.A. Detecting and Browsing Events in Unstructured Text

[16] Hassan Saif,1 Miriam Fernandez,1 Yulan He,2 Harith Alani1, On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter

[17] Heather S. Packer, Sina Samangooei, Jonathon S. Hare Event Detection using Twitter and Structured Semantic Query Expansion

[18] http://www.iamsterdam.com

[19] http://www.amsterdam.info

[20] http://www.holland.com

Παράρτημα Α

Στο παράρτημα αυτό επισυνάπτεται το API (Application Programming Interface) του Twitter για REST και Streaming APIs

Title	Resource family	Requests / 15-min window (user auth)	Requests / 15-min window (app auth)
GET application/rate_limit_status	application	180	180
GET favorites/list	favorites	15	15
GET followers/ids	followers	15	15
GET followers/list	followers	15	30
GET friends/ids	friends	15	15
GET friends/list	friends	15	30
GET friendships/show	friendships	180	15
GET help/configuration	help	15	15
GET help/languages	help	15	15
GET help/privacy	help	15	15
GET help/tos	help	15	15
GET lists/list	lists	15	15
GET lists/members	lists	180	15
GET lists/members/show	lists	15	15
GET lists/ownerships	lists	15	15
GET lists/show	lists	15	15
GET lists/statuses	lists	180	180
GET lists/subscribers	lists	180	15
GET lists/subscribers/show	lists	15	15
GET lists/subscriptions	lists	15	15
GET search/tweets	search	180	450
GET statuses/lookup	statuses	180	60
GET statuses/oembed	statuses	180	180
GET statuses/retweeters/ids	statuses	15	60
GET statuses/retweets/:id	statuses	15	60
GET statuses/show/:id	statuses	180	180
GET statuses/user_timeline	statuses	180	300
GET trends/available	trends	15	15
GET trends/closest	trends	15	15
GET trends/place	trends	15	15
GET users/lookup	users	180	60

Παράρτημα Β

Στο παράρτημα αυτό επισυνάπτεται script στην προγραμματιστική γλώσσα Python το οπίο χρησιμοποιείται για συλλογή δεδομένων σε json μορφή .

```
__author__ = 'Kasiani'
import oauth2
import time
import urllib2
import json
import tweepy

endpoint1 = "https://api.twitter.com/1.1/search/tweets.json"
#add more params to call
params = {
"oauth_version": "1.0",
"oauth_nonce": oauth2.generate_nonce(), #random string
"oauth_timestamp": int(time.time())
}

from auth import TwitterAuth

class TwitterAuth:consumer =
oauth2.Consumer(key="3I306T4BdMnBFWOZBggFgz4ya",secret="eErhCeeeCzoIN7
QF8iGHnEKxFrZZEfixU6jVYOomwrX8N87d4U")
   token = oauth2.Token(key="823701890-
5n9AZXRfT6fryPvLaNmhwHtF74qBTbVRirgQ7nsM",secret="5dpFve1CYnA2Kf7E1
SxmDTT4weJw523m1B06Fvyhno13F")
params["oauth_consumer_key"] = TwitterAuth.consumer.key
params["oauth_token"] = TwitterAuth.token.key

from tweepy.parsers import RawParser
rawParser = RawParser()
```

```
api = tweepy.API(auth_handler=TwitterAuth.consumer, parser=rawParser)

prev_id="520924571755491328"
words=["event","earthquake"]

for word in words:
    url = endpoint1
    params["q"] = word #searched word
    params["count"] = 100 #returned results per call
    params["max_id"] = str(prev_id)

    params["result_type"] = "mixed"

    req = oauth2.Request(method="GET", url=url, parameters=params)
    signature_method = oauth2.SignatureMethod_HMAC_SHA1()
    req.sign_request(signature_method, TwitterAuth.consumer, TwitterAuth.token )
    headers = req.to_header()
    url = req.to_url()
    response = urllib2.Request(url)
    data = json.load(urllib2.urlopen(response))

    if data["statuses"] != []:
        prev_id = int(data["statuses"][-1]["id"]) - 1

    f = open("outfile_" + str(word) + ".json", "w")
    json.dump(data["statuses"], f)
    print data
    print word
    f.close()
    time.sleep(5)
```