ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Αλγόριθμοι εξόρυξης δεδομένων και εφαρμογές

Data mining algorithms and applications

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Παπαγεωργίου Καλλιόπη

Βόλος, 2017

Επιβλέποντες καθηγητές:

Χούστης Ηλίας

Ομότιμος Καθηγητής Π.Θ.

(Υπογραφή)

.....

Χούστης Ηλίας

Ομότιμος Καθηγητής Π.Θ.

Βάβαλης Εμμανουήλ

Καθηγητής Π.Θ.

(Υπογραφή)

.....

Βάβαλης Εμμανουήλ

Καθηγητής Π.Θ.

(Υπογραφή)

.....

Παπαγεωργίου Καλλιόπη

Διπλωματούχος Μηχανικός Ηλεκτρονικών Υπολογιστών, Τηλεπικοινωνιών και Δικτύων του τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστημίου Θεσσαλίας

Στην οικογένειά μου

Ευχαριστίες

Θα ήθελα να ευχαριστήσω αρχικά όλους όσους με στήριξαν και με βοήθησαν όλα αυτά τα χρόνια κατά τη διάρκεια της φοίτησής μου στη σχολή. Πιο συγκεκριμένα θέλω να ευχαριστήσω τον καθηγητή μου κ. Χούστη Ηλία για την καθοδήγησή του, τη βοήθειά του και την εμπιστοσύνη που μου έδειξε. Επιπλέον θέλω να ευχαριστήσω τον καθηγητή μου κ. Βάβαλη Εμμανουήλ για την υποστήριξή του στη διπλωματική μου εργασία. Όλα αυτά συνέβαλαν έτσι ώστε να καταφέρω να ολοκληρώσω με επιτυχία τη διπλωματική μου εργασία.

Ένα μεγάλο ευχαριστώ στους φίλους μου που στάθηκαν δίπλα μου καθόλη τη διάρκεια των σπουδών μου και στα πλαίσια της φοίτησής μου στη σχολή και επιπλέον σε προσωπικό επίπεδο, τα οποία βέβαια θεωρώ ότι ειναι συνυφασμένα.

Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου και την αδερφή μου που όλα αυτά τα χρόνια στέκονται δίπλα μου σε όλα. Οι συμβουλές, οι αξίες και οι αρχές που μου δίδαξαν είναι για μένα πολύτιμα εφόδια για την μετέπειτα πορεία μου. Η σημερινή μου αυτή επιτυχία είναι μέρος και της δικής τους προσωπικής επιτυχίας. Κλείνοντας ευχαριστώ ιδιαίτερα τον αρραβωνιαστικό μου Σάκη, πλέον μέλος της οικογένειάς μου, που ήταν ένα πολύτιμο στήριγμα για μένα τα τελευταία χρόνια.

Παπαγεωργίου Καλλιόπη,
Βόλος 2017

Περίληψη

Οι άνθρωποι ανεξάρτητα από την καταγωγή τους καθώς και την εποχή που ζουν νιώθουν την ανάγκη να εκδηλώσουν τα συναισθήματά τους και τις απόψεις τους σχετικά με την επικαιρότητα σε πολιτικό, οικονομικό, κοινωνικό και προσωπικό επίπεδο. Ήδη εδώ και χρόνια δίνεται η ευκαιρία στον κόσμο να εκφράσει αυτά που πιστεύει και αισθάνεται μέσα από τα Μέσα Μαζικής Ενημέρωσης, εφημερίδες, ραδιόφωνο, τηλεόραση. Βέβαια πρέπει να σημειωθεί ότι υπάρχουν και άλλοι τρόποι έκφρασης των ανθρώπων πέρα από αυτά που προαναφέρθηκαν. Τα τελευταία χρόνια όμως έχει αναπτυχθεί ραγδαία η χρήση των κοινωνικών δικτύων ως μέσο επικοινωνίας αλλά και διατύπωσης απόψεων και εκδήλωσης συναισθημάτων. Ένα από αυτά τα κοινωνικά δίκτυα είναι και το Twitter το οποίο και θα αναλυθεί στη συνέχεια. Αρχικά, θα φανεί μέσα σε αυτή τη διπλωματική, ότι αναλύονται οι πληροφορίες που θα συλλεχθούν από το Twitter με σκοπό να απαντηθούν συγκεκριμένα ερωτήματα που μπορεί να υπάρχουν στο μυαλό μας. Ιδιαίτερα, θα εξεταστεί το ζήτημα των προσφυγικών ροών και συγκεκριμένα τα συναισθήματα του κόσμου, που εκφράζονται μέσα απο τα Tweets, πάνω σε αυτό το ζήτημα. Αυτό θα επιτευχθεί μέσα από κάποιες μεθόδους αρχικά εξόρυξης δεδομένων και έπειτα ανάλυσης του συνόλου δεδομένων.

Στο πρώτο και στο δεύτερο κεφάλαιο γίνεται εισαγωγή στα κοινωνικά δίκτυα και στην εξόρυξη δεδομένων και τις μεθόδους της αντίστοιχα.

Στο τρίτο κεφάλαιο εκτελείται σύμφωνα με τα απαραίτητα βήματα η διαδικασία της εξόρυξης δεδομένων από το Twitter με χρήση της προγραμματιστικής γλώσσας R.

Στο τέταρτο κεφάλαιο απαντώνται κάποια ερωτήματα που υπάρχουν σχετικά με το Twitter και αυτό γίνεται αναλύοντας τις πληροφορίες που έχουν συλλεχθεί μέσα από τη διαδικασία εξόρυξης δεδομένων.

Τέλος, στο πέμπτο και τελευταίο κεφάλαιο γίνεται μια μελέτη πάνω στα tweets, σχετικά με το θέμα των προσφύγων, όσον αφορά τη συχνότητα λέξεων που αυτά περιέχουν και επίσης σχετικά με τα συναισθήματα που μπορεί να δημιουργούνται μέσω αυτών των tweets.

Abstract

People regardless of their origin and the era that live feel the need to express their feelings and opinions about timeliness in political, economic, social and personal level. Already for years, people have the opportunity to express what they believe and feel through the media, newspapers, radio, television. Of course it should be noted that there are additional ways of human expression beyond those mentioned above. In recent years, however, it has developed rapidly, the use of social networks as a means of communication and expression of opinions and emotions. One of these social networks is Twitter, which will be analyzed below. Initially, it will be seen in this thesis, that information collected from Twitter are analyzed in order to answer specific questions that may be exist in our minds. Particularly, will be examined the issue of refugee flows, specifically the feelings of people, expressed through the Tweets, on this matter. This will be achieved, initially, through some data mining methods and then through some methods of dataset analysis.

In the first and second chapter there is an introduction to social networks and in data mining and its methods respectively.

In the third chapter is performed through the necessary steps, the data mining process from Twitter using the programming language R.

In the fourth chapter are answered some questions that exist about Twitter, and this is done by analyzing the information that have collected through the data mining process.

Finally, in the fifth and final chapter there is a study on the tweets, on the subject of refugees, concerning the frequency of words they contain and also about the feelings can be created through these tweets.

Περιεχόμενα

1. Εισαγωγή στα Κοινωνικά Δίκτυα

Τα Κοινωνικά Δίκτυα είναι ένα σύνολο αλληλεπιδράσεων και διαπροσωπικών σχέσεων. Ο όρος στις μέρες μας βέβαια χρησιμοποιείται επίσης για να περιγράψει ιστοσελίδες οι οποίες επιτρέπουν τη διεπαφή ανάμεσα στους χρήστες. Οι πιο γνωστές από αυτές τις ιστοσελίδες είναι το Facebook, Twitter, Instagram και Linkedin τα οποία αποτελούν εικονικές κοινότητες όπου οι χρήστες μπορούν να επικοινωνούν και να αναπτύσσουν διαπροσωπικές σχέσεις μέσα από αυτές. Ένα κοινωνικό δίκτυο αποτελεί μια κοινωνική δομή και είναι μια πλατφόρμα που όπως αναφέρθηκε δημιουργεί κοινωνικές σχέσεις μεταξύ των ανθρώπων με κοινά ενδιαφέροντα ή δραστηριότητες. Βασικές υπηρεσίες που παρέχουν και μάλιστα δωρεάν είναι η δημιουργία προφίλ, το ανέβασμα εικόνων και βίντεο, ο σχολιασμός σε ενέργειες που γίνονται από άλλα μέλη του δικτύου, η άμεση ανταλλαγή μηνυμάτων και πολλά άλλα. Λόγω των δυνατοτήτων των κοινωνικών δικτύων και ειδικότερα της έκφρασης των απόψεων και συναισθημάτων των χρηστών τους τα οποία και θα αναλυθούν ειδικότερα στην παρούσα διπλωματική εργασία, είναι φυσικό να υπάρχει εξάπλωση της χρήσης τους με τα χρόνια και ανά τον κόσμο. Οι άνθρωποι επιλέγουν να εκφραστούν μέσα από αυτά και αυτό έχει ως αποτέλεσμα την ανάγκη για κατανόηση της συμπεριφοράς των χρηστών όταν συνδέονται στα κοινωνικά δίκτυα και για εξαγωγή συμπερασμάτων από τα δεδομένα που συλλέγονται. Οι μελέτες της συμπεριφοράς των χρηστών επιτρέπουν την αξιολόγηση της απόδοσης των υπαρχόντων συστημάτων με κύριο στόχο τον καλύτερο σχεδιασμό των ιστοσελίδων και της τοποθέτησης διαφημίσεων. Τα μοντέλα συμπεριφοράς των χρηστών στα κοινωνικά δίκτυα είναι ζωτικής σημασίας καθώς η εξόρυξη γνώσης από τα δεδομένα τους μπορεί να αποκαλύψει “κρυφές” κοινωνικές τάσεις και να προκύψουν σημαντικά ευρήματα σχετικά με τη συμπεριφορά των ανθρώπων.

Σαν αντιπροσωπευτικό παράδειγμα κοινωνικού δικτύου σε αυτή τη διπλωματική εργασία χρησιμοποιείται το Twitter το οποίο δημιουργήθηκε στις 21 Μαρτίου 2006 από τον Τζακ Ντόρσεϊ και δημοσιεύθηκε τον Ιούλιο του ίδιου χρόνου. Είναι ένας ιστοχώρος κοινωνικής δικτύωσης που επιτρέπει στους χρήστες του να στέλνουν και να διαβάζουν σύντομα μηνύματα (μέχρι 140 χαρακτήρες), που ονομάζονται tweets. Τα μηνύματα μπορούν να αναγνωστούν και από μη συνδεδεμένους χρήστες, αλλά μόνο οι συνδεδεμένοι χρήστες μπορούν να δημοσιεύσουν κείμενα. Το Twitter απασχολεί ένα μοντέλο κοινωνικού δικτύου που ονομάζεται “following”, στο οποίο κάθε χρήστης μπορεί να διαλέξει όποιον θέλει να “ακολουθεί” (follow) χωρίς κάποια έγκριση από εκείνον και επίσης μπορεί να λαμβάνει tweets από αυτόν χωρίς να απαιτείται κάποια άδεια.

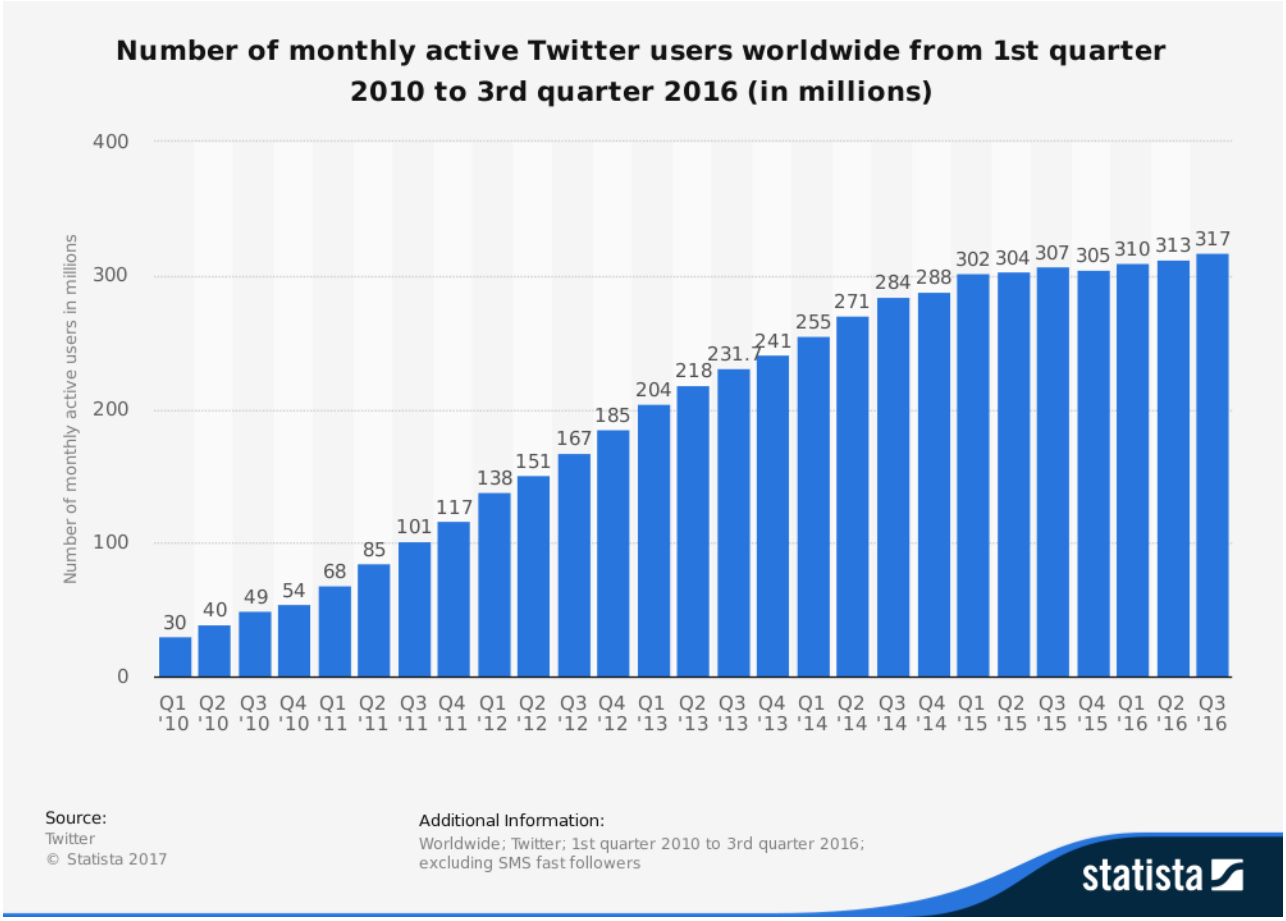
Είναι αξιοσημείωτο ότι τα τελευταία χρόνια όλο και περισσότεροι χρήστες του διαδικτύου κάνουν χρήση των κοινωνικών δικτύων. Ενδεικτικά στις 1 Σεπτεμβρίου του 2016, σύμφωνα με έρευνα που διεξήχθει ο αριθμός των χρηστών των κοινωνικών δικτύων υπολογίζεται στα 3.868.500.000 παγκοσμίως (σύμφωνα με την ιστοσελίδα http://www.statisticbrain.com). Σημαντικό κοινωνικό δίκτυο με βασικό γνώρισμα την έκφραση απόψεων είναι το Twitter, το οποίο μέχρι 1 Σεπτεμβρίου του 2016 σύμφωνα με στατιστικά απο την ίδια την εταιρεία του Twitter αριθμεί 695.750.000 καταγγεγραμμένους χρήστες. Για να πάρουμε μια ιδέα σχετικά με τον όγκο της πληροφορίας που παράγεται καθημερινά από το Twitter αναφέρουμε οτι ο μέσος αριθμός των tweets κάθε ημέρα ανέρχεται στα 58 εκατομμύρια. Εύκολα συμπεραίνει κανείς ότι αυτή η τάση είναι ικανή να δημιουργήσει τεράστιο όγκο πληροφορίας γύρω από αυτά. Μερικοί από τους λόγους που ωθούν τους ανθρώπους στην χρήση των μέσων κοινωνικής δικτύωσης είναι η δυνατότητα έκφρασης προσωπικής άποψης,η επικοινωνία με άλλους χρήστες, η ενημέρωση για θέματα που τους αφορούν και η διαφήμιση. Στα σχήματα 1.1, 1.2 και 1.3 απεικονίζονται στατιστικά στοιχεία από έρευνες σχετικά με το ποσοστό των χρηστών στα κοινωνικά δίκτυα, συγκεκριμένα στο Twitter καθώς και σχεδιάγραμμα σχετικά με την αύξηση των ενεργών χρηστών του Twitter αντίστοιχα.

Largest Socials Networks in the World by Number of Users	Number of Users
Facebook	1,374,000,000
QZone	635,000,000
Google+	347,000,000
LinkedIn	336,000,000
Instagram	302,000,000
Twitter	289,000,000
Tumblr	237,000,000
Sina Weibo	162,000,000
Snapchat	113,000,000
Pinterest	73,500,000

Σχήμα 1.1 Αριθμός χρηστών στα κοινωνικά δίκτυα

Twitter Company Statistics	Data
Total number of registered Twitter users	695,750,000
Total number of active Twitter users	342,000,000
Number of new Twitter users signing up every day	135,000
Number of unique Twitter site visitors every month	195 million
Average number of tweets per day	58 million
Number of Twitter search engine queries every day	2.1 billion
Percent of Twitter users who use their phone to tweet	43 %
Percent of tweets that come from third party applicants	60%
Number of people that are employed by Twitter	2,500
Number of active Twitter users every month	115 million
Percent of Twitters who don't tweet but watch other people tweet	40%
Number of days it takes for 1 billion tweets	5 days
Number of tweets that happen every second	9,100

Σχήμα 1.2 Στατιστικά στοιχεία σχετικά με το Twitter



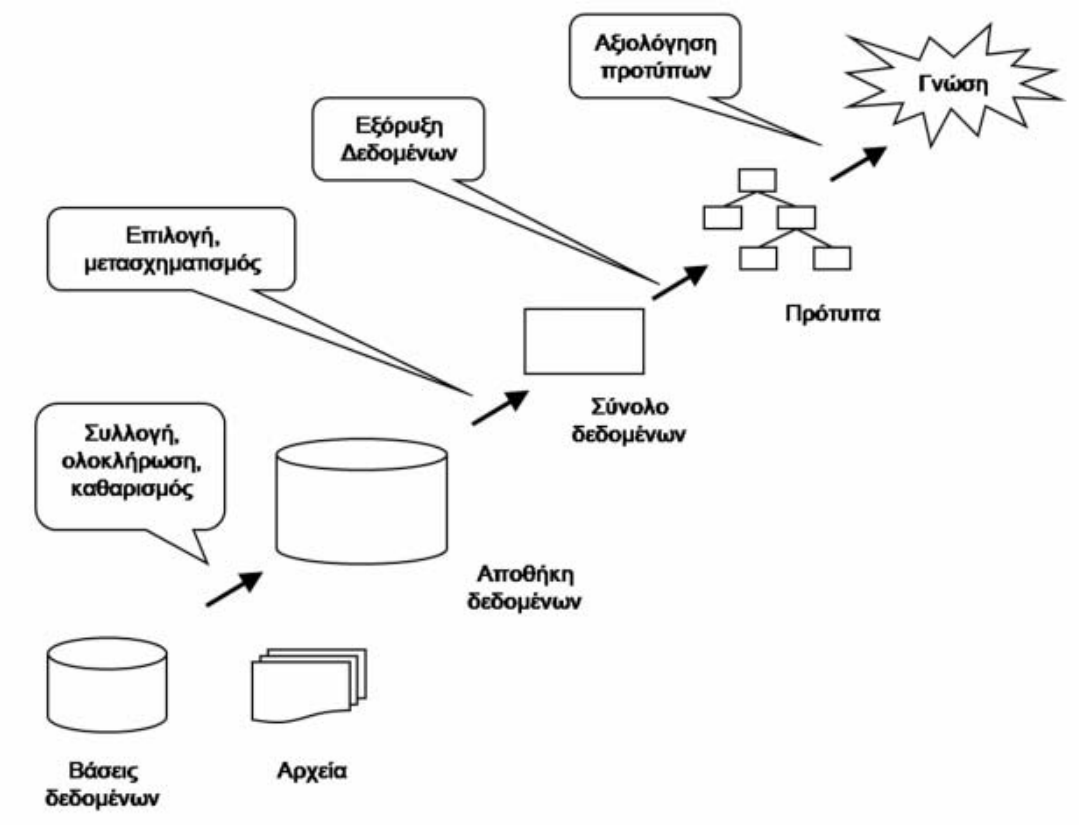
Σχήμα 1.3 Αριθμός των ενεργών χρηστών στο Twitter 2010-2016

Μπορεί να γίνει εύκολα αντιληπτό οτι αυτός ο κατακλυσμός των κοινωνικών δικτύων από δεδομένα διαφόρων μορφών δημιουργεί την ανάγκη εξόρυξης, επεξεργασίας και ανάλυσης αυτών έτσι ώστε να μετατραπούν σε πληροφορίες που θα μας επιτρέψουν να εξάγουμε χρήσιμα συμπεράσματα ανάλογα με το εκάστοτε θέμα ενδιαφέροντος. Το μεγαλύτερο κομμάτι της πληροφορίας που διαχέεται στο διαδίκτυο έχει τη μορφή κειμένου. Για αυτό το λόγο η εξόρυξη κειμένου είναι πολύ σημαντική στις μέρες μας και θα αναλυθεί σε επόμενο κεφάλαιο.

2. Εισαγωγή στην Εξόρυξη Δεδομένων

Με τον όρο Εξόρυξη Δεδομένων ορίζεται η ανακάλυψη πληροφορίας ή προτύπων από βάσεις δεδομένων με χρήση αλγορίθμων συσταδοποίησης ή κατηγοριοποίησης καθώς και των αρχών της στατιστικής, της τεχνητής νοημοσύνης, της μηχανικής μάθησης και των συστημάτων βάσεων δεδομένων. Στόχος της εξόρυξης δεδομένων είναι η πληροφορία που θα εξαχθεί και τα πρότυπα που θα προκύψουν να έχουν δομή κατανοητή προς τον άνθρωπο έτσι ώστε να τον βοηθήσουν να πάρει τις κατάλληλες αποφάσεις. Πιο αναλυτικά η εξόρυξη δεδομένων συνήθως απαρτίζεται από τα εξής στάδια όπως φαίνονται και στο σχήμα 2.1:

(a) Συλλογή, ολοκλήρωση και καθαρισμός των δεδομένων
(b) Η επιλογή των δεδομένων και ο μετασχηματισμός τους
(c) Η εξόρυξη των δεδομένων
(d) Η αξιολόγηση των προτύπων και η ανακάλυψη της γνώσης



Σχήμα 2.1 Στάδια Εξόρυξης Δεδομένων

Στο παραπάνω σχήμα φαίνεται ότι τα αρχικά δεδομένα πρέπει να συλλεχτούν από τις διάφορες πηγές, να ομογενοποιηθούν και να καθαριστούν. Έπειτα για να διεξαχθεί επιτυχώς η εξόρυξη δεδομένων, πρέπει καταρχήν να δημιουργηθεί το κατάλληλο σύνολο δεδομένων (data set). Επιλογή δεδομένων σημαίνει καταρχάς επιλογή των κατάλληλων γνωρισμάτων ή χαρακτηριστικών. Η επιλογή των γνωρισμάτων είναι άμεσα συναρτημένη με την εργασία που εκτελεί ο αναλυτής. Ορισμένα γνωρίσματα μπορεί να είναι χρήσιμα για μια εργασία, ενώ ορισμένα άλλα για κάποια άλλη. Ο αναλυτής αρχικά επιλέγει τα γνωρίσματα τα οποία θεωρεί ότι περιέχουν ουσιαστική πληροφορία που σχετίζεται με την ανάλυση του. Βέβαια η αρχική υποκειμενική επιλογή χαρακτηριστικών δεν είναι αρκετή καθώς στο αρχικό στάδιο, ο αναλυτής αποκλείει τα γνωρίσματα που εμφανώς δεν σχετίζονται με την ανάλυση του. Στο στάδιο αυτό γίνεται και ο μετασχηματισμός των δεδομένων που γίνεται συνήθως για να προσαρμοστούν τα δεδομένα σε απαιτήσεις των μεθόδων ανάλυσης. Το τελικό αποτέλεσμα αυτού του σταδίου είναι ένα σύνολο δεδομένων που θα χρησιμοποιηθεί για την εξαγωγή προτύπων δηλαδή στο στάδιο της εξόρυξης δεδομένων. Ο αναλυτής πρέπει να επιλέξει το είδος εργασίας εξόρυξης δεδομένων που θα εφαρμόσει. Υπάρχουν διάφορες εργασίες εξόρυξης δεδομένων και χωρίζονται σε εργασίες επιβλεπόμενης μάθησης και εργασίες μη επιβλεπόμενης μάθησης. Οι κυριότερες διαδικασίες εξόρυξης δεδομένων είναι οι ακόλουθες:

(a) Η Κατηγοριοποίηση η οποία στοχεύει στην εκτίμηση των τιμών ενός γνωρίσματος-στόχου με ονομαστικές τιμές, το οποίο ορίζει την κατηγορία των αντικειμένων.
(b) Η Παλινδρόμηση η οποία μοιάζει με την Κατηγοριοποίηση αλλά το γνώρισμα-στόχος έχει αριθμητικές τιμές.
(c) Η Ανάλυση Συστάδων αλλιώς Συσταδοποίηση που επιμερίζει ένα σύνολο αντικειμένων σε ομάδες βάσει ομοιότητας και χωρίς την ύπαρξη προκαθορισμένων κατηγοριών.
(d) Η Ανάλυση Κανόνων Συσχέτισης που ανακαλύπτει σχέσεις μεταξύ τιμών των γνωρισμάτων, οι οποίες εμφανίζονται συχνά μαζί.
(e) Τέλος, η Ανίχνευση Ανωμαλιών που εντοπίζει και αναλύει περιπτώσεις οι οποίες αποκλίνουν από το κανονικό ή το συνηθισμένο ή και ακόμα παρουσιάζουν λάθη στα δεδομένα και απαιτούν περαιτέρω έρευνα.

Ο αναλυτής σύμφωνα με τις προτεραιότητές του θα επιλέξει τη μέθοδο ή τις μεθόδους εξόρυξης δεδομένων που θα εφαρμόσει. Τέλος, θα αξιολογήσει τα πρότυπα δεδομένων που προέκυψαν από το προηγούμενο στάδιο. Αν τα αποτελέσματα δεν είναι τα επιθυμητά θα επανέλθει σε προηγούμενα στάδια είτε τροποποιώντας το σύνολο δεδομένων είτε χρησιμοποιώντας διαφορετικές μεθόδους εξόρυξης και θα επαναλάβει τα βήματα. Μετά την ολοκλήρωση των παραπάνω σταδίων έχει ολοκληρωθεί η διαδικασία εξόρυξης και έχουν προκύψει τα συμπεράσματα της ανάλυσης και έπεται η λήψη αποφάσεων.

Στη σύγχρονη εποχή, η γνώση αποτελεί πολύτιμο κεφάλαιο και η εξόρυξη δεδομένων είναι το εργαλείο για την ανάκτησή της. Εκτός από τις μεθόδους Συσταδοποίησης που θα χρησιμοποιηθούν στην παρούσα εργασία, θα αναλυθούν επίσης και άλλες δύο σύγχρονες τεχνικές οι οποίες είναι η Εξόρυξη Κειμένου και η Ανάλυση Συναισθήματος.

2.1. Εξόρυξη Κειμένου

Τον τελευταίο καιρό προσελκύουν το ενδιαφέρον τεχνικές που αφορούν την επεξεργασία δεδομένων κειμένου. Ακαδημαϊκοί ερευνητές και πάροχοι λογισμικού συντονίζουν τις ενέργειες τους και προτείνουν μεθόδους ικανές να αναλύσουν κείμενα, να εξάγουν χρήσιμα συμπεράσματα και να ολοκληρώσουν τα αποτελέσματα τους με αυτά άλλων μεθόδων που επεξεργάζονται δομημένα δεδομένα. Η τάση αυτή οφείλεται στο γεγονός ότι τα δεδομένα αυτού του είδους είναι πολύ περισσότερα. Τα δεδομένα αυτά περιέχουν πολύτιμη πληροφορία και η ανάκτηση τους μπορεί να αποδειχθεί εξαιρετικά χρήσιμη. Το text mining δηλαδή η εξόρυξη κειμένου είναι διαδικασία αυτόματης εξαγωγής γνώσης από διαφορετικούς γραπτούς πόρους καθώς επίσης και η όσο το δυνατόν καλύτερη οργάνωση αυτής της γνώσης για οποιαδήποτε μελλοντική αναφορά. Η εξόρυξη κειμένου είναι μια νέα περιοχή έρευνας η οποία ασχολείται με προβλήματα επεξεργασίας εγγράφων κειμένων και εξαγωγή γνώσης από τα κείμενα επεξεργασίας. Χρησιμοποιεί διάφορες μεθόδους όπως η κατηγοριοποίηση κειμένων, η ταξινόμηση και η ομαδοποίηση κειμένων, η επεξεργασία φυσικής γλώσσας, η δημιουργία περιλήψεων κειμένων, η εύρεση προτύπων συσχέτισης, η αναγνώριση θεμάτων και η εξαγωγή πληροφορίας. Ο κύριος στόχος της εξόρυξης κειμένου είναι να βοηθήσει τους χρήστες να εξάγουν πληροφορίες από μεγάλα κείμενα. Ένα παράδειγμα εξόρυξης δεδομένων κειμένου είναι η χειροκίνητη δρομολόγηση των μηνυμάτων στον κατάλληλο αποδέκτη για παράδειγμα σε μια εταιρεία, η οποία είναι εξαιρετικά αργή. Πολλές φορές η ανάγνωση του θέματος δεν είναι αρκετή και χρειάζεται να γίνει ανάγνωση του σώματος του μηνύματος. Με τη βοήθεια των εργαλείων εξόρυξης κειμένου είναι δυνατόν να γίνει κατανοητό το περιεχόμενο του μηνύματος έτσι ώστε να δρομολογηθεί στον αποδέκτη με αυτόματο τρόπο. Με τη χρήση των τεχνικών εξόρυξης κειμένου είναι δυνατή η ανάλυση τάσεων, η αναγνώριση δηλαδή της μεταβολής των απόψεων ανθρώπων με την πάροδο του χρόνου.

2.2. Αλγόριθμοι Συσταδοποίησης

Συστάδα θεωρούμε μια συλλογή από στοιχεία τα οποία είναι όμοια μεταξύ τους και έχουν διαφορές από στοιχεία που ανήκουν σε άλλες συστάδες. Η ανάλυση σε συστάδες αποσκοπεί στο διαχωρισμό μιας συλλογής από στοιχεία σε υποσύνολα έτσι ώστε να υπάρχει ομοιογένεια μέσα σε ένα υποσύνολο και ανομοιογένεια μεταξύ των στοιχείων που ανήκουν σε διαφορετικά υποσύνολα. Επιπλέον μπορεί να αποσκοπεί στην ιεραρχική οργάνωση των συστάδων με τη διαδοχική ομαδοποίηση αυτών, έτσι ώστε σε κάθε στάδιο της ιεραρχίας, οι συστάδες που ανήκουν στην ίδια ομάδα να είναι πιο όμοιες μεταξύ τους από αυτές που ανήκουν σε άλλη ομάδα. Μια πολύ βασική έννοια για την ανάλυση κατά συστάδες είναι οι έννοιες της απόστασης και της ομοιότητας. Εύκολα διαπιστώνεται πως αυτές οι δύο έννοιες είναι αντίθετες, στοιχεία που είναι όμοια θα έχουν μεγάλη ομοιότητα και μικρή απόσταση. Οι έννοιες αυτές είναι πολύ χρήσιμες καθώς μας επιτρέπουν να μετρήσουμε πόσο μοιάζουν τα στοιχεία μεταξύ τους και επομένως να τα τοποθετήσουμε στην ίδια ομάδα. Η ανάλυση συστάδων πραγματοποιείται με τη χρήση πολυάριθμων αλγορίθμων με τελείως διαφορετικές ιδιότητες μεταξύ τους ως προς τον τρόπο λειτουργίας και το βαθμό απόδοσής τους και επεξεργάζεται συστάδες οι οποίες εννοιολογικά σημαίνουν αποστάσεις μεταξύ των στοιχείων, πυκνές περιοχές με σημεία στο χώρο, ειδικές κατανομές στοιχείων κτλ. Πέραν της προσεκτικής επιλογής ενός αλγόριθμου ακολουθεί και η ρύθμιση ορισμένων παραμέτρων, όπως ο τύπος μέτρησης των αποστάσεων, κάποιο αριθμητικό όριο στοιχείων που πρέπει να έχει μία συστάδα ή ο επιτρεπτός αριθμός των συστάδων τελικής αποδοχής. Υπάρχουν διάφορα μέτρα απόστασης που μπορούν να χρησιμοποιηθούν, όπως η ευκλείδια απόσταση, η απόσταση Manhatan, η απόσταση Chebychev, ο συντελεστής συσχέτισης του Pearson κά. Η επιλογή της απόστασης έχει να κάνει με τη μέθοδο που θα χρησιμοποιήσω αλλά και τον τύπο των δεδομένων μου καθώς και τα δεδομένα. Στην ανάλυση κατά συστάδες υπάρχει ένα μεγάλο πλήθος από αλγόριθμους που έχουν προταθεί και ο καθένας τους βασίζεται σε διαφορετική φιλοσοφία. Σχεδόν όλοι τους δέχονται ένα σύνολο παραμέτρων που μπορεί να είναι το πλήθος των ομάδων, διανύσματα αρχικοποίησης που απαιτούνται από τον αλγόριθμο κάποιες υποθέσεις για την πυκνότητα των διανυσμάτων στο χώρο και άλλες διάφορες παραμέτρους. Η επιστημονική βιβλιογραφία περιλαμβάνει ένα μεγάλο αριθμό μεθόδων ανάλυσης συστάδων. Οι τρεις βασικές κατηγορίες είναι οι εξής:

● Ιεραρχικές μέθοδοι. Οι ιεραρχικές μέθοδοι (hierarchical methods) δημιουργούν μια ιεραρχία από συστάδες. Στο κατώτατο επίπεδο της ιεραρχίας βρίσκονται τα μεμονωμένα στοιχεία. Στο ανώτατο επίπεδο βρίσκεται μια υπερσυστάδα, η οποία περιλαμβάνει όλα τα

στοιχεία. Κάθε ενδιάμεσο επίπεδο ορίζει ένα σύνολο συστάδων. Η ιεραρχία προκύπτει από μια διαδικασία διαδοχικών συγχωνεύσεων ή διασπάσεων συστάδων. Οι σχετικές τεχνικές αντιστοίχως χωρίζονται σε συσσωρευτικές και διαιρετικές.

(a) Οι συσσωρευτικές (agglomerative) μέθοδοι αρχικά θεωρούν κάθε ξεχωριστό στοιχείο ως μια συστάδα. Τα πιο όμοια στοιχεία επιλέγονται και συγχωνεύονται, δημιουργώντας μια νέα συστάδα. Από τις συστάδες που προκύπτουν, επιλέγονται οι πιο όμοιες και συγχωνεύονται. Η διαδικασία επαναλαμβάνεται μέχρι να ενταχθούν όλα τα στοιχεία σε μια ενιαία συστάδα. Οι συσσωρευτικές μέθοδοι έχουν ως αφετηριακό σημείο το κατώτερο επίπεδο της ιεραρχίας των διαδοχικών συγχωνεύσεων, και σταδιακά ανέρχονται τα επίπεδα. Υιοθετούν δηλαδή μια προσέγγιση «από κάτω προς τα επάνω» (bottom up).

(b) Οι διαιρετικές (divisive) μέθοδοι αρχικά θεωρούν όλα τα στοιχεία ως μέλη μιας ενιαίας συστάδας. Η αρχική αυτή συστάδα διαιρείται σε δύο υποομάδες. Η διάσπαση γίνεται με τέτοιον τρόπο, ώστε οι υποομάδες οι οποίες θα προκύψουν θα έχουν τη μεγαλύτερη ανομοιότητα. Η διαδικασία των διαδοχικών διασπάσεων επαναλαμβάνεται μέχρι κάθε στοιχείο να αποτελεί μια ξεχωριστή υποομάδα. Οι διαιρετικές μέθοδοι έχουν αφετηριακό σημείο το ανώτατο επίπεδο της ιεραρχίας και ακολουθούν μια προσέγγιση «από επάνω προς τα κάτω» (top down).

Για την επιλογή των συστάδων δημιουργείται ένας πίνακας ανομοιότητας. Εάν τα δεδομένα περιέχουν N σημεία, τότε ο πίνακας είναι διαστάσεων NXN. Κάθε εγγραφή του πίνακα είναι ένα μέτρο ανομοιότητας ή απόστασης μεταξύ δύο σημείων. Ο πίνακας ανομοιότητας έχει την ακόλουθη μορφή:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \dots & \dots & \dots & 0 & \\ d(N,1) & \dots & \dots & d(N,N-1) & 0 \end{bmatrix}$$

όπου d(x1,x2) είναι η απόσταση μεταξύ των σημείων x1 και x2 . Εφόσον η απόσταση κάθε σημείου από τον εαυτό του είναι μηδενική (d(xi,xi)=0) οι εγγραφές της διαγωνίου από επάνω και αριστερά προς κάτω και δεξιά έχουν μηδενικές τιμές. Επειδή η απόσταση μεταξύ δύο σημείων είναι συμμετρική (d(xi,xj)= d(xj,xi)), η διαγώνιος χωρίζει τον πίνακα σε δύο κατοπτρικά μέρη, οπότε διατηρούνται μόνο οι εγγραφές οι οποίες βρίσκονται κάτω από τη διαγώνιο.

Στην ιεραρχική μέθοδο δημιουργείται μια ιεραρχία, η οποία περιλαμβάνει ένα σύνολο από δυνατές συστάδες. Κάθε επίπεδο της ιεραρχίας περιγράφει ένα συγκεκριμένο τρόπο διαμοιρασμού των στοιχείων σε συστάδες. Αποτελεί αρμοδιότητα του χρήστη να αποφασίσει πιο είναι το κατάλληλο επίπεδο, το οποίο περιγράφει έναν φυσικό τρόπο διαμοιρασμού των στοιχείων, δηλαδή ποιες είναι οι συστάδες, οι οποίες είναι επαρκώς όμοιες μεταξύ τους. Εάν στα δεδομένα υπάρχουν N σημεία, τότε και στις δύο κατηγορίες μεθόδων υπάρχουν N-1 επίπεδα.

● Διαχωριστικές μέθοδοι. Οι διαχωριστικές μέθοδοι (partitioning methods) επιμερίζουν τα στοιχεία σε k συστάδες. Τυπικά το πλήθος των συστάδων προκαθορίζεται από τον χρήστη. Στις μεθόδους αυτής της κατηγορίας εφαρμόζεται μια επαναληπτική διαδικασία, κατά την οποία τα στοιχεία μετακινούνται από μια συστάδα σε μια άλλη. Η ποιότητα της κάθε λύσης ενδεχόμενων συστάδων μετράται με τη βοήθεια ενός κριτηρίου. Σε κάθε επανάληψη και με τη μετακίνηση των σημείων, η τιμή του κριτηρίου μειώνεται. Στόχος είναι να δημιουργηθούν συστάδες, οι οποίες να περιέχουν όμοια στοιχεία, ενώ τα στοιχεία διαφορετικών συστάδων να είναι ανόμοια. Οι διαχωριστικές μέθοδοι δημιουργούν ένα σύνολο συστάδων, σε αντίθεση με τις ιεραρχικές μεθόδους, οι οποίες δημιουργούν μια ιεραρχική δομή διαδοχικών επιπέδων, όπου κάθε επίπεδο ορίζει ένα σύνολο συστάδων. Επίσης, είναι υπολογιστικά λιγότερο ακριβές από τις ιεραρχικές μεθόδους, και για τον λόγο αυτό μπορούν να εφαρμοστούν σε μεγαλύτερα σύνολα δεδομένων. Ο πιο γνωστός αλγόριθμος της διαχωριστικής ανάλυσης συστάδων είναι ο k-Means. Στόχος της είναι να κατανείμει ένα σύνολο στοιχείων σε έναν προκαθορισμένο αριθμό συστάδων, με τρόπο που να αυξάνει την ομοιότητα εντός των συστάδων. Ο αλγόριθμος περιλαμβάνει μια επαναληπτική διαδικασία, όπου σε κάθε επανάληψη υπολογίζεται το κέντρο της συστάδας (centroid). Τα αντικείμενα εντάσσονται στη συστάδα με το πλησιέστερο κέντρο. Αναλυτικότερα, ο αλγόριθμος της μεθόδου k-Means έχει ως ακολούθως:

(1) Διάλεξε τον αριθμό των συστάδων.
(2) Επέλεξε τυχαία k κεντρικά σημεία (συχνά σημεία του συνόλου δεδομένων).
(3) Αντιστοίχισε κάθε σημείο του συνόλου δεδομένων στο κοντινότερο κεντρικό σημείο.
(4) Μετακίνησε το κεντρικό σημείο στο μέσο όλων των σημείων που αντιστοιχούν σε αυτό.
(5) Επανέλαβε τα βήματα 3-4 μέχρι τα κεντρικά σημεία να σταματήσουν να μετακινούνται.

● Μέθοδοι βασισμένες σε μοντέλα. Οι βασισμένες σε μοντέλα μεθόδους (model based methods), θεωρούν ότι κάθε μια από τις συστάδες περιγράφεται από ένα μαθηματικό μοντέλο και εντοπίζουν τα στοιχεία που ανήκουν σε κάθε συστάδα, με στόχο να ικανοποιείται το αντίστοιχο μοντέλο. Μια πολύ διαδεδομένη μέθοδος αυτής της κατηγορίας είναι ένα ειδικός τύπος νευρωνικών δικτύων, που ονομάζονται Αυτοοργανούμενοι Χάρτες (Self Organizing Maps).

2.3. Ανάλυση Συναισθήματος

Η ανάλυση συναισθήματος από κείμενο (sentiment analysis) είναι μια εφαρμογή του data mining που στοχεύει κυρίως στην εξαγωγή της άποψης από κείμενο και στην ταξινόμησή της ως αρνητική ή θετική. Συγκεκριμένα αποσκοπεί στην κατηγοριοποίηση με βάση την συναισθηματική κατάσταση των ανθρώπων (πχ. χαρά, λύπη, θυμός) η οποία ανήκει στον τομέα του sentiment analysis. Η ανάπτυξη του διαδικτύου και η ανταλλαγή τεραστίων ποσοτήτων πληροφορίας μεταξύ των χρηστών σε όλο τον κόσμο καθιστά επιτακτική τη μελέτη και ανάλυση αλγορίθμων που συμπεραίνουν αυτοματοποιημένα τα συναισθήματα, τις επιθυμίες και τις πεποιθήσεις των ανθρώπων με βάση το κείμενο. Καθημερινά, αναπαράγονται κριτικές και απόψεις για διάφορα πολιτικά, κοινωνικά, αθλητικά ή άλλα γεγονότα, προϊόντα, ταινίες κτλ. με αποτέλεσμα ο όγκος της πληροφορίας που αναπτύσσεται να είναι αδύνατο να επεξεργαστεί μόνο από τον άνθρωπο χωρίς τη βοήθεια του υπολογιστή. Έτσι γίνεται εύκολα αντιληπτό γιατί η επιστημονική αλλά και η βιομηχανική κοινότητα έχει δείξει έντονο ενδιαφέρον στον τομέα αυτό. Καινοτόμες επιχειρήσεις αχολούνται με την εξόρυξη γνώσης μέσα από τις αξιολογήσεις χρηστών σε ηλεκτρονικά καταστήματα και κοινωνικά δίκτυα με βάση την ανάλυση συναισθήματος. Ένα πρόβλημα που προσπαθεί να επιλύσει η ανάλυση συναισθήματος είναι το εξής, ένας υπολογιστής δεν είναι απαραίτητο να καταλαβαίνει πλήρως την σημασιολογία της κάθε λέξης αλλά θα πρέπει να εντοπίζει τη συνολική στάση του ανθρώπου που γράφει ένα σχόλιο ή μία κριτική. Οι άνθρωποι εκφράζουν τη γνώμη τους με κριτικές και σχόλια σε ιστοθέσεις όπως το Amazon, σε tweets, σε blogs, σε ιστοτόπους κοινωνικής δικτύωσης και σε emails. Το υλικό αυτό είναι τεράστιο σε όγκο, καταγράφει απόψεις εκατομμυρίων ή και δισεκατομμυρίων ανθρώπων και ανανεώνεται καθημερινά. Η αξιοποίηση του υλικού αυτού δίνει πρωτόγνωρες δυνατότητες. Όπως προκύπτει από την παραπάνω ανάλυση, μία από τις πιο κοινές εφαρμογές ανάλυσης συναισθήματος είναι η παρακολούθηση των συμπεριφορών και των συναισθημάτων στο διαδίκτυο, συγκεκριμένα όσον αφορά τα προϊόντα, τις υπηρεσίες ή ακόμη και τους ανθρώπους, και να καθοριστεί αν αυτά ταξινομούνται ως θετικά ή αρνητικά.

3. Εξόρυξη στο Twitter με χρήση της R

Όπως αναφέρθηκε σε προηγούμενο κεφάλαιο σαν αντιπροσωπευτικό παράδειγμα κοινωνικού δικτύου σε αυτή τη διπλωματική εργασία χρησιμοποιείται το Twitter το οποίο δημιουργήθηκε στις 21 Μαρτίου 2006 από τον Τζακ Ντόρσεϊ και δημοσιεύθηκε τον Ιούλιο του ίδιου χρόνου. Είναι ένας ιστοχώρος κοινωνικής δικτύωσης που επιτρέπει στους χρήστες του να στέλνουν και να διαβάζουν σύντομα μηνύματα (μέχρι 140 χαρακτήρες), που ονομάζονται tweets. Τα μηνύματα μπορούν να αναγνωστούν και από μη συνδεδεμένους χρήστες, αλλά μόνο οι συνδεδεμένοι χρήστες μπορούν να δημοσιεύσουν κείμενα. Το Twitter απασχολεί ένα μοντέλο κοινωνικού δικτύου που ονομάζεται “following”, στο οποίο κάθε χρήστης μπορεί να διαλέξει όποιον θέλει να “ακολουθεί” (follow) χωρίς κάποια έγκριση από εκείνον και επίσης μπορεί να λαμβάνει tweets από αυτόν χωρίς να απαιτείται κάποια άδεια. Ο περιορισμός σχετικά με των αριθμό των χαρακτήρων είναι που οδηγεί τους χρήστες να περάσουν το μήνυμα τους σε όσο το δυνατόν πιο συμπυκνωμένη μορφή. Σε αυτά τα κείμενα του Twitter συνηθίζεται να αναφέρεται κάποιο νέο, μια είδηση, κάτι που συνέβη στο χρήστη ή κάτι που αυτός σκέφτηκε, γενικώς μια πληροφορία. Πέρα των χαρακτήρων το tweet μπορεί να περιλαμβάνει φωτογραφίες, βίντεο, συνδέσμους και hashtags τα οποία έχουν ως σκοπό την κατηγοριοποίηση συζητήσεων και δημοσιεύσεων. Ένα hashtag, δηλαδή, μπορεί να προσδώσει νόημα σε μία δημοσίευση η οποία μπορεί υπό άλλες συνθήκες να μην είχε νόημα, διότι δείχνει σε τι αναφέρεται. Η βασική ιδέα του Twitter ως Κοινωνικό δίκτυο είναι ότι ένας χρήστης(user) ακολουθεί κάποιους άλλους χρήστες για τους οποίους ενδιαφέρεται να δεί τι θα πουν και τι θα κάνουν. Αντίστοιχα οι άλλοι χρήστες πού ενδιαφέρονται για αυτόν μπορούν με τη σειρά τους να τον ακολουθήσουν. Όσοι χρήστες έχουν επιλέξει να ακολουθήσουν κάποιον χρήστη θα ενημερωθούν για το εκάστοτε κείμενο που αυτός θα δημοσιεύσει και θα μπορούν να το διαβάσουν, να απαντήσουν (reply) σε αυτό και να το αναδημοσιεύσουν. Υπολογίζεται πως κάθε μέρα δημοσιεύονται περισσότερα από 58 εκατομμύρια tweets.

Το περιορισμένο μέγεθος των tweets περιορίζει το φάσμα χρήσης τους για επιστημονικούς σκοπούς. Παρόλους τους περιορισμούς, πολλοί ερευνητές κάνουν χρήση εργαλείων εξόρυξης κειμένου για ανάλυση των δημοσιεύσεων που γίνονται στο Twitter. Αυτό συμβαίνει για πολλούς λόγους, με κάποιους εκ των οποίων να φαίνονται παρακάτω:

(a) Το Twitter είναι εξαιρετικά δημοφιλής πλατφόρμα για τα μέσα ενημέρωσης γι’αυτό και παρέχει περισσότερο χώρο για έρευνα.
(b) Με το Twitter είναι εύκολο να ακολουθήσεις τη ροή μιας συζήτησης.

(c) Το Twitter κάνει χρήση hashtags που κάνουν πιο εύκολη τη συλλογή, ταξινόμηση, και την επέκταση των αναζητήσεων κατά τη συλλογή των δεδομένων.
(d) Τα επιθυμητά δεδομένα μπορούν εύκολα να ανακτηθούν αφου σημαντικά συμβάντα, ειδήσεις και εκδηλώσεις στο Twitter τείνουν να επικεντρώνονται γύρω από κάποιο hashtag.
(e) Τα APIs του Twitter είναι πιο ανοιχτά και προσβάσιμα σε σύγκριση με τα αντίστοιχα που παρέχουν άλλες πλατφόρμες κοινωνικών μέσων μαζικής ενημέρωσης, γεγονός που καθιστά το Twitter ευνοϊκότερη επιλογή για τους προγραμματιστές που επιζητούν πρόσβαση σε δεδομένα. Αυτό αυξάνει, κατά συνέπεια, και τη διαθεσιμότητα των εργαλείων για τους ερευνητές και δημιουργεί μια κοινότητα επικοινωνίας που διευκολύνει κατά πολύ το έργο τους.
(f) Στο Twitter πολλές φορές συγκεντρώνονται ποικίλες απόψεις πάνω σε κάποιο γεγονός της επικαιρότητας, γεγονός που κάνει την ομαδοποίηση τους σε θέματα συζήτησης ιδανική ώστε ο τελικός χρήστης να δεί το περιεχόμενο που συμπίπτει περισσότερο με τα δικά του ενδιαφέροντα ή προτιμήσεις.

Το Twitter παρόλο που δεν επιβάλλει κάποιο όριο ως προς το πλήθος των δεδομένων που παρέχει στους προγραμματιστές μέσω των APIs του, θέτει περιορισμούς σχετικά με τη λήψη των δεδομένων αυτών εντός σύντομων χρονικών διαστημάτων. Οι περιορισμοί αυτοί εφαρμόζονται ανα λογαριασμό χρήστη. Για παράδειγμα τα ερωτήματα αναζήτησης (search) δεν μπορούν να ξεπερνούν τα 180 σε χρονικό παράθυρο 15 λεπτών.

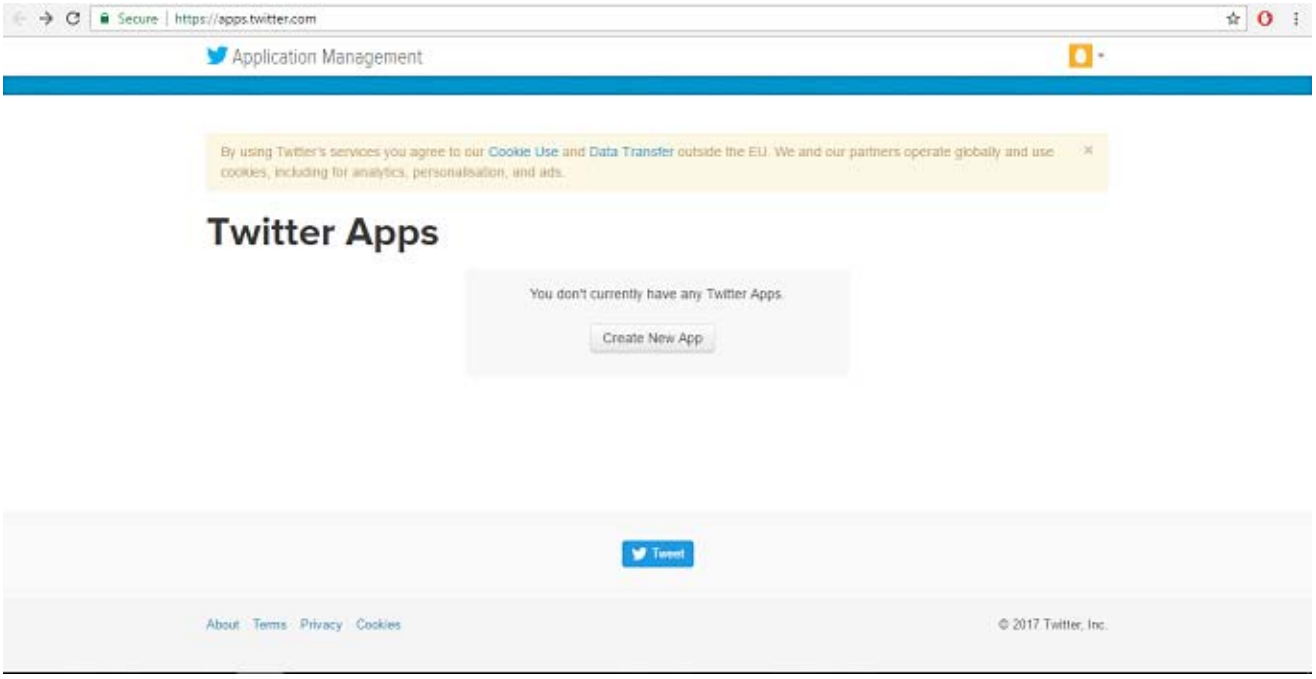
Στην παρούσα εργασία θα χρησιμοποιήσουμε ένα Twitter API για να εξάγουμε δεδομένα από αυτό. Τα κυριότερα βήματα που θα πραγματοποιηθούν στην προσπάθεια για την εξόρυξη δεδομένων από το Twitter είναι τα εξής:

(1) Εξαγωγή δεδομένων από το Twitter
(2) Καθαρισμός και επεξεργασία των δεδομένων
(3) Εξόρυξη κειμένου
(4) Συσταδοποίηση
(5) Ανάλυση συναισθήματος

Για την εξόρυξη των δεδομένων θα χρησιμοποιηθεί η γλώσσα προγραμματισμού R, μια γλώσσα που έχει φτιαχτεί ειδικά για εφαρμογές στη Στατιστική και στην επεξεργασία δεδομένων και που χρησιμοποιείται ευρύτατα ειδικά σε ακαδημαϊκό περιβάλλον. Η γλώσσα R είναι δωρεάν software και τρέχει σε όλα τα ευρέως χρησιμοποιούμενα λειτουργικά συστήματα για desktop υπολογιστές (Windows, Mac OS, Linux). Επίσης χρησιμοποιείται Γραφικό περιβάλλον χρήσης (GUI) της R 3.3.2.

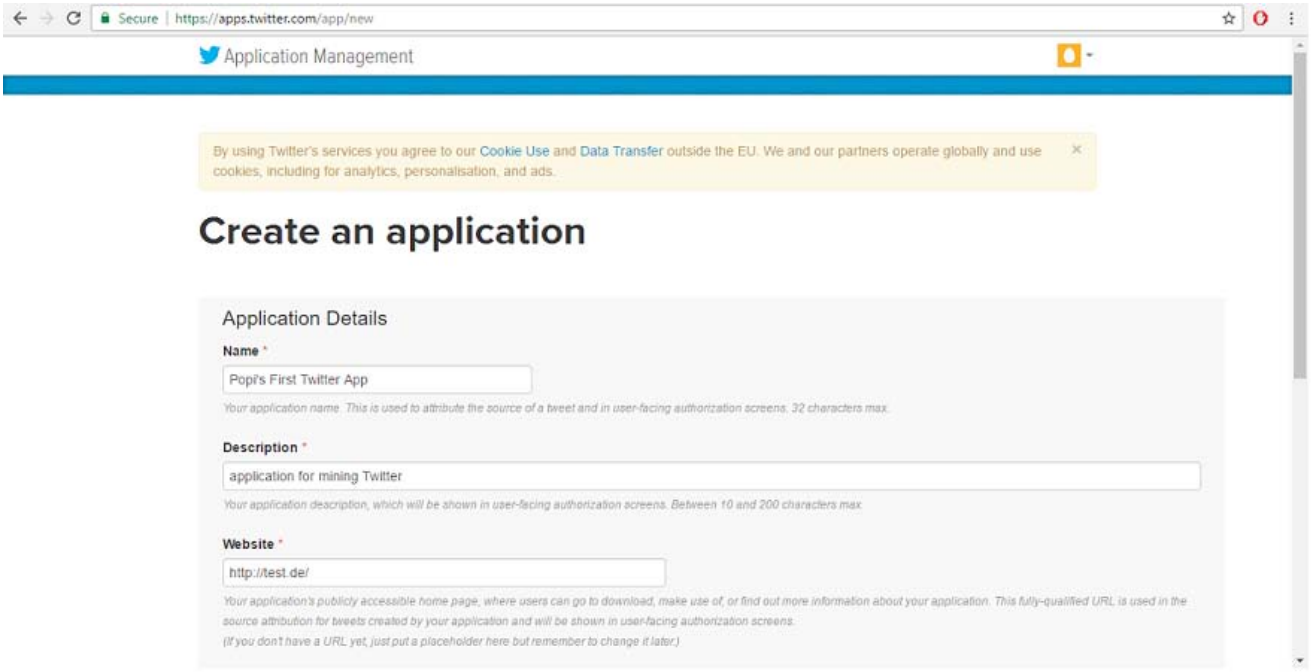
Για να ξεκινήσει η διαδικασία της εξόρυξης δεδομένων από το Twitter θα πρέπει πρώτα να ακολουθηθεί μια διαδικασία εξουσιοδότησης για χρήση του API αναζήτησης, καθώς το Twitter πλέον απαιτεί εξουσιοδότηση. Αφού ολοκληρωθούν τα βήματα για τη διαδικασία εξουσιοδότησης θα είναι δυνατόν να πραγματοποιηθεί οποιαδήποτε αναζήτηση. Παρακάτω παρουσιάζονται αναλυτικά τα βήματα για την διαδικασία εξουσιοδότησης.

Βήμα 1ο: Είσοδος στο διαθέσιμο λογαριασμό στη σελίδα https://apps.twitter.com/. Αν υπάρχει λογαριασμός στο Twitter μπορεί να χρησιμοποιηθεί αν όχι πρέπει να δημιουργηθεί λογαριασμός στο Twitter έτσι ώστε να είναι δυνατόν να δημιουργηθεί εφαρμογή. Θα πρέπει να δείχνει αυτό που απεικονίζεται στο σχήμα 3.1

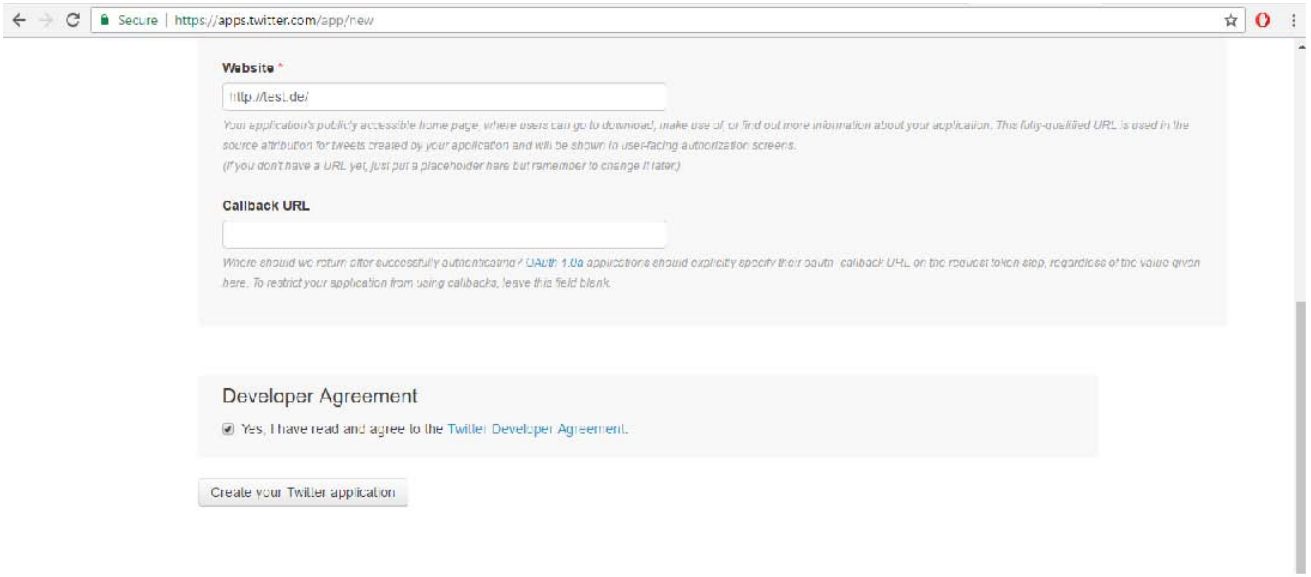


Σχήμα 3.1 Είσοδος στο Twitter Apps

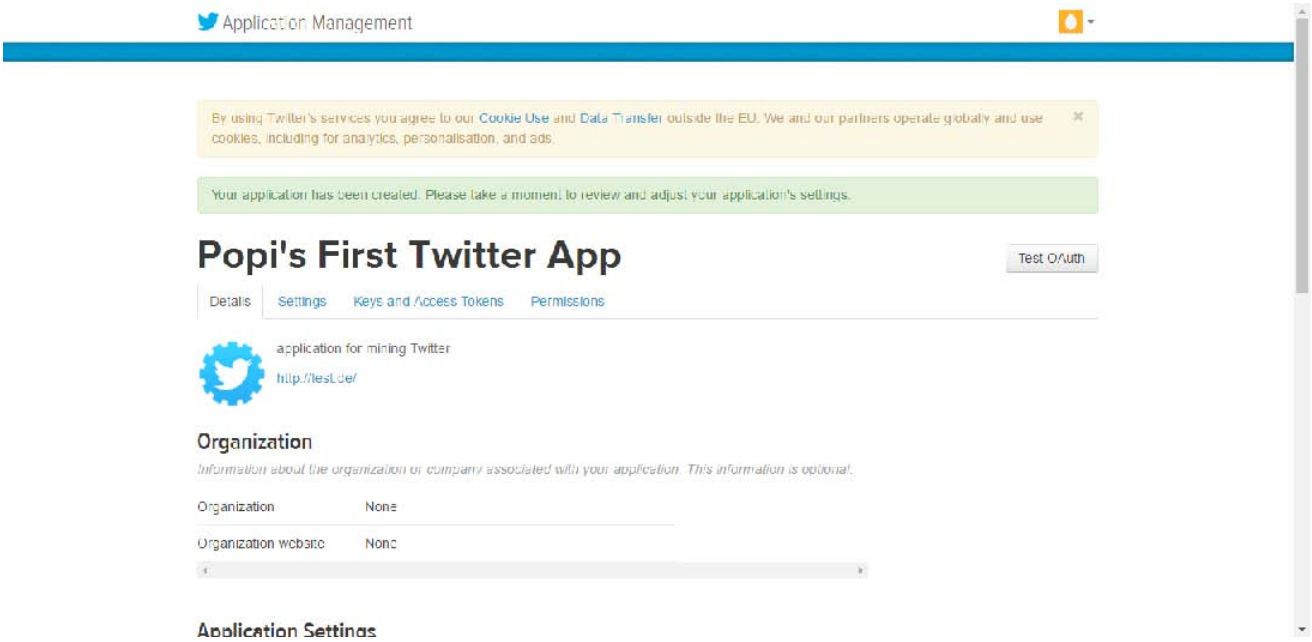
Βήμα 2ο: Επιλογή του κουμπιού Create New App που οδηγεί στην παρακάτω φόρμα σύμφωνα με τα σχήματα 3.2 και 3.3. Στην φόρμα πρέπει να συμπληρωθούν τα εξής: όνομα για την εφαρμογή και περιγραφή της εφαρμογής. Έπειτα στο πεδίο Website πρέπει να συμπληρωθεί κάποιο έγκυρο url, οπότε αν δεν υπάρχει κάποιο μπορεί να χρησιμοποιηθεί το http://test.de/. Τέλος, στο πεδίο Callback URL συμπληρώνεται το http://127.0.0.1:1410 και αφού γίνει επιλογή στο Developer Agreement πρέπει να πατηθεί το κουμπί “Create your Twitter application” που θα οδηγήσει με τη σειρά του στη σελίδα όπως φαίνεται στο σχήμα 3.4.



Σχήμα 3.2 Φόρμα συμπλήρωσης για την εφαρμογή



Σχήμα 3.3 Φόρμα συμπλήρωσης για την εφαρμογή



Σχήμα 3.4 Σελίδα διαχείρισης της εφαρμογής

Στη σελίδα αυτή και συγκεκριμένα στην καρτέλα Keys and Access Tokens υπάρχουν δύο πεδία που ονομάζονται Consumer Key και Consumer Secret τα οποία περιέχουν δύο κωδικούς. Αυτά χρειάζονται μετά στο αρχείο R που θα χρησιμοποιηθεί για να εξουσιοδοτηθεί η χρήση του Twitter API από την R. Παρακάτω στην ίδια σελίδα υπάρχει μια επιλογή “Create my access token” που θα δημιουργήσει δύο κωδικούς που ονομάζονται Access Token και Access Token Secret και που θα χρειαστούν σε επόμενο βήμα για να εξουσιοδοτήσουν την R έτσι ώστε να έχει πρόσβαση στο Twitter API.

Βήμα 3ο: Πρέπει να φορτωθούν οι εξής βιβλιοθήκες:

(a) twitteR
(b) ROAuth
(c) RCurl

Παρακάτω φαίνεται ο κώδικας που χρησιμοποιείται. Στο σημείο αυτό θα πρέπει να σημειωθεί ότι κατά τη διάρκεια του τρεξίματος του κώδικα μπορεί να χρειαστεί να γίνει εγκατάσταση και σε κάποια packages που μπορεί να χρειάζονται και να μην είναι εγκατεστημένα και πάντα θα πρέπει οι εκδόσεις που χρησιμοποιούμε σε όλα να συμφωνούν.

```
# authorisation
if (!require('pacman')) install.packages('pacman')
pacman::p_load(twitteR, ROAuth, RCurl)

api_key = 'YsPu7vYhIzmrJzd9KSa8Kwxtg'
api_secret = 'mblvbqYFdxzGkf9aDgM2PsoGe0nFe3tdql1ub2y8XukDMwnKeE'
access_token = '832621652286377985-NLYclrXYAJN710mdwA2K48kdLUeOBss'
access_token_secret = 'h3ZHEhgKPSCoRYlpDGwWAzuWu1WK83KAkQF21MLUNIlZF'

# Set SSL certs globally
options(RCurlOptions = list(cainfo = system.file('CurlSSL', 'cacert.pem', package = 'RCurl')))

# set up the URLs
reqURL = 'https://api.twitter.com/oauth/request_token'
accessURL = 'https://api.twitter.com/oauth/access_token'
authURL = 'https://api.twitter.com/oauth/authorize'

twitCred = OAuthFactory$new(consumerKey = api_key, consumerSecret = api_secret,
requestURL = reqURL, accessURL = accessURL, authURL = authURL)

twitCred$handshake(cainfo = system.file('CurlSSL', 'cacert.pem', package = 'Rcurl'))
```

Βήμα 4ο: Εγκαθίσταται το sentiment package που θα χρειαστεί.

```
if (!require('pacman')) install.packages('pacman&')
pacman::p_load(devtools, installr)
install.Rtools()
install_url('http://cran.r-project.org/src/contrib/Archive/Rstem/Rstem_0.4-1.tar.gz')
install_url('http://cran.r-project.org/src/contrib/Archive/sentiment/sentiment_0.2.tar.gz')
```

Βήμα 5ο: Ρύθμιση των παραμέτρων που χρειάζονται για την εξουσιοδότηση σχετικά με τη δημιουργία script αρχείων της R.

if (!require('pacman')) install.packages('pacman')
pacman::p_load(twitteR, sentiment, plyr, ggplot2, wordcloud, RColorBrewer, httpuv, RCurl, base64enc)

options(RCurlOptions = list(cainfo = system.file('CurlSSL', 'cacert.pem', package = 'RCurl')))

api_key = 'YsPu7vYhIzmrJzd9KSa8Kwxtg'
api_secret = 'mblvbqYFdxzGkf9aDgM2PsoGe0nFe3tdql1ub2y8XukDMwnKeE'
access_token = '832621652286377985-NLYclrXYAJN710mdwA2K48kdLUeOBss'
access_token_secret = 'h3ZHEhgKPSCoRYlpDGwWAzuWu1WK83KAkQF21MLUNIlZF'

setup_twitter_oauth(api_key,api_secret,access_token,access_token_secret)

Μετά την ολοκλήρωση όλων των απαραίτητων ενεργειών είναι όλα έτοιμα για το επόμενο βήμα που είναι η εξαγωγή των δεδομένων από το Twitter.

3.1. Απόκτηση δεδομένων

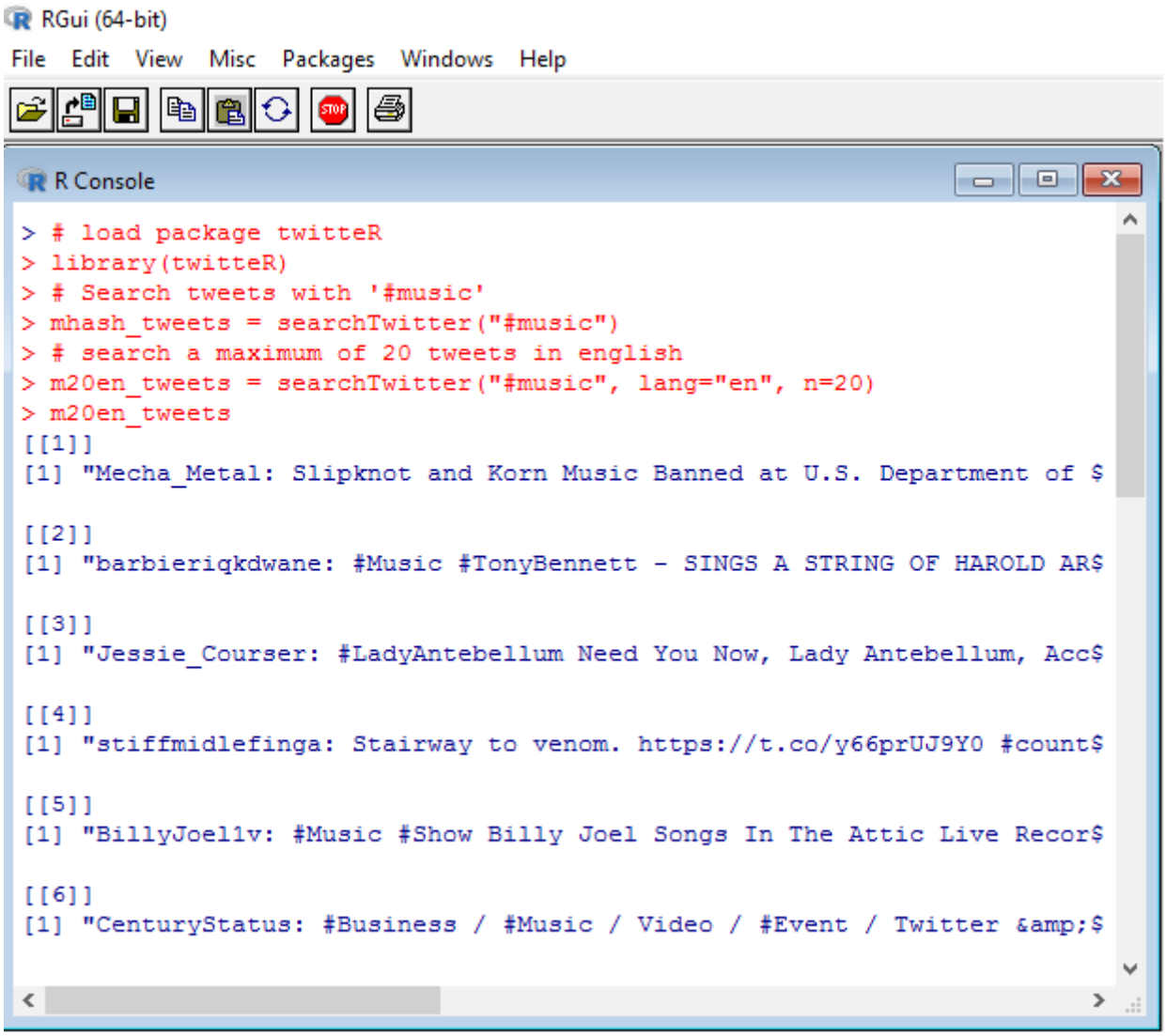
Στο στάδιο αυτό μπορεί να γίνει αναζήτηση στα δεδομένα του Twitter. Ο πιο απλός και γρήγορος τρόπος για να εξαχθούν τα δεδομένα, ο οποίος θα χρησιμοποιηθεί στην παρούσα εργασία, είναι με τη χρήση των λειτουργιών του πακέτου της R “twitteR”. Ένας άλλος τρόπος είναι μέσω του πακέτου “XML” της R.

Κάποια παραδείγματα αναζήτησης που μπορούν να υλοποιηθούν με τη χρήση του twitteR είναι τα εξής:

1. δημόσια tweets
2. θέματα μεγάλου ενδιαφέροντος
3. tweets που περιέχουν ένα συγκερκιμένο hashtag
4. tweets που περιέχουν μια συγκεκριμένη λέξη
5. tweets από ένα συγκεκριμένο χρήστη

Στη συνέχεια παρατίθενται κάποια παραδείγματα κώδικα που εξάγουν δεδομένα με διάφορους από τους τρόπους αναζήτησης που προαναφέρθηκαν. Αξίζει να σημειωθεί ότι στους τρόπους αναζήτησης τα αποτελέσματα μπορούν να περιοριστούν σύμφωνα με τον αριθμό τους, με τη γλώσσα που είναι γραμμένα, με την ημερομηνία που καταχωρήθηκαν ακόμα και με τη γεωγραφική θέση την ώρα που αναρτήθηκαν.

Παράδειγμα 1ο: Λήψη των tweets, με χρήση της searchTwitter, που περιέχουν ένα συγκεκριμένο hashtag(#music) με χρήση των εξής περιορισμών, του αριθμού τους(20) και της γλώσσας(αγγλικά).



Μέρος των αποτελεσμάτων για την αναζήτηση με βάση το hashtag #music φαίνεται κι εδώ:

[[1]]
[1] "Mecha_Metal: Slipknot and Korn Music Banned at U.S. Department of Defense Command Post https://t.co/3uoVRBRCYQ #metal #music"

[[2]]
[1] "barbieriqkdwane: #Music #TonyBennett - SINGS A STRING OF HAROLD ARLEN - COLUMBIA LP - 6 EYED LOGO https://t.co/0xYFr2Ff3R #onsale… https://t.co/HV5o9vEDif"

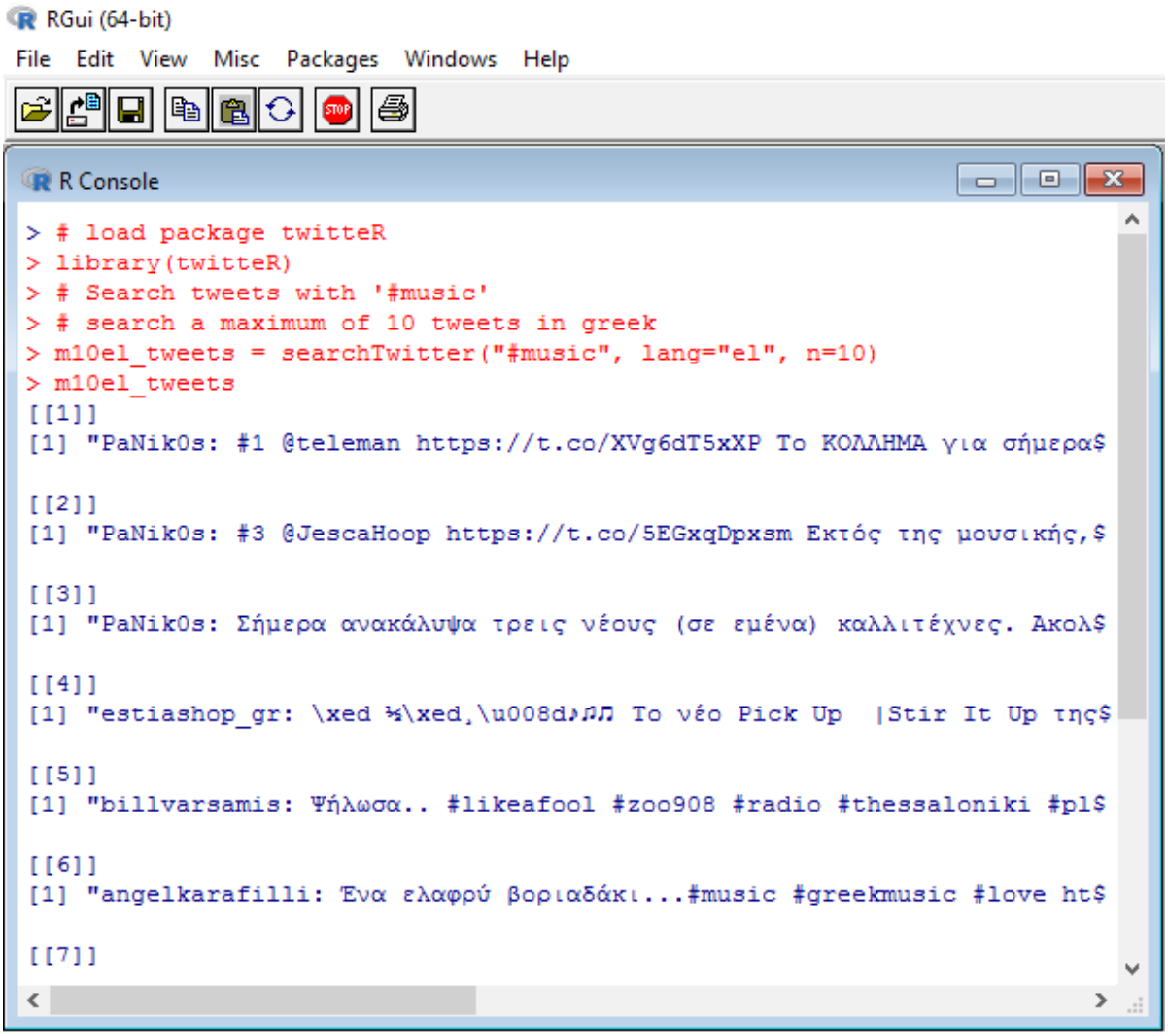
[[3]]
[1] "Jessie_Courser: #LadyAntebellum Need You Now, Lady Antebellum, Acceptable https://t.co/zc7orGfxhh #CountryMusic #Music"

[[4]]
[1] "stiffmidlefinga: Stairway to venom. https://t.co/y66prUJ9Y0 #counterculture #hippy #poetry #music #art #viral"

[[5]]
[1] "BillyJoel1v: #Music #Show Billy Joel Songs In The Attic Live Recording LP Vinyl Record 1981 Gatefold_Photo… https://t.co/5qVTcBGsoA"

[[6]]
[1] "CenturyStatus: #Business / #Music / Video / #Event / Twitter & Website Promotion https://t.co/2Or72S6jB7"

Παράδειγμα 2ο: Λήψη των tweets, με χρήση της searchTwitter, που περιέχουν ένα συγκεκριμένο hashtag(#music) με χρήση των εξής περιορισμών, του αριθμού τους(10) και της γλώσσας(ελληνικά).



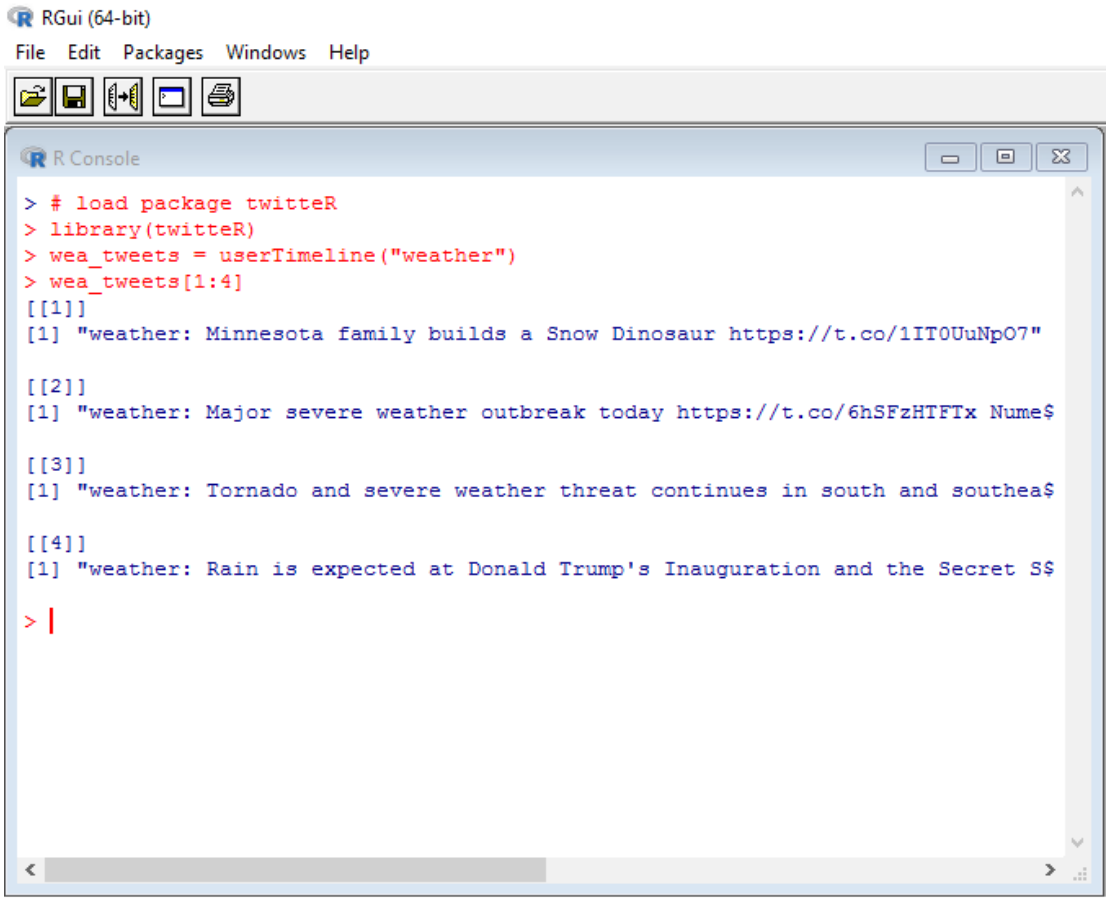
Μέρος των αποτελεσμάτων για την αναζήτηση με βάση το hashtag #music φαίνεται κι εδώ:

[[1]]
[1] "PaNik0s: #1 @teleman https://t.co/XVg6dT5xXP Το ΚΟΛΛΗΜΑ για σήμερα! Όταν το ανακάλυψα, το άκουσα 7-8 φορές συνεχόμενα #music"

[[2]]
[1] "PaNik0s: #3 @JescaHoop https://t.co/5EGxqDpxsm Εκτός της μουσικής, πολύ ωραίο video clip #music"

[[3]]
[1] "PaNik0s: Σήμερα ανακάλυψα τρεις νέους (σε εμένα) καλλιτέχνες. Ακολουθεί δείγμα από την δουλειά τους, με αντίστροφη σειρά κολλήματος: #music"

Παράδειγμα 3ο: Λήψη των tweets, με χρήση της userTimeline(), που περιέχουν μια συγκεκριμένη λέξη το “weather”. Μετά τον κώδικα παρατίθενται κάποια αποτελέσματα.



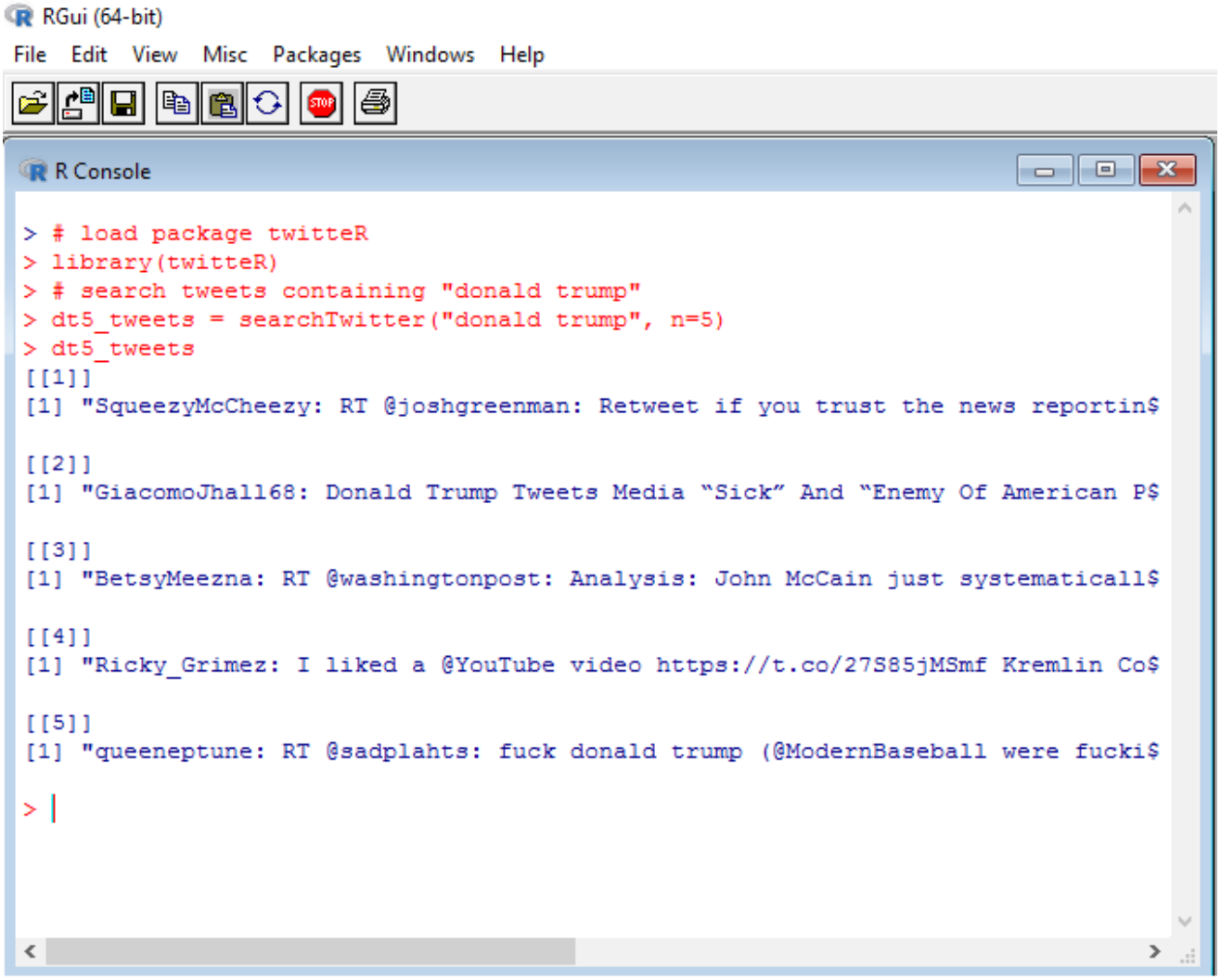
[[1]]
[1] "weather: Minnesota family builds a Snow Dinosaur https://t.co/1IT0UuNpO7"

[[2]]
[1] "weather: Major severe weather outbreak today https://t.co/6hSFzHTFTx Numerous tornado warnings already in Alabama and Florida https://t.co/WdNEVAW0z3"

[[3]]
[1] "weather: Tornado and severe weather threat continues in south and southeast tonight and Sunday https://t.co/MlNzwbBTSQ"

[[4]]
[1] "weather: Rain is expected at Donald Trump's Inauguration and the Secret Service has banned umbrellas https://t.co/XYBbRkM4l3 https://t.co/KBGcHo56oR"

Παράδειγμα 4ο: Λήψη των tweets, με χρήση της searchTwitter, που περιέχει μία συγκεκριμένη λέξη(donald trump) με περιορισμό ως προς τον αριθμό τους(5). Μετά τον κώδικα παρατίθενται κάποια αποτελέσματα.



[[1]]
[1] "SqueezyMcCheezy: RT @joshgreenman: Retweet if you trust the news reporting in the New York Times more than you trust the words out of Donald Trump's mouth…."

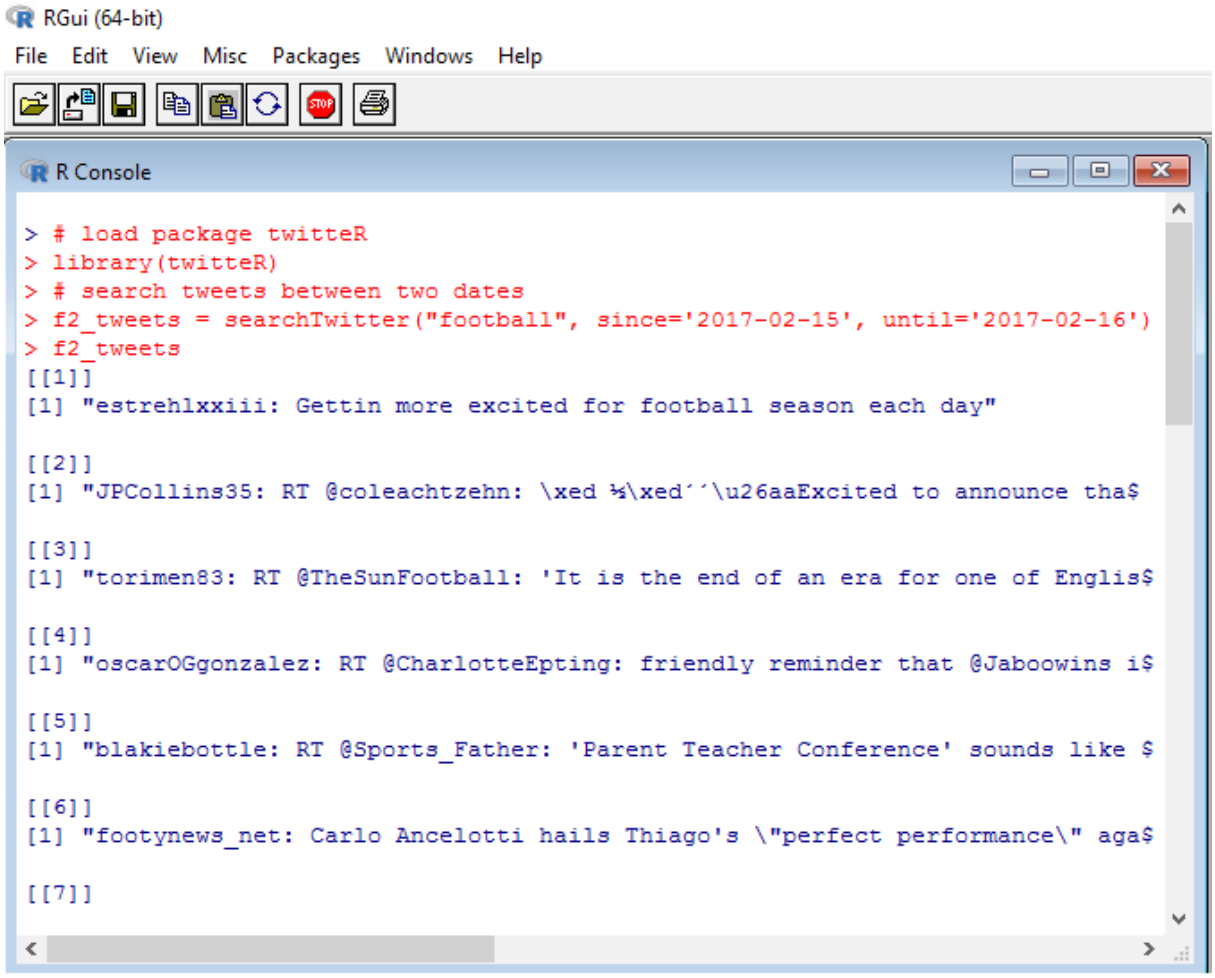
[[2]]
[1] "GiacomoJhall68: Donald Trump Tweets Media “Sick” And “Enemy Of American People”(A MADMAN IS TERRORIZING AMERICAN PEOPLE) | Deadline https://t.co/gonZdTOImX"

[[3]]
[1] "BetsyMeezna: RT @washingtonpost: Analysis: John McCain just systematically dismantled Donald Trump’s entire worldview https://t.co/h4S5zED3GU"

[[4]]
[1] "Ricky_Grimez: I liked a @YouTube video https://t.co/27S85jMSmf Kremlin Cooling Its Enthusiasm For President Donald Trump | Andrea Mitchell | MSNBC"

[[5]]
[1] "queeneptune: RT @sadplahts: fuck donald trump (@ModernBaseball were fuckin sick what an ace night) https://t.co/VILKXtgEul"

Παράδειγμα 5ο: Λήψη των tweets, με χρήση της searchTwitter, που περιέχει μία συγκεκριμένη λέξη(football) ανάμεσα σε δύο ημερομηνίες. Μετά τον κώδικα παρατίθενται κάποια αποτελέσματα.



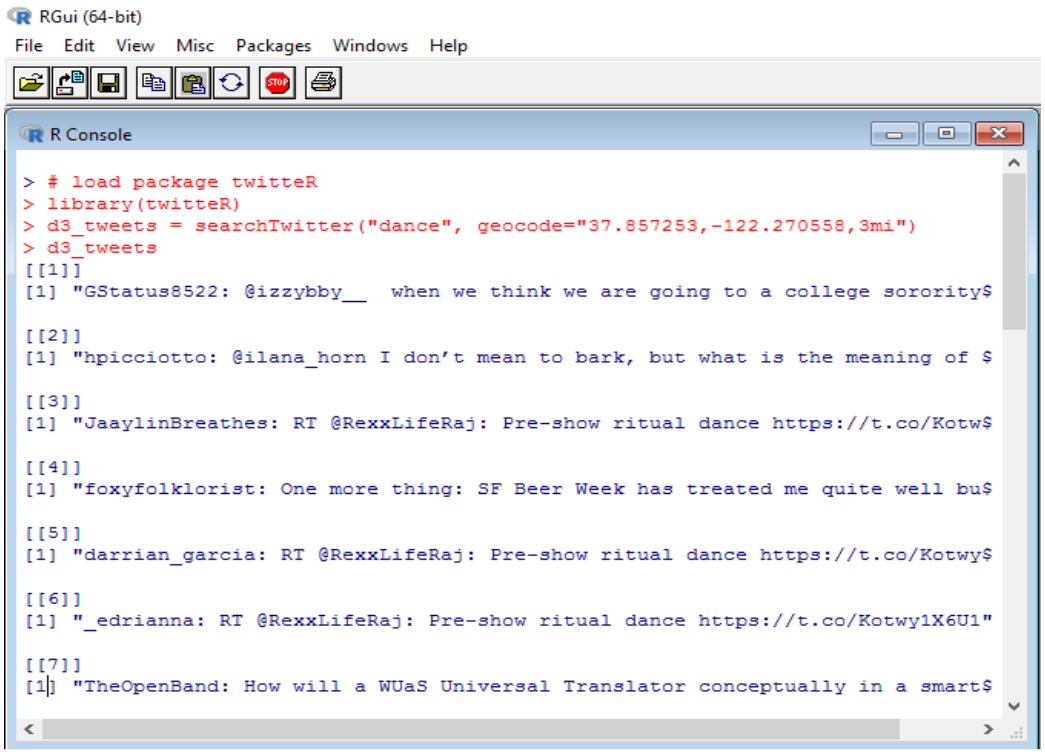
[[1]]
[1] "estrehlxxiii: Gettin more excited for football season each day"

[[2]]
[1] "JPCollins35: RT @coleachtzehn: \xed ½\xed´´\u26aaExcited to announce that I will be playing football at Hastings College in Nebraska next year! #CodeRed \u26aa\xed ½\xed´´ https://t.…"

[[3]]
[1] "torimen83: RT @TheSunFootball: 'It is the end of an era for one of English football's historic clubs' - @neilashton_ on Arsenal https://t.co/saJO2XybH5"

[[4]]
[1] "oscarOGgonzalez: RT @CharlotteEpting: friendly reminder that @Jaboowins is still a serial rapist and still playing football without any consequences\xed ½\xed±\u008d\xed ¼\xed¿»"

Παράδειγμα 6ο: Λήψη των tweets, με χρήση της searchTwitter, που περιέχει μία συγκεκριμένη λέξη(dance) σε συγκεκριμένη γεωγραφική θέση 3 μίλια γεωγραφικού πλάτους/ μήκους. Μετά τον κώδικα παρατίθενται κάποια αποτελέσματα.



[[1]]
[1] "GStatus8522: @izzybby__ when we think we are going to a college sorority party @ a club but we actually going to a high school dance where we can't yike"

[[2]]
[1] "hpicciotto: @ilana_horn I don’t mean to bark, but what is the meaning of the dance?"

[[3]]
[1] "JaaylinBreathes: RT @RexxLifeRaj: Pre-show ritual dance https://t.co/Kotwy1X6U1"

[[4]]
[1] "foxyfolklorist: One more thing: SF Beer Week has treated me quite well but having to choose between beer events & dance events is just cruel."

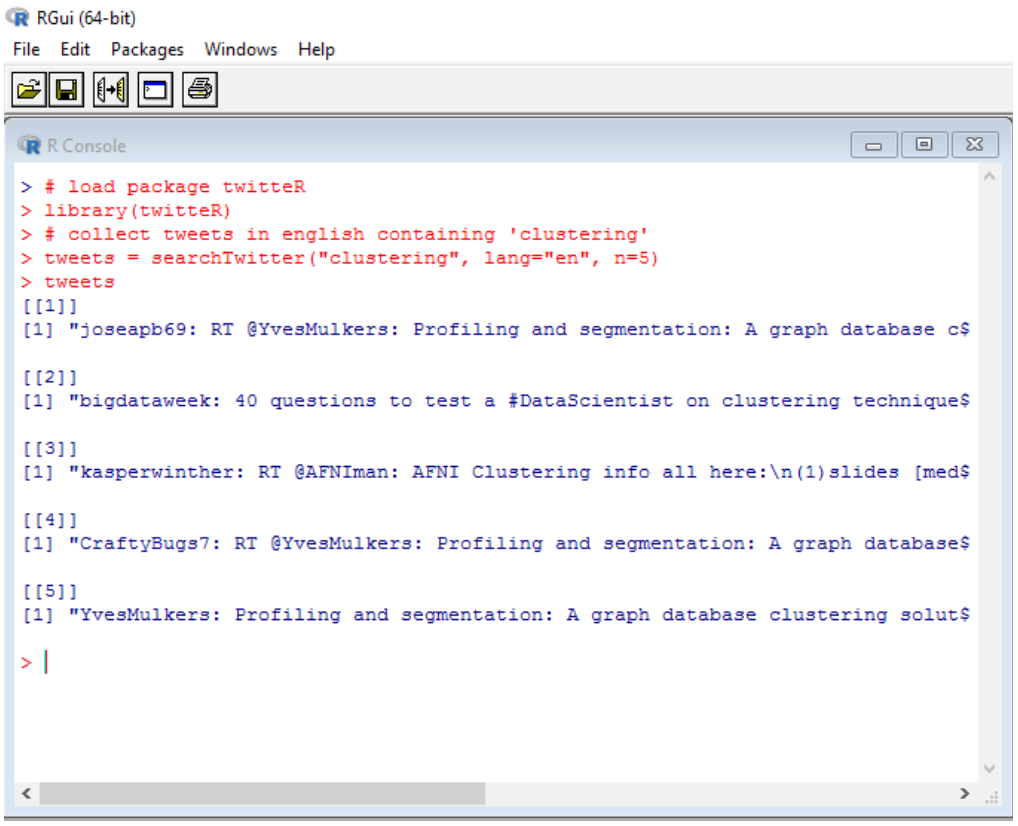
[[5]]
[1] "darrian_garcia: RT @RexxLifeRaj: Pre-show ritual dance https://t.co/Kotwy1X6U1"

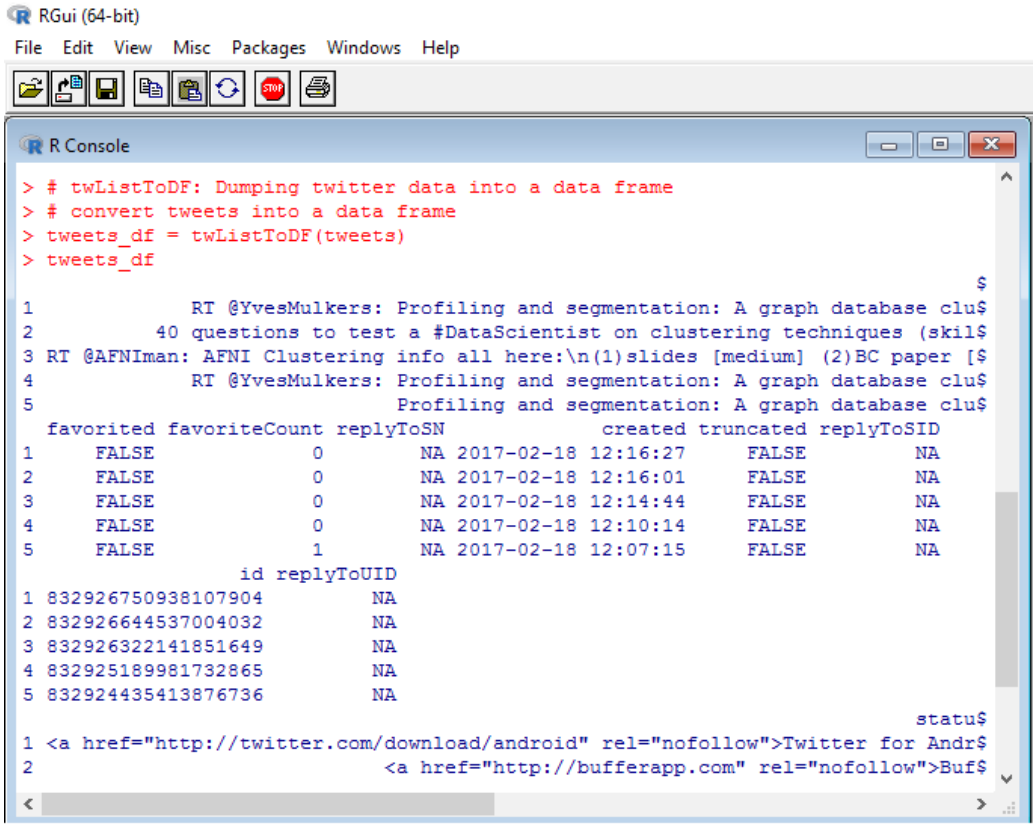
3.2. Επεξεργασία Δεδομένων

Αφού συλλέχθηκαν τα δεδομένα, το επόμενο βήμα περιλαμβάνει κάποιου είδους επεξεργασίας, χειρισμό των δεδομένων, καθαρισμό, μορφοποίηση και φιλτράρισμα. Όπως και στην προηγούμενη ενότητα, υπάρχουν δύο τρόποι για να γίνει αυτό οι οποίοι είναι επεξεργασία (α)με χρήση του πακέτου “twitteR” και (β)με χρήση του πακέτου “XML”. Από τη στιγμή που που πριν επιλέχθηκε χρήση του “twitteR” και σε αυτό το στάδιο θα χρησιμοποιηθεί η ίδια μέθοδος.

Μετά τη συλλογή των επιθυμητών tweets, πρέπει να εξαχθεί το περιεχόμενό τους. Ο πιο απλός τρόπος για να εξαχθούν τα δεδομένα που σχετίζονται με την αναζήτηση από τα tweets που συλλέχθηκαν, είναι με τη χρήση της συνάρτησης twListToDF που βάζει τα πάντα σε ένα πλαίσιο δεδομένων.

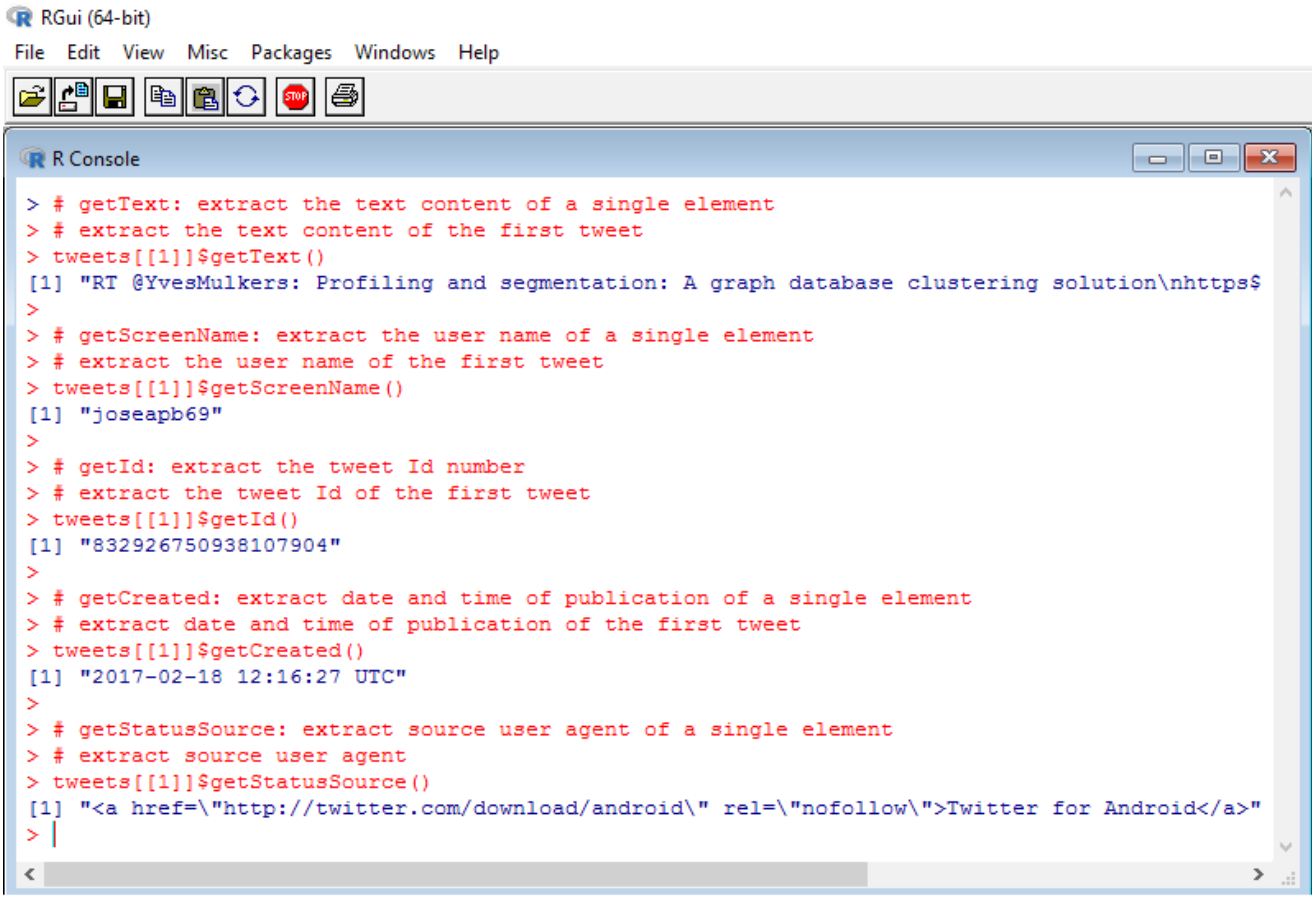
Στις δύο επόμενες εικόνες φαίνεται η παραπάνω διαδικασία.


RGui (64-bit)
File Edit Packages Windows Help
R Console
> # load package twitteR
> library(twitteR)
> # collect tweets in english containing 'clustering'
> tweets = searchTwitter("clustering", lang="en", n=5)
> tweets
[[1]]
[1] "joseapb69: RT @YvesMulkers: Profiling and segmentation: A graph database c$
[[2]]
[1] "bigdataweek: 40 questions to test a #DataScientist on clustering technique$
[[3]]
[1] "kasperwinther: RT @AFNIman: AFNI Clustering info all here:\n(1)slides [med$
[[4]]
[1] "CraftyBugs7: RT @YvesMulkers: Profiling and segmentation: A graph database$
[[5]]
[1] "YvesMulkers: Profiling and segmentation: A graph database clustering solut$
>

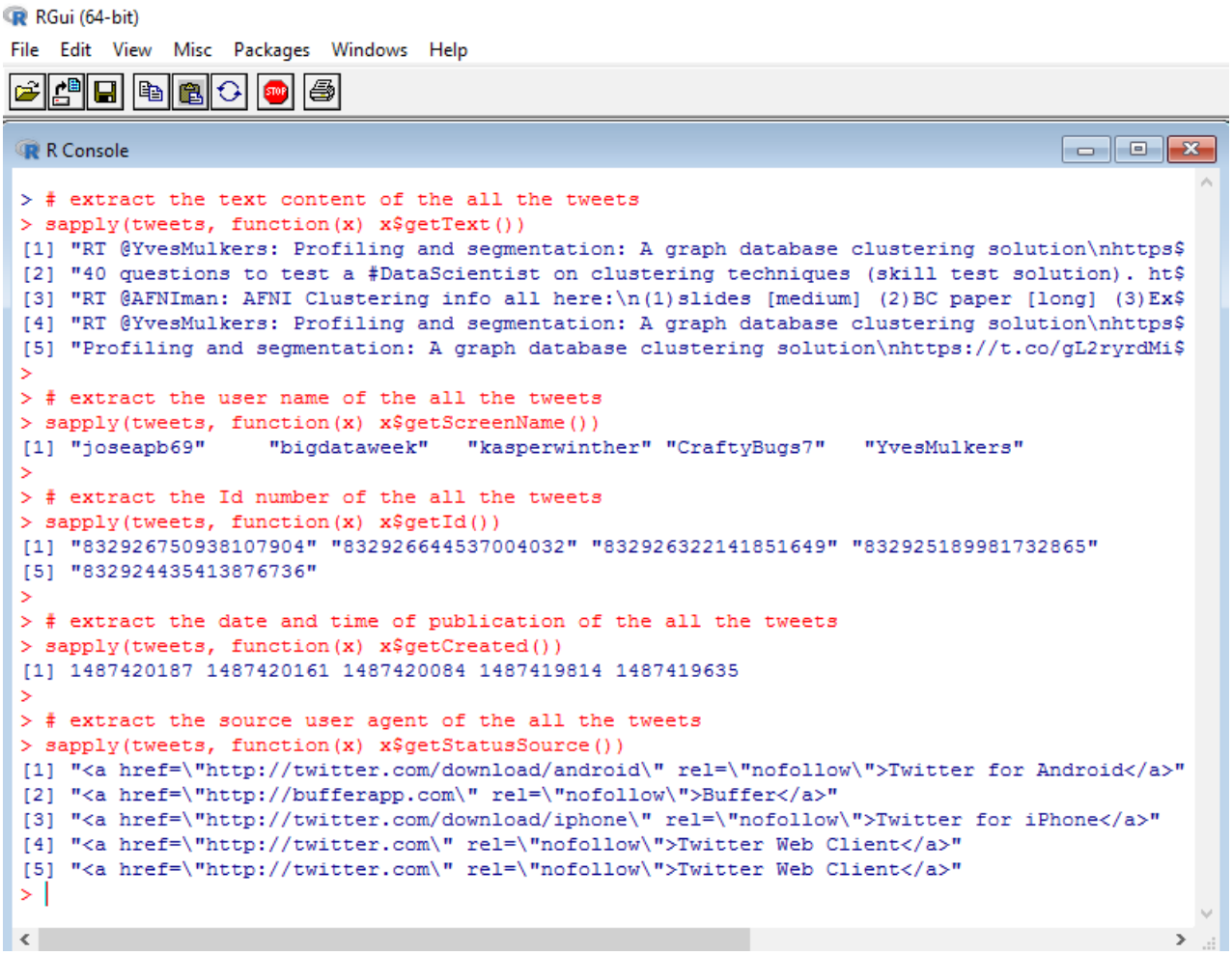

RGui (64-bit)
File Edit View Misc Packages Windows Help
R Console
> # twListToDF: Dumping twitter data into a data frame
> # convert tweets into a data frame
> tweets_df = twListToDF(tweets)
> tweets_df
$
1 RT @YvesMulkers: Profiling and segmentation: A graph database clu$
2 40 questions to test a #DataScientist on clustering techniques (skil$
3 RT @AFNIman: AFNI Clustering info all here:\n(1)slides [medium] (2)BC paper [$
4 RT @YvesMulkers: Profiling and segmentation: A graph database clu$
5 Profiling and segmentation: A graph database clu$
favorited favoriteCount replyToSN created truncated replyToSID
1 FALSE 0 NA 2017-02-18 12:16:27 FALSE NA
2 FALSE 0 NA 2017-02-18 12:16:01 FALSE NA
3 FALSE 0 NA 2017-02-18 12:14:44 FALSE NA
4 FALSE 0 NA 2017-02-18 12:10:14 FALSE NA
5 FALSE 1 NA 2017-02-18 12:07:15 FALSE NA
id replyToUID
1 832926750938107904 NA
2 832926644537004032 NA
3 832926322141851649 NA
4 832925189981732865 NA
5 832924435413876736 NA
statu$
1 <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Andr$
2 <a href="http://bufferapp.com" rel="nofollow">Buf$

Εναλλακτικά, αντί της συνάρτησης twListToDF, μπορούν να χρησιμοποιηθούν οι μέθοδοι get για να εξαχθούν τα επιθυμητά πεδία από ένα στοιχείο μιας λίστας από το Twitter. Οι μέθοδοι θα δοκιμαστούν στο πρώτο tweet που είναι το εξής:

[1] "joseapb69: RT @YvesMulkers: Profiling and segmentation: A graph database clustering solution\nhttps://t.co/gL2ryrdMiN https://t.co/tvNdCTIdcz"



Τις περισσότερες φορές, χρειάζεται να εξαχθούν συγκεκριμένες πληροφορίες από όλα τα tweets που συλλέχθηκαν. Αυτό μπορεί να γίνει με διάφορους τρόπους και ένας από αυτούς είναι με τη χρήση της συνάρτησης sapply. Συνεχίζοντας στο ίδιο παράδειγμα, δηλαδή στην αναζήτηση σχετικά με τη λέξη clustering, βλέπουμε παρακάτω τα αποτελέσματα της χρήσης της sapply. Μετά την εικόνα από τον κώδικα βλέπουμε ποια ήταν τα tweets στα οποία έγινε η επεξεργασία.



[[1]]
[1] "joseapb69: RT @YvesMulkers: Profiling and segmentation: A graph database clustering solution\nhttps://t.co/gL2ryrdMiN https://t.co/tvNdCTIdcz"
[[2]]
[1] "bigdataweek: 40 questions to test a #DataScientist on clustering techniques (skill test solution). https://t.co/LJwlKHO92v https://t.co/MLK9f09P9J"
[[3]]
[1] "kasperwinther: RT @AFNIman: AFNI Clustering info all here:\n(1)slides [medium] (2)BC paper [long] (3)Exec summary [short] (4)PNAS letter [short]\nhttps://t.…"
[[4]]
[1] "CraftyBugs7: RT @YvesMulkers: Profiling and segmentation: A graph database clustering solution\nhttps://t.co/gL2ryrdMiN https://t.co/tvNdCTIdcz"
[[5]]
[1] "YvesMulkers: Profiling and segmentation: A graph database clustering solution\nhttps://t.co/gL2ryrdMiN https://t.co/tvNdCTIdcz"

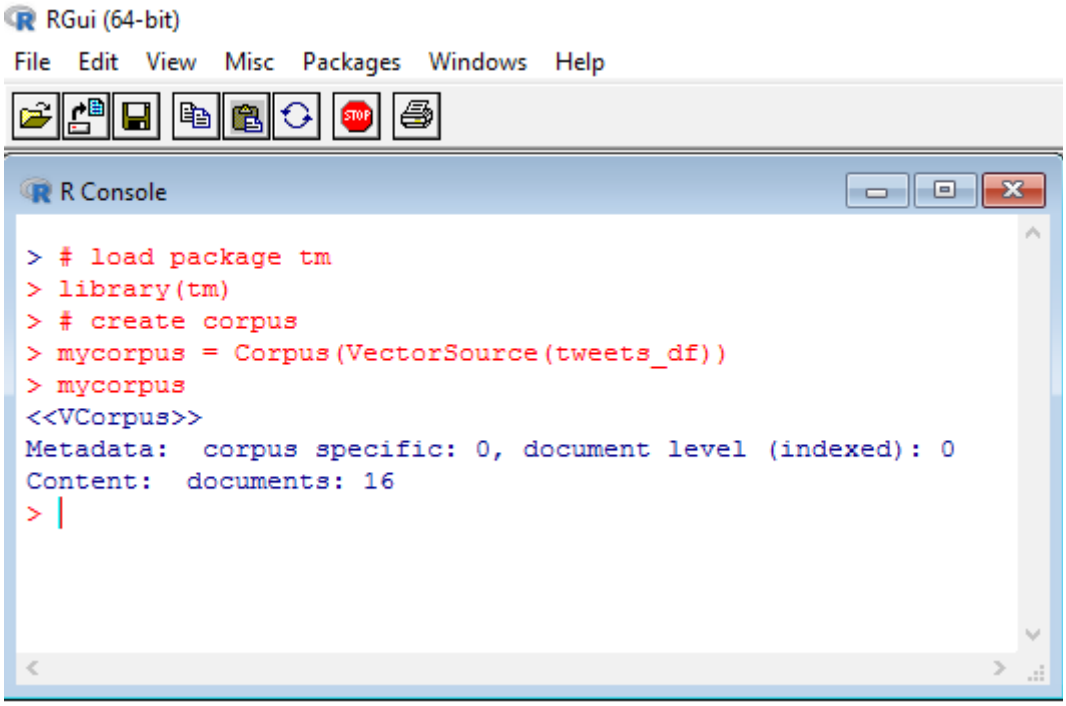
3.3. Εξόρυξη κειμένου

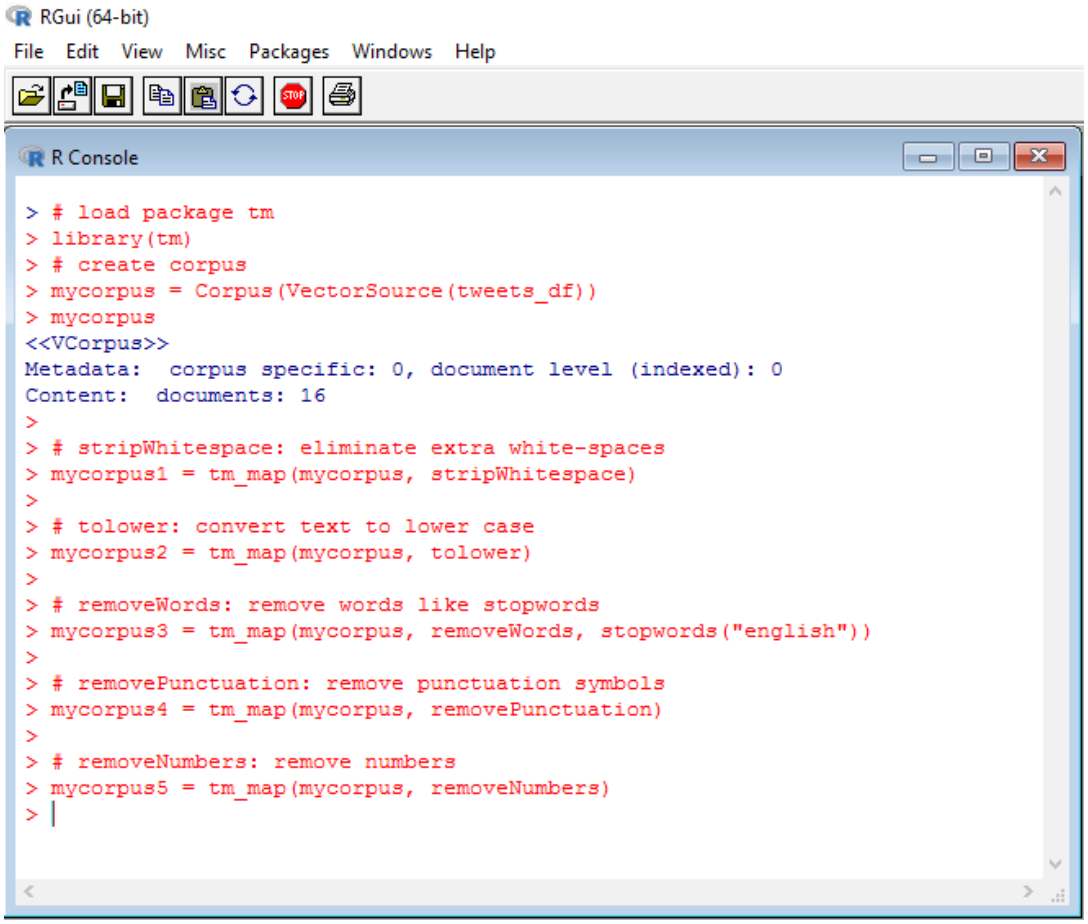
Το βασικό πακέτο για την εκτέλεση εργασιών εξόρυξης κειμένου στην R είναι το πακέτο tm. Η κύρια δομή για τη διαχείριση των εγγράφων στο tm είναι το λεγόμενο (Lexical)Corpus που αντιπροσωπεύει μια συλλογή εγγράφων κειμένου. Εάν τα δεδομένα κειμένου είναι σε ένα διάνυσμα αντικειμένου, που θα είναι συνήθως κατά την εξαγωγή πληροφοριών από το Twitter, ο τρόπος για να δημιουργηθεί ένα corpus είναι:

mycorpus = Corpus(VectorSource(object))

όπου στη θέση του object θα βρίσκεται το tweets_df που δημιουργήθηκε πριν με την εξής εντολή: *tweets_df = twListToDF(tweets)* όπου tweets τα αποτελέσματα αναζήτησης της λέξης “clustering”.

Από τη στιγμή που θα δημιουργηθεί το corpus, θα τροποποιηθούν τα έγγραφα που περιέχει κάνοντας κάποιες αφαιρέσεις σε λέξεις, σύμβολα ακόμα και σε κάποια επιπλέον κενά κ.α. Αυτές οι λειτουργίες μπορούν να εκτελεστούν από το πακέτο tm με τους λεγόμενους μετασχηματισμούς με τη χρήση της συνάρτησης tm_map.





Μια κοινή προσέγγιση στον τομέα της εξόρυξης κειμένου είναι να δημιουργηθεί ένας πίνακας term-document από ένα corpus με τη χρήση των εξής συναρτήσεων:

1) DocumentTermMatrix: δημιουργεί ένα πίνακα με τα έγγραφα(documents) ως σειρές και τους όρους(terms) ως στήλες
2) TermDocumentMatrix: δημιουργεί ένα πίνακα με τους όρους(terms) ως γραμμές και τα έγγραφα(documents) ως στήλες

Κάθε ένας από αυτούς τους δύο τύπους των πινάκων είναι στην πραγματικότητα πολύ σημαντικός για το μεγαλύτερο μέρος της ανάλυσης στην R, διότι έτσι εφαρμόζονται τα εξής: η διαδικασία κατηγοριοποίησης, η ανάλυση συστάδας, η ανάλυση κανόνων συσχέτισης, και ούτω καθεξής. Η μέθοδος αυτή θα χρησιμοποιηθεί σε επόμενο κεφάλαιο.

3.4. Συσταδοποίηση

Υπάρχουν πολλά παραδείγματα εξόρυξης δεδομένων στα οποία μπορούν να χρησιμοποιηθούν οι αλγόριθμοι συσταδοποίησης που έχουν αναφερθεί σε προηγούμενο κεφάλαιο. Οπότε θα πρέπει να επιλεχθεί κάποιο έτσι ώστε να γίνει ορατή η χρησιμότητα των αλγόριθμων αυτών στην εξόρυξη δεδομένων από το Twitter. Ένα wordcloud μπορεί να είναι ένα από τα καλύτερα εργαλεία που επιτρέπει τις απεικονίσεις στις περισσότερες από τις λέξεις και όρους που περιέχονται στα tweets. Αν και η κύρια χρήση του είναι για διερευνητικούς σκοπούς, έχει το πλεονέκτημα να είναι κατανοητό από τους περισσότερους χρήστες, και να είναι οπτικά ελκυστικό με τα ανθρώπινα μάτια (αν γίνει σε μικρό αριθμό δειγμάτων). Ένα word graph είναι κατά κάποιο τρόπο παρόμοιο με ένα wordcloud αν και δεν είναι το ίδιο πράγμα. Παρακάτω δίνεται ένα παράδειγμα δημιουργίας ενός word graph χρησιμοποιώντας tweets που έχουν εξαχθεί από τον χρήστη @Greenpeace, με απώτερο σκοπό τη χρήση αλγορίθμων συσταδοποίησης έτσι ώστε να ταξινομηθούν τα αποτελέσματα σε κατηγορίες για να βελτιωθεί το γράφημα. Συγκεκριμένα θα χρησιμοποιηθεί ο k-means αλγόριθμος. Ο αλγόριθμος k-means είναι μία από τις πολλές τεχνικές που μπορούν να χρησιμοποιηθούν για την ανάλυση κειμένου. Η ομαδοποίηση κ-μέσων είναι μια μέθοδος διαχωρισμού δεδομένων σε ‘k’ υποσύνολα, όπου κάθε στοιχείο από τα δεδομένα έχει εκχωρηθεί στη πιο κοντινή συστάδα με βάση την απόσταση του στοιχείου δεδομένων από το κέντρο της συστάδας. Για να χρησιμοποιηθεί ο αλγόριθμος k-means σε δεδομένα κειμένου, πρέπει να γίνουν κάποιες μετατροπές στα δεδομένα αυτά. Ευτυχώς, η R παρέχει διάφορα πακέτα για να απλοποιηθεί η διαδικασία.

Κώδικας:

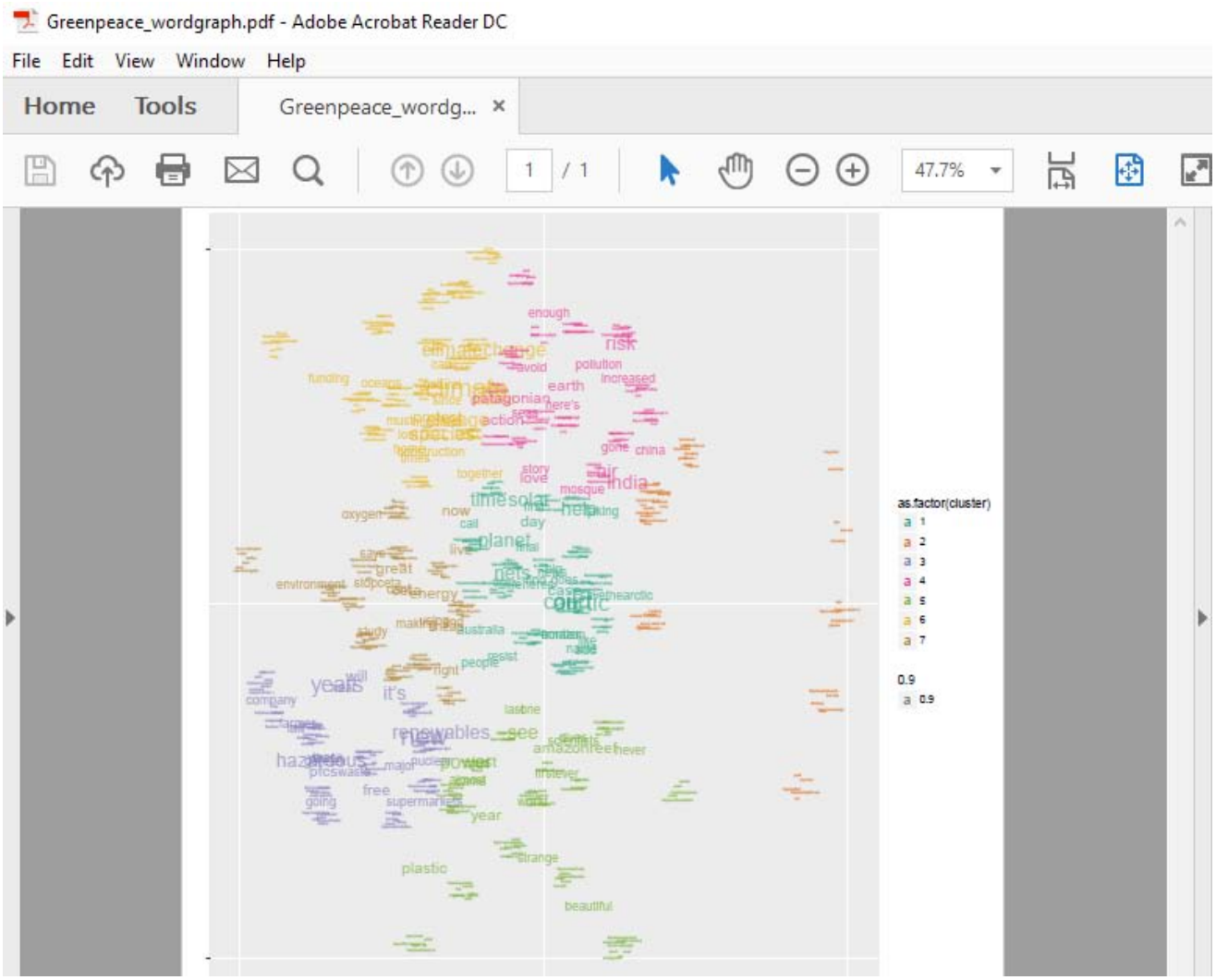
```
# required packages
require(tm)
require(igraph)
require(ggplot)
require(RColorBrewer)

# get tweets from @Greenpeace
gp_tweets = userTimeline("Greenpeace", n=1000)

# extract text
gp_text = sapply(gp_tweets, function(x) x$getText())
```

```
# create a corpus via VectorSource
gp_corpus1 = Corpus(VectorSource(gp_text))

# define list of transformations
gp_stopwords = unique(c(stopwords(), "greenpeace", "via"))
# list of transformations
trans = list(weighting=weightTf, stopwords=gp_stopwords)

gp_corpus <- tm_map(gp_corpus1, stripWhitespace)
gp_corpus <- tm_map(gp_corpus, removeNumbers)
gp_corpus <- tm_map(gp_corpus, removePunctuation)
gp_corpus <- tm_map(gp_corpus, content_transformer(tolower))

# create a term-document matrix
gp_tdm = TermDocumentMatrix(gp_corpus, control=trans)

# Remove sparse terms from matrix
gp_clean = removeSparseTerms(gp_tdm, .995)

# convert as matrix
gp_clean = as.matrix(gp_clean)

# first create a word affiliations matrix
affi_matrix = gp_clean %*% t(gp_clean)

# then create an adjacency matrix with zeroes in its diagonal
adja_matrix = affi_matrix
diag(adja_matrix) = 0

# Create a graph
gp_graph = graph.adjacency(adja_matrix, weighted=TRUE, mode="undirected",
                add.rownames=TRUE)
# coordinates for visualization
posi_matrix=layout.fruchterman.reingold(gp_graph)
posi_matrix = cbind(V(gp_graph)$name, posi_matrix)
```

```
# create a data frame
gp_df = data.frame(posi_matrix, stringsAsFactors=FALSE)
names(gp_df) = c("word", "x", "y")
gp_df$x = as.numeric(gp_df$x)
gp_df$y = as.numeric(gp_df$y)

# let's make a first attempt
# size effect
se = diag(affi_matrix) / max(diag(affi_matrix))
# plot
par(bg = "gray15")
with(gp_df, plot(x, y, type="n", xaxt="n", yaxt="n", xlab="", ylab="", bty="n"))
with(gp_df, text(x, y, labels=word, cex=log10(diag(affi_matrix)),
col=hsv(0.95, se, 1, alpha=se)))

# To improve our graph, we can perform a k-means cluster analysis to find groups
# k-means with 7 clusters
words_km = kmeans(cbind(as.numeric(posi_matrix[,2]), as.numeric(posi_matrix[,3])), 7)

# add frequencies and clusters in a data frame
gp_df = transform(gp_df, freq=diag(affi_matrix), cluster=as.factor(words_km$cluster))
row.names(gp_df) = 1:nrow(gp_df)

# here's the final plot
# graphic with ggplot
gp_words = ggplot(gp_df, aes(x=x, y=y)) +
geom_text(aes(size=freq, label=gp_df$word, alpha=.90, color=as.factor(cluster))) +
labs(x="", y="") +
scale_size_continuous(breaks = c(10,20,30,40,50,60,70,80,90), range = c(1,8)) +
scale_colour_manual(values=brewer.pal(8, "Dark2")) +
scale_x_continuous(breaks=c(min(gp_df$x), max(gp_df$x)), labels=c("","")) +
scale_y_continuous(breaks=c(min(gp_df$y), max(gp_df$y)), labels=c("",""))

# save the image in pdf format
ggsave(plot=gp_words, filename="Greenpeace_wordgraph.pdf", height=10, width=10)
```

Τα αποτελέσματα του παραπάνω κώδικα φαίνονται στις παρακάτω εικόνες. Στην πρώτη εικόνα φαίνεται το γράφημα που δημιουργήθηκε από τον κώδικα(word graph). Στο γράφημα βέβαια δεν είναι ξεκάθαρες οι λέξεις καθώς χρησιμοποιήθηκαν 1000 tweets από τον χρήστη.

Στη δεύτερη εικόνα φαίνεται το γράφημα μετά την εφαρμογή της μεθόδου συσταδοποίησης k-means το οποίο επιλέχθηκε να αποθηκευτεί σε ένα αρχείο pdf. Εδώ φαίνεται καθαρά η κατηγοριοποίηση σε συστάδες που έγινε στο σύνολο δεδομένων που χρησιμοποιήθηκαν για να δημιουργηθεί το αρχικό γράφημα. Κάθε συστάδα απεικονίζεται στο γράφημα με διαφορετικό χρώμα.



4. Εξόρυξη στις Ερωτήσεις για το Twitter

Η συλλογή πληροφοριών από τα tweets δεν οφελούν μόνο το ίδιο το Twitter. Αντίθετα, τα δεδομένα συγκεντρώνονται από το Twitter, γιατί οδηγούμαστε από τα κίνητρα του κάθε έργου εξόρυξης δεδομένων με σκοπό να ανακαλυφθούν χρήσιμες, έγκυρες, απρόσμενες και κατανοητές πληροφορίες από τα δεδομένα. Οι πληροφορίες από το Twitter αναλύονται επειδή δίνουν απαντήσεις σε συγκεκριμένες ερωτήσεις που υπάρχουν στο μυαλό των ανθρώπων. Μπορεί να είναι μια πολύ απλή ερώτηση όπως ποιος είναι ο μέσος αριθμός των χαρακτήρων στα tweets σχετικά με τον #Antetokounmpo, ή μπορεί να είναι ένα πιο σύνθετο ζήτημα, για παράδειγμα, τι είδους συσχετισμοί μπορούν να δημιουργηθούν μεταξύ δύο δεδομένων hashtags όπως #PabloPicasso και #SalvadorDali, βέβαια αν υπάρχουν. Υπάρχουν κυριολεκτικά σε όλους δεκάδες ερωτήματα σχετικά με αγαπημένα θέματα, και τη στιγμή που θα ξεκινήσει η απάντησή τους, αργά ή γρήγορα, περισσότερες ερωτήσεις θα προκύψουν από αυτές τις απαντήσεις. Μερικά από τα πιο συνηθισμένα ερωτήματα εστιάζουν γύρω από τις τρεις εξής ρωτήσεις:

(1) Τι συζητά ο κόσμος;
(2) Ποιες είναι οι συχνότητες στα δεδομένα;
(3) Ποιες σχέσεις μπορούν να εξαχθούν από τα tweets;
(4) Ποια είναι η ψυχολογία / γνώμη των ανθρώπων;

Φυσικά υπάρχουν περισσότερες ερωτήσεις, αλλά η παρούσα εργασία θα επικεντρωθεί στα παραπάνω ερωτήματα. Στις επόμενες ενότητες θα αναλυθούν περαιτέρω οι ερωτήσεις που προαναφέρθηκαν και θα παρουσιαστούν κάποια παραδείγματα κώδικα.

4.1. Δημοφιλή Θέματα Συζήτησης

Ποια είναι τα δημοφιλή θέματα συζήτησης; Τι συζητούν οι άνθρωποι σχετικά με κάποιο #hashtag; Τι συζητούν οι άνθρωποι για κάποιο συγκεκριμένο όρο; Για ποια θέματα μιλάει ένας συγκεκριμένος χρήστης; Ποια είναι τα θέματα συζήτησης για κάποιους συγκεκριμένους χρήστες; Αυτά είναι κάποια από τα πολλά ερωτήματα που προβληματίζουν πολλές φορές κάποιον. Οι πληροφορίες από το Twitter αναλύονται επειδή δίνουν απαντήσεις σε συγκεκριμένες ερωτήσεις που υπάρχουν στο μυαλό των ανθρώπων όπως προαναφέρθηκε. Ίσως το κύριο ερώτημα που τίθεται προς απάντηση, μέσα από μια ανάλυση στο Twitter, σχετίζεται με την ανακάλυψη των απόψεων των ανθρώπων σχετικά με κάποιο συγκεκριμένο θέμα, ή ακόμα και όσον αφορά τα θέματα συζήτησης κάποιου συγκεκριμένου χρήστη.

Συνήθως ο πιο γρήγορος τρόπος για να ανακαλυφθούν περισσότερα σχετικά με το τι λένε οι άνθρωποι για κάποιο συγκεκριμένο θέμα είναι μέσα από την απεικόνιση των πιο συχνών λέξεων και όρων που περιέχονται στα tweets. Αυτό μπορεί να υλοποιηθεί με τη δημιουργία ενός barplot με τους πιο συχνούς όρους ή επιλέγοντας μια περισσότερο ελκυστική οπτικά επιλογή που αποτελείται από τη χρήση κάποιου είδους wordcloud που είναι, όπως αναφέρθηκε σε προηγούμενο κεφάλαιο, ένα από τα καλύτερα εργαλεία που επιτρέπει τις απεικονίσεις στις περισσότερες από τις λέξεις και όρους που περιέχονται στα tweets. Και οι δύο επιλογές είναι καλές, αλλά έχουν ένα βασικό περιορισμό, δεν δείχνουν πώς σχετίζονται οι λέξεις. Απλά αντικατοπτρίζουν τις πιο δημοφιλείς λέξεις στα tweets. Για να υπάρξει κάποια εικόνα για τις πιθανές σχέσεις των λέξεων, η συνιστώμενη συμβουλή είναι να χρησιμοποιηθεί κάποιο είδος γραφήματος (ή δικτύου) για να φανούν αυτές οι σχέσεις. Ας δούμε μερικά απλά παραδείγματα.

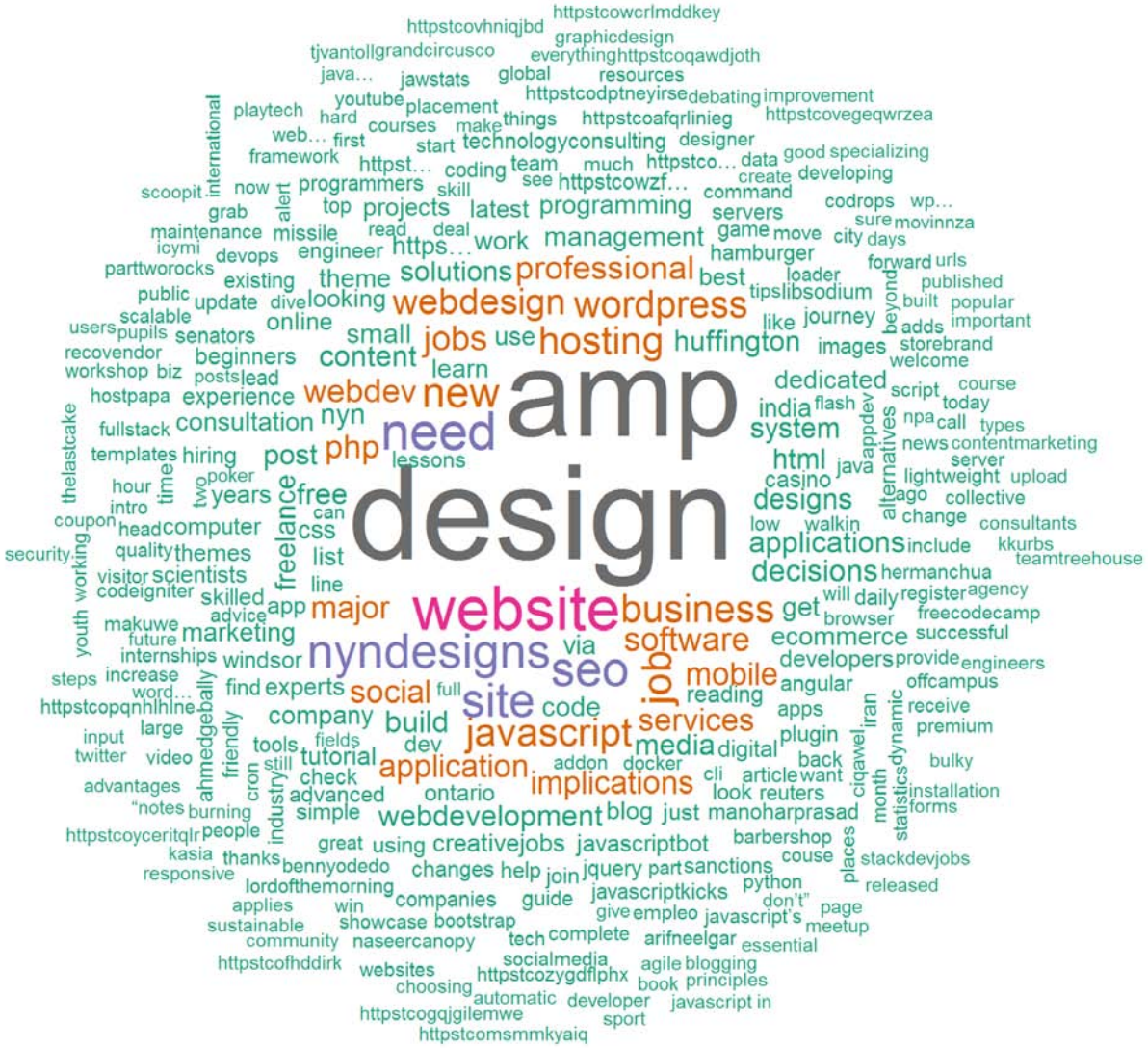
Παράδειγμα 1ο: Simple Wordcloud (given topic:web development)

Κώδικας:

```
# Load all the required packages
library(twitteR)
library(tm)
library(wordcloud)
library(RColorBrewer)

# Let's get some tweets in english containing the words "web development"
mach_tweets = searchTwitter("web development", n=500, lang="en")

# Extract the text from the tweets in a vector
mach_text = sapply(mach_tweets, function(x) x$getText())

# create a corpus
mach_corpus = Corpus(VectorSource(mach_text))

# create document term matrix applying some transformations
wd_stopwords = unique(c(stopwords(), "web", "development"))

trans = list(weighting=weightTf, stopwords=wd_stopwords)
```

```
wd_corpus <- tm_map(mach_corpus, stripWhitespace)

wd_corpus <- tm_map(wd_corpus, removeNumbers)

wd_corpus <- tm_map(wd_corpus, removePunctuation)

wd_corpus <- tm_map(wd_corpus, content_transformer(tolower))

# create a term-document matrix
tdm = TermDocumentMatrix(wd_corpus, control=trans)

# Obtain words and their frequencies
# define tdm as matrix
m = as.matrix(tdm)
# get word counts in decreasing order
word_freqs = sort(rowSums(m), decreasing=TRUE)
# create a data frame with words and their frequencies
dm = data.frame(word=names(word_freqs), freq=word_freqs)

# plot wordcloud
wordcloud(dm$word, dm$freq, random.order=FALSE, colors=brewer.pal(8, "Dark2"))

# save the image in png format
png("WebDevelopmentCloud.png", width=12, height=8, units="in", res=300)
wordcloud(dm$word, dm$freq, random.order=FALSE, colors=brewer.pal(8, "Dark2"))
dev.off()
```

Στην παρακάτω εικόνα φαίνεται το αποτέλεσμα που προκύπτει μετά την εφαρμογή του κώδικα. Εξάγοντας τα tweets τα οποία περιέχουν τις λέξεις web development, δημιουργήθηκε το wordcloud που αποτελείται από τις πιο συνηθισμένες λέξεις ή όρους που περιέχουν τα tweets αυτά (εκτός φυσικά από τις ίδιες τις λέξεις web development). Το wordcloud που δημιουργήθηκε, αποθηκεύτηκε σε μορφή εικόνας (png format).



Wordcloud for web development

Παράδειγμα 2ο: Comparison Wordcloud (given users:PC Companies)

Ακόμα ένας τύπος γραφήματος στο πακέτο του wordcloud είναι και το comparison wordcloud (wordcloud σύγκρισης).

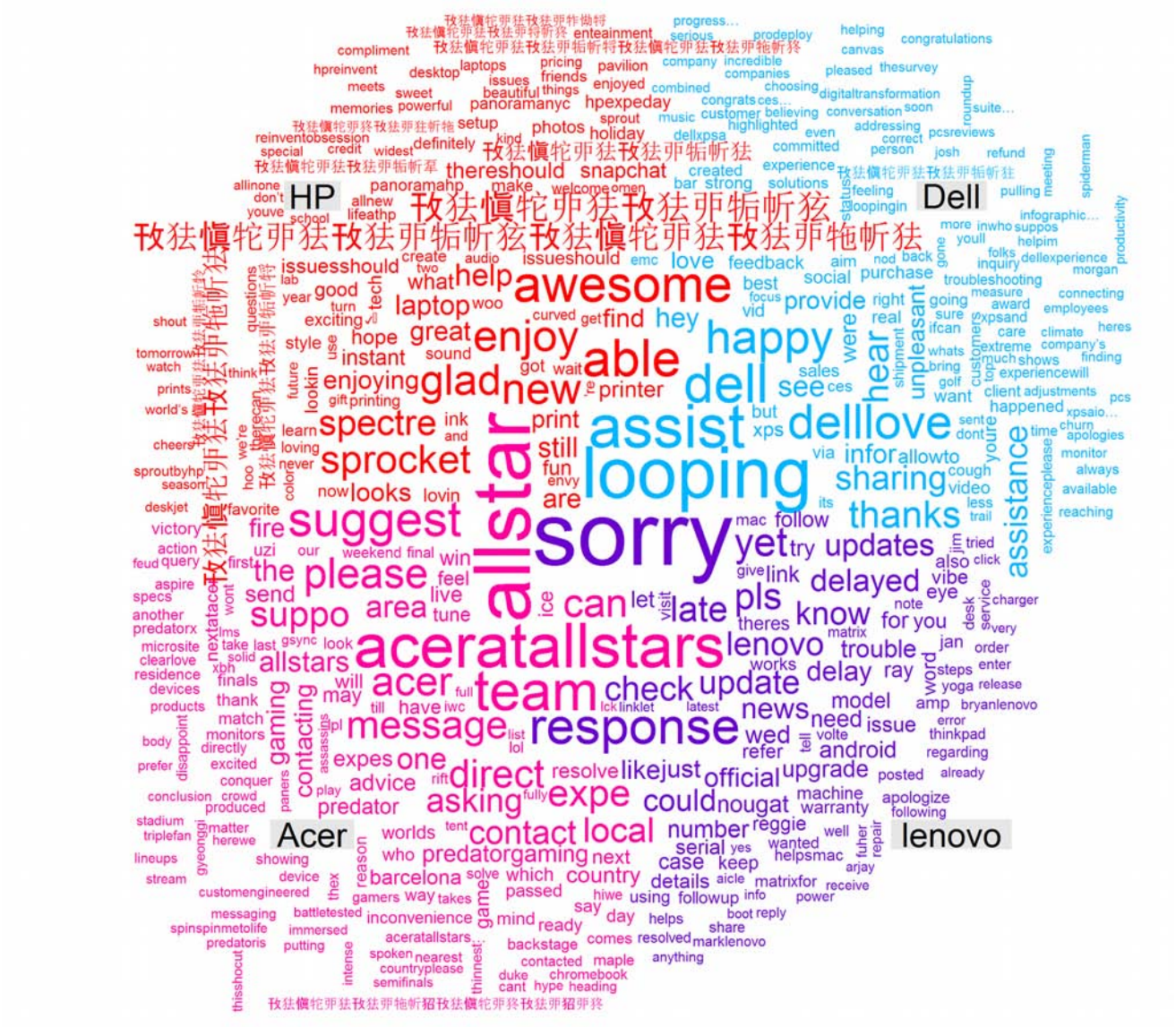
Κώδικας:

```
# Load all the required packages
library(twitteR)
library(tm)
library(wordcloud)
library(RColorBrewer)

# collect tweets from pc companies
# dell tweets
dell_tweets = userTimeline("Dell", n=1000)

# hp tweets
hp_tweets = userTimeline("HP", n=1000)

# acer tweest
acer_tweets = userTimeline("Acer", n=1000)

# lenovo tweets
lenovo_tweets = userTimeline("lenovo", n=1000)

# get text
dell_txt = sapply(dell_tweets, function(x) x$getText())
hp_txt = sapply(hp_tweets, function(x) x$getText())
acer_txt = sapply(acer_tweets, function(x) x$getText())
lenovo_txt = sapply(lenovo_tweets, function(x) x$getText())

# clean text
clean.text = function(x)
{
  # remove rt
  x = gsub("rt", "", x)
  # remove at
```

```
  x = gsub("@\\w+", "", x)
  # remove punctuation
  x = gsub("[[:punct:]]", "", x)
  # remove numbers
  x = gsub("[[:digit:]]", "", x)
  # remove links http
  x = gsub("http\\w+", "", x)
  # remove tabs
  x = gsub("[ |\t]{2,}", "", x)
  # remove blank spaces at the beginning
  x = gsub("^ ", "", x)
  # remove blank spaces at the end
  x = gsub(" $", "", x)
  return(x)
}

# clean texts
dell_clean = clean.text(dell_txt)
hp_clean = clean.text(hp_txt)
acer_clean = clean.text(acer_txt)
lenovo_clean = clean.text(lenovo_txt)

# Join texts in a vector for each company
dell = paste(dell_clean, collapse=" ")
hp = paste(hp_clean, collapse=" ")
acer = paste(acer_clean, collapse=" ")
lenovo = paste(lenovo_clean, collapse=" ")

# put everything in a single vector
all = c(dell, hp, acer, lenovo)

# remove stop-words
all = removeWords(all,
c(stopwords("english"), "dell", "hp", "acer", "lenovo"))

# create corpus
corpus = Corpus(VectorSource(all))
```

```
corpus <- tm_map(corpus, content_transformer(tolower))

# create term-document matrix
tdm = TermDocumentMatrix(corpus)

# convert as matrix
tdm = as.matrix(tdm)

# add column names
colnames(tdm) = c("Dell", "HP", "Acer", "lenovo")

# Plot comparison wordcloud
# comparison cloud
comparison.cloud(tdm, random.order=FALSE,
colors = c("#00B2FF", "red", "#FF0099", "#6600CC"),
title.size=1.5, max.words=500)

# save the image in png format
png("PcCompaniesComparisonCloud.png", width=12, height=8, units="in", res=300)
comparison.cloud(tdm, random.order=FALSE,
colors = c("#00B2FF", "red", "#FF0099", "#6600CC"),
title.size=1.5, max.words=500)
dev.off()

# Plot commonality cloud
# commonality cloud
commonality.cloud(tdm, random.order=FALSE,
colors = brewer.pal(8, "Dark2"),
title.size=1.5)

# save the image in png format
png("PcCompaniesCommonalityCloud.png", width=8, height=5, units="in", res=300)
commonality.cloud(tdm, random.order=FALSE,
colors = brewer.pal(8, "Dark2"),
title.size=1.5)
dev.off()
```

Στις παρακάτω εικόνες φαίνεται το αποτέλεσμα που προκύπτει μετά την εφαρμογή του κώδικα. Εξάγοντας τα 1000 πιο πρόσφατα tweets από τον καθένα από τους χρήστες του Twitter @Dell, @HP, @Acer, @lenovo, που είναι εταιρείες κυρίως ηλεκτρονικών υπολογιστών, δημιουργήθηκε το comparison wordcloud στην πρώτη εικόνα που αποτελείται από λέξεις που περιέχουν τα tweets αυτά και είναι χωρισμένες στους τέσσερις χρήστες από τους οποίους προήλθαν. Το wordcloud που δημιουργήθηκε, αποθηκεύτηκε σε μορφή εικόνας (png format).



Comparison Wordcloud for PC Companies

Στη δεύτερη εικόνα δημιουργήθηκε το commonality wordcloud που περιέχει τις κοινές λέξεις από όλα τα tweets των τεσσάρων χρηστών. Το wordcloud που δημιουργήθηκε, αποθηκεύτηκε σε μορφή εικόνας (png format) και φαίνεται παρακάτω.

Commonality Wordcloud for PC Companies

Παράδειγμα 3ο: Word Graph (given topic:economy and politics)

Επιλέγοντας ένα θέμα για παράδειγμα «οικονομία» και «πολιτική», στόχος είναι να ερευνηθούν οι βασικοί όροι που χρησιμοποιούνται σε tweets σχετικά με αυτό το θέμα καθώς και η πιθανή σχέση τους. Μια ενδιαφέρουσα επιλογή για την επίτευξη αυτού του στόχου είναι η χρήση ενός γραφήματος λέξεων (word graph), όπως στο παρακάτω παράδειγμα.

Κώδικας:

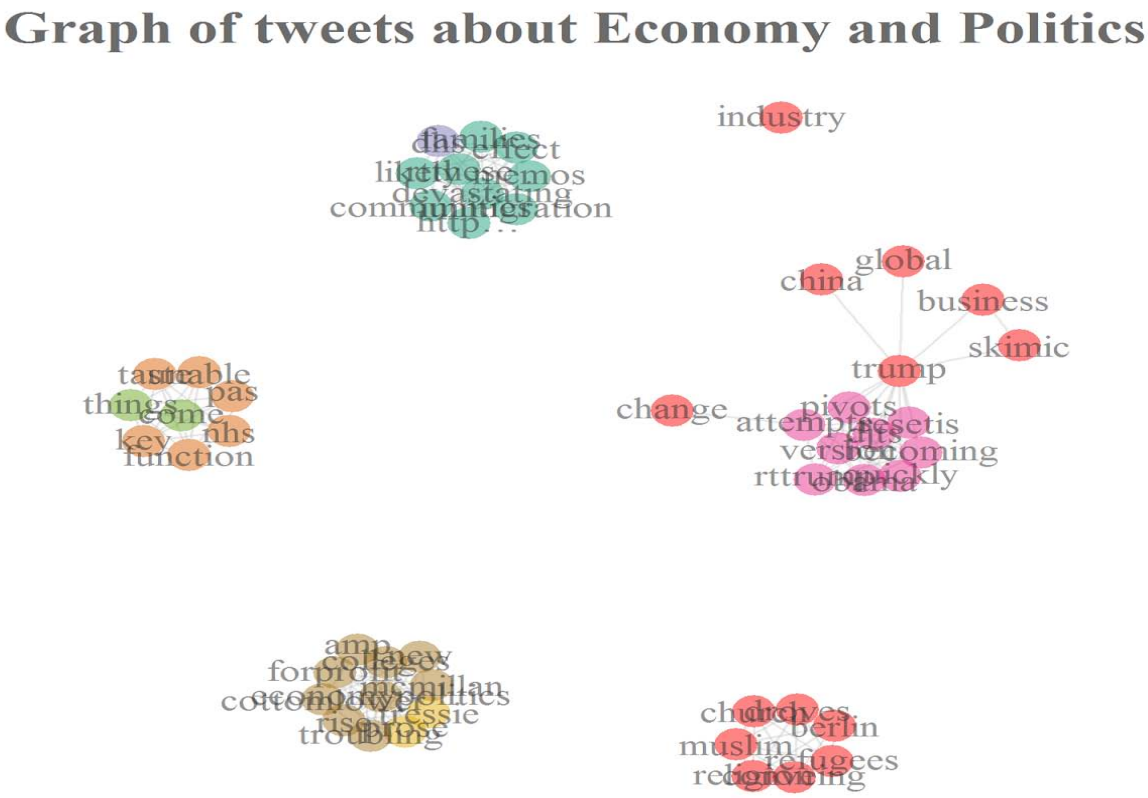
```
# Load all the required packages
library(twitteR)
library(tm)
library(igraph)
library(RColorBrewer)

# Let's get some tweets in english containing the words "economy and politics"
eap_tweets = searchTwitter("economy and politics", n=80, lang="en")

# Extract the text from the tweets in a vector
eap_text = sapply(eap_tweets, function(x) x$getText())

# clean text
clean.text = function(x)
{
  # remove rt
  x = gsub("rt", "", x)
  # remove at
  x = gsub("@\\w+", "", x)
  # remove punctuation
  x = gsub("[[:punct:]]", "", x)
  # remove numbers
  x = gsub("[[:digit:]]", "", x)
  # remove links http
  x = gsub("http\\w+", "", x)
  # remove tabs
  x = gsub("[ |\t]{2,}", "", x)
```

```
  # remove blank spaces at the beginning
  x = gsub("^ ", "", x)
  # remove blank spaces at the end
  x = gsub(" $", "", x)
  return(x)
}

# clean texts
eap_clean = clean.text(eap_text)

# applying some transformations
eap_stopwords = unique(c(stopwords(), "economy", "and", "politics"))
trans = list(weighting=weightTf, stopwords=eap_stopwords)

# create a corpus
eap_corpus = Corpus(VectorSource(eap_clean))

# applying some transformations and create document term matrix
eap_corpus <- tm_map(eap_corpus, content_transformer(tolower))
tdm = TermDocumentMatrix(eap_corpus, control=trans)

# Obtain words and their frequencies
# define tdm as matrix
m = as.matrix(tdm)

# word counts
wc = rowSums(m)

# get those words above the 3rd quantile
lim = quantile(wc, probs=0.5)
good = m[wc > lim,]

# remove columns (docs) with zeroes
good = good[,colSums(good)!=0]

# adjacency matrix
M = good %*% t(good)
```

```
# set zeroes in diagonal
diag(M) = 0

# graph
g = graph.adjacency(M, weighted=TRUE, mode="undirected",
add.rownames=TRUE)
# layout
glay = layout.fruchterman.reingold(g)

# let's superimpose a cluster structure with k-means clustering
kmg = kmeans(M, centers=8)
gk = kmg$cluster

# create nice colors for each cluster
gbrew = c("red", brewer.pal(8, "Dark2"))
gpal = rgb2hsv(col2rgb(gbrew))
gcols = rep("", length(gk))
for (k in 1:8) {
gcols[gk == k] = hsv(gpal[1,k], gpal[2,k], gpal[3,k], alpha=0.5)
}

# prepare ingredients for plot
V(g)$size = 10
V(g)$label = V(g)$name
V(g)$degree = degree(g)
#V(g)$label.cex = 1.5 * log10(V(g)$degree)
V(g)$label.color = hsv(0, 0, 0.2, 0.55)
V(g)$frame.color = NA
V(g)$color = gcols
E(g)$color = hsv(0, 0, 0.7, 0.3)

# plot
plot(g, layout=glay)
title("\nGraph of tweets about Economy and Politics",
col.main="gray40", cex.main=1.5, family="serif")
```

```
# save the image in png format
png("Economy_and_Politics_wordgraph.png", width=10, height=6, units="in", res=300)
plot(g, layout=glay)
title("\nGraph of tweets about Economy and Politics",
col.main="gray40", cex.main=1.5, family="serif")
dev.off()
```

Στην παρακάτω εικόνα φαίνεται το αποτέλεσμα που προκύπτει μετά την εφαρμογή του κώδικα. Εξάγοντας τα tweets τα οποία περιέχουν το θέμα “economy and politics”, δημιουργήθηκε το wordgraph που αποτελείται από τις πιο συνηθισμένες λέξεις ή όρους που περιέχουν τα tweets αυτά (εκτός φυσικά από τις ίδιες τις λέξεις economy and politics). Το wordgraph που δημιουργήθηκε, αποθηκεύτηκε σε μορφή εικόνας (png format).



Wordgraph for “economy and politics”

4.2. Εξέταση Συχνότητας Δεδομένων

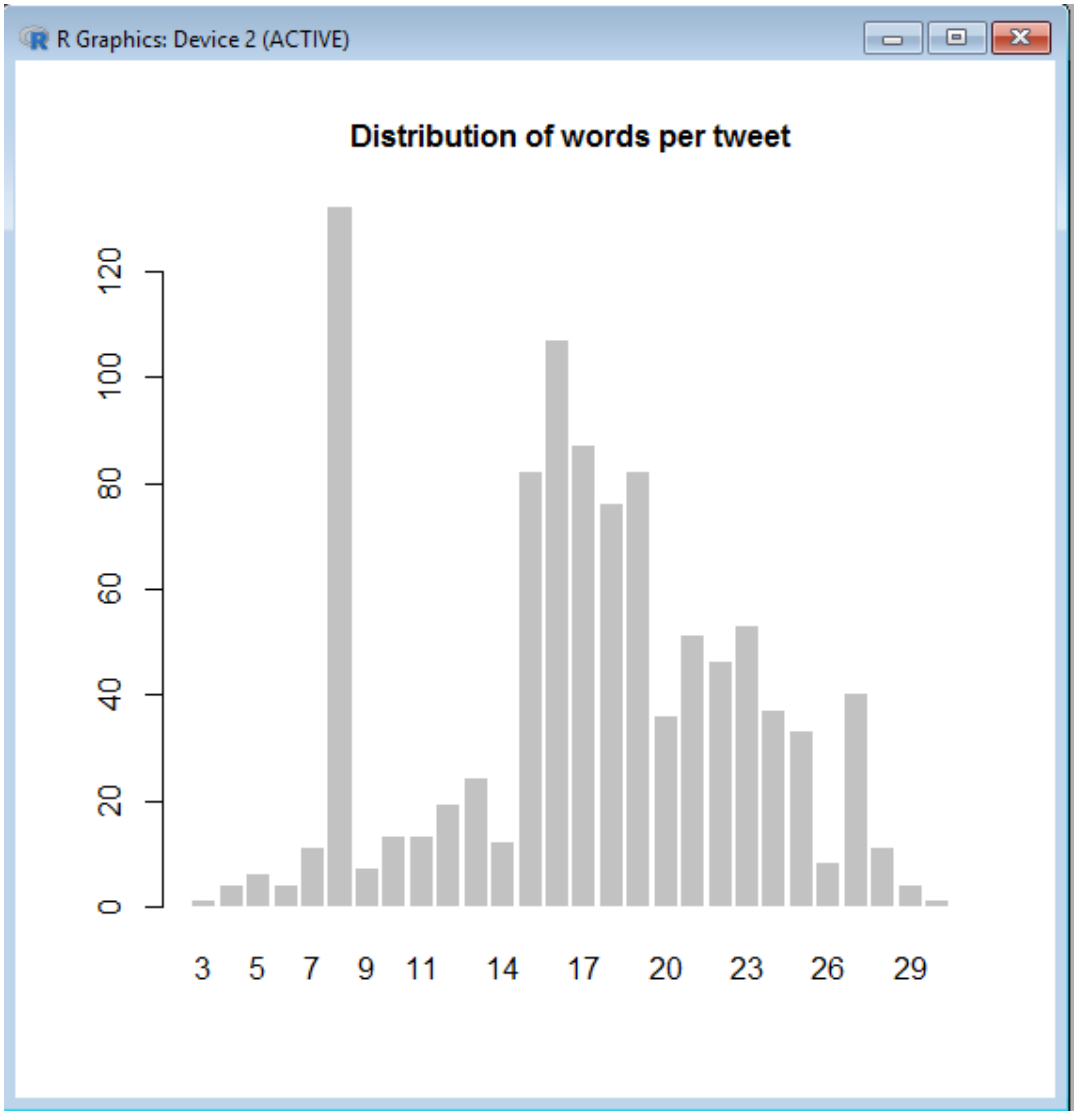
Κάθε αξιοπρεπές σχέδιο ανάλυσης δεδομένων απαιτεί μια καλή γνώση των δεδομένων, τον υπολογισμό των συνολικών στατιστικών στοιχείων, τον έλεγχο των κατανομών, και την εκτέλεση διερευνητικής ανάλυσης. Αναλύοντας τα δεδομένα από το Twitter πρέπει να αντιμετωπιστούν ερωτήσεις όπως, ποιος είναι ο μέσος αριθμός των λέξεων ανά tweet; ποιο είναι το μέσο μήκος μιας λέξης; ποιος είναι ο αριθμός των hashtags ανά tweet; ποια η λεξιλογική ποικιλομορφία των tweets; ποιες είναι οι πιο συχνές λέξεις / όροι; Μία από τις πιο απλές τεχνικές που μπορούν να εφαρμοστούν για να απαντηθούν τα ερωτήματα αυτά, είναι βασική ανάλυση συχνοτήτων. Στο παράδειγμα παρακάτω θα χρησιμοποιηθούν tweets σχετικά με τις λέξεις “talent show”. Για να διατηρηθούν τα πράγματα απλά, θα πραγματοποιηθεί ανάλυση συχνότητας στα tweets που έχουν εξαχθεί χωρίς να γίνει κανένας καθαρισμός δεδομένων.

Κώδικας:

```
# Load all the required packages
library(twitteR)
library(tm)
library(ggplot2)
library(RColorBrewer)

# Let's get some tweets in english containing the words "talent show"
ts_tweets = searchTwitter("talent show", n=1000, lang="en")

# Extract the text from the tweets in a vector
ts_text = sapply(ts_tweets, function(x) x$getText())

# characters per tweet
chars_per_tweet = sapply(ts_text, nchar)
summary(chars_per_tweet)

# how many words per tweets
# split words
words_list = strsplit(ts_text, " ")
```

```
# words per tweet
words_per_tweet = sapply(words_list, length)
# barplot
barplot(table(words_per_tweet), border=NA,
   main="Distribution of words per tweet", cex.main=1)
```

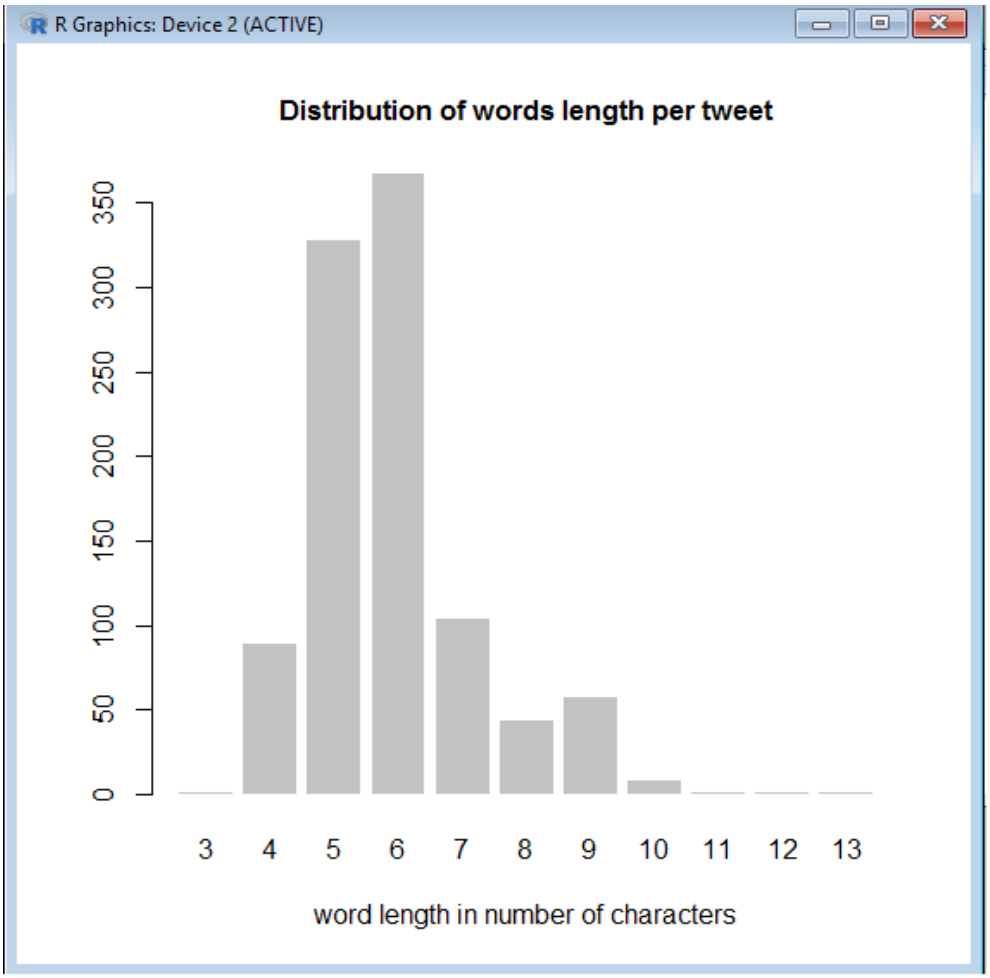
Στην παρακάτω εικόνα φαίνεται το αποτέλεσμα που προκύπτει μετά την εφαρμογή του κώδικα. Εξάγοντας τα tweets τα οποία περιέχουν το θέμα “talent show”, δημιουργήθηκε το γράφημα που δείχνει την κατανομή των λέξεων, που προήλθαν από τα εξαγόμενα tweet, ανά tweet.



Συνεχίζοντας τον προηγούμενο κώδικα παράγονται τα παρακάτω.
Κώδικας(συνέχεια του προηγούμενου):

```
# length of words per tweet
wsize_per_tweet = sapply(words_list, function(x) mean(nchar(x)))
# barplot
barplot(table(round(wsize_per_tweet)), border=NA,
  xlab = "word length in number of characters",
  main="Distribution of words length per tweet", cex.main=1)
```

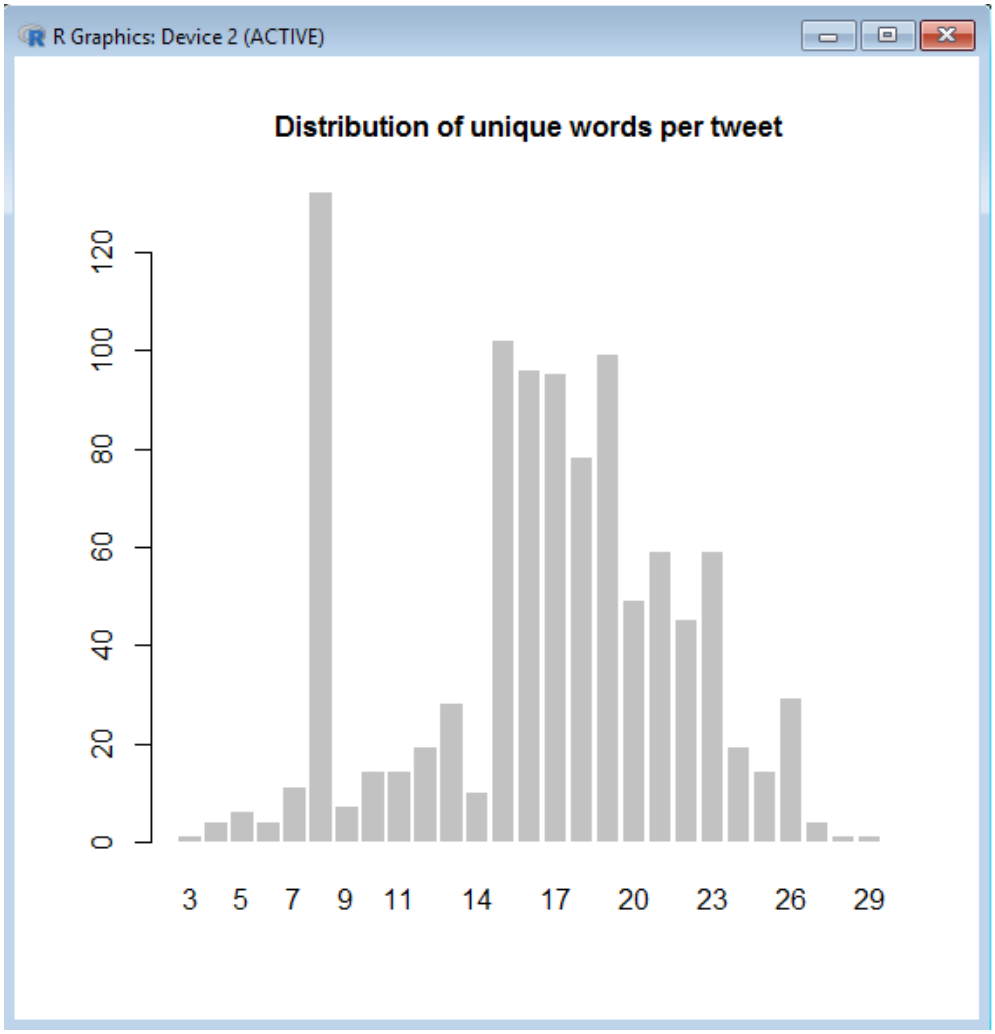
Στην παρακάτω εικόνα φαίνεται το αποτέλεσμα που προκύπτει μετά την εφαρμογή του κώδικα. Εξάγοντας τα tweets τα οποία περιέχουν το θέμα “talent show”, δημιουργήθηκε το γράφημα που δείχνει την κατανομή του μήκους των λέξεων, που προήλθαν από τα εξαγόμενα tweet, ανά tweet.



Συνεχίζοντας τον προηγούμενο κώδικα παράγονται τα παρακάτω.
Κώδικας(συνέχεια του προηγούμενου):

```
# how many unique words per tweet
uniq_words_per_tweet = sapply(words_list, function(x) length(unique(x)))
# barplot
barplot(table(uniq_words_per_tweet), border=NA,
   main="Distribution of unique words per tweet", cex.main=1)
```

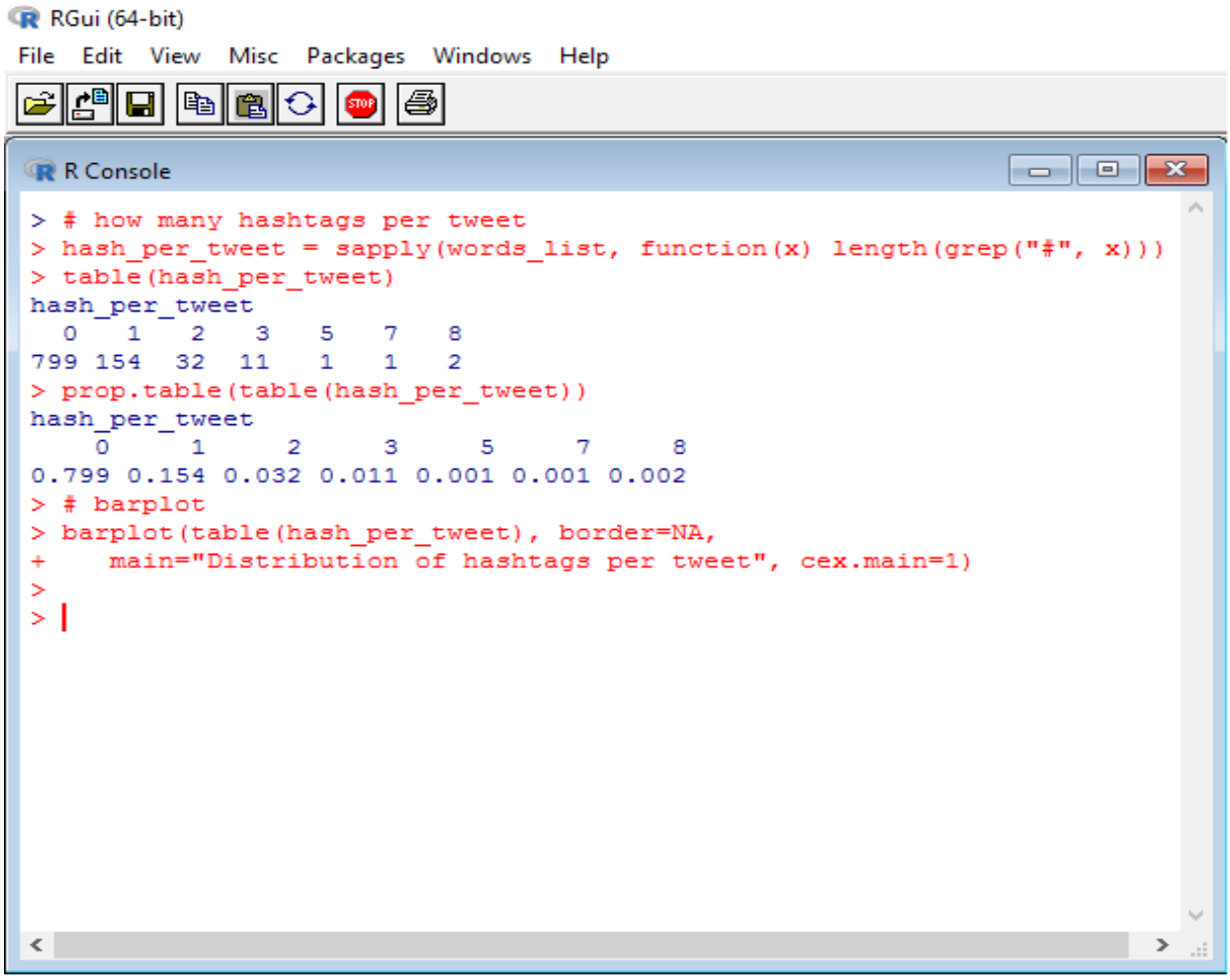
Στην παρακάτω εικόνα φαίνεται το αποτέλεσμα που προκύπτει μετά την εφαρμογή του κώδικα. Εξάγοντας τα tweets τα οποία περιέχουν το θέμα “talent show”, δημιουργήθηκε το γράφημα που δείχνει την κατανομή των μοναδικών λέξεων, που προήλθαν από τα εξαγόμενα tweet, ανά tweet.



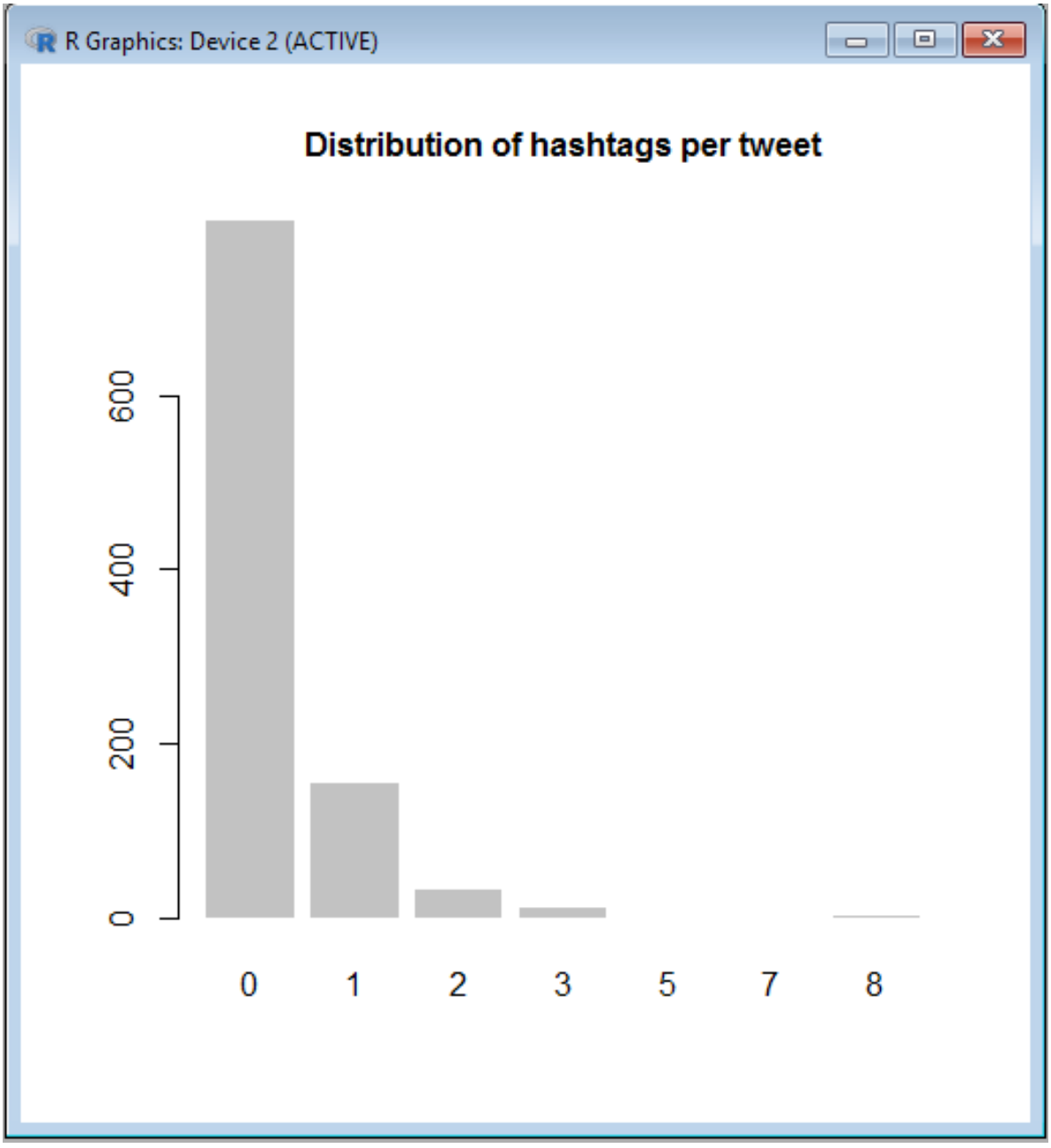
Συνεχίζοντας τον προηγούμενο κώδικα παράγονται τα παρακάτω.
Κώδικας(συνέχεια του προηγούμενου):

```
# how many hashtags per tweet
hash_per_tweet = sapply(words_list, function(x) length(grep("#", x)))
table(hash_per_tweet)
prop.table(table(hash_per_tweet))
# barplot
barplot(table(hash_per_tweet), border=NA,
  main="Distribution of hashtags per tweet", cex.main=1)
```

Στην παρακάτω εικόνα φαίνεται το αποτέλεσμα που προκύπτει μετά την εφαρμογή του κώδικα.



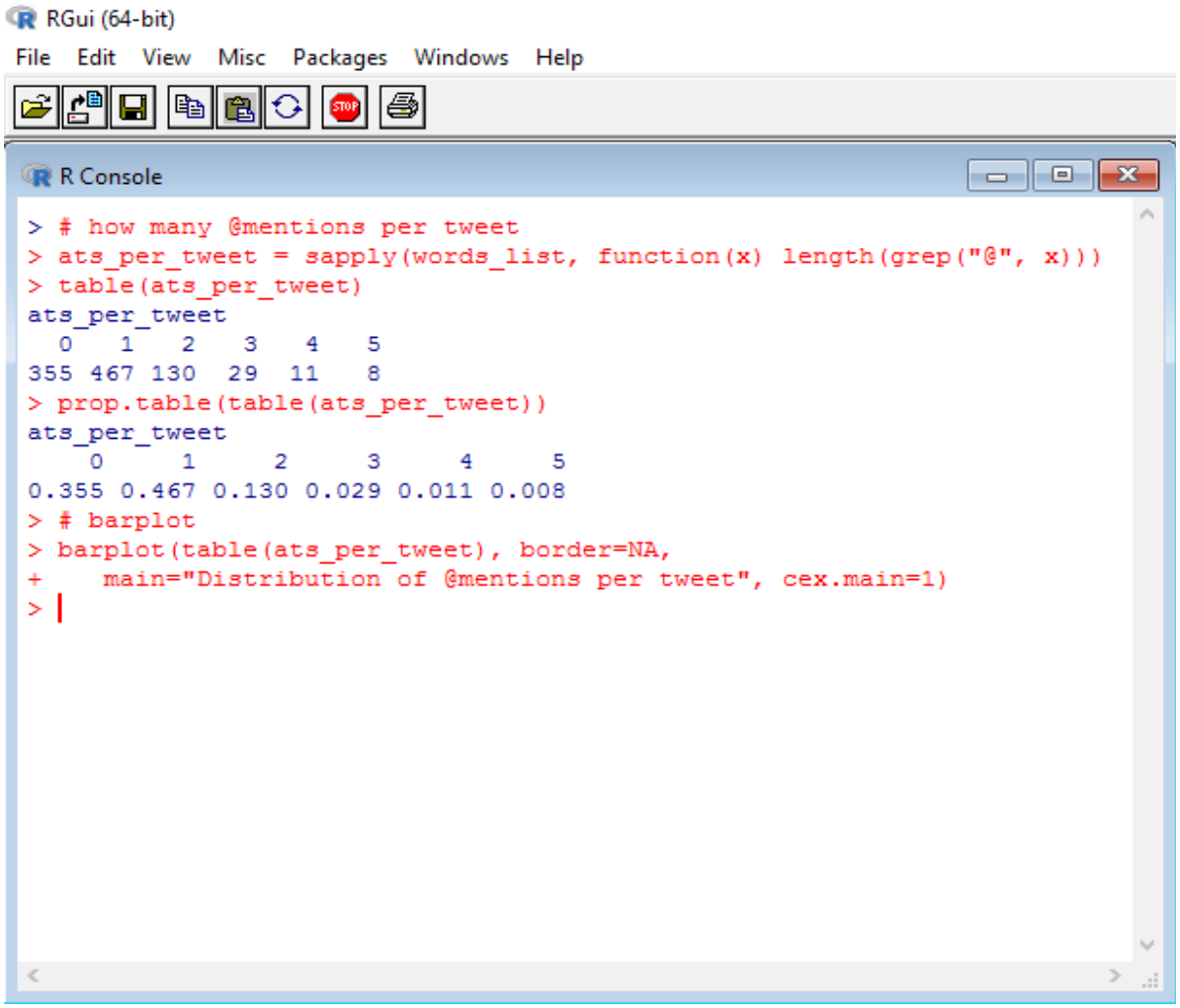
Εξάγοντας τα tweets τα οποία περιέχουν το θέμα “talent show”, δημιουργήθηκε το γράφημα που δείχνει την κατανομή των hashtags, που προήλθαν από τα εξαγόμενα tweet, ανά tweet.



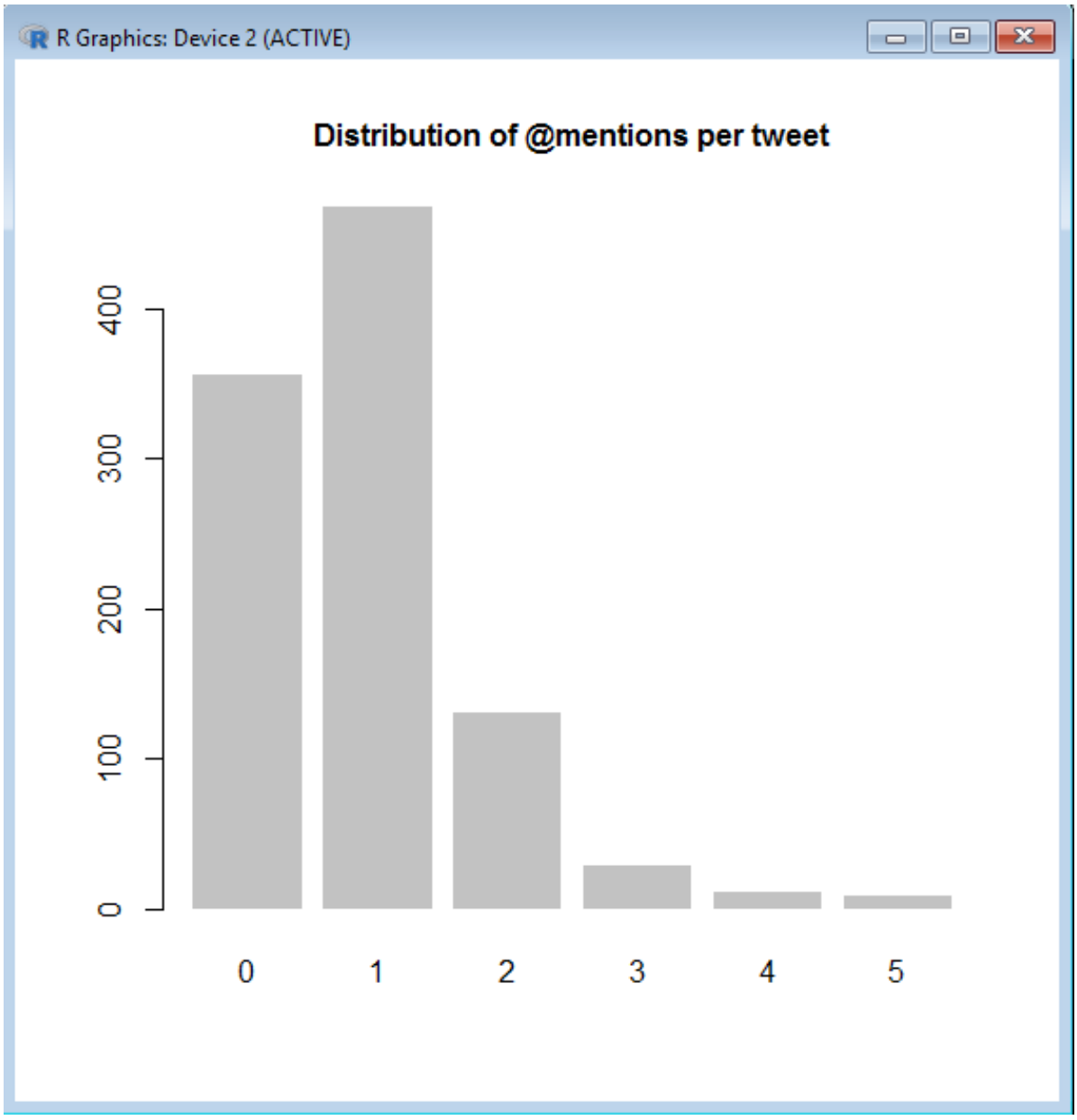
Συνεχίζοντας τον προηγούμενο κώδικα παράγονται τα παρακάτω.
Κώδικας(συνέχεια του προηγούμενου):

```
# how many @mentions per tweet
ats_per_tweet = sapply(words_list, function(x) length(grep("@", x)))
table(ats_per_tweet)
prop.table(table(ats_per_tweet))
# barplot
barplot(table(ats_per_tweet), border=NA,
  main="Distribution of @mentions per tweet", cex.main=1)
```

Στην παρακάτω εικόνα φαίνεται το αποτέλεσμα που προκύπτει μετά την εφαρμογή του κώδικα.



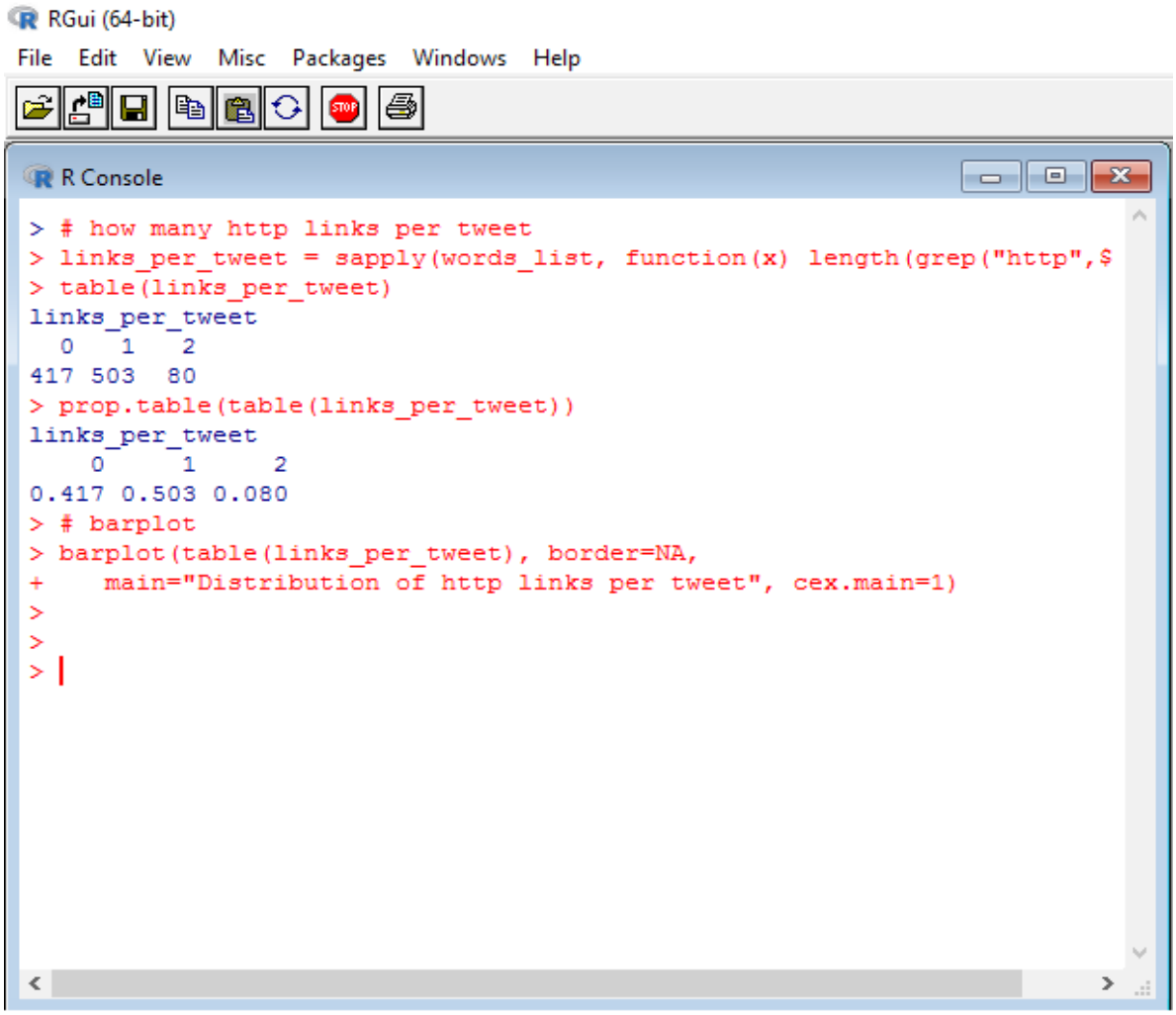
Εξάγοντας τα tweets τα οποία περιέχουν το θέμα “talent show”, δημιουργήθηκε το γράφημα που δείχνει την κατανομή των @mentions, που προήλθαν από τα εξαγόμενα tweet, ανά tweet.



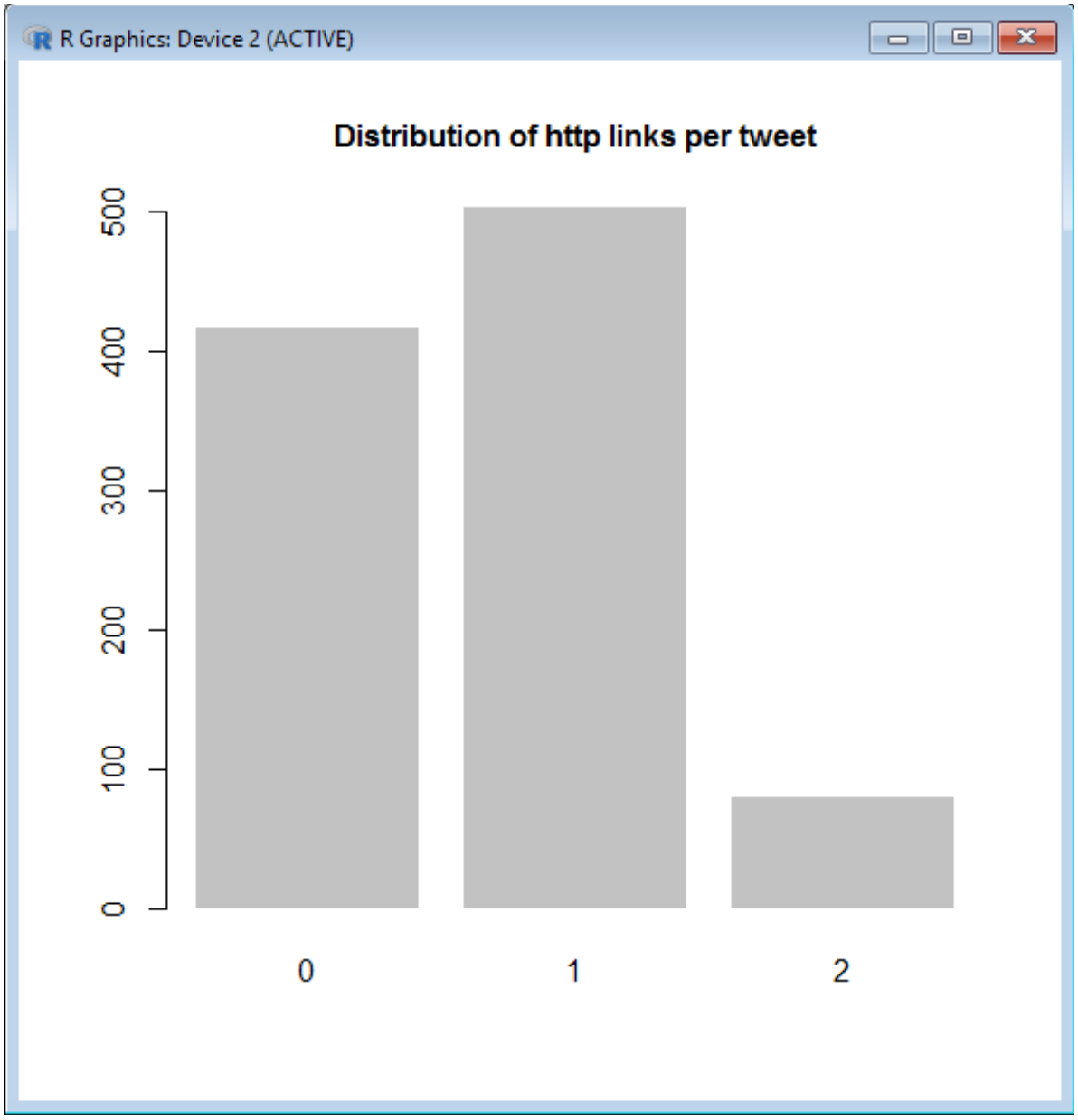
Συνεχίζοντας τον προηγούμενο κώδικα παράγονται τα παρακάτω.
Κώδικας(συνέχεια του προηγούμενου):

```
# how many http links per tweet
links_per_tweet = sapply(words_list, function(x) length(grep("http", x)))
table(links_per_tweet)
prop.table(table(links_per_tweet))
# barplot
barplot(table(links_per_tweet), border=NA,
  main="Distribution of http links per tweet", cex.main=1)
```

Στην παρακάτω εικόνα φαίνεται το αποτέλεσμα που προκύπτει μετά την εφαρμογή του κώδικα.



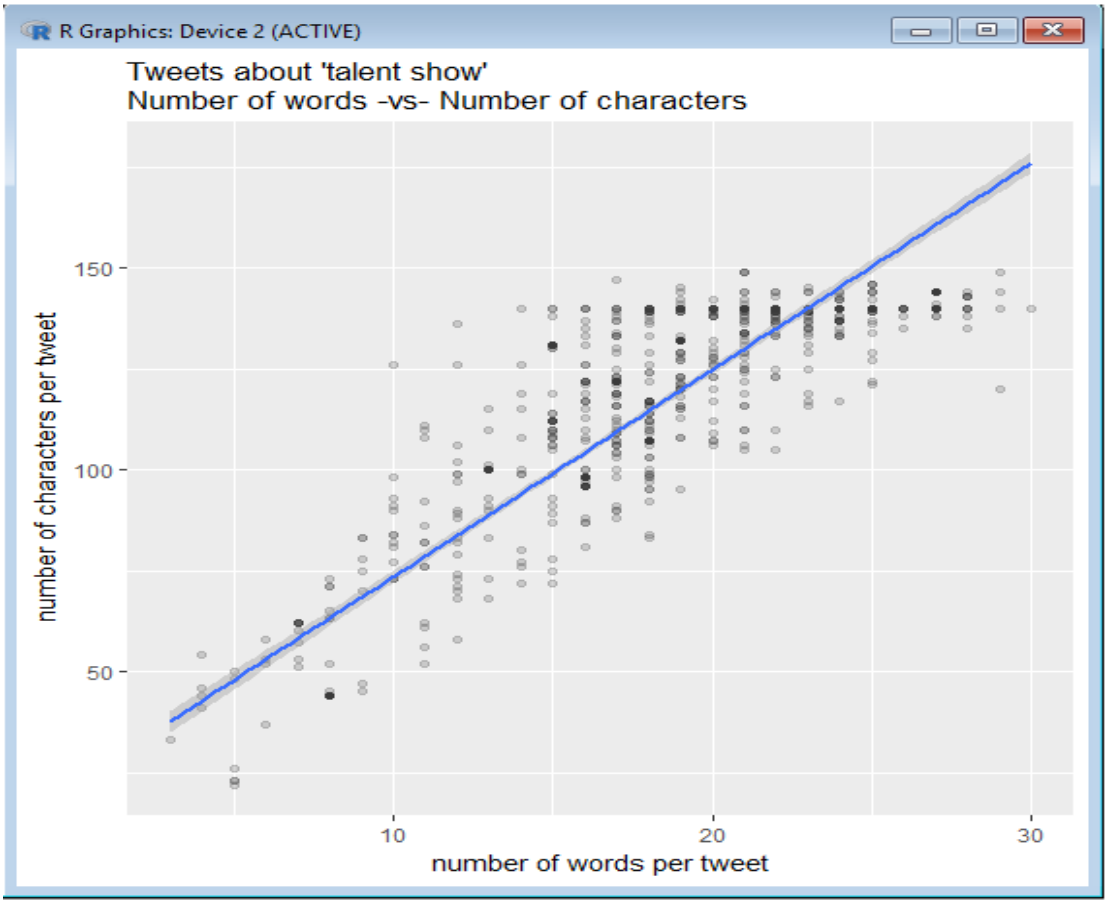
Εξάγοντας τα tweets τα οποία περιέχουν το θέμα “talent show”, δημιουργήθηκε το γράφημα που δείχνει την κατανομή των http links, που προήλθαν από τα εξαγόμενα tweet, ανά tweet.



Συνεχίζοντας τον προηγούμενο κώδικα παράγονται τα παρακάτω.
Κώδικας(συνέχεια του προηγούμενου):

```
# The more words in a tweet, the more characters per word
# words -vs- chars
ggplot(tsdf, aes(x=words, y=chars)) +
geom_point(colour="gray20", alpha=0.2) +
stat_smooth(method="lm") +
labs(x="number of words per tweet", y="number of characters per tweet") +
labs(title="Tweets about 'talent show' \nNumber of words -vs- Number of characters")
```

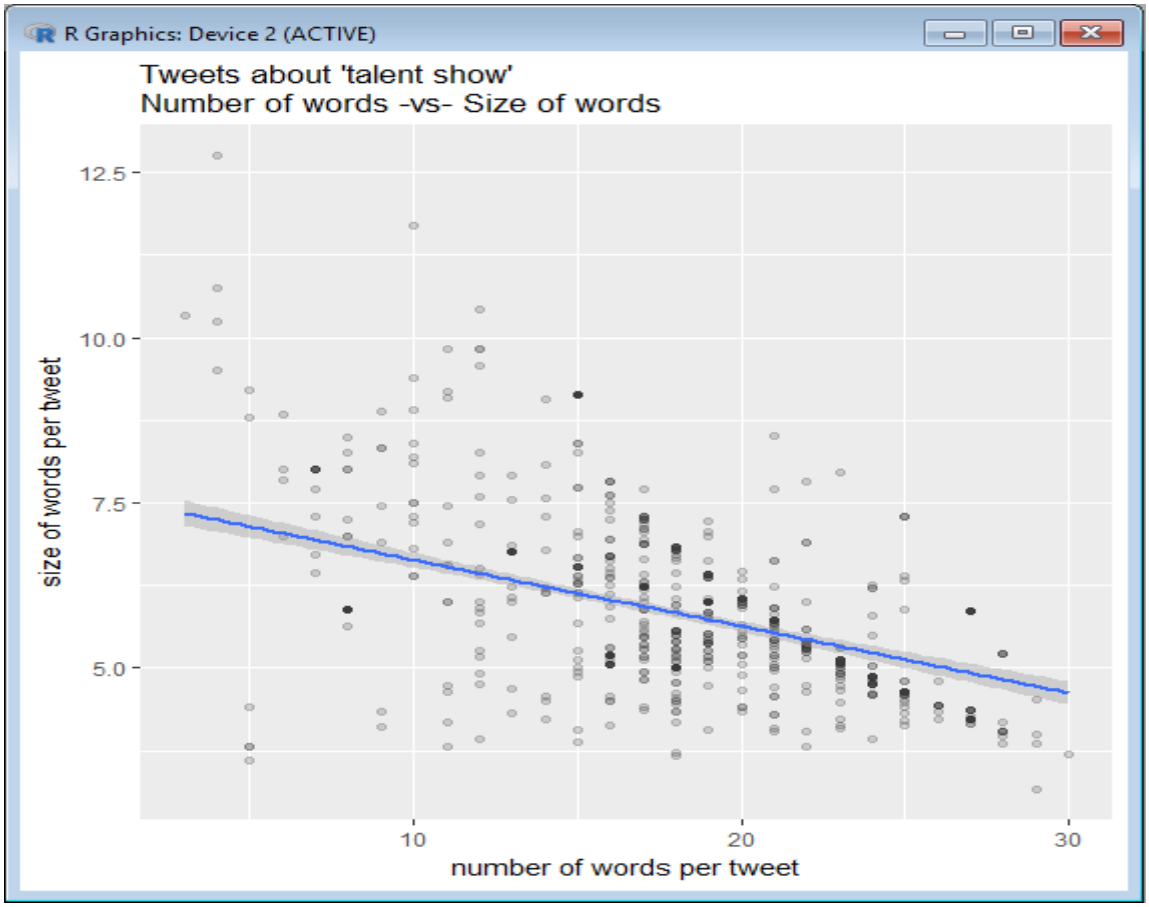
Στην παρακάτω εικόνα φαίνεται το αποτέλεσμα που προκύπτει μετά την εφαρμογή του κώδικα. Εξάγοντας τα tweets τα οποία περιέχουν το θέμα “talent show”, δημιουργήθηκε το γράφημα που δείχνει την κατανομή των λέξεων σε σχέση με τους χαρακτήρες μιας λέξης, που προήλθαν από τα εξαγόμενα tweet, ανά tweet. Στο γράφημα φαίνεται ότι όσο περισσότερες λέξεις σε ένα tweet, τόσο περισσότεροι χαρακτήρες σε μια λέξη.



Συνεχίζοντας τον προηγούμενο κώδικα παράγονται τα παρακάτω.
Κώδικας(συνέχεια του προηγούμενου):

```
# The more words in a tweet, the shorter the words
# words -vs- word length
ggplot(tsdf, aes(x=words, y=lengths)) +
geom_point(colour="gray20", alpha=0.2) +
stat_smooth(method="lm") +
labs(x="number of words per tweet", y="size of words per tweet") +
labs(title="Tweets about 'talent show' \nNumber of words -vs- Size of words")
```

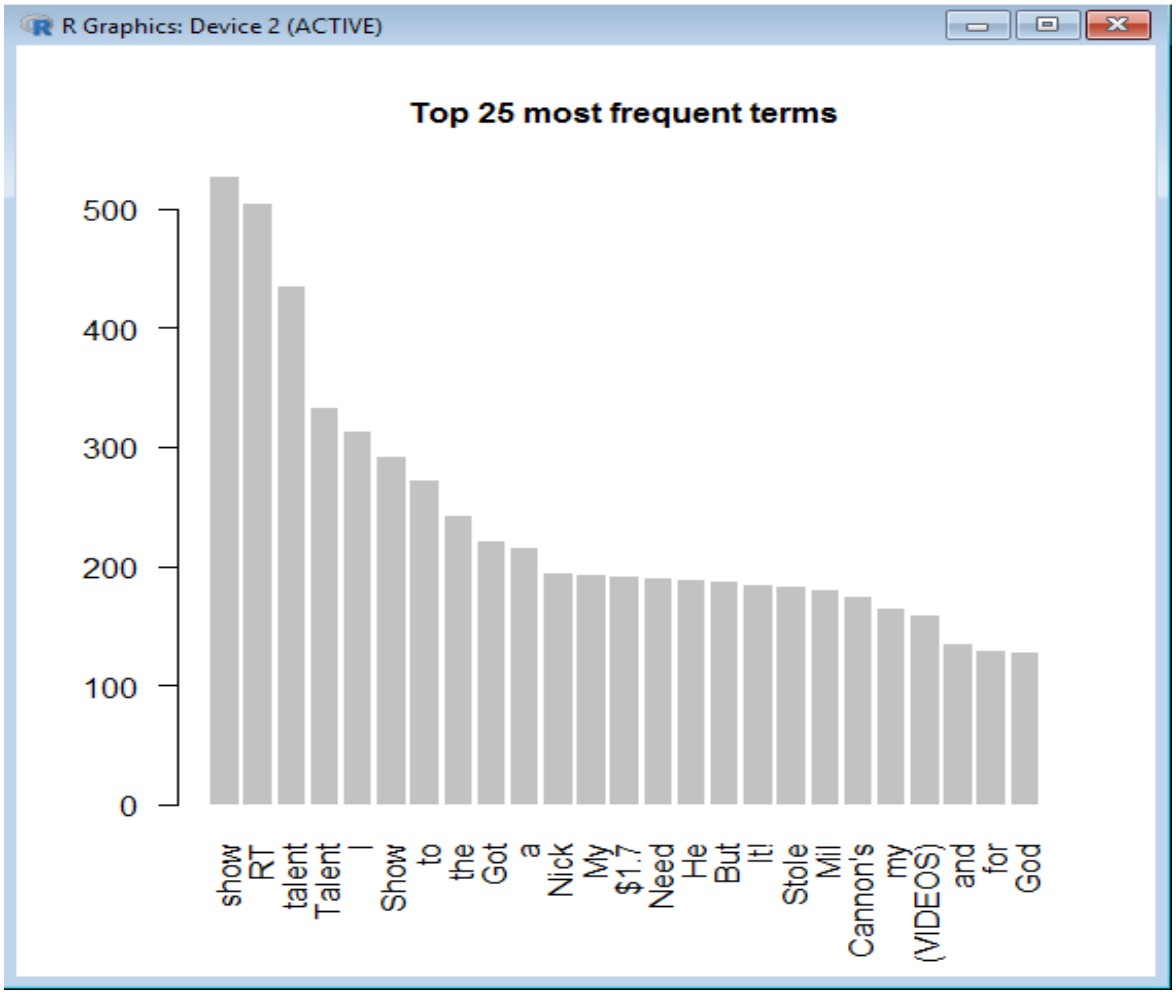
Στην παρακάτω εικόνα φαίνεται το αποτέλεσμα που προκύπτει μετά την εφαρμογή του κώδικα. Εξάγοντας τα tweets τα οποία περιέχουν το θέμα “talent show”, δημιουργήθηκε το γράφημα που δείχνει την κατανομή των λέξεων σε σχέση με το μέγεθός τους, που προήλθαν από τα εξαγόμενα tweet, ανά tweet. Στο γράφημα φαίνεται ότι όσο περισσότερες λέξεις σε ένα tweet, τόσο μικρότερες είναι οι λέξεις.



Συνεχίζοντας τον προηγούμενο κώδικα παράγονται τα παρακάτω.
Κώδικας(συνέχεια του προηγούμενου):

```
# most frequent words
mfw = sort(table(unlist(words_list)), decreasing=TRUE)
# top-25 most frequent
top25 = head(mfw, 25)
# barplot
barplot(top25, border=NA, las=2, main="Top 25 most frequent terms", cex.main=1)
```

Στην παρακάτω εικόνα φαίνεται το αποτέλεσμα που προκύπτει μετά την εφαρμογή του κώδικα. Εξάγοντας τα tweets τα οποία περιέχουν το θέμα “talent show”, δημιουργήθηκε το γράφημα που δείχνει τις 25 πιο συχνές λέξεις που εμφανίζονται στα εξαγόμενα tweets.



4.3. Ανάλυση των Tweets των Χρηστών

Ένα άλλο είδος ερωτήσεων κατά την εξέταση των δεδομένων από το twitter σχετίζεται με την αξιοποίηση οντοτήτων των tweet όπως @mentions, #hashtags, και συνδέσεις http, για παράδειγμα, ποιες οντότητες εμφανίζονται στα tweets ενός χρήστη; ποιες είναι οι οντότητες που εμφανίζονται πιο συχνά στα tweets ενός χρήστη; πόσα από τα tweets ενός χρήστη περιέχουν #hashtags; πόσα από τα tweets ενός χρήστη περιέχουν τουλάχιστον μία οντότητα;

Είναι ευρέως αποδεκτό ότι τα tweets που περιέχουν #hashtags είναι πιο πολύτιμα από αυτά που δεν έχουν επειδή κάποιος σκόπιμα ενσωματώνει συγκεντρωτικά πληροφορίες σε αυτά τα tweets. Εφαρμόζοντας τις έννοιες της ανάλυσης συχνότητας, μπορούν να υπολογιστούν πράγματα όπως ο μέσος αριθμός των hashtags ανά tweet ή το μέσο μήκος των hashtags. Θα υλοποιηθεί ένα απλό παράδειγμα, σχετικά με τα tweets τριών χρηστών και συγκεκριμένα ποια #hashtags χρησιμοποιούν, στο οποίο θα χρησιμοποιηθούν και πάλι wordclouds.

Στο παράδειγμα θα αναλυθούν τα tweets από τους λογαριασμούς τριών καναλιών της τηλεόρασης:

(1) @ALPHA_TV
(2) @StarChannelGr
(3) @ANT1TV

Κώδικας:

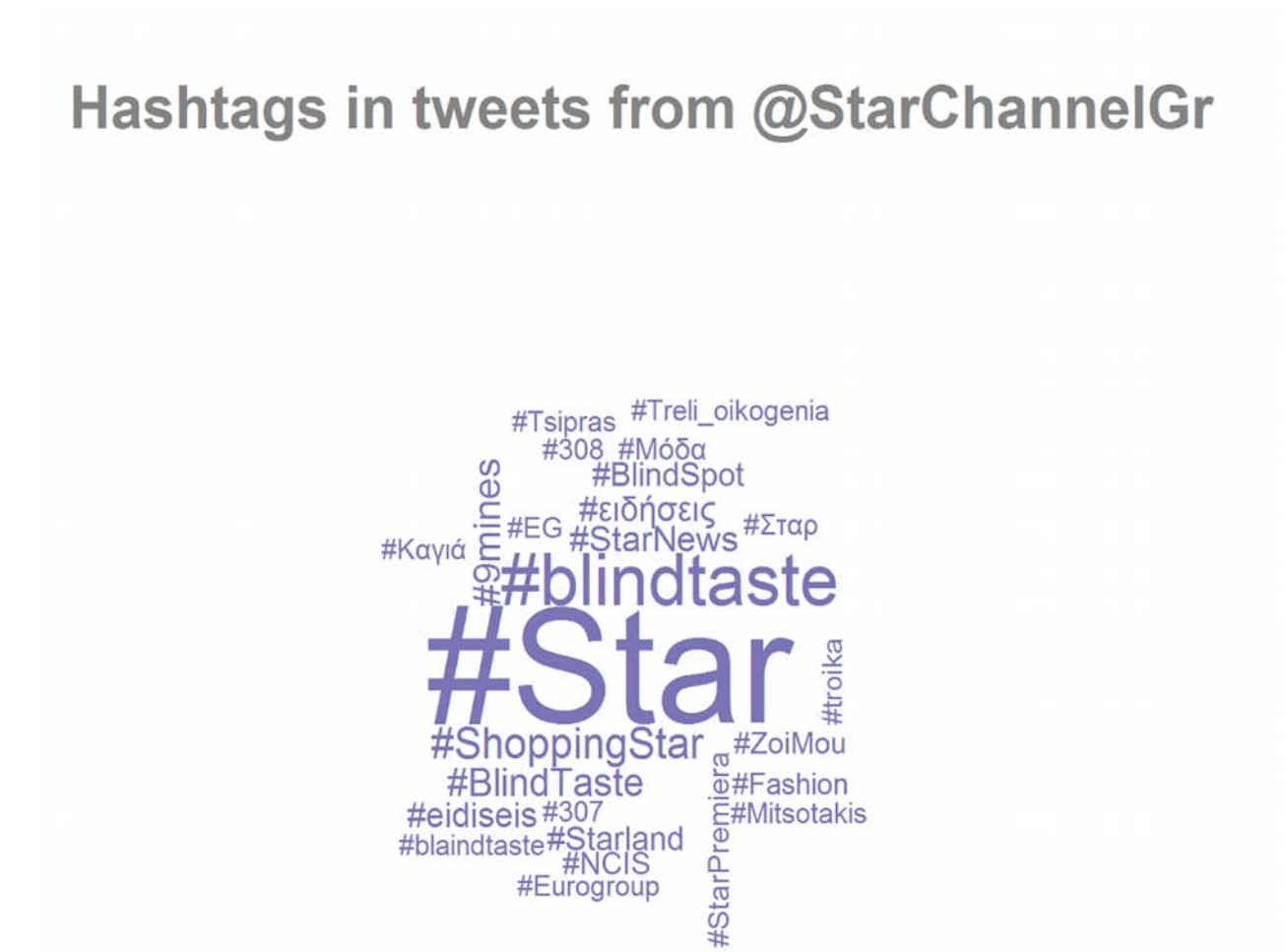
```
# Load the required packages
library(twitteR)
library(tm)
library(stringr)
library(wordcloud)

# harvest tweets from each user
alpha_tweets = userTimeline("ALPHA_TV", n=500)
star_tweets = userTimeline("StarChannelGr", n=500)
ant1_tweets = userTimeline("ANT1TV", n=500)
```

```
# dump tweets information into data frames
alpha_df = twListToDF(alpha_tweets)
star_df = twListToDF(star_tweets)
ant1_df = twListToDF(ant1_tweets)

# get the hashtags
alpha_hashtags = str_extract_all(alpha_df$text, "#\\w+")
star_hashtags = str_extract_all(star_df$text, "#\\w+")
ant1_hashtags = str_extract_all(ant1_df$text, "#\\w+")

# put tags in vector
alpha_hashtags = unlist(alpha_hashtags)
star_hashtags = unlist(star_hashtags)
ant1_hashtags = unlist(ant1_hashtags)

# calculate hashtag frequencies
alpha_tags_freq = table(alpha_hashtags)
star_tags_freq = table(star_hashtags)
ant1_tags_freq = table(ant1_hashtags)

# put all tags in a single vector
all_tags = c(alpha_tags_freq, star_tags_freq, ant1_tags_freq)

# Let's plot wordclouds for each user
# ALPHA_TV hashtags wordcloud
wordcloud(names(alpha_tags_freq), alpha_tags_freq, random.order=FALSE,
   colors="#1B9E77")
title("\n\nHashtags in tweets from @ALPHA_TV",
   cex.main=1.5, col.main="gray50")

png("AlphaTvCloud.png", width=12, height=8, units="in", res=300)
wordcloud(names(alpha_tags_freq), alpha_tags_freq, random.order=FALSE,
   colors="#1B9E77")
title("\n\nHashtags in tweets from @ALPHA_TV",
   cex.main=1.5, col.main="gray50")
dev.off()
```

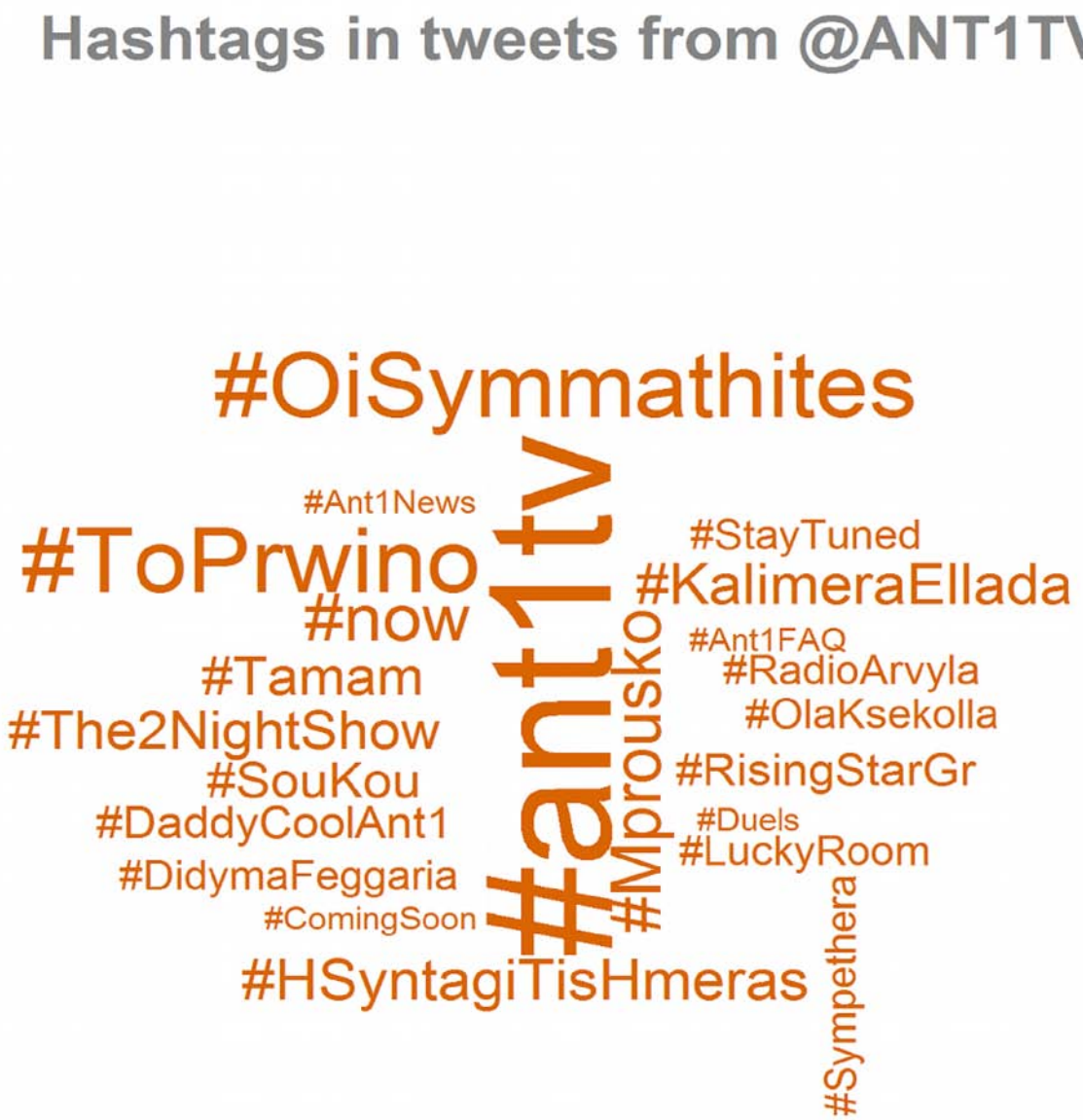
```
# StarChannelGr hashtags wordcloud
wordcloud(names(star_tags_freq), star_tags_freq + 7, random.order=FALSE,
   colors="#7570B3")
title("\nHashtags in tweets from @StarChannelGr",
   cex.main=1.5, col.main="gray50")

png("StarChannelGrCloud.png", width=12, height=8, units="in", res=300)
wordcloud(names(star_tags_freq), star_tags_freq + 7, random.order=FALSE,
   colors="#7570B3")
title("\nHashtags in tweets from @StarChannelGr",
   cex.main=1.5, col.main="gray50")
dev.off()


# ANT1TV hashtags wordcloud
wordcloud(names(ant1_tags_freq), ant1_tags_freq, random.order=FALSE,
   colors="#D95F02")
title("\n\nHashtags in tweets from @ANT1TV",
   cex.main=1.5, col.main="gray50")

png("Ant1TvCloud.png", width=12, height=8, units="in", res=300)
wordcloud(names(ant1_tags_freq), ant1_tags_freq, random.order=FALSE,
   colors="#D95F02")
title("\n\nHashtags in tweets from @ANT1TV",
   cex.main=1.5, col.main="gray50")
dev.off()
```

Στις παρακάτω εικόνες φαίνεται το αποτέλεσμα που προκύπτει μετά την εφαρμογή του κώδικα. Εξάγοντας τα 500 πιο πρόσφατα tweets από τον καθένα από τους χρήστες του Twitter @ALPHA_TV, @StarChannelGr, @ANT1TV, που είναι κανάλια της τηλεόρασης, δημιουργήθηκαν τα παρακάτω wordcloud που αποτελούνται από τα συχνότερα #hashtags που χρησιμοποιούν στα tweets τους. Τα wordcloud που δημιουργήθηκαν, αποθηκεύτηκαν σε μορφή εικόνας (png format).

Wordcloud for @ALPHA_TV



Wordcloud for @StarChannelGr



Wordcloud for @ANT1TV

Συνεχίζοντας τον προηγούμενο κώδικα παράγονται τα παρακάτω.
Κώδικας(συνέχεια του προηγούμενου):

```
# Now let's plot one single wordcloud
# vector of colors
cols = c(
   rep("#1B9E77", length(alpha_tags_freq)),
   rep("#7570B3", length(star_tags_freq)),
   rep("#D95F02", length(ant1_tags_freq))
)

# wordcloud
wordcloud(names(all_tags), all_tags, random.order=FALSE, min.freq=1,
   colors=cols, ordered.colors=TRUE)
mtext(c("@ALPHA_TV", "@StarChannelGr", "@ANT1TV"), side=3,
   line=2, at=c(0.20, 0.54, 0.85), col=c("#1B9E77", "#7570B3", "#D95F02"),
   family="serif", font=2, cex=1.5)

png("AllChannelsCloud.png", width=12, height=8, units="in", res=300)
wordcloud(names(all_tags), all_tags, random.order=FALSE, min.freq=1,
   colors=cols, ordered.colors=TRUE)
mtext(c("@ALPHA_TV", "@StarChannelGr", "@ANT1TV"), side=3,
   line=2, at=c(0.20, 0.54, 0.85), col=c("#1B9E77", "#7570B3", "#D95F02"),
   family="serif", font=2, cex=1.5)
dev.off()
```

Στην παρακάτω εικόνα φαίνεται το αποτέλεσμα που προκύπτει μετά την εφαρμογή του κώδικα. Εξάγοντας τα 500 πιο πρόσφατα tweets από τον καθένα από τους χρήστες του Twitter @ALPHA_TV, @StarChannelGr, @ANT1TV, που είναι κανάλια της τηλεόρασης, δημιουργήθηκε το παρακάτω wordcloud που αποτελείται από τα συχνότερα #hashtags, που προέρχονται από όλα τα κανάλια, που χρησιμοποιούν στα tweets τους. Το wordcloud που δημιουργήθηκε, αποθηκεύτηκε σε μορφή εικόνας (png format).

@ALPHA_TV @StarChannelGr @ANT1TV

Wordcloud for all the Channels

4.4. Ανάλυση Συναισθήματος Ανθρώπων

Η ανάλυση συναισθήματος, που αναφέρεται επίσης ως εξόρυξη γνώμης, συνεπάγεται την εξόρυξη απόψεων και των συναισθημάτων στο κείμενο. Μία από τις πιο κοινές εφαρμογές της ανάλυσης συναισθήματος είναι ο εντοπισμός των συμπεριφορών και των συναισθημάτων στο διαδίκτυο, ειδικά όσον αφορά τη στερέωση των προϊόντων, των υπηρεσιών, των εμπορικών σημάτων ή ακόμη και των ανθρώπων. Η κύρια ιδέα είναι να καθοριστεί εάν αντιμετωπίζονται θετικά ή αρνητικά από ένα δεδομένο κοινό. Μια ενδιαφέρουσα επιλογή που μπορεί να χρησιμοποιηθεί για να πραγματοποιηθεί ανάλυση συναισθήματος είναι με τη χρήση του πακέτου sentiment της R. Αυτό το πακέτο περιέχει δύο χρήσιμες συναρτήσεις που είναι οι εξής:

classify_emotion: Η συνάρτηση αυτή βοηθά να αναλυθεί κάποιο κείμενο και να κατηγοριοποιηθεί σε διαφορετικούς τύπους συναισθημάτων όπως θυμός, αηδία, φόβος, χαρά, λύπη και έκπληξη. Η ταξινόμηση μπορεί να γίνει με τη χρήση δύο αλγορίθμων, ο ένας είναι ο ταξινομητής Bayes λειτουργώντας πάνω στο “λεξικό συναισθημάτων” των Carlo Strapparava και Alessandro Valitutti. Ο άλλος είναι μια απλή διαδικασία ψηφοφορίας.

classify_polarity: Σε αντίθεση με την ταξινόμηση των συναισθημάτων, η συνάρτηση classify_polarity επιτρέπει την ταξινόμηση κάποιου κειμένου ως θετικό ή αρνητικό. Σε αυτήν την περίπτωση, η ταξινόμηση μπορεί να γίνει με τη χρήση του αλγόριθμου του Bayes λειτουργώντας πάνω στο λεξικό υποκειμενικότητας του Janyce Wiebe ή με ένα απλό αλγόριθμο ψηφοφορίας.

Θα υλοποιηθεί ένα απλό παράδειγμα, σχετικά με τα tweets που χρησιμοποιούν τον όρο “capital controls”.

Κώδικας:

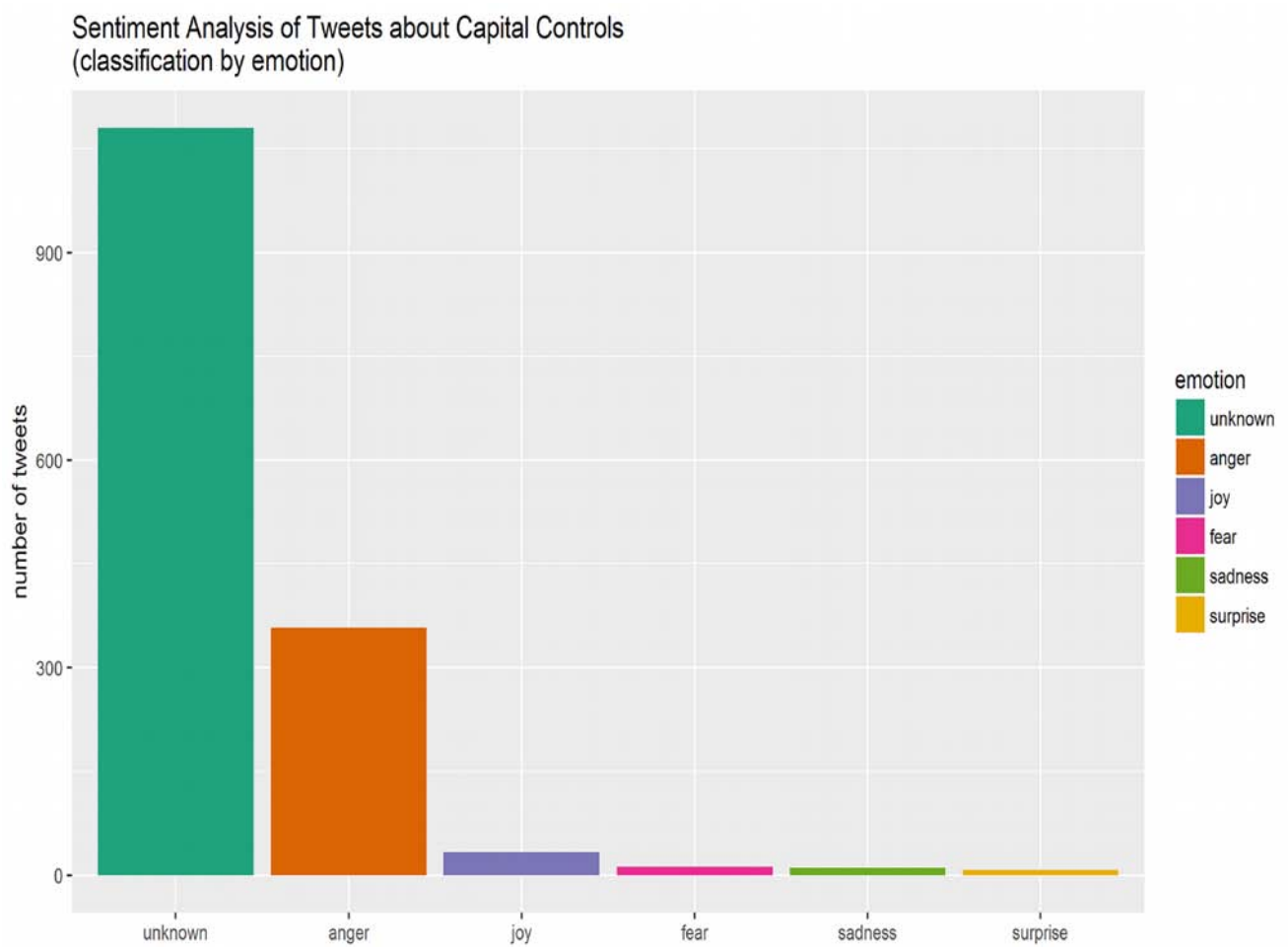
```
# Load the necessary packages
library(twitteR)
library(sentiment)
library(plyr)
library(ggplot2)
library(wordcloud)
library(RColorBrewer)
```

```
# Let's collect some tweets containing the term "capital controls"
# harvest some tweets
some_tweets = searchTwitter("capital controls", n=1500, lang="en")

# get the text
some_txt = sapply(some_tweets, function(x) x$getText())

# Prepare the text for sentiment analysis
# remove retweet entities
some_txt = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", some_txt)
# remove at people
some_txt = gsub("@\\w+", "", some_txt)
# remove punctuation
some_txt = gsub("[[:punct:]]", "", some_txt)
# remove numbers
some_txt = gsub("[[:digit:]]", "", some_txt)
# remove html links
some_txt = gsub("http\\w+", "", some_txt)
# remove unnecessary spaces
some_txt = gsub("[ \t]{2,}", "", some_txt)
some_txt = gsub("^\\s+|\\s+$", "", some_txt)

# define "tolower error handling" function
try.error = function(x)
{
  # create missing value
  y = NA
  # tryCatch error
  try_error = tryCatch(tolower(x), error=function(e) e)
  # if not an error
  if (!inherits(try_error, "error"))
  y = tolower(x)
  # result
  return(y)
}
# lower case using try.error with sapply
```

```
some_txt = sapply(some_txt, try.error)

# remove NAs in some_txt
some_txt = some_txt[!is.na(some_txt)]
names(some_txt) = NULL

# Perform Sentiment Analysis
# classify emotion
class_emo = classify_emotion(some_txt, algorithm="bayes", prior=1.0)
# get emotion best fit
emotion = class_emo[,7]
# substitute NA's by "unknown"
emotion[is.na(emotion)] = "unknown"

# classify polarity
class_pol = classify_polarity(some_txt, algorithm="bayes")
# get polarity best fit
polarity = class_pol[,4]

# Create data frame with the results and obtain some general statistics
# data frame with results
sent_df = data.frame(text=some_txt, emotion=emotion,
polarity=polarity, stringsAsFactors=FALSE)

# sort data frame
sent_df = within(sent_df,
  emotion <- factor(emotion, levels=names(sort(table(emotion), decreasing=TRUE))))

# Let's do some plots of the obtained results
# plot distribution of emotions
ggplot(sent_df, aes(x=emotion)) +
geom_bar(aes(y=..count.., fill=emotion)) +
scale_fill_brewer(palette="Dark2") +
labs(x="emotion categories", y="number of tweets") +
labs(title = "Sentiment Analysis of Tweets about Capital Controls\n(classification by
emotion)")
```

```
# save the image in png format
png("Capital_Controls_plot.png", width=10, height=6, units="in", res=300)
ggplot(sent_df, aes(x=emotion)) +
geom_bar(aes(y=..count.., fill=emotion)) +
scale_fill_brewer(palette="Dark2") +
labs(x="emotion categories", y="number of tweets") +
labs(title = "Sentiment Analysis of Tweets about Capital Controls\n(classification by emotion)")
dev.off()
```

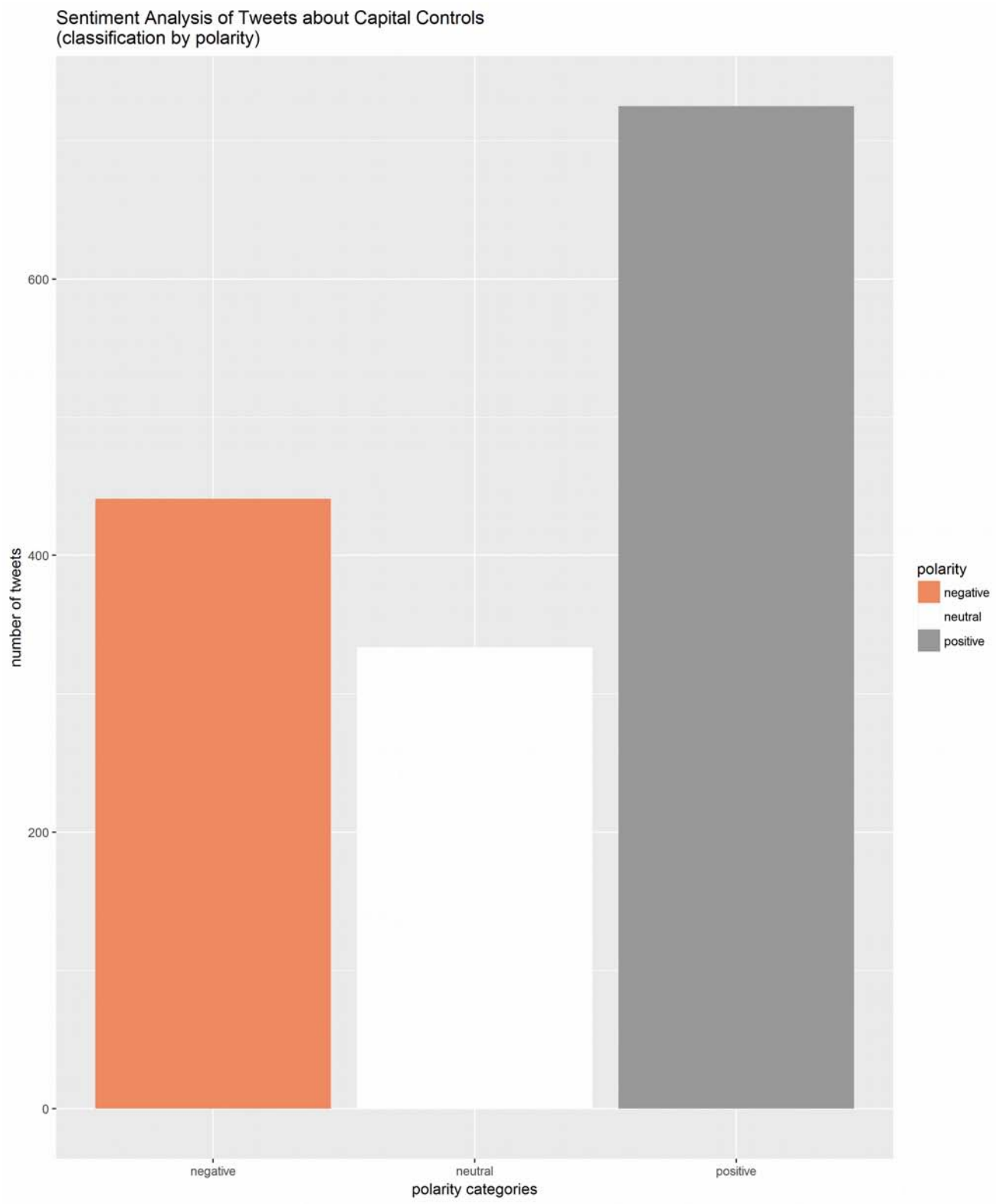
Στην παρακάτω εικόνα φαίνεται το αποτέλεσμα που προκύπτει μετά την εφαρμογή του κώδικα, εξάγοντας 1500 tweets που χρησιμοποιούν τον όρο “capital controls”. Μετά τη χρήση της ανάλυσης συναισθήματος κατηγοριοποιήθηκαν σύμφωνα με τα συναισθήματα των ανθρώπων. Το γράφημα που δημιουργήθηκε, αποθηκεύτηκε σε μορφή εικόνας (png format).



Συνεχίζοντας τον προηγούμενο κώδικα παράγονται τα παρακάτω.
Κώδικας(συνέχεια του προηγούμενου):

```
# plot distribution of polarity
ggplot(sent_df, aes(x=polarity)) +
geom_bar(aes(y=..count.., fill=polarity)) +
scale_fill_brewer(palette="RdGy") +
labs(x="polarity categories", y="number of tweets") +
labs(title = "Sentiment Analysis of Tweets about Capital Controls\n(classification by emotion)")

# save the image in png format
png("Capital_Controls_classification.png", width=10, height=12, units="in", res=300)
ggplot(sent_df, aes(x=polarity)) +
geom_bar(aes(y=..count.., fill=polarity)) +
scale_fill_brewer(palette="RdGy") +
labs(x="polarity categories", y="number of tweets") +
labs(title = "Sentiment Analysis of Tweets about Capital Controls\n(classification by emotion)")
dev.off()
```

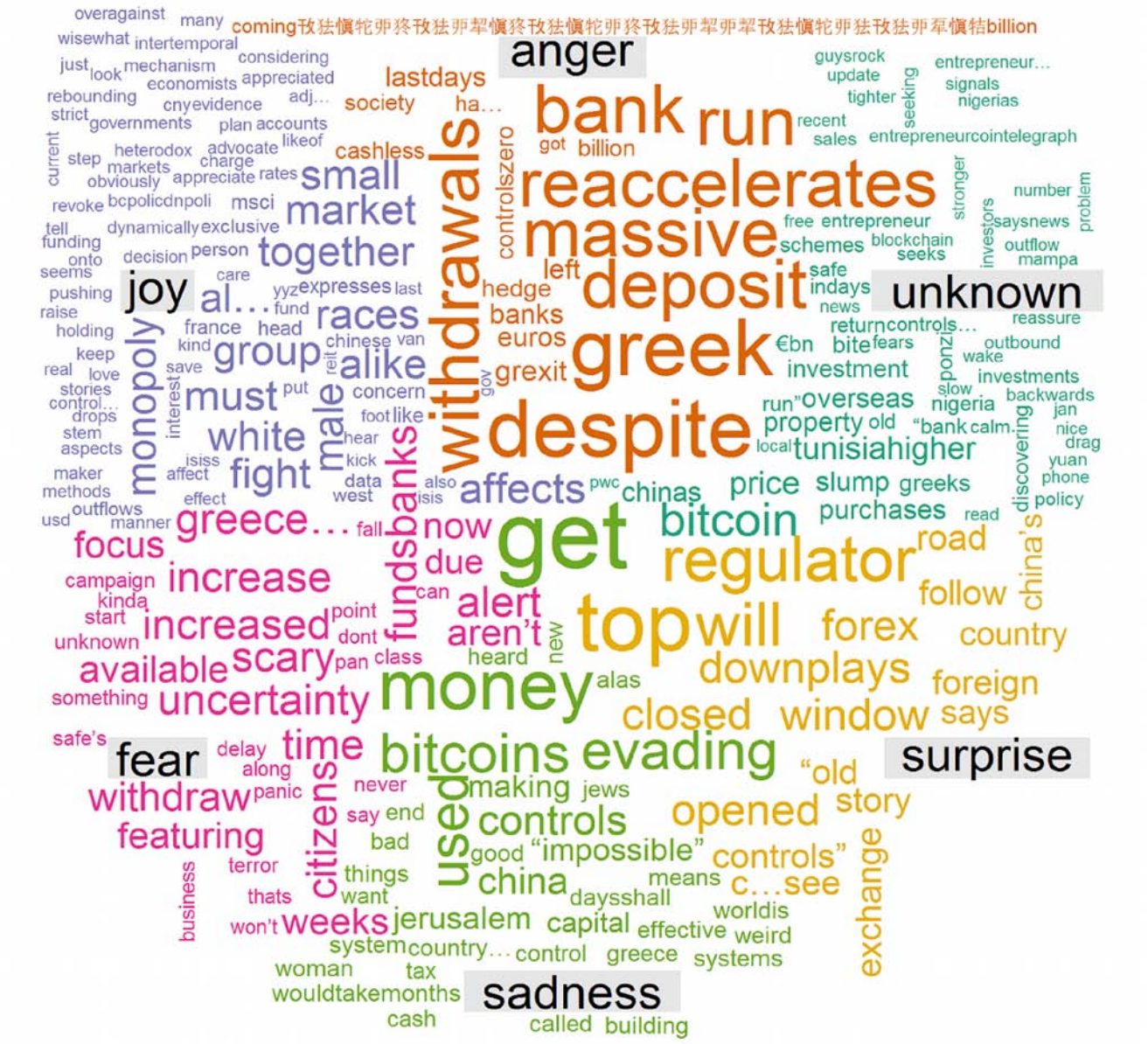
Στην παρακάτω εικόνα φαίνεται το αποτέλεσμα που προκύπτει μετά την εφαρμογή του κώδικα, εξάγοντας 1500 tweets που χρησιμοποιούν τον όρο “capital controls”. Μετά τη χρήση της ανάλυσης συναισθήματος κατηγοριοποιήθηκαν σύμφωνα με την πολικότητα σε θετικά, αρνητικά και ουδέτερα. Το γράφημα που δημιουργήθηκε, αποθηκεύτηκε σε μορφή εικόνας (png format).


Sentiment Analysis of Tweets about Capital Controls
(classification by polarity)
number of tweets
600
400
200
0
negative
neutral
positive
polarity categories
polarity
negative
neutral
positive

Συνεχίζοντας τον προηγούμενο κώδικα παράγονται τα παρακάτω.
Κώδικας(συνέχεια του προηγούμενου):

```
# Separate the text by emotions and visualize the words with a comparison cloud
# separating text by emotion
emos = levels(factor(sent_df$emotion))
nemo = length(emos)
emo.docs = rep("", nemo)
for (i in 1:nemo)
{
  tmp = some_txt[emotion == emos[i]]
  emo.docs[i] = paste(tmp, collapse=" ")
}

# remove stopwords
emo.docs = removeWords(emo.docs, stopwords("english"))
# create corpus
corpus = Corpus(VectorSource(emo.docs))
tdm = TermDocumentMatrix(corpus)
tdm = as.matrix(tdm)
colnames(tdm) = emos

# comparison word cloud
comparison.cloud(tdm, colors = brewer.pal(nemo, "Dark2"),
  scale = c(3,.5), random.order = FALSE, title.size = 1.5)

# save the image in png format
png("CapitalControlsCloud.png", width=8, height=6, units="in", res=300)
comparison.cloud(tdm, colors = brewer.pal(nemo, "Dark2"),
  scale = c(3,.5), random.order = FALSE, title.size = 1.5)
dev.off()
```

Στην παρακάτω εικόνα φαίνεται το αποτέλεσμα που προκύπτει μετά την εφαρμογή του κώδικα, εξάγοντας 1500 tweets που χρησιμοποιούν τον όρο “capital controls”. Μετά τη χρήση της ανάλυσης συναισθήματος, δημιουργήθηκε το παρακάτω comparison wordcloud που αποτελείται από τις λέξεις των tweets οι οποίες έχουν χωριστεί ανά συναίσθημα. Το wordcloud που δημιουργήθηκε, αποθηκεύτηκε σε μορφή εικόνας (png format).



Wordcloud for Capital Controls

4.5. Συσταδοποίηση Δεδομένων

Υπάρχουν πολλά παραδείγματα εξόρυξης δεδομένων στα οποία μπορούν να χρησιμοποιηθούν οι αλγόριθμοι συσταδοποίησης όπως έχει αναφερθεί σε προηγούμενο κεφάλαιο. Οπότε θα πρέπει να επιλεχθεί κάποιο έτσι ώστε να γίνει ορατή η χρησιμότητα των αλγόριθμων αυτών στην εξόρυξη δεδομένων από το Twitter. Στο παρακάτω παράδειγμα θα γίνει εξόρυξη των tweets ενός συγκεκριμένου χρήστη και έπειτα ανάλυσή τους με τη χρήση αλγορίθμων συσταδοποίησης. Ο χρήστης που θα χρησιμοποιηθεί είναι ο @BillGates, ο οποίος είναι ένας Αμερικανός επιχειρηματίας, προγραμματιστής υπολογιστών, και εφευρέτης.

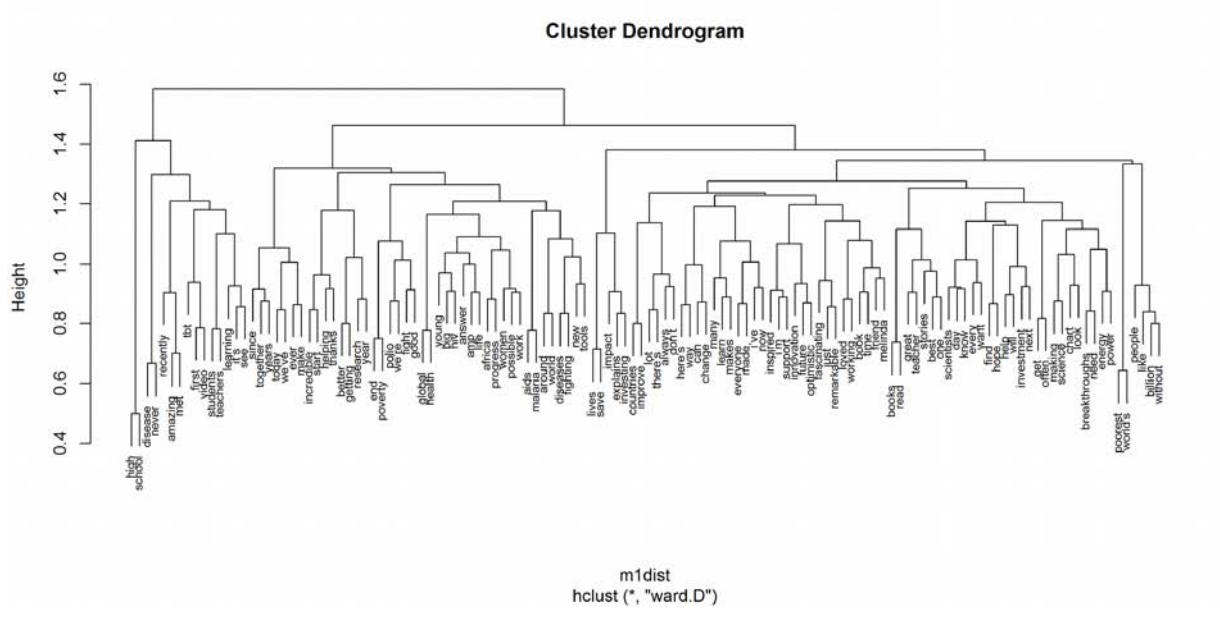
Κώδικας:

```
# Load the necessary packages
library(twitteR)
library(tm)
library(cluster)
library(FactoMineR)
library(RColorBrewer)
library(ggplot2)

# harvest tweets from BillGates
bg_tweets = userTimeline("BillGates", n=500)

# dump tweets information into a data frame
bg_df = twListToDF(bg_tweets)

# get the text
bg_txt = bg_df$text

# Let's do some text cleaning
# remove retweet entities
bg_clean = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", bg_txt)
# remove Atpeople
bg_clean = gsub("@\\w+", "", bg_clean)
# remove punctuation symbols
```

```
bg_clean = gsub("[[:punct:]]", "", bg_clean)
# remove numbers
bg_clean = gsub("[[:digit:]]", "", bg_clean)
# remove links
bg_clean = gsub("http\\w+", "", bg_clean)

# Create Corpus, apply transformations, and get term-document matrix
# corpus
bg_corpus = Corpus(VectorSource(bg_clean))

# remove stoprwords
bg_stopwords = unique(c(stopwords("english"), "billgates"))
trans = list(weighting=weightTf, stopwords=bg_stopwords)

# remove extra white-spaces
bg_corpus = tm_map(bg_corpus, stripWhitespace)

# convert to lower case
bg_corpus <- tm_map(bg_corpus, content_transformer(tolower))

# term-document matrix
tdm = TermDocumentMatrix(bg_corpus, control=trans)

# convert as matrix
m = as.matrix(tdm)

# We need to keep most frequent terms
# For instance, let's keep those words that have a frequency > 90 percentile
# remove sparse terms (word frequency > 90% percentile)
wf = rowSums(m)
m1 = m[wf>quantile(wf,probs=0.9), ]

# remove columns with all zeros
m1 = m1[,colSums(m1)!=0]

# for convenience, every matrix entry must be binary (0 or 1)
m1[m1 > 1] = 1
```

```
# Let's keep exploring by applying a cluster analysis
# This will let us discover more about groups of words
# distance matrix with binary distance
m1dist = dist(m1, method="binary")

# cluster with ward method
clus1 = hclust(m1dist, method="ward")

# plot dendrogram
plot(clus1, cex=0.7)

# save the image in png format
png("BillGatesDendrogram.png", width=12, height=6, units="in", res=300)
plot(clus1, cex=0.7)
dev.off()
```

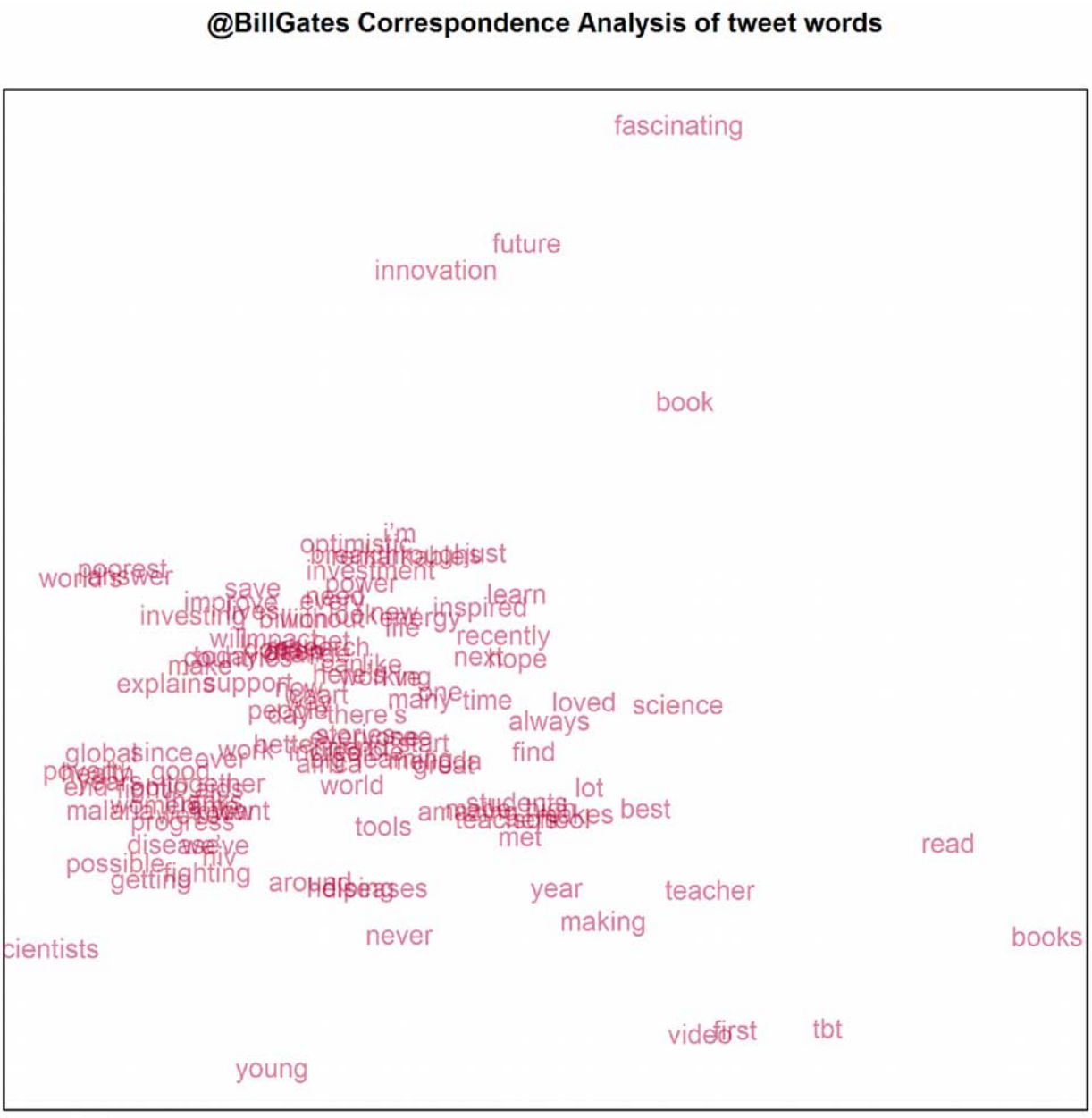
Στην παρακάτω εικόνα φαίνεται το αποτέλεσμα που προκύπτει μετά την εφαρμογή του κώδικα, εξάγοντας 500 tweets από τον χρήστη @BillGates. Μετά τη χρήση της ανάλυσης συστάδων στο σύνολο των πιο συχνών λέξεων δημιουργήθηκε το δενδρόγραμμα το οποίο ομαδοποιεί τις λέξεις αυτές. Το δενδρόγραμμα που δημιουργήθηκε, αποθηκεύτηκε σε μορφή εικόνας (png format).



Συνεχίζοντας τον προηγούμενο κώδικα παράγονται τα παρακάτω.
Κώδικας(συνέχεια του προηγούμενου):

```
#For a better visualization, we can apply a Correspondence Analysis
#(using package FactoMineR)
# correspondance analysis
bg_ca = CA(m1, graph=FALSE)

# default plot of words
plot(bg_ca$row$coord, type="n", xaxt="n", yaxt="n", xlab="", ylab="")
text(bg_ca$row$coord[,1], bg_ca$row$coord[,2], labels=rownames(m1),
   col=hsv(h=.95, s=1, v=.7, alpha=0.5))
title(main="@BillGates Correspondence Analysis of tweet words", cex.main=1)

# save the image in png format
png("BillGatesCorrespondenceAnalysis.png", width=8, height=8, units="in", res=300)
plot(bg_ca$row$coord, type="n", xaxt="n", yaxt="n", xlab="", ylab="")
text(bg_ca$row$coord[,1], bg_ca$row$coord[,2], labels=rownames(m1),
   col=hsv(h=.95, s=1, v=.7, alpha=0.5))
title(main="@BillGates Correspondence Analysis of tweet words", cex.main=1)
dev.off()
```

Στην παρακάτω εικόνα φαίνεται το αποτέλεσμα που προκύπτει μετά την εφαρμογή του κώδικα, εξάγοντας 500 tweets από τον χρήστη @BillGates. Μετά τη χρήση της ανάλυσης συστάδων στο σύνολο των πιο συχνών λέξεων, για την καλύτερη απεικόνιση θα εφαρμοστεί η correspondence analysis χρησιμοποιώντας το πακέτο FactoMineR. Το γράφημα που δημιουργήθηκε, αποθηκεύτηκε σε μορφή εικόνας (png format).

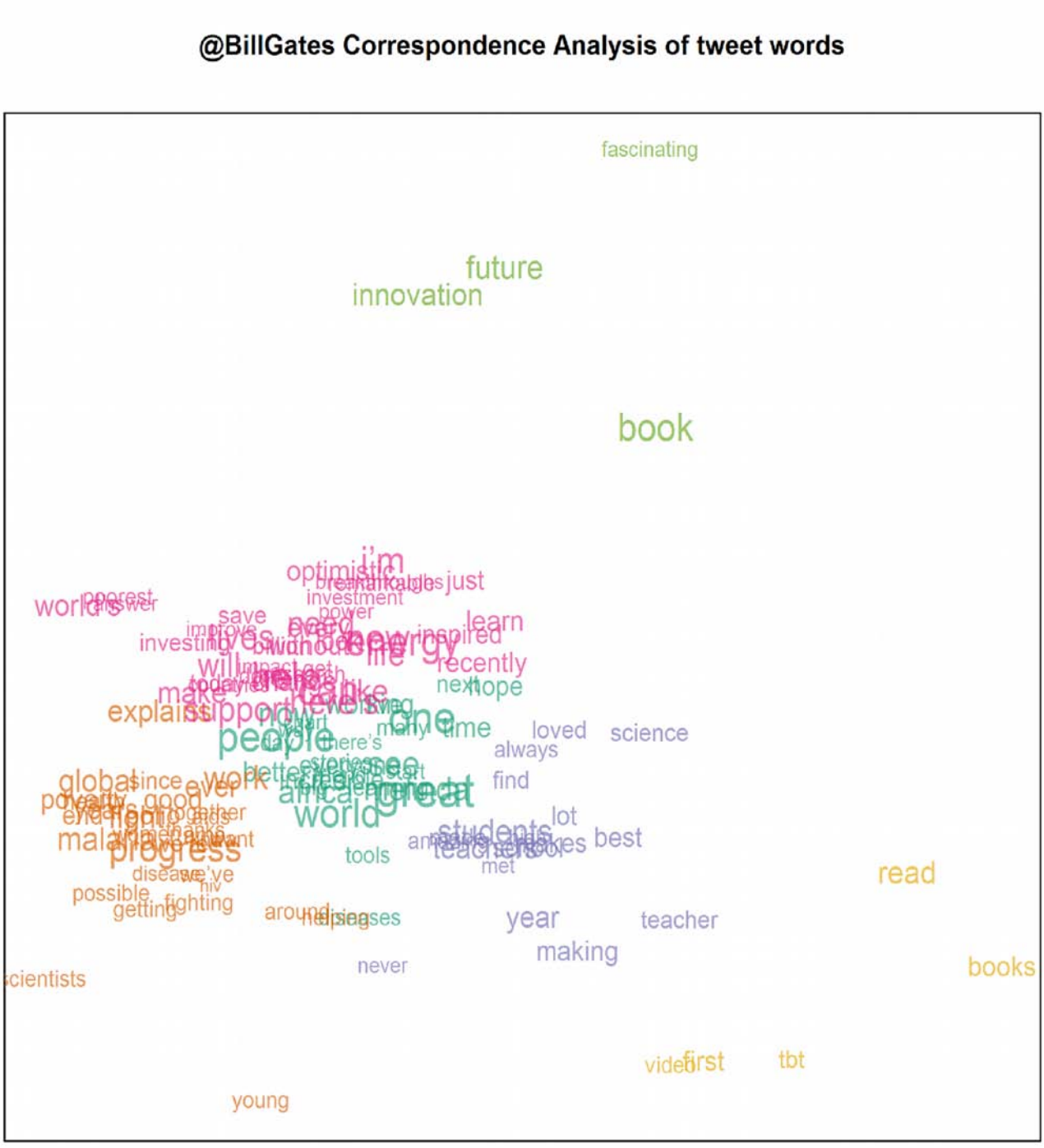

@BillGates Correspondence Analysis of tweet words
fascinating
future
innovation
book
learn
inspired
recently
next
hope
loved
science
always
find
lot
best
read
year
teacher
making
never
books
cientists
video
first
tbt
young

Συνεχίζοντας τον προηγούμενο κώδικα παράγονται τα παρακάτω.
Κώδικας(συνέχεια του προηγούμενου):

```
# To improve the correspondance analysis plot, we can apply a clustering method
# like k-means or partitioning around medoids (pam)
# partitioning around medoids with 6 clusters
k = 6
# pam clustering
bg_pam = pam(bg_ca$row$coord[,1:2], k)

# get clusters
clusters = bg_pam$clustering

# Let's try to get a nicer plot
# first we need to define a color palette
gbrew = brewer.pal(8, "Dark2")

# I like to use hsv encoding
gpal = rgb2hsv(col2rgb(gbrew))

# colors in hsv (hue, saturation, value, transparency)
gcols = rep("", k)
for (i in 1:k) {
  gcols[i] = hsv(gpal[1,i], gpal[2,i], gpal[3,i], alpha=0.65)
}

# plot with frequencies
wcex = log10(rowSums(m1))
plot(bg_ca$row$coord, type="n", xaxt="n", yaxt="n", xlab="", ylab="")
title("@BillGates Correspondence Analysis of tweet words", cex.main=1)
for (i in 1:k)
{
  tmp <- clusters == i
  text(bg_ca$row$coord[tmp,1], bg_ca$row$coord[tmp,2],
      labels=rownames(m1)[tmp], cex=wcex[tmp],
      col=gcols[i])
}
```

```
# save the image in png format
png("BillGatesCorrespondenceAnalysis2.png", width=8, height=8, units="in", res=300)
plot(bg_ca$row$coord, type="n", xaxt="n", yaxt="n", xlab="", ylab="")
title("@BillGates Correspondence Analysis of tweet words", cex.main=1)
for (i in 1:k)
{
  tmp <- clusters == i
  text(bg_ca$row$coord[tmp,1], bg_ca$row$coord[tmp,2],
      labels=rownames(m1)[tmp], cex=wcex[tmp],
      col=gcols[i])
}
dev.off()
```

Στην παρακάτω εικόνα φαίνεται το αποτέλεσμα που προκύπτει μετά την εφαρμογή του κώδικα, εξάγοντας 500 tweets από τον χρήστη @BillGates. Μετά τη χρήση της ανάλυσης συστάδων στο σύνολο των πιο συχνών λέξεων, και μετά την εφαρμογή της correspondence analysis, για να βελτιωθεί η correspondence analysis, μπορεί να εφαρμοστεί μια μέθοδος συσταδοποίησης όπως για παράδειγμα ο αλγόριθμος k-means ή ο αλγόριθμος partitioning around medoids (PAM). Το γράφημα που δημιουργήθηκε, αποθηκεύτηκε σε μορφή εικόνας (png format).


@BillGates Correspondence Analysis of tweet words
fascinating
future
innovation
book
i'm
optimistic
just
investment
power
poorest
worlds
save
need
learn
inspired
recently
hope
life
energy
make
explains
support
new
one
people
time
loved
science
always
find
global
work
ever
africa
great
world
good
poverty
progress
malaria
lot
best
students
teachers
tools
met
read
possible
getting
fighting
year
teacher
making
never
books
scientists
video
first
tbt
young

5. Εξεταζόμενη περίπτωση

Σε αυτή τη μελέτη, θα γίνει μια προσπάθεια για να διερευνηθεί πως η χρήση του Twitter αντανακλά τις εκφράσεις του κοινού σχετικά με την τρέχουσα άποψή τους για το θέμα των προσφυγικών ροών. Το έργο της εξόρυξης και ανάλυσης των δεδομένων είναι πιο αποδοτικό και αποτελεσματικό, καθώς γίνεται χρήση των πακέτων ανοικτού κώδικα της γλώσσας R, για την εύκολη ανάκτηση, προ-επεξεργασία, ανάλυση και οπτικοποίηση των δεδομένων του Twitter. Στόχος είναι να πραγματοποιηθεί το έργο της ανάλυσης και μέσα από αυτή να αποκαλυφθούν γνώσεις σχετικά με το θέμα των προσφυγικών ροών που δεν θα ήταν εύκολο, αν όχι ανέφικτο, να γίνει με τις παραδοσιακές προσεγγίσεις. Με την εφαρμογή των τεχνικών ανάλυσης συναισθήματος και εξόρυξης κειμένων στα tweets που αφορούν το θέμα των προσφύγων, μπορούν να προκύψουν κάποια συμπεράσματα όσον αφορά τη συχνότητα λέξεων που χρησιμοποιούνται στα περισσότερα tweets, και επίσης σχετικά με τα συναισθήματα που μπορεί να δημιουργούνται μέσα από αυτά τα tweets.

Ήδη από τον Απρίλιο του 2015 το προσφυγικό ζήτημα έχει εξελιχθεί σε μείζον και σε φλέγον πολιτικό ζήτημα όχι μόνον για τα επιμέρους κράτη-μέλη της Ευρωπαϊκής ένωσης όπως είναι η Ελλάδα και η Ιταλία τα οποία υποδέχονται τους πρόσφυγες, αλλά και για όλη την Ευρώπη ακόμα και για άλλα κράτη εκτός Ευρώπης. Σήμερα, αναφορικά με το προσφυγικό ζήτημα, ανοίγεται μέσα στον κόσμο ένα μεγάλο μέτωπο με δύο αντιπάλους, αυτούς που κρίνουν θετικά τις προσφυγικές ροές θεωρώντας ότι πρέπει να υπάρξει βοήθεια και στήριξη στους ανθρώπους αυτούς και από την άλλη είναι και οι άνθρωποι που κρίνουν αρνητικά την κατάσταση αυτή σκεπτόμενοι ότι δεν είναι οι αρμόδιοι για να βοηθήσουν και να στηρίξουν ουσιαστικά τους πρόσφυγες. Θα μπορούσε να πει κανείς πως είναι τρεις οι κατηγορίες των απόψεων των ανθρώπων καθώς υπάρχουν και εκείνοι οι οποίοι διατηρούν ουδέτερη στάση. Θα ακολουθήσει παραδείγμα με τον απαιτούμενο κώδικα και τα αποτελέσματα που θα παραχθούν, από τα οποία θα προκύψουν κάποια συμπεράσματα σχετικά με το θέμα αυτό.

Στο παράδειγμα που θα ακολουθήσει θα χρησιμοποιηθεί ο όρος “refugees” για την εξαγωγή των tweets των χρηστών στα οποία θα γίνουν οι απαραίτητες αναλύσεις.

Κώδικας:

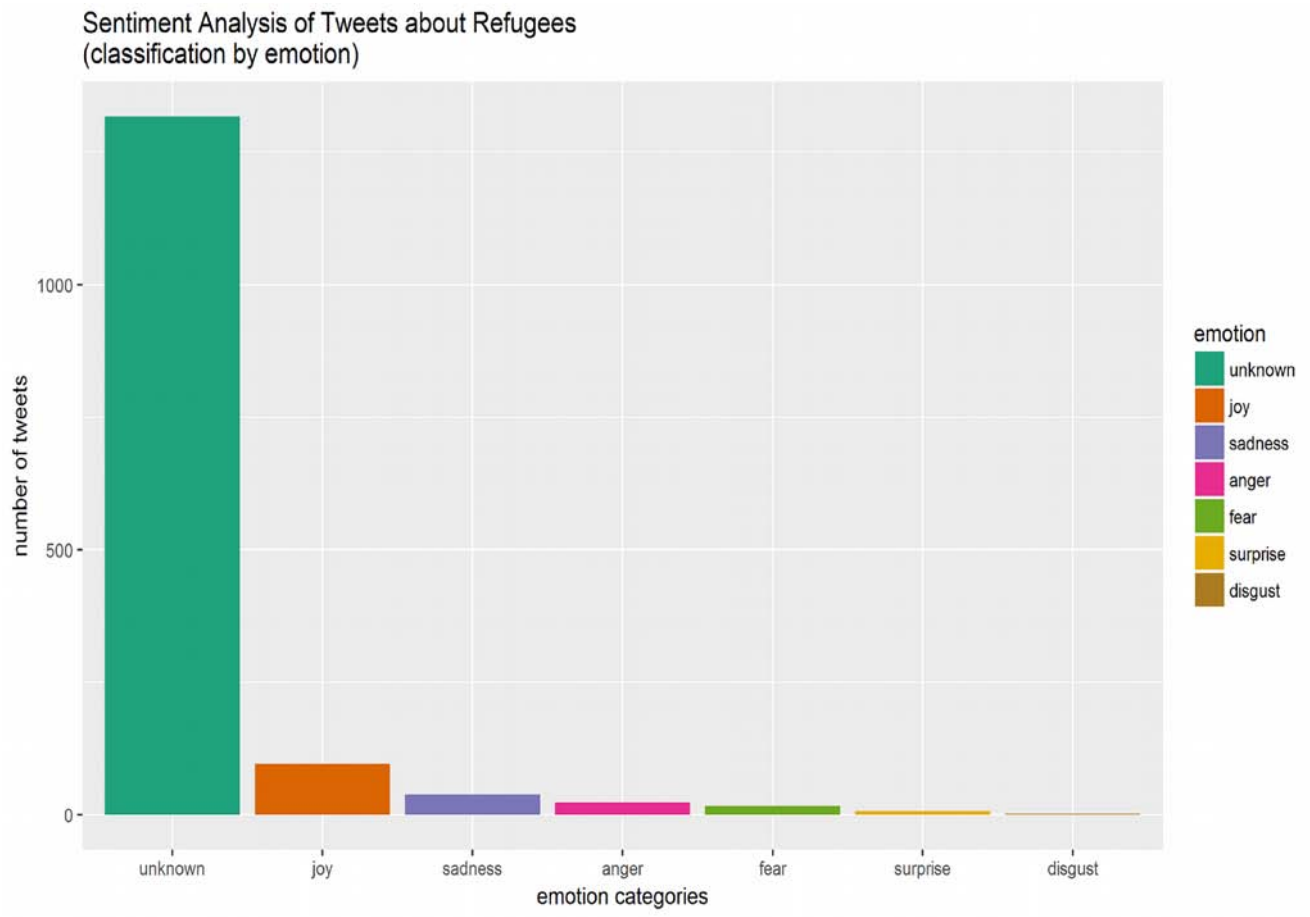
```
# Load the necessary packages
library(twitteR)
library(sentiment)
library(plyr)
library(ggplot2)
library(wordcloud)
library(RColorBrewer)

# Let's collect some tweets containing the term "refugees"
# harvest some tweets
some_tweets = searchTwitter("refugees", n=1500, lang="en")

# get the text
some_txt = sapply(some_tweets, function(x) x$getText())

# Prepare the text for sentiment analysis
# remove retweet entities
some_txt = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", some_txt)
# remove at people
some_txt = gsub("@\\w+", "", some_txt)
# remove punctuation
some_txt = gsub("[[:punct:]]", "", some_txt)
# remove numbers
some_txt = gsub("[[:digit:]]", "", some_txt)
# remove html links
some_txt = gsub("http\\w+", "", some_txt)
# remove unnecessary spaces
some_txt = gsub("[ \t]{2,}", "", some_txt)
some_txt = gsub("^\\s+|\\s+$", "", some_txt)

# define "tolower error handling" function
try.error = function(x)
{
  # create missing value
  y = NA
```

```
  # tryCatch error
  try_error = tryCatch(tolower(x), error=function(e) e)
  # if not an error
  if (!inherits(try_error, "error"))
  y = tolower(x)
  # result
  return(y)
}
# lower case using try.error with sapply
some_txt = sapply(some_txt, try.error)

# remove NAs in some_txt
some_txt = some_txt[!is.na(some_txt)]
names(some_txt) = NULL

# Perform Sentiment Analysis
# classify emotion
class_emo = classify_emotion(some_txt, algorithm="bayes", prior=1.0)
# get emotion best fit
emotion = class_emo[,7]
# substitute NA's by "unknown"
emotion[is.na(emotion)] = "unknown"

# classify polarity
class_pol = classify_polarity(some_txt, algorithm="bayes")
# get polarity best fit
polarity = class_pol[,4]

# Create data frame with the results and obtain some general statistics
# data frame with results
sent_df = data.frame(text=some_txt, emotion=emotion,
polarity=polarity, stringsAsFactors=FALSE)

# sort data frame
sent_df = within(sent_df,
  emotion <- factor(emotion, levels=names(sort(table(emotion), decreasing=TRUE))))
```

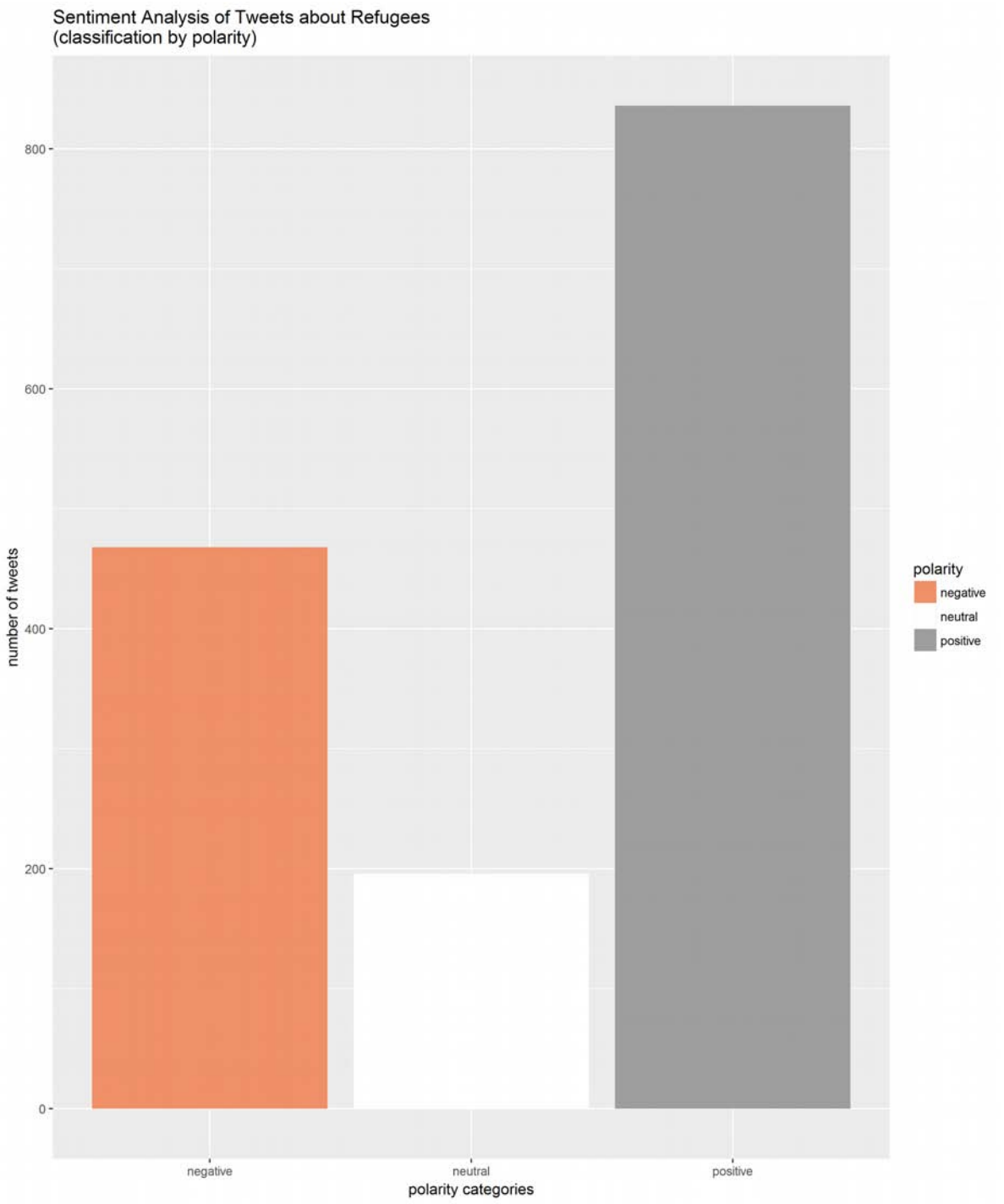
```
# Let's do some plots of the obtained results
# plot distribution of emotions
ggplot(sent_df, aes(x=emotion)) +
geom_bar(aes(y=..count.., fill=emotion)) +
scale_fill_brewer(palette="Dark2") +
labs(x="emotion categories", y="number of tweets") +
labs(title = "Sentiment Analysis of Tweets about Refugees\n(classification by emotion)")

# save the image in png format
png("Refugees_plot.png", width=10, height=6, units="in", res=300)
ggplot(sent_df, aes(x=emotion)) +
geom_bar(aes(y=..count.., fill=emotion)) +
scale_fill_brewer(palette="Dark2") +
labs(x="emotion categories", y="number of tweets") +
labs(title = "Sentiment Analysis of Tweets about Refugees\n(classification by emotion)")
dev.off()


# plot distribution of polarity
ggplot(sent_df, aes(x=polarity)) +
geom_bar(aes(y=..count.., fill=polarity)) +
scale_fill_brewer(palette="RdGy") +
labs(x="polarity categories", y="number of tweets") +
labs(title = "Sentiment Analysis of Tweets about Refugees\n(classification by polarity)")

# save the image in png format
png("Refugees_classification.png", width=10, height=12, units="in", res=300)
ggplot(sent_df, aes(x=polarity)) +
geom_bar(aes(y=..count.., fill=polarity)) +
scale_fill_brewer(palette="RdGy") +
labs(x="polarity categories", y="number of tweets") +
labs(title = "Sentiment Analysis of Tweets about Refugees\n(classification by polarity)")
dev.off()


# Separate the text by emotions and visualize the words with a comparison cloud
# separating text by emotion
```

```
emos = levels(factor(sent_df$emotion))
nemo = length(emos)
emo.docs = rep("", nemo)
for (i in 1:nemo)
{
  tmp = some_txt[emotion == emos[i]]
  emo.docs[i] = paste(tmp, collapse=" ")
}

# remove stopwords
emo.docs = removeWords(emo.docs, stopwords("english"))
# create corpus
corpus = Corpus(VectorSource(emo.docs))
tdm = TermDocumentMatrix(corpus)
tdm = as.matrix(tdm)
colnames(tdm) = emos

# comparison word cloud
comparison.cloud(tdm, colors = brewer.pal(nemo, "Dark2"),
  scale = c(3,.5), random.order = FALSE, title.size = 1.5)

# save the image in png format
png("RefugeesCloud.png", width=8, height=6, units="in", res=300)
comparison.cloud(tdm, colors = brewer.pal(nemo, "Dark2"),
  scale = c(3,.5), random.order = FALSE, title.size = 1.5)
dev.off()
```

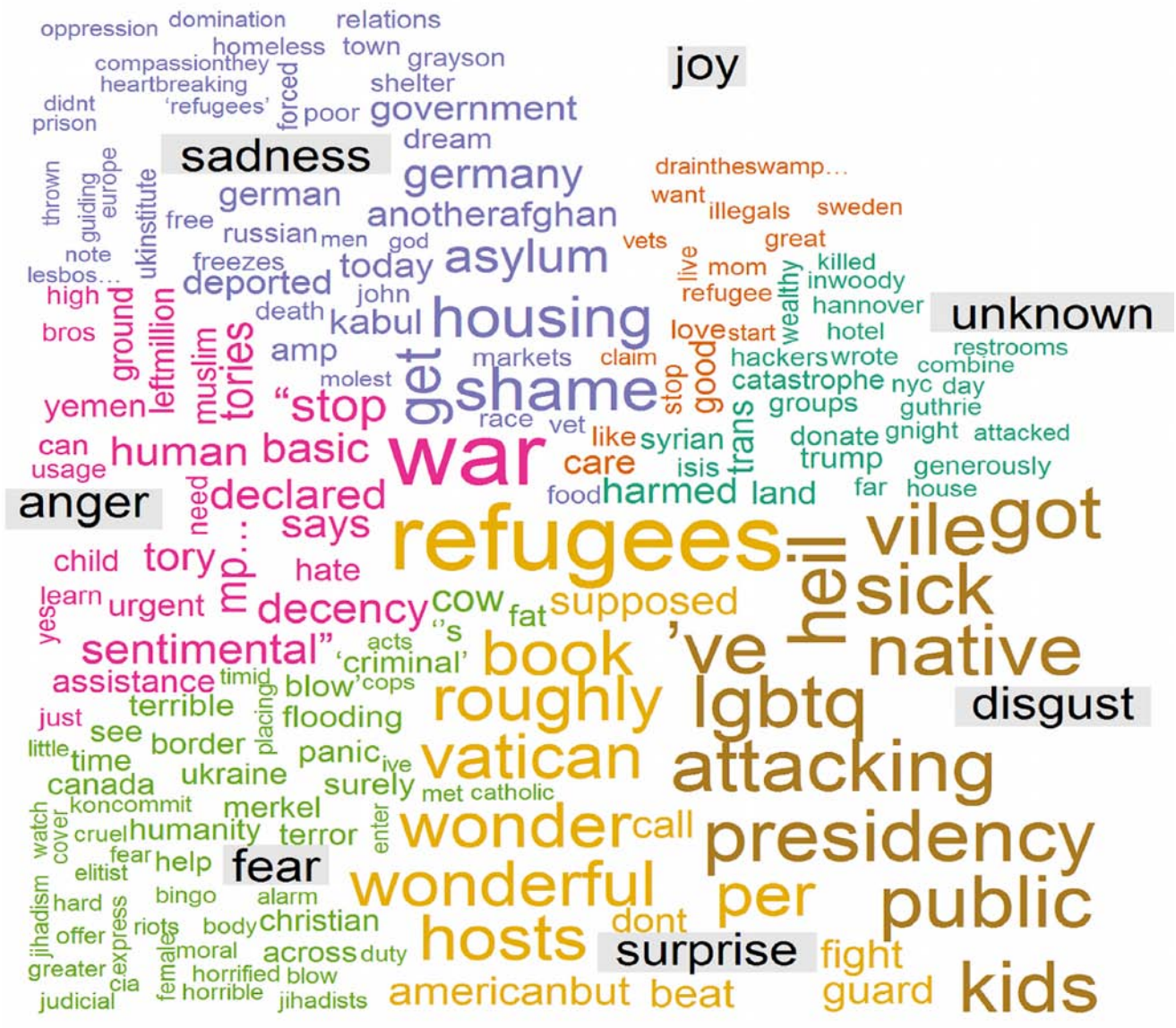
Στις παρακάτω εικόνες φαίνονται τα αποτελέσματα που προκύπτουν μετά την εφαρμογή του κώδικα, εξάγοντας 1500 tweets που είναι γραμμένα στα αγγλικά και χρησιμοποιούν τον όρο “refugees”. Μετά τη χρήση των μεθόδων για τον καθαρισμό των δεδομένων, πραγματοποιείται η ανάλυση συναισθήματος, έτσι ώστε τα tweets αυτά να κατηγοριοποιηθούν σύμφωνα με τα συναισθήματα των ανθρώπων. Τα γραφήματα που δημιουργήθηκαν, αποθηκεύτηκαν σε μορφή εικόνας (png format).



Μετά την ανάλυση το κείμενο κατηγοριοποιήθηκε αρχικά σε διαφορετικούς τύπους συναισθημάτων όπως θυμός, αηδία, φόβος, χαρά, λύπη, έκπληξη και άγνωστο συναίσθημα, όπως φαίνεται στο παραπάνω ιστόγραμμα. Η ταξινόμηση έγινε με τη χρήση του ταξινομητή Bayes λειτουργώντας πάνω στο “λεξικό συναισθημάτων” των Carlo Strapparava και Alessandro Valitutti. Συμπερασματικά τα περισσότερα από τα tweets έχουν άγνωστο συναισθηματικό περιεχόμενο. Λίγα είναι τα tweets εκείνα που εκφράζουν κάποιο είδος συναισθήματος από αυτά που αναφέρθηκαν προηγουμένως. Στη συνέχεια δοκιμάστηκε να γίνει κατηγοριοποίηση που να διευκρινίζει αν το κάθε tweet είναι θετικό ή αρνητικό όπως φαίνεται στο επόμενο απλούστερο γράφημα.



Είναι σαφές ότι τα περισσότερα από τα tweets είναι θετικά, περίπου 830, αν και ο αριθμός των αρνητικών tweet είναι σχετικά μεγάλος, περίπου 470, ενώ ένα μέρος των tweets διατηρεί ουδέτερη στάση, περίπου 200. Τέλος, χρησιμοποιώντας τις λέξεις στα tweets, δημιουργήθηκε ένα comparison wordcloud που χρησιμοποιεί τα συναισθήματα των λέξεων για να καθορίσει τις θέσεις τους μέσα σε αυτό.



Ένα wordcloud μπορεί να δώσει μια πιο διαισθητική αίσθηση σχετικά με τα tweets των ανθρώπων. Το συγκεκριμένο comparison wordcloud διανέμει στα συναισθήματα που αναφέρθηκαν νωρίτερα τους όρους που χρησιμοποιήθηκαν συχνότερα στα tweets που συλλέχθηκαν και σχετίζονται με το συναίσθημα αυτό. Το μέγεθος της γραμματοσειράς των λέξεων ποικίλει ανάλογα με τη σχετική συχνότητα της λέξης (όσο μεγαλύτερο μέγεθος τόσο συχνότερη η λέξη στο συγκεκριμένο συναίσθημα). Κατ 'αρχάς, στην περιοχή θυμού υπάρχουν αρκετά ουσιαστικά όπως war(πόλεμος), hate(μίσος), που εμφανίζουν σίγουρα το συναίσθημα του θυμού καθώς και human(άνθρωπος), child(παιδί), assistance(βοήθεια) που δείχνουν πιθανόν ότι υπάρχει θυμός σχετικά με τη βοήθεια που ίσως να χρειάζονται πολλοί

άνθρωποι, ενήλικες και παιδιά, και δεν υπάρχει όπως θα έπρεπε. Θα μπορούσε επίσης να σημαίνει θυμό και για τους ανθρώπους που προκαλούν τις άσχημες καταστάσεις. Στη ζώνη αηδίας παρατηρούνται φράσεις όπως sick(άρρωστος), vile(αχρείος) και attacking(επίθεση) τα οποία μπορεί να δείχνουν την απέχθεια που αισθάνονται οι άνθρωποι σχετικά με τις επιθέσεις που γίνονται από ανθρώπους που τους θεωρούν αχρείους και κατ'επέκταση υπάρχουν άρρωστοι άνθρωποι. Μια παρατήρηση είναι ότι η ζώνη αηδίας δεν είναι πολύ πυκνοκατοικημένη και αυτό μπορεί να σημαίνει ότι ο όρος πρόσφυγες δεν δημιουργεί κατά κύριο λόγο απωθητικά συναισθήματα, αλλά περισσότερο θλίψης ή φόβου, όπως φαίνεται στο wordcloud, σχετικά με την κατάσταση που επικρατεί. Στη ζώνη θλίψης προέκυψαν οι εκφράσεις housing(στέγαση), asylum(άσυλο), shame(ντροπή), germany(Γερμανία), government(κυβέρνηση). Είναι βέβαιο ότι επικρατούν συναισθήματα λύπης και ντροπής παράλληλα, σχετικά με τη στέγαση των προσφύγων και τα άσυλα επειδή ίσως δεν είναι αρκετά ή κάποια δεν είναι σε καλή κατάσταση και επίσης κυριαρχεί θλίψη λόγω των κυβερνήσεων και θλίψη σε σχέση με την Γερμανία. Εντός της ζώνης του φόβου κάποιοι όροι που παρατηρούνται είναι blow(πλήγμα), flooding(πλημμύρα) terror(τρόμος), horrible(φρικτός), panic(πανικός), criminal(εγκληματίας) που είναι λογικές εκφράσεις του φόβου στον πληθυσμό. Στη ζώνη φόβου επίσης προέκυψαν κάποιες όχι τόσο σαφείς εκφράσεις, με κάποιες από αυτές να είναι οι εξής christian(Χριστιανός), merkel(Μέρκελ), jihadists(τζιχαντιστές) και border(σύνορο) που μπορεί να δείχνουν ότι κυριαρχεί φόβος για τους τζιχαντιστές και επίσης και σε θέματα θρησκείας, σχετικά με την Μέρκελ και φόβος σχετικά με τα σύνορα των χωρών. Στις ζώνες χαράς και έκπληξης δεν προέκυψαν τόσο σαφείς εκφράσεις αλλά και ο αριθμός τους είναι περιορισμένος. Αυτό ίσως συμβαίνει γιατί τα δύο αυτά συναισθήματα δεν εκφράζονται και τόσο συχνά γιατί δεν κυριαρχούν οπότε δεν δημιουργείται κάποια ξεκάθαρη εικόνα. Τέλος, υπάρχει και ένα πλήθος λέξεων που δεν μπορούν να κατηγοριοποιηθούν οπότε εντάσσονται στην κατηγορία του άγνωστου συναισθήματος.

Τελικά, φάνηκε ότι μέσω της ανάλυσης συναισθήματος προκύπτουν κάποια συμπεράσματα σε σχέση με τα συναισθήματα των ανθρώπων και κάποιο θέμα συζήτησης ή κάποιον όρο, τα οποία πολλές φορές βέβαια μπορεί να μην είναι ξεκάθαρα. Σε κάθε περίπτωση είναι ένα επιπλέον εργαλείο για την εξαγωγή συμπερασμάτων σε ευρύτερο επίπεδο σχετικά με κάποιο θέμα. Στο θέμα των προσφύγων είναι γεγονός ότι μέσα από την ανάλυση φάνηκε η ποικιλομορφία των συναισθημάτων των ανθρώπων σχετικά με αυτό καθώς και πως αυτά τα συναισθήματα συνδέονται με κάποιες λέξεις ή εκφράσεις που χρησιμοποιούν οι άνθρωποι για να εκφραστούν μέσα από τα tweets που δημοσιεύουν.

6. Συμπεράσματα και προτάσεις για μελλοντική εργασία

Συμπερασματικά στόχος της συγκεκριμένης διπλωματικής εργασίας ήταν η αναλυτική παρουσίαση των κοινωνικών δικτύων καθώς και των μεθόδων εξόρυξης δεδομένων που μπορούν να χρησιμοποιηθούν για την εξαγωγή συμπερασμάτων. Στα πλαίσια του στόχου αυτού παρουσιάστηκαν όλα τα βήματα που οδηγούν στην εξόρυξη γνώσης συγκεκριμένα από το Twitter. Τα βήματα αυτά ξεκινούν από τη σύνδεση με το Twitter και την εξαγωγή δεδομένων από αυτό, συνεχίζουν με την επεξεργασία και τον μετασχηματισμό τους, την ανακάλυψη πληροφορίας μέσω της εξόρυξης κειμένου και τη χρήση αλγορίθμων συσταδοποίησης για την κατηγοριοποίηση των πληροφοριών και τέλος την ανάλυση συναισθήματος που όπως φάνηκε οδηγεί στην καλύτερη γνώση και εικόνα σχετικά με την άποψη των ανθρώπων για διάφορα θέματα. Οι γνώσεις και οι πληροφορίες που προκύπτουν από τέτοιες αναλύσεις μπορούν να χρησιμοποιηθούν από πολλές επιστήμες όπως η κοινωνιολογία, η εγκληματολογία, το μάρκετινγκ. Ουσιαστικά μέσα από τα μοντέλα που δημιουργήθηκαν απαντήθηκαν ερωτήσεις σχετικά με το Twitter που μπορεί να υπήρχαν στο μυαλό όλων σχετικά με τα δημοφιλή θέματα συζήτησης, με τη συχνότητα που εμφανίζουν τα δεδομένα, με τις συσχετίσεις που μπορεί να παρουσιάζουν, με τα συναισθήματα των ανθρώπων κ.α.

Υπάρχει, ωστόσο, μια ποικιλία από μεθόδους ανάλυσης κειμένων που δεν ήταν δυνατόν να αναφερθούν σε ένα σχετικά σύντομο κείμενο όπως η παρούσα εργασία, αλλά αυτά θα μπορούσαν να διερευνηθούν περαιτέρω μελλοντικά. Επιπροσθέτως, υπάρχουσες δημοσκοπήσεις και έρευνες που έχουν γίνει, αποτελούν άλλη μια χρήσιμη πηγή πληροφοριών και θα ήταν πολύ ενδιαφέρον να δούμε το βαθμό στον οποίο ταιριάζουν τα αποτελέσματά τους, με τα αποτελέσματα της ανάλυσης συναισθήματος που διεξάγονται στα κοινωνικά δίκτυα. Θα πρέπει ακόμα να τονιστεί ότι τα εργαλεία που αναπτύσσονται από την ανάλυση των κοινωνικών δικτύων, βρίσκονται σε πρώιμα στάδια, και είναι απίθανο να αντικαταστήσουν τις πιο παραδοσιακές μεθόδους έρευνας όπως η δειγματοληπτική έρευνα, τουλάχιστον στο εγγύς μέλλον. Παρ 'όλα αυτά οι δυνατότητες για τα δεδομένα που εξάγονται από τα κοινωνικά δίκτυα είναι σημαντικές και στο μέλλον, παρόμοιες τέτοιες εφαρμογές μπορεί να γίνουν πιο ισχυρές και χρήσιμες, αν οι αναλυτές επιλέγουν να χρησιμοποιήσουν ιστοσελίδες κοινωνικών δικτύων για την παροχή στατιστικών στοιχείων με σκοπό τη διεξαγωγή κάποιας ανάλυσης. Επιπλέον, θα πρέπει να σημειωθεί ότι η προτεινόμενη προσέγγιση θα μπορούσε να εφαρμοστεί και σε άλλες κοινωνικές πλατφόρμες στο διαδίκτυο, όπως στο Facebook. Τέλος, θα ήταν πολύ χρήσιμο να μπορεί να εφαρμοστεί ένας αποτελεσματικός τρόπος για την εξάλειψη των άσχετων tweets, καθώς τα περισσότερα από αυτά οδηγούν σε ψευδή αποτελέσματα και επηρεάζουν την ακρίβεια της ανάλυσης.

Βιβλιογραφία

[1] Κοινωνικό δίκτυο: https://el.wikipedia.org/wiki/%CE%9A%CE%BF%CE%B9%CE%BD%CF%89%CE%BD%CE%B9%CE%BA%CF%8C_%CE%B4%CE%AF%CE%BA%CF%84%CF%85%CE%BF#cite_note-1

[2] Twitter: https://el.wikipedia.org/wiki/Twitter

[3] http://www.statisticbrain.com/

[4] https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[5] Εξόρυξη δεδομένων: https://el.wikipedia.org/wiki/%CE%95%CE%BE%CF%8C%CF%81%CF%85%CE%BE%CE%B7_%CE%B4%CE%B5%CE%B4%CE%BF%CE%BC%CE%AD%CE%BD%CF%89%CE%BD

[6] Κύρκος, Ε. (2015). Εξόρυξη Γνώσης από Δεδομένα. [Κεφάλαιο 6]. Στο Κύρκος, Ε. 2015. *Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων*. [ηλεκτρ. βιβλ.] Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. κεφ 6.

[7] Παπαδάκης, Ε. (2016). Ανάλυση Συναισθήματος από Κείμενο με Τεχνικές Μηχανικής Μάθησης και Χρήση Λεξικού (Διπλωματική εργασία). ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ, Αθήνα.

[8] Πανταζής, Ο. (2016). Εξόρυξη Δεδομένων από το Twitter και Εφαρμογή Αλγορίθμων Μη-Επιβλεπόμενης Μηχανικής Μάθησης για Συσταδοποίηση Κειμένων (Διπλωματική εργασία). ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ, Αθήνα.

[9] Ταράτσα, Ν. (2011). Εξόρυξη γνώσης σε κοινωνικά δίκτυα (Μεταπτυχιακή εργασία). Πανεπιστήμιο Πειραιώς, Πειραιάς.

[10] Tutorial: Using R and Twitter to Analyse Consumer Sentiment: https://colinpriest.com/2015/07/04/tutorial-using-r-and-twitter-to-analyse-consumer-sentiment/

[11] Mining twitter with R: https://sites.google.com/site/miningtwitter/home

[12] Nirmala, C.R., Roopa G.M., Naveen Kumar K.R. (2015). Twitter data analysis for unemployment crisis, 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 29-31 Oct. 2015 (pp:420 – 423). Davangere, Karnataka, India: IEEE.

[13] Bright, J., Margetts, H., Hale, S., Yasseri, T. (2014). The Use of Social Media for Research and Analysis: A Feasibility Study. DWP ad hoc research report no. 13. United Kingdom: Corporate Document Services. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/387591/use-of-social-media-for-research-and-analysis.pdf

[14] Russell, M.A. (2013). Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More. United States of America: O'Reilly Media. https://books.google.gr/books?hl=en&lr=&id=_VkrAQAAQBAJ&oi=fnd&pg=PR4&dq=data+mining+twitter&ots=JqqylyWulH&sig=aphFBn6G23x24GRBvLO1zCvXmLg&redir_esc=y#v=onepage&q=data%20mining%20twitter&f=false

[15] Miner, G.D., Elder IV, J.F., Nisbet, R.A., Delen, D., Fast, A., Hill, T. (2011). Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. United states of America: Academic Press. https://books.google.gr/books?hl=en&lr=&id=-B6amxqygTMC&oi=fnd&pg=PP2&dq=data+mining+twitter&ots=FOhi2q0S2D&sig=7coPW_odpZZNiGEX3VVi6j633Oo&redir_esc=y#v=onepage&q=data%20mining%20twitter&f=false

[16] Han, J., Kamber, M., Pei, J. (2011). Data Mining: Concepts and Techniques Third Edition. United States of America: Morgan Kaufmann Publishers. https://books.google.gr/books?hl=en&lr=&id=pQws07tdpjoC&oi=fnd&pg=PP1&dq=data+mining+twitter&ots=tyLyXUoA-Z&sig=zwMYVztmbcyDyv4eJInbo8CU0ps&redir_esc=y#v=onepage&q=data%20mining%20twitter&f=false

[17] Witten, I.H., Frank, E., Hall, M.A., Pal, C.J. (2016). Data Mining: Practical Machine Learning Tools and Techniques Fourth Edition. United States of America: Morgan Kaufmann Publishers. https://books.google.gr/books?hl=en&lr=&id=1SylCgAAQBAJ&oi=fnd&pg=PP1&dq=data+mining+twitter&ots=8HHQqemxza&sig=zoJ_pEf77P8biSJGbQ-BrWVJJL8&redir_esc=y#v=onepage&q=data%20mining%20twitter&f=false