ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ

Διπλωματική Εργασία

"Forecasting trends in internet media impact on users' information"

-

"Πρόβλεψη τάσεων στην επίδραση των διαδικτυακών ΜΜΕ στη διαμόρφωση της πληροφόρησης των χρηστών"

Φοιτητής:
Νιόζας Ιωάννης

Επιβλέπουσα:
Δασκαλοπούλου Ασπασία

Συνεπιβλέπων:
Βασιλακόπουλος Μιχάλης

Περίληψη

Τα τελευταία χρόνια η καθημερινή ενημέρωση από διαδικτυακούς τόπους έχει μπει ολοένα και περισσότερο στη ζωή του κοινού. Τα μεγαλύτερα Μ.Μ.Ε πλέον εστιάζουν σε αυτό το πολυμορφικό μέσο και όλες οι έντυπες εφημερίδες είναι πλέον και σε ηλεκτρονική μορφή ενώ ακόμα και τα τηλεοπτικά κανάλια δίνουν μεγάλο βάρος στην ενημέρωση βάση της ιστοσελίδας τους. Στη διαδικτυακή ενημέρωση επίσης σημαντικό ρόλο αποτελούν πλέον και τα μέσα κοινωνικής δικτύωσης τα οποία επιτρέπουν στον ίδιο τον αναγνώστη να γίνει πομπός ή αναμεταδότης ειδήσεων.

Στόχος της παρούσας εργασίας είναι η ανάλυση δεδομένων από ειδησεογραφικά site αλλά και μέσα κοινωνικής δικτύωσης για την εξεύρεση της τυχόν επιρροής τους ενός μέσου στο άλλο. Για την ακρίβεια θα αναζητήσουμε ομοιότητες στη δημοφιλία συγκεκριμένων λέξεων-κλειδιών στην ημερήσια ατζέντα των μεγαλύτερων(σε επισκεψιμότητα) ηλεκτρονικών εφημερίδων με τις συζητήσεις μελών του Twitter στο Ελληνικό διαδίκτυο. Αυτό έχει σαν σκοπό να δημιουργήσουμε ένα μοντέλο που αναπαριστά τη συσχέτιση των δύο μέσων και συγκεκριμένα το πως οι “τάσεις”(trends) στη θεματολογία των άρθρων των ΜΜΕ επηρεάζουν τις τάσεις στη θεματολογία των Tweets.

Το μοντέλο αυτό στη συνέχεια θα μας βοηθήσει να προβλέψουμε την επίπτωση που θα έχει ένα ειδησεογραφικό trend που ξεκίνησε από τα news sites στο Twitter και κατά πόσο τα tweets που αναφέρονται σε αυτό θα αυξηθούν ή θα μειωθούν ανάλογα με το πόσο ασχολούνται τα sites.

Όταν αναφερόμαστε σε ένα ειδησεογραφικό trend εννοούμε την απότομη αλλαγή στη δημοφιλία μια συγκεκριμένης λέξης-κλειδιού ή μιας θεματολογίας η οποία μπορεί να σχετίζεται με γεγονότα της επικαιρότητας.

Περιεχόμενα

Κατάλογος Εικόνων

1. Εισαγωγή

Τα Μέσα Μαζικής Ενημέρωσης ή Επικοινωνίας (ΜΜΕ) είναι όλα τα διαθέσιμα μέσα με τα οποία μπορεί να ενημερωθεί για προηγούμενα και τρέχοντα συμβάντα ένα μεγάλο πλήθος ανθρώπων. .

Σύμφωνα με έρευνα που παρουσίασε το Εθνικό Κέντρο Κοινωνικών Ερευνών το 84,6% των Ελλήνων ενημερώνονται μέσω διαδικτύου, ενώ μόλις το 44,5% επέλεξε ως κύρια μορφή ενημέρωσης τις εφημερίδες. Το 47,5% επέλεξε το ραδιόφωνο και το 45% την τηλεόραση, την άλλοτε «βασίλισσα» της ενημέρωσης.

Τα ΜΜΕ παίζουν ένα σημαντικό ρόλο στη διαμόρφωση κοινής γνώμης πάνω σε ένα πλήθος από θέματα που αφορούν την επικαιρότητα και όχι μόνο. Ο στόχος της εργασίας έχει ιδιαίτερο ενδιαφέρον καθώς επιχειρεί να επιβεβαιώσει την επιρροή των ΜΜΕ όχι μόνο στην κοινή γνώμη πάνω σε ένα θέμα αλλά και στο ίδιο θέμα της «κοινής συζήτησης» κάθε φορά όπως αυτό αντικατοπτρίζεται στα μέσα κοινωνικής δικτύωσης.

Το διαδίκτυο έχει καταστήσει την πληροφορία προσβάσιμη σε όποιον διαθέτει συσκευή με σύνδεση σε αυτό με σχεδόν μηδαμινό κόστος. Ο όγκος της πληροφορίας που αναφέρεται σε ειδήσεις είναι τεράστιος καθώς η δημιουργία και η αναπαραγωγή της δεν απαιτεί σχεδόν καθόλου πόρους. Επίσης η ανάγνωση ειδήσεων λόγω του διαδικτύου έχει γίνει ευκολότερη από ποτέ, ενώ το γεγονός της εμφάνισης φορητών ηλεκτρικών συσκευών με πρόσβαση στο internet έχει αυξήσει κατακόρυφα την κατανάλωση τους.

Η ειδήσεις είναι τα προιόντα μιας ηλεκτρονική εφημερίδας καθώς όσο περισσότερη επισκεψιμότητα έχουν τα άρθρα της τόσο περισσότερες διαφημίσεις(η οποίες είναι η κύρια πηγή χρηματοδότησης) προσελκύονται στον ιστότοπο. ¨Ετσι η ανάδειξη συγκεκριμένων ειδήσεων από τις εφημερίδες αλλά και η έμφαση που δίνεται σε αυτές ή η χρονική διάρκεια στην οποία προβάλονται έχει να κάνει με το αγοραστικό τους κοινό. Τα κοινωνικά δίκτυα και συγκεκριμένα το Twitter είναι ένα τόπος όπου οι ειδήσεις συζητιόνται και αναμεταδίδονται από τους χρήστες τους ενώ οι γνώμες τους καθώς και το ποσοστό αναπαραγωγής της είδησης δείχνει κατά πόσο αυτή είναι δημοφιλής ανάμεσα στους αναγνώστες.

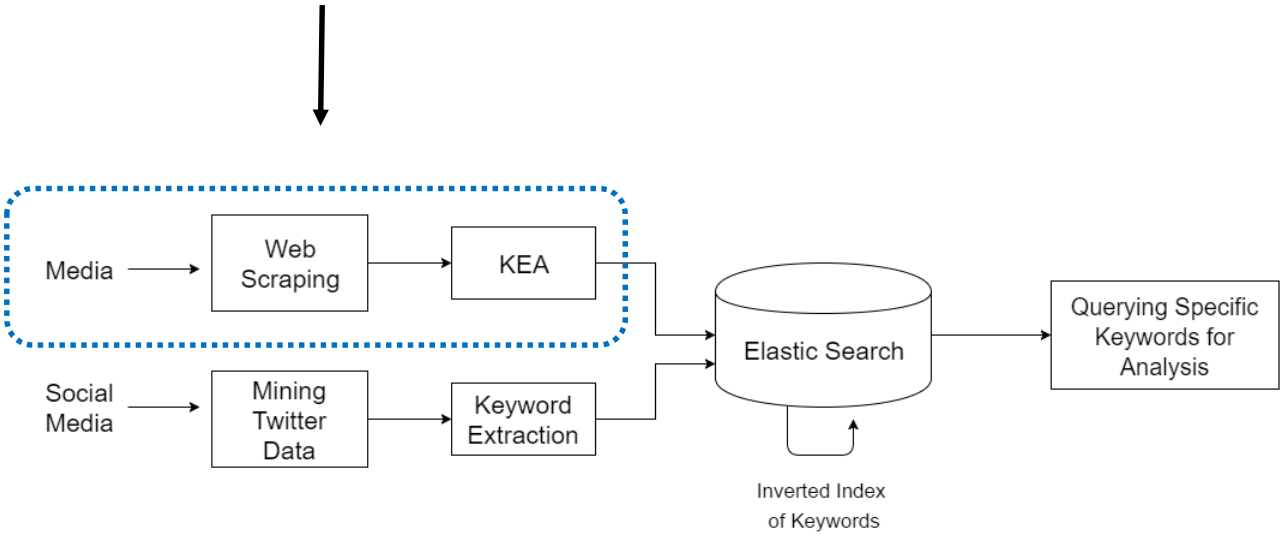
¨Στο τομέα του Data mining υπάρχει ξεχωριστός κλάδος που ασχολείται με την ανάλυση δεδομένων από τα κοινωνικά δίκτυα με σκοπό το marketing και την εξαγωγή συμπερασμάτων για τις προτιμήσεις των χρηστών τους.

Από την άλλη ο όγκος της πληροφορίας που διακινείται στα ενημερωτικά sites είναι πολύ μεγάλος και η αρχειοθέτηση της με συγκεκριμένες μεθόδους είναι απαραίτητη για την ευκολότερη αναζήτηση και μελέτη των ειδήσεων.
Η διαδικασία αυτή ονομάζεται **αποδελτοποίηση τύπου** ενώ υπάρχει ξεχωριστός κλάδος ο οποίος ασχολείται με την αρχειοθέτηση και ανάλυση της πληροφορίας από τα Μ.Μ.Ε.
Στην παρούσα εργασία ασχοληθήκαμε με την εξαγωγή πληροφοριών από τα άρθρα των δημοφιλέστερων ενημερωτικών ιστοστελίδων γραμμένων στα Ελληνικά καθώς και των tweets με την ίδια γλώσσα. Στη συνέχεια χρησιμοποιήσαμε σαν αντικείμενο μελέτης τις λέξεις κλειδιά τους και μελετήσαμε πιθανές συσχετίσει ανάμεσα στα ποσοστά εμφάνισής τους ανά ημέρα και τέλος συγκρίναμε μοντέλα πρόβλεψης για τη συσχέτιση αυτή.

2. Ανάλυση-Σχεδίαση Συστήματος για άρθρα

Η εξαγωγή των άρθρων από τα τα news sites έγινε με **web crawler** τον οποίο προγραμματίσαμε με τη βοήθεια βιβλιοθηκών της java, τα δεδομένα τα αποθηκεύσαμε στην βάση δεδομένων **elasticsearch**, ενώ η επεξεργασία των δεδομένων έγινε με εργαλεία machine learning.

Παρακάτω εξηγούνται αναλυτικά οι έννοιες και τα εργαλεία που χρησιμοποιήσαμε.



Εικόνα 1 - Απεικόνιση Συστήματος (Πρώτο στάδιο)

2.1. Συλλογή δεδομένων από Media

2.1.1. WebCrawler

Είναι ένα internet bot το οποίο συστηματικά πλοηγείται στο παγκόσμιο ιστό για να αντλεί πληροφορία και να την αποθηκεύει. Επίσης χρησιμοποιείται για indexing των ιστοσελίδων σε βάσεις δεδομένων έτσι ώστε να γίνεται αναζήτηση σε αυτές πιο γρήγορα και εύκολα.

Περιορισμοί

Οι crawlers έχουν περιορισμούς από τα site που επισκέπτονται καθώς καταναλώνουν πόρους από το bandwindth της σελίδας. Στο αρχείο robots.txt περιέχονται πληροφορίες σχετικά με τα δικαιώματα που έχουν οι crawlers στο να αντλούν πληροφορίες.

2.1.2. WebScraping

Είναι η διαδικασία εξαγωγής πληροφορορικών από ιστοσελίδες από αυτοματοποιημένο λογισμικό(web crawler). Χρησιμοποιείται για data minining πληροφοριών όπως ο καιρός, τιμές προϊόντων, απόψεις χρηστών κτλ.

Οι ιστοσελίδες είναι φτιαγμένες συνήθως σε HTML και περιέχουν πολλά χρήσιμα δεδομένα σε μορφή κειμένου. Παρόλα αυτά είναι φτιαγμένες για τους χρήστες και δεν είναι εύκολο να προσπελαστούν από αυτόματους crawlers. [2]

Τεχνικές

Human copy-and-paste

Μερικές φορές ακόμα και το καλύτερο εργαλείο web scraping δεν μπορεί να αντικαταστήσει το "ανθρώπινο χέρι" όσον αφορά τη συλλογή πληροφοριών από ιστοσελίδες ενώ σε συγκεκριμένες περιπτώσεις είναι αναπόφευκτο ειδικά όταν οι σελίδες χρησιμοποιούνε φράγματα στην αυτοματοποίηση.

Text pattern matching

Μία απλή προσέγγιση εξαγωγής πληροφοριών από ιστοσελίδες είναι το pattern matching, όπου χρησιμοποιούνται κανονικές εκφράσεις που να ταιριάζουν σε μοτίβα κειμένων.

HTTP programming

Στατικές και Δυναμικές ιστοσελίδες μπορούν να ανακτηθούν χρησιμοποιώντας HTTP requests στον απομακρυσμένο web server με χρήση socket programming.

HTML parsing

Πολλές ιστοσελίδες έχουν μια μεγάλη συλλογή από σελίδες οι οποίες δημιουργούνται δυναμικά από τη βάση δεδομένων. Στο data mining ένα πρόγραμμα αναγνωρίζει τέτοιες σελίδες και εξάγει το περιεχόμενο το μεταφράζει σε μία σχετική μορφή η οποία ονομάζεται wrapper.

2.1.3. Web Scraper

Το κομμάτι της υλοποίησης του web scraper αποτελείται από τα εξής βασικά μέρη [3]

• **Fetch**: Προσκόμιση δεδομένων από το διαδίκτυο για επεξεργασία, όπως ιστοσελίδες σε μορφή HTML, ροές RSS σε μορφή XML.
• **Parse**: Ανάλυση των στοιχείων του κάθε δεδομένου από μια ροή RSS.
• **Scrap**: Εντοπισμός και εξαγωγή του μέρους της πληροφορίας που μας ενδιαφέρει μέσα από ένα σύνολο δεδομένων HTML. Αφαίρεση των πιο κοινών σημείων στίξης και μετατροπή του σε καθαρό κείμενο λέξεων με μοναδικό διαχωριστικό το κενό.
• **Search**: Αναζήτηση μέσα στο εξαγμένο κομμάτι πληροφορίας για προκαθορισμένα αλλά και νέα άγνωστα keywords.
• **Store**: Αποθήκευση στην βάση δεδομένων των στοιχείων που εντοπίστηκαν, καθώς και διάφορα ακόμη στοιχεία για στατιστικούς κυρίως λόγους όπως ημερομηνία που δημοσιεύθηκε η αγγελία, τον τίτλο και τη διεύθυνση (url) της.

2.1.4 Τεχνολογία RSS

Η **RSS** είναι μία τεχνολογία την οποία ο χρήστης χρησιμοποιεί για να πάρει και να διαβάσει πληροφορίες που έχουν σταλεί σε αυτόν αντί να επισκεφθεί μόνος, τον κατάλληλο ιστότοπο, για να την αναζητήσει και να την προσπελάσει. Η RSS είναι υπεύθυνη για την αυτοματοποιημένη λήψη στην επιφάνεια εργασίας ειδήσεων, συζητήσεων, podacasts, videocasts και μουσικής από διάφορους δικτυακούς τόπους. Είναι μία οικογένεια προτύπων ανταλλαγής και διανομής περιεχομένου που βασίζονται στη γλώσσα XML. Ένα κανάλι τροφοδοσίας **RSS** (RSS feed) αποτελείται από μία λίστα στοιχείων που περιέχουν ένα τίτλο καθώς και το σύνδεσμο προς την αντίστοιχη ιστοσελίδα ή αρχείο. Η τεχνική RSS επιτρέπει σε κάποιον όχι μόνο να συνδεθεί μέσω συνδέσμου (link) με μία ιστοσελίδα, αλλά και να γίνει συνδρομητής σε αυτή, με πλήρη ενημέρωσή του για κάθε αλλαγή της σελίδας. Αυτή η κατάσταση ονομάζεται “incremental web”(αυξητικό δίκτυο) ή “live web”(ζωντανό δίκτυο).

Η RSS είναι ένας εναλλακτικός τρόπος ενημέρωσης των χρηστών. Επιτρέπει στο χρήστη να βλέπει πότε ανανεώθηκε το περιεχόμενο των δικτυακών τόπων που τον ενδιαφέρουν. Μπορεί να λαμβάνει κατευθείαν στον υπολογιστή του τους τίτλους των τελευταίων ειδήσεων και των άρθρων που επιθυμεί, ή ακόμα και εικόνων ή βίντεο, αμέσως μόλις αυτά γίνουν διαθέσιμα, χωρίς να είναι απαραίτητο να επισκέπτεται καθημερινά τους αντίστοιχους δικτυακούς τόπους. Για να μπορεί ο χρήστης να κάνει χρήση της RSS τεχνικής θα πρέπει να προμηθευτεί ένα πρόγραμμα ανάγνωσης ειδήσεων (RSS reader). Το πρόγραμμα αυτό είναι ένα ειδικό λογισμικό στο οποίο προσθέτει τις σελίδες RSS που τον ενδιαφέρουν και αυτό με τη σειρά του ελέγχει τις σελίδες αυτές και τον ενημερώνει διαρκώς για οτιδήποτε νέο. Αφού επιλέξει πρόγραμμα ανάγνωσης, θα πρέπει να αποφασίσει ποιο περιεχόμενο θέλει να λαμβάνει. Ο χρήστης θα πρέπει να αναζητήσει στο Διαδίκτυο και στους αγαπημένους του δικτυακούς τόπους τις σελίδες RSS που τον ενδιαφέρουν και να γραφτεί συνδρομητής σε αυτές [4].

Παράδειγμα xml ενός RSS feed

```
<?xml version="1.0" encoding="UTF-8" ?>
<rss version="2.0">
<channel>
 <title>RSS Title</title>
 <description>This is an example of an RSS feed</description>
 <link>http://www.example.com/main.html</link>
 <lastBuildDate>Mon, 06 Sep 2010 00:01:00 +0000 </lastBuildDate>
 <pubDate>Sun, 06 Sep 2009 16:20:00 +0000</pubDate>
 <ttl>1800</ttl>

 <item>
  <title>Example entry</title>
  <description>Here is some text containing an interesting description.</description>
  <link>http://www.example.com/blog/post/1</link>
  <guid isPermaLink="true">7bd204c6-1655-4c27-aeee-53f933c5395f</guid>
  <pubDate>Sun, 06 Sep 2009 16:20:00 +0000</pubDate>
 </item>

</channel>
</rss>
```

Εικόνα 2 - Παράδειγμα XML αρχείου για RSS feed(wikipedia)

2.1.5. META TAGS

Τα meta tags είναι στοιχεία της HTML και η χρήση τους αποσκοπεί στην παροχή κάποιων συγκεκριμένων κατηγοριών δεδομένων (metadata) για μία ιστοσελίδα. Τοποθετούνται στον κώδικα της ιστοσελίδας, στο head, και οι πληροφορίες που περιέχουν παρέχονται από τον webmaster. Οι βασικές κατηγορίες δεδομένων, για τις

οποίες θα μιλήσουμε παρακάτω, είναι ο τίτλος της ιστοσελίδας (title), η περιγραφή της (meta description), καθώς και οι σχετικές λέξεις κλειδιά (meta keywords). Τα δεδομένα αυτά, με εξαίρεση τον τίτλο της ιστοσελίδας, δεν είναι ορατά από τον επισκέπτη της ιστοσελίδας, καθώς η λειτουργικότητα τους αφορά κατά κύριο λόγο στις μηχανές αναζήτησης. Οι μηχανές αναζήτησης χρησιμοποιούν τις πληροφορίες που εμπεριέχονται στα meta tags, ως συμπληρωματικές της ανάλυσης που κάνουν στο περιεχόμενο κάθε ιστοσελίδας, κατά την αξιολόγηση της για την κατάταξη στα αποτελέσματα αναζήτησης. Φυσικά τα meta tags δεν είναι από μόνα τους αρκετά για να επιτύχετε υψηλές θέσεις κατάταξης, καθώς οι μηχανές αναζήτησης λαμβάνουν υπόψη δεκάδες κριτήρια [5]. Κάθε υποσελίδα του site, εφόσον έχει διαφορετικό θέμα, καλό είναι να έχει και τα δικά της meta tags, τη δική της περιγραφή και τα δικά της keywords τα οποία θα εστιάζουν στη λίστα με τις λέξεις-κλειδιά που δημιουργήσαμε στο προηγούμενο κεφάλαιο.

Παράδειγμα κώδικα html

```
<meta name="GENERATOR" content="Netvolution WCM" />
<meta name="PageHandler" content="Netvolution.Site.Engine.PageHandler" />
<meta name="DESCRIPTION" content="Οι επιστήμονες του Πανεπιστημίου .... "/>
<meta name="KEYWORDS" content="dna,υπολογιστή,υπολογιστή dna,ένα..." />
```

Εικόνα 3 - Παράδειγμα κώδικα html για meta tags

2.2. Εξαγωγή Λέξεων-Φράσεων Κλειδιών από Κείμενα (Keyword-Keyphrase Extraction)

Η Εξαγωγή Λέξεων-Φράσεων Κλειδιών από κείμενα αποτελεί ένα ξεχωριστό και ραγδαίως αναπτυσσόμενο κλάδο της ερευνητικής κοινότητας τα τελευταία χρόνια, καθώς μεγιστοποιείται ολοένα και περισσότερο η σημασία της γρηγορότερης και ορθότερης διαχείρισης του τεράστιου όγκου δεδομένων κειμένου Online, ο οποίος μεγαλώνει εκθετικά παράλληλα με τον Παγκόσμιο Ιστό.

Οι μέθοδοι εξαγωγής λέξεων-φράσεων κλειδιών από κείμενα παρουσιάζουν μεγάλη ποικιλομορφία και πολλές διαφορετικές προσεγγίσεις από απλοϊκές με τη χρήση ενός μόνο αλγόριθμου, μέχρι και συνδυασμούς αλγορίθμων, λεξιλογίων, ομαδοποίησης και μηχανικής μάθησης[6].

Μερικά από αυτά τα εργαλεία με τη μεγαλύτερη αποτελεσματικότητα είναι τα εξής[7]:

Carrot2: Χρησιμοποιεί δύο αλγόριθμους STC και Lingo για την αναζήτηση πλήρεις φράσεων – κλειδιών με κάποιους περιορισμούς

KEA: Είναι ένας πρότυπος αλγόριθμος για εξόρυξη λέξεων-φράσεων κλειδιών. Παρέχει πρόβλεψη της μάθησης από το λεξικό RDF (σε μορφή SKOS). Το λεξικό περιέχει ιεραρχική ταξινόμηση. Δίνει επίσης επιλογές για Machine Learning μέσω Weka.

Maui: Χρησιμοποιεί σαν βασικό εργαλείο το KEA, αλλά δίνει επιλογές για την ενίσχυση του λεξιλογίου από την Wikipedia.

wikiFier: Μοιάζει με τον αλγόριθμο Maui. Χρησιμοποιεί επίσης wikipedia για την εξαγωγή φράσεων κλειδιών .

Stanford: Χρησιμοποιεί LDA για την εκμάθηση. Παρέχει επίσης επιλογές για Machine Learning.

Από αυτά επιλέχτηκε η χρήση του OpenSource εργαλείου **KEA Algorithm** για την εξαγωγή λέξεων κλειδιών για αυτή τη διπλωματική εργασία.

2.2.1. Αλγόριθμος KEA

Ο αλγόριθμος KEA – Keyphrases Extraction Algorithm είναι απλός και αποτελεσματικός και χρησιμοποιεί τον αλγόριθμο μηχανικής μάθησης Naive Bayes για

εκπαίδευση και εξόρυξη φράσεων κλειδιά. Χρησιμοποιείται για αυτόματη εξόρυξη λέξεων-φράσεων κλειδιών από κείμενα. Έχει την δυνατότητα να μπορεί να αναγνωρίζει υποψήφιες φράσεις κλειδιά, να υπολογίζει τιμές χαρακτηριστικών για κάθε υποψήφια φράση κλειδί και χρησιμοποιεί έναν αλγόριθμο μηχανικής μάθησης προκειμένου να προβλέψει ποιες υποψήφιες λέξεις-φράσεις κλειδιά είναι οι ιδανικές και χαρακτηρίζουν το έγγραφο που τις περιέχει. Το σχήμα μηχανικής μάθησης χτίζει ένα μοντέλο πρόβλεψης χρησιμοποιώντας έγγραφα εκπαίδευσης με γνωστές φρά-σεις κλειδιά και στη συνέχεια το μοντέλο αυτό χρησιμοποιείται για να εντοπίσει φράσεις κλειδιά σε νέα έγγραφα. Ο αλγόριθμος KEA είναι απλός, ισχυρός με πολύ καλά αποτελέσματα και ελεύθερα διαθέσιμος.

Η υλοποίηση του αλγορίθμου είναι διαθέσιμη στην ηλεκτρονική Ψηφιακή βιβλιοθήκη της Νέας Ζηλανδίας http://www.nzdl.org. Το έργο Ψηφιακή βιβλιοθήκη της Νέας Ζηλανδίας είναι ένα ερευνητικό πρόγραμμα του πανεπιστημίου του Waikato, στόχος του οποίου είναι να αναπτυχθεί η υπάρχουσα τεχνολογία για τις ψηφιακές βιβλιοθήκες και να διατεθεί στο κοινό προκειμένου να χρησιμοποιηθεί από άλλους για την δημιουργία νέων συλλογών.
Ο στόχος του αλγόριθμου KEA είναι να παρέχει χρήσιμα μεταδεδομένα που δεν υπήρχαν πριν [8].

2.2.2. Στάδια αλγόριθμου KEA

Ο αλγόριθμος KEA έχει δύο στάδια:
1. *Εκπαίδευση*: Το στάδιο αυτό περιλαμβάνει την δημιουργία ενός μοντέλου για τον εντοπισμό λέξεων-φράσεων κλειδιά χρησιμοποιώντας έγγραφα εκπαίδευσης για τα οποία είναι γνωστές οι λέξεις-φράσεις κλειδιά.
2. *Εξόρυξη*: Το στάδιο αυτό περιλαμβάνει την εξαγωγή φράσεων ≪κλειδιά≫ από νέα έγγραφα χρησιμοποιώντας το μοντέλο του σταδίου 1.
Και στα δύο στάδια ο αλγόριθμος KEA επιλέγει μια συλλογή από υποψήφιες λέξεις κλειδιά και υπολογίζει τις τιμές συγκεκριμένων χαρακτηριστικών που θα αναλυθούν στη συνέχεια[8].

2.2.3. Επιλογή Υποψήφιων φράσεων

Ο αλγόριθμος KEA επιλέγει τις υποψήφιες φράσεις σε τρία στάδια. Αρχικά "καθαρίζει" το κείμενο που του δίνεται σαν είσοδος, μετά αναγνωρίζει τις υποψήφιες φράσεις και στη συνέχεια κάνει χρήση ενός στελεχωτή και συμπτύσσει τα πεζά και κεφαλαία γράμματα.

Η διαδικασία "καθαρισμού" του κειμένου εισόδου περιλαμβάνει τη χρήση φίλτρου κανονικοποίησης και αναγνώριση των αρχικών ορίων των φράσεων. Το κείμενο εισόδου διασπάται σε σύμβολα (tokens) και υποβάλλεται σε επεξεργασία.

Η επεξεργασία περιλαμβάνει:

1. Αντικατάσταση των σημείων στίξεως, παρενθέσεων και αριθμών με τα όρια φράσεων.

2. Αφαίρεση των απόστροφων.

3. Διάσπαση των ενωμένων με μεσαία παύλα λέξεων.

4. Διαγραφή των υπολοιπόμενων χαρακτήρων καθώς και των συμβόλων (tokens) που δεν περιέχουν γράμματα.

Το αποτέλεσμα της διαδικασίας καθαρισμού του κειμένου είναι ένα σύνολο από γραμμές κάθε μια από τις οποίες περιέχει μια ακολουθία συμβόλων κάθε ένα από τα οποία περιέχει τουλάχιστον ένα γράμμα.

Μετά την ολοκλήρωση του σταδίου "καθαρισμού" του κειμένου, ο αλγόριθμος KEA εξετάζει όλες τις υποακολουθίες λέξεων (n-grams) για ένα προκαθορισμένο μήκος το οποίο ορίζει ο χρήστης και αποφασίζει ποιές από αυτές είναι κατάλληλες υποψήφιες φράσεις.

Προκειμένου να αποφασίσει την καταλληλόλητα των υποψήφιων φράσεων, ακολουθεί τους παρακάτω απλούς και συνάμα αποτελεσματικούς **κανόνες**:

1. Οι υποψήφιες φράσεις έχουν ένα προκαθορισμένο ανώτατο όριο μήκους το οποίο μετριέται σε πλήθος λέξεων.

2. Οι υποψήφιες φράσεις δεν μπορεί να είναι κύρια ονόματα (λέξεις που εμφανίζονται πάντα με αρχικό κεφαλαίο γράμμα).

3. Οι υποψήφιες φράσεις δεν μπορούν να αρχίζουν ή να τελειώνουν με μια λέξη που ανήκει στη λίστα απαγορευμένων λέξεων.

Για παράδειγμα αν στην γραμμή εισόδου είχαμε σαν είσοδο το κομμάτι μιας πρότασης «η μέθοδος της εξαντλητικής ανάλυσης», οι υποψήφιες παραγόμενες φράσεις θα είναι:

1. «μέθοδος»

2. «εξαντλητικής»

3. «ανάλυσης»

4. «μέθοδος της εξαντλητικής ανάλυσης»

5. «εξαντλητικής ανάλυσης»

Οι φράσεις «η μέθοδος της εξαντλητικής ανάλυσης» και «της εξαντλητικής Ανάλυσης» δεν είναι υποψήφιες επειδή ξεκινούν με το άρθρο «η» το οποίο ανήκει στη λίστα απαγορευμένων λέξεων(stopwords). Επομένως δεν ισχύει ο τρίτος από τους κανόνες που αναφέραμε παραπάνω.

Το τελευταίο στάδιο της διαδικασίας επιλογής υποψήφιων φράσεων είναι η χρήση στελεχωτή(stemmer) και η σύμπτυξη πεζών και κεφαλαίων γραμμάτων.Ο στελεχωτής που χρησιμοποιείται είναι ο Lovins ο οποίος αφαιρεί την κατάληξη από κάθε σύμβολο (token) και συνεχίζει την διαδικασία στο αποτέλεσμα μέχρι να μην είναι δυνατή κάποια επιπλέον αλλαγή.

Η διαδικασία λειτουργίας του στελεχωτή Lovins μπορεί να περιγραφεί ως εξής:

1. Αφαίρεσε την κατάληξη από ένα σύμβολο (token).

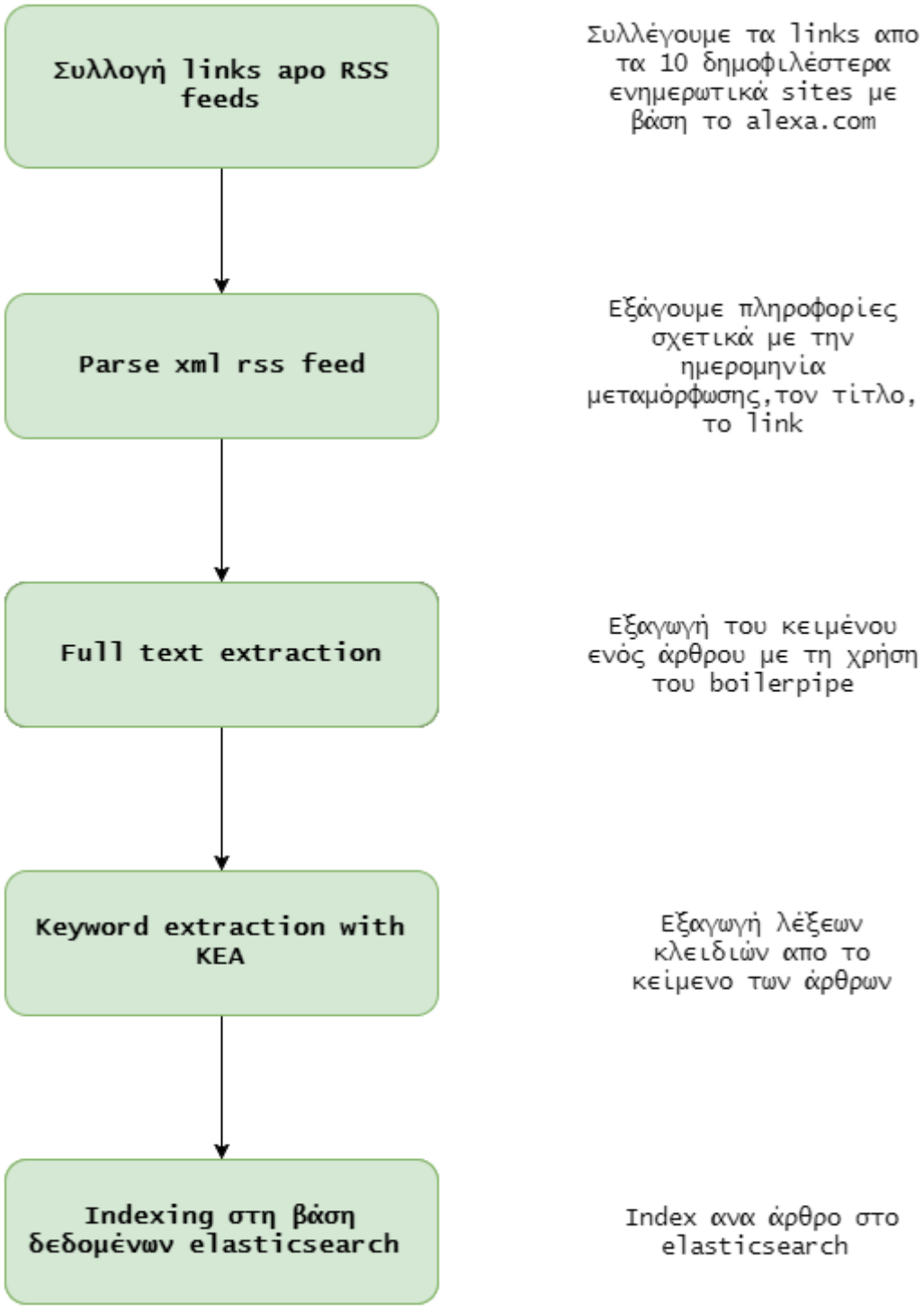
2. Αν δύναται να επιπλέον αλλαγή στο αποτέλεσμα του βήματος 1 τότε επανεκτέλεσε το βήμα 1; διαφορετικά επέστρεψε το αποτέλεσμα.

Η στελέχωση και η σύμπτυξη πεζών και κεφαλαίων επιτρέπει τον όμοιο χειρισμό των διαφορετικών παραλλαγών μιας φράσης. Για παράδειγμα οι φράσεις «Οικονομικό Πανεπιστήμιο Αθηνών» και «Οικονομικού Πανεπιστημίου Αθηνών» είναι ίδιες. Όμως χωρίς την χρήση στελεχωτή θα αναγνωρίζονταν από τον αλγόριθμο ΚΕΑ σαν διαφορετικές φράσεις.

Η πραγματική μορφή των συμβόλων (tokens) πριν την χρήση στελεχωτή διατηρείται από τον αλγόριθμο ΚΕΑ σε μια δομή δεδομένων. Όταν έχει ολοκληρωθεί η επιλογή

των φράσεων κλειδιά επιστρέφονται σαν αποτέλεσμα οι φράσεις στην αρχική τους μορφή και όχι με την μορφή που προέκυψε μετά τη χρήση στελεχωτή [8].

2.3. Υλοποίηση Scraper



Εικόνα 4 - Στάδια υλοποίησης του Scraper

2.3.1. Συλλογή Links από RSS

Ο στόχος του προγράμματος ήταν να εξάγουμε κείμενα και τις λέξεις κλειδιά τους έτσι ώστε να τις κατατάξουμε ανάλογα με τη συχνότητα και την ημερομηνία εμφάνισης τους(inverted index, postings list).

Αρχικά συγκεντρώνουμε τα links από τα **rss newsfeed** των site που θέλουμε να ελέγξουμε και τα αποθηκεύουμε σε ένα αρχείο .txt. Σε κάθε γραμμή του αρχείου υπάρχει ένα link με το xml αρχείο του rss.

2.3.2. Parse xml RSS feed

Διατρέχουμε το αρχείο ανά γραμμή και για κάθε link κάνουμε parsing το xml αρχείο του. Οι πληροφορίες που παίρνουμε είναι ο τίτλος του κάθε άρθρου, η ημερομηνία μεταφόρτωσης, το link του και ένα σύντομο description.

Το parsing αυτό γίνεται χρησιμοποιώντας τη βιβλιοθήκη ROME της java η οποία επιστρέφει μια λίστα με στοιχεία SyndEntry η οποία διαπερνάται μέσω ενός for loop. Για κάθε **SyndEntry** entry τιμή της μεταβλητή καλούμε τις συναρτήσεις entry.getLink(),entry.getPublishedDate() και entry.getDescription() οι οποίες επιστρέφουν τις αντίστοιχες τιμές.

Γράψαμε ένα πρόγραμμα το οποίο τρέχει μια φορά στο τέλος της μέρας και αναγνωρίζει μέσω του Published Date τα καινούργια άρθρα μιας λίστας από τα δημοφιλέστερα news sites (σύμφωνα με το alexa) και κάνουμε scraping τα δεδομένα τους.

2.3.3. Full text extraction

Στη συνέχεια έπρεπε να εξάγουμε το κείμενο από τα άρθρα και αυτό έγινε χρησιμοποιώντας τη βιβλιοθήκη **boilerpipe** [12] η οποία επιστρέφει ένα String με το κείμενο

```
String content;
final HTMLDocument htmlDoc = HTMLFetcher.fetch(new URL(url));
final TextDocument doc = new BoilerpipeSAXInput(
htmlDoc.toInputSource()).getTextDocument();
content = CommonExtractors.ARTICLE_EXTRACTOR.getText(doc);
```

Εικόνα 5 - Κώδικας ενσωμάτωσης της βιβλιοθήκης boilerpipe στην εφαρμογή

2.3.4. Μετατροπή KEA και Εκπαίδευση στα Ελληνικά για την εξαγωγή λέξεων κλειδιών από τα άρθρα που συλλέξαμε

Δεν υπήρχε έκδοση του που να δουλεύει με ελληνικά κείμενα επομένως τροποποιήσαμε τον αλγόριθμο του KEA για την εργασία μας.

Μετατροπή

Για να λειτουργήσει φτιάξαμε ένα αρχείο με ελληνικά stopwords, και αλλάξαμε την κωδικοποίηση σε UTF-8.

Εκπαίδευση

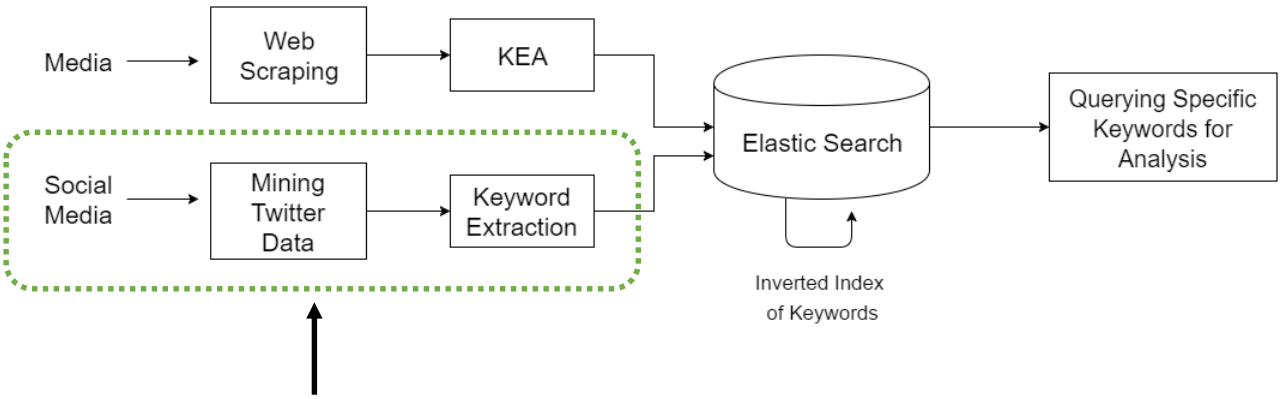
Όσον αφορά την εκπαίδευση του αλγορίθμου έπρεπε να χρησιμοποιήσουμε ένα ιδανικό δείγμα άρθρων γραμμένων στα ελληνικά με έτοιμα keywords. Ένα τέτοιο δείγμα δεν υπήρχε διαθέσιμο online, συνεπώς έπρεπε να το δημιουργήσουμε.
Για τη δημιουργία του χρησιμοποιήσαμε ειδησεογραφικά άρθρα που είχαν αναρτημένα meta tags, τα οποία ήταν πολύ αντιπροσωπευτικά, και τα χρησιμοποιήσαμε ως keywords.
Αυτό επιτεύχθηκε με ένα script σε γλώσσα Python και τη βιβλιοθήκη Goose. Συλλέξαμε ένα σύνολο, 1010 άρθρων με αντίστοιχα keywords για το κάθε άρθρο. Κατηγοριοποιήσαμε τα άρθρα ανάλογα με τη θεματολογία τους (πολιτικά, αθλητικά, κοινωνικά, διεθνή, οικονομικά και lifestyle) . Στη συνέχεια κάναμε training τον αλγόριθμο KEA (KEAmodelBuilder) με όλες τις θεματολογίες ξεχωριστά δημιουργώ-

ντας διαφορετικά settings για τον KeywordExtractor ο οποίος ανάλογα με τις ρυθμίσεις εξάγει από 1 έως 5 keywords ξεχωριστά για κάθε άρθρο.

2.3.5. Καταγραφή δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν για έρευνα στη παρούσα εργασία, αφορούν τον Νοέμβριο του 2016. Περιέχουν άρθρα από τα 10 δημοφιλέστερα news portals, από την 1η Νοεμβρίου του 2016 στις 00.00 μέχρι και την 30η του ίδιου μήνα στις 23.59. Ο συνολικός αριθμός τους είναι 1.233.474 άρθρα.

3. Ανάλυση-Σχεδίαση Συστήματος για Social Media Posts(Tweets)



Εικόνα 6 - Απεικόνιση Συστήματος (Δεύτερο στάδιο)

3.1 Συλλογή δεδομένων από Social Media (Twitter)

Τα Social Media μετέτρεψαν ριζικά την μονολογία των παραδοσιακών Μέσων Μαζικής Ενημέρωσης, σε έναν ευρύτερο διάλογο. Μείωσαν την νοητή απόσταση επικοινωνίας μεταξύ των ανθρώπων, προσφέροντας άμεση αλληλεπίδραση και διαδραστικότητα, κάνοντας την ανταλλαγή απόψεων θέμα γλώσσας και όχι γεωγραφικής θέσης.

Εξέγειραν και ενθάρρυναν την έκφραση άποψης, σχολιασμού και αναμετάδοσης των γεγονότων.

Με αυτό τον τρόπο ενίσχυσαν την αντικειμενικότητα της μετάδοσης των ειδήσεων, καταστώντας την ενημέρωση ορθότερη μέσω αυτών, αφού τα συγκεκριμένα γεγονότα που θα μεταδοθούν άμεσα ή θα αναμεταδοθούν είναι επιλογή του συνόλου (των χρηστών).

Πλέον δεν αποτελεί υπερβολή η εξίσωση της δυναμικής ενός μέσου κοινωνικής δικτύωσης, ως πηγή ενημέρωσης, με ένα παραδοσιακό Μέσο Μαζικής Ενημέρωσης

όπως η τηλεόραση ή οι εφημερίδες[9].

3.1.1. Η ψυχολογική τάση των ανθρώπων να εκφράζονται μέσω των Κοινωνικών Δικτύων

Η ανάγκη της έκφρασης είναι χαρακτηριστικό της ανθρώπινης φύσης. Ο άνθρωπος αναζητούσε πάντα την εξήγηση και τη διερεύνηση από οτιδήποτε τον περιέβαλλε. Επομένως, μέσω της επικοινωνίας, της ανταλλαγή ιδεών, απόψεων και εμπειριών του εκπληρώνεται η επιθυμία του να συνδεθεί με άλλους, πράγμα που δίνει αξία και σημασία στην ύπαρξή του.

Τα μέσα κοινωνικής δικτύωσης παρέχουν τη δυνατότητα αυτή· ενισχύοντας το κοινωνικό κεφάλαιο με άμεσο και γρήγορο τρόπο μέσω της τεχνολογίας τους. Αυτός είναι ο κύριος λόγος που οι περισσότεροι άνθρωποι τα χρησιμοποιούν σε τόσο μεγάλο βαθμό μεταφέροντας και καταγράφοντας καθημερινά τις σκέψεις τους [9].

3.2.2. Σημαντικότερα Μέσα Κοινωνικής Δικτύωσης

Οι δύο κυριότεροι εκφραστές των μέσων κοινωνικής δικτύωσης αυτή τη στιγμή είναι το Facebook και το Twitter, σύμφωνα με τον αριθμό των χρηστών τους και τη δημοφιλία τους.

Και οι δύο εταιρείες προσφέρουν μέσω των API τους τη δυνατότητα στους γνώστες προγραμματισμού να εξάγουν πληροφορία, μέσα από αυτά, προσαρμοσμένη στο ενδιαφέρον τους. Στοχεύοντας συγκεκριμένα χαρακτηριστικά όπως λέξεις κλειδιά ή χρήστες με κοινά γνωρίσματα.

Από τα δύο αυτά μέσα, αν και το Facebook αυτή τη στιγμή έχει τους περισσότερους χρήστες παγκοσμίως, επιλέχτηκε το Twitter για την εξαγωγή πληροφορίας.

Κυρίως γιατί το Facebook δίνει τη δυνατότητα στους χρήστες του να κάνουν δημοσιεύσεις (posts) φιλτράροντας το κοινό το οποίο μπορεί να τις δει (public, friends,

adjusted κλπ). Δυνατότητα που χρησιμοποιεί το μεγαλύτερο ποσοστό του. Εν αντιθέσει στο Twitter αν και παρέχεται κάτι παρόμοιο στους χρήστες του (να κάνουν την προσωπική τους σελίδα "προστατευμένη" δηλαδή μη ορατή σε χρήστες που δεν τους ακολουθούν) αυτό δεν έχει ευρεία απήχηση.

Ενώ σύμφωνα με τη δημοσιοποίηση των στοιχείων του, το ποσοστό όλο και μειώνεται (2009 είχε 10% προστατευμένους λογαριασμούς, ενώ το 2012 μόλις 2.3%) [13].

Το αποτέλεσμα είναι οποιαδήποτε δημοσίευση (post) κάνει ένας χρήστης να είναι προσβάσιμη από οποιονδήποτε (public), ακόμη και από κάποιον που δεν είναι εγγεγραμένος χρήστης στο εν λόγο κοινωνικό δίκτυο.

Η σημαντική αυτή διαφορά του Twitter μας δίνει τη δυνατότητα να έχουμε ένα πιο αντιπροσωπευτικό δείγμα, συνεπώς ασφαλέστερα και ορθότερα αποτελέσματα σε σχέση με το Facebook, όσον αφορά την αντίληψη της κοινής γνώμης [9].

3.1.3. Χαρακτηριστικά του Twitter

Το Twitter είναι μία υπηρεσία microblogging, η οποία επιτρέπει στους χρήστες (users) του να δημοσιεύουν την άποψη τους, με λακωνικό τρόπο. Δηλαδή μέσα σε ένα μικρό πλαίσιο 140 χαρακτήρων, τα λεγόμενα tweets.

Ένα τόσο μικρό πλαίσιο πολύ δύσκολα αντιστοιχεί σε σκέψεις ή ιδέες, με αποτέλεσμα ο χρήστης να καλείται να περιορίσει την ουσία του νοήματος από αυτό που θέλει να πει μέσα σε μετρημένες λέξεις. Ως εκτούτου δίνει μια ιδιαίτερη δυναμική και καθοριστική σημασία στη κάθε λέξη που χρησιμοποιείται.

Το συγκεκριμένο δίκτυο ενδείκνυται για σχολιασμό της επικαιρότητας.

Μία τέτοια βάση χρηστών είναι οι βαθύτερες δυναμικές δικτύου κάτω από το Twitter, οι οποίες και το καθιστούν τόσο σημαντικό. Ο δίαυλος επικοινωνίας που επιτρέπει στους χρήστες να μοιράζονται σύντομα αποφθέγματα με ταχύτητα και ταυτόχρονα το ακολουθητικό (following) μοντέλο του Twitter το κάνει να μοιάζει περισσότερο ως ένα γράφημα ενδιαφέροντος παρά ως κοινωνικό δίκτυο. Έτσι, ενώ κάποιες ιστοσελίδες κοινωνικής δικτύωσης, όπως το Facebook και το LinkedIn, απαιτούν την αμοιβαία αποδοχή της σύνδεσης μεταξύ των χρηστών (υπονοώντας μία σύνδεση

στον πραγματικό κόσμο), το σχεσιακό μοντέλο του Twitter επιτρέπει σε ένα χρήστη να παρακολουθεί τις τελευταίες ενέργειες οποιουδήποτε άλλου χρήστη, ακόμα και αν αυτός ο άλλος χρήστης δεν επιλέξει να ακολουθεί τον πρώτο (follow back). Το Twitter μοιάζει πιο πολύ με μία εφαρμογή που συγκεντρώνει όλα τα blogs που ενδιαφέρουν κάποιον, και τον ενημερώνει για οτιδήποτε νέο γράφτηκε σε αυτά. Παράλληλα φιλοξενεί και τη δική του σελίδα blog. Είτε πρόκειται για ενδιαφέρον σε ένα συγκεκριμένο πολιτικό ή κοινωνικό θέμα, για ενημέρωση πάνω στις τελευταίες τεχνολογικές εξελίξεις ή σε κάποια αθλητική ομάδα, για κουτσομπολιό ή επιθυμία για επικοινωνία με κάποιον άλλο, το Twitter προσφέρει απεριόριστες ευκαιρίες για να ικανοποιηθεί οποιαδήποτε εκφραστική τάση της επικαιρότητας [9].

3.1.4. Στοιχεία και πληροφορίες μέσα σε ένα tweet

Οι χρήστες ενημερώνουν τη κατάστασή τους (status update), με *tweets* (τιτιβίσματα), τα οποία εμφανίζονται στο *timeline* (χρονοδιάγραμμα) του κάθε χρήστη. Τα tweets μπορούν να περιλαμβάνουν μία ή περισσότερες οντότητες (*entities*) ανάμεσα στους 140 χαρακτήρες περιεχομένου τους και να αναφέρουν μία ή περισσότερες θέσεις (*places*) που αντιστοιχούν σε τοποθεσίες στον πραγματικό κόσμο.

Τα tweets είναι η ουσία του Twitter και ενώ θεωρητικά είναι οι 140 χαρακτήρες περιεχομένου που σχετίζονται με την ενημέρωση κατάστασης του χρήστη, στην πραγματικότητα υπάρχουν πολλά περισσότερα μεταδεδομένα (metadata) εντός τους.

Εκτός από το περιεχόμενο κειμένου του ίδιου του tweet, τα tweets συνοδεύονται από δύο επιπλέον κομμάτια μεταδεδομένων ιδιαίτερης σημασίας: οντότητες και θέσεις. Οι οντότητες ενός tweet είναι ουσιαστικά αναφορές χρηστών, hashtags, διευθύνσεις URL, και δεδομένα (media) που μπορεί να σχετίζονται με ένα tweet. Ενώ, οι θέσεις είναι τοποθεσίες στον πραγματικό κόσμο που μπορεί να συνδέονται με ένα tweet, είτε η πραγματική θέση στην οποία είχε συγγραφεί, είτε μία αναφορά σε κάποια θέση που περιγράφεται στο περιεχόμενό του.

Παρακάτω εξετάζεται ένα tweet με το εξής περιεχόμενο :

Don't believe the main stream (fake news) media @nytimes & @washingtonpost.The White House is running VERY WELL. I inherited a MESS and am in the process of fixing it.

#fakenews

45.wh.gov/XYQXNw

Το tweet αυτό αποτελείται από 140 χαρακτήρες και περιέχει τις εξής τρεις οντότητες:

· αναφορά στο χρήστη @nytimes @washingtonpost

· το hashtag #fakenews

· και το http URL: 45.wh.gov/XYQXNw

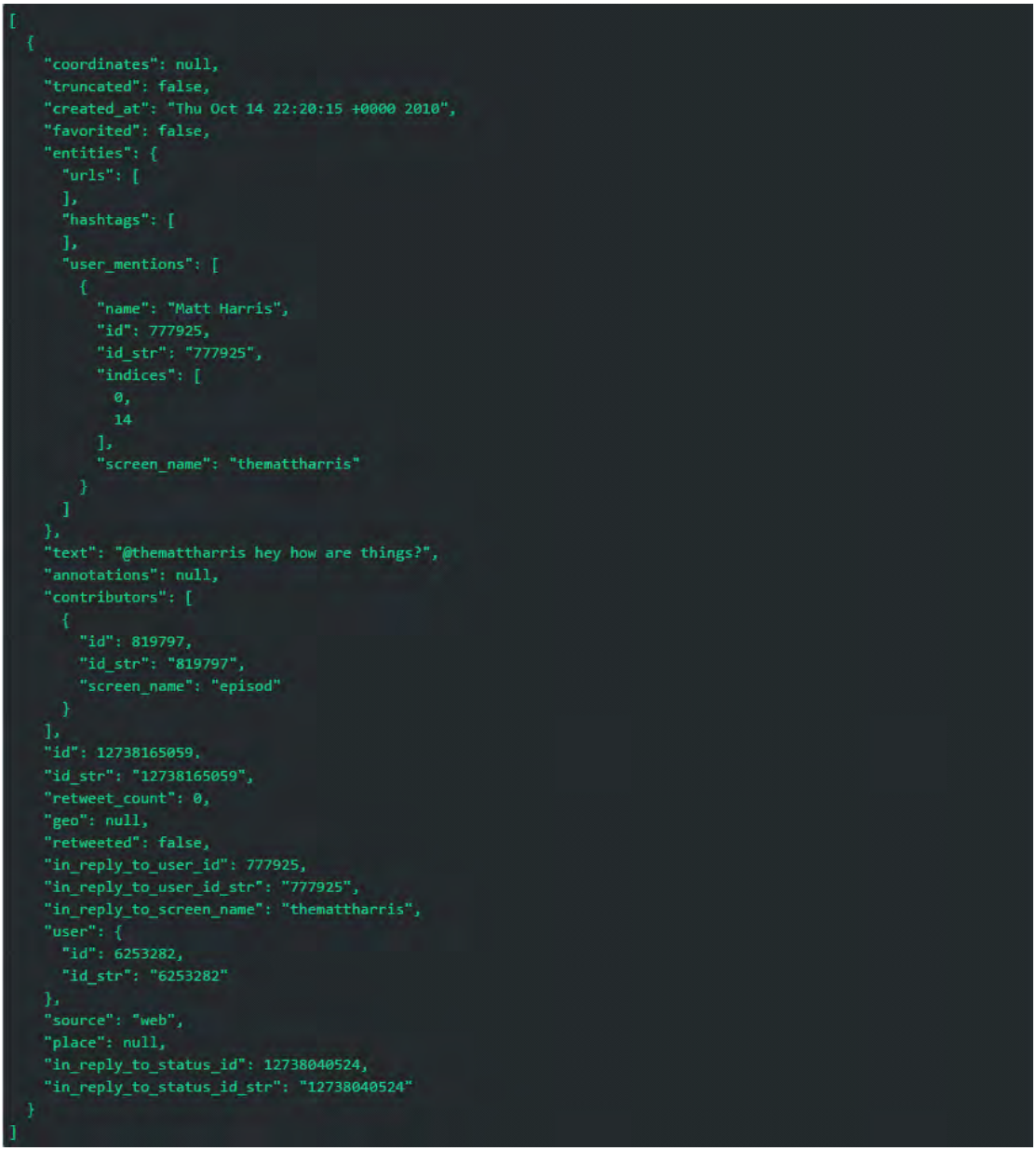
Όλα αυτά είναι αρκετά μεταδεδομένα που εμπεριέχονται σε κάτι λιγότερο από 140 χαρακτήρες και δείχνουν ακριβώς πόσο ισχυρό μπορεί να είναι ένα μικρό tweet. Το οποίο όπως φάνηκε μπορεί να αναφέρεται σε πολλούς άλλους χρήστες του Twitter, συνδέσεις σε ιστοσελίδες και παραπομπές σε διάφορα θέματα μέσω hashtags που δρουν ως σημεία συγκέντρωσης των tweets που αναφέρονται σε αυτό το θέμα σε όλο το Twitter, για εύκολη αναζήτηση.

Τέλος, τα timelines είναι χρονολογικά ταξινομημένες συλλογές από tweets. Με την αφηρημένη έννοια, ένα timeline είναι οποιαδήποτε συλλογή από tweets που εμφανίζονται με χρονολογική σειρά, ωστόσο, δύο μόνο timelines είναι σημαντικά. Το *home timeline,* το οποίο εμφανίζεται μόλις ένας χρήστης συνδέεται στο λογαριασμό του και περιέχει όλα τα tweets από τους χρήστες που ακολουθεί, και το *user timeline,* που είναι μία συλλογή από tweets ενός ορισμένου χρήστη.

Ενώ τα timelines είναι συλλογές από tweets με σχετικά χαμηλή ταχύτητα, τα *streams* είναι δείγματα δημόσιων tweets που ρέουν μέσω του Twitter σε πραγματικό χρόνο και μπορεί συγκεντρώσει εκατοντάδες χιλιάδες tweets ανά λεπτό κατά τη διάρκεια εκδηλώσεων με ιδιαίτερα μεγάλο ενδιαφέρον, όπως προεδρικές συζητήσεις, εκλογές, μεγάλα αθλητικά γεγονότα κ.α.. Όλος αυτός ο όγκος δεδομένων παρουσιάζει ενδιαφέρουσες τεχνολογικές (engineering) προκλήσεις και είναι ένας σημαντικός λόγος που διάφορες εταιρίες έχουν συνεργαστεί με το Twitter για να μετατρέψουν αυτόν τον όγκο σε μία πιο καταναλωτική μορφή[9].

3.1.5. JSON Format (Snowflake)

Το κάθε tweet είναι αποθηκευμένο και γραμμένο σύμφωνα με τη μορφή JSON Object [10].
Η JSON η αναλυτικότερα JavaScript Object Notation, είναι ένας απλός τρόπος για αποθήκευση πληροφοριών με καλή οργάνωση για ευκολότερη πρόσβαση στα στοιχεία του.
Το κάθε tweet έχει το δικό του ξεχωριστό ID. Αυτό υλοποιείται από το Snowflake μια υπηρεσία που χρησιμοποιείται για να δημιουργήσει μοναδικά αναγνωριστικά για τα αντικείμενα στο Twitter (Tweets, άμεσα μηνύματα, Χρήστες, Συλλογές, κατάλογοι κ.λπ.). Αυτά τα αναγνωριστικά είναι μοναδικοί 64-bit unsigned ακέραιοι, οι οποίοι βασίζονται στην ώρα που δημιουργούνται, αντί να είναι διαδοχικοί.
Το JSON format ενός Tweet φαίνεται στο σχήμα.



```
[
  {
    "coordinates": null,
    "truncated": false,
    "created_at": "Thu Oct 14 22:20:15 +0000 2010",
    "favorited": false,
    "entities": {
      "urls": [
      ],
      "hashtags": [
      ],
      "user_mentions": [
        {
          "name": "Matt Harris",
          "id": 777925,
          "id_str": "777925",
          "indices": [
            0,
            14
          ],
          "screen_name": "themattharris"
        }
      ]
    },
    "text": "@themattharris hey how are things?",
    "annotations": null,
    "contributors": [
      {
        "id": 819797,
        "id_str": "819797",
        "screen_name": "episod"
      }
    ],
    "id": 12738165059,
    "id_str": "12738165059",
    "retweet_count": 0,
    "geo": null,
    "retweeted": false,
    "in_reply_to_user_id": 777925,
    "in_reply_to_user_id_str": "777925",
    "in_reply_to_screen_name": "themattharris",
    "user": {
      "id": 6253282,
      "id_str": "6253282"
    },
    "source": "web",
    "place": null,
    "in_reply_to_status_id": 12738040524,
    "in_reply_to_status_id_str": "12738040524"
  }
]
```


Εικόνα 7 - Tweet Json Format(dev.twitter.com)

3.2. Υλοποίηση Script που να εξάγει τα tweets που θέλουμε

Για τη δημιουργία οποιασδήποτε εφαρμογής που έχει να κάνει με το API του Twitter, απαιτείται η απόκτηση ενός πρωτοκόλλου εξουσιοδότησης που αντιστοιχεί στην εφαρμογή. Το πρωτόκολλο αυτό αποκτάται απλά με τη δημιουργία λογαριασμού χρήστη στο Twitter, και ονομάζεται OAuth [9].

3.2.1. OAuth

Το OAuth είναι ένα πρωτόκολλο εξουσιοδότησης που επιτρέπει την αποστολή ασφαλών εξουσιοδοτημένων αιτήσεων προς το Twitter API. Οι χρήστες δεν χρειάζεται να μοιράζονται τους κωδικούς τους με άλλες εφαρμογές, αυξάνοντας έτσι την ασφάλεια του λογαριασμού τους. Επιπλέον, υπάρχουν πολλές συμβατές βιβλιοθήκες με την υλοποίηση του OAuth στο Twitter, που το καθιστούν πρότυπο πρωτόκολλο ασφαλείας.

3.2.2. API

Η Διεπαφή Προγραμματισμού Εφαρμογών, ή API (Application Programming Interface) είναι ένα εργαλείο το οποίο διευκολύνει την αλληλεπίδραση με προγράμματα ηλεκτρονικών υπολογιστών και υπηρεσίες web.
Πολλές υπηρεσίες web παρέχουν APIs στους προγραμματιστές προκειμένου να αλληλεπιδρούν με τις υπηρεσίες τους και να έχουν πρόσβαση σε δεδομένα με ένα προγραμματιστικό τρόπο.
Συγκεκριμένα το Twitter παρέχει αρκετά διαφορετικά APIs για διαφορετικές χρήσεις. Στη παρούσα εργασία μας απασχόλησαν μόνο τα δύο από αυτά.

3.2.3. REST API

Η αρχική προσέγγιση της εξαγωγής των tweets ήταν η εύρεση όλων των ελληνικών λογαριασμών του Twitter. Η αποθήκευση όλων των tweets που είχαν γράψει μεχρι τώρα και ο διαχωρισμός τους ανάλογα με την ημερομηνία, ταξινομώντας τα σε διαφορετικές μέρες του ενός μήνα. Η εν λόγο συλλογιστική ήταν εφικτή μόνο με τη χρήση του REST API.
Το REST (Representational State Transfer) APIs παρέχει για ανάγνωση και εγγραφή δεδομένων του Twitter προσδιορίζοντας εφαρμογές και χρήστες του Twitter χρησιμοποιώντας το πρωτόκολλο ασφαλείας OAuth. Επιτρέπει την αναζήτηση και

εξαγωγή ενός υπάρχοντος συνόλου δεδομένων που έχει δημιουργηθεί από τα tweets που έχουν ήδη δημοσιευθεί.

Μέσα από τη χρήση του μπορεί κανείς να ζητήσει tweets που ταιριάζουν σε κάποια συγκεκριμένα κριτήρια αναζήτησης, όπως hashtags, ονόματα χρηστών, τοποθεσίες, θέσεις κλπ.

Με το REST API του Twitter, οι προγραμματιστές αναζητούν ή τραβούν μόνο ένα συγκεκριμένο αριθμό από tweets που έχουν ήδη δημοσιευθεί, ο οποίος περιορίζεται από τα όρια ταχύτητας (Rate Limits) του Twitter. Για έναν μεμονωμένο χρήστη, ο μέγιστος αριθμός των tweets που μπορεί να λάβει είναι τα τελευταία 3.200 tweets, ανεξάρτητα από τα κριτήρια αναζήτησης. Επιπλέον, υπάρχει περαιτέρω περιορισμός στον αριθμό των αιτήσεων που μπορούν να γίνουν σε ένα ορισμένο χρονικό διάστημα (180 αιτήσεις σε διάστημα 15 λεπτών).

Εξαιτίας των ορίων αυτών, η έγκαιρη εφαρμογή της παραπάνω συλλογιστικής ήταν πολύ δύσκολη.

3.2.4. STREAMING API

Επομένως αποφασίστηκε η αλλαγή του σεναρίου, με τη χρήση του Streaming API.

Το Streaming API του Twitter επιτρέπει την αναζήτηση δεδομένων, σε σχεδόν πραγματικό χρόνο, από tweets τα οποία μόλις έχουν δημοσιευθεί, χρησιμοποιώντας είτε βασικά είτε OAuth πρωτόκολλα ασφαλείας. Με το Streaming API, οι χρήστες καταχωρούν συγκεκριμένα κριτήρια (hashtags, ονόματα χρηστών, τοποθεσίες, θέσεις κλπ.) και όσο δημοσιεύονται tweets που ταιριάζουν με τα κριτήρια αυτά, ωθούνται απευθείας στο χρήστη μαζί με πληροφορίες για το συγγραφέα του εκάστοτε tweet.

Είναι περισσότερο ώθηση δεδομένων από το Twitter, παρά τράβηγμα των δεδομένων

από τον τελικό χρήστη.

Το Streaming API έχει το μειονέκτημα πως το πραγματικό ποσοστό του συνόλου των tweets, που λαμβάνουν οι χρήστες, ποικίλλει σε μεγάλο βαθμό με βάση τα κριτήρια αναζήτησης και τις τρέχουσες κυκλοφοριακές συνθήκες.

Παρ’ όλα αυτά τα οφέλη της ύπαρξης μίας πραγματικού χρόνου ροής δεδομένων Twitter, καθιστούν ιδιαίτερα σημαντική την ενσωμάτωση του Streaming API σε διάφορους τύπους εφαρμογών.

Μία σύνδεση με το Streaming API απαιτεί τη διατήρηση μίας μόνιμης ανοιχτής HTTP σύνδεσης. Όπως φαίνεται στην Εικόνα 5, η διαδικασία Streaming λαμβάνει τα tweets, εκτελεί ανάλυση, φιλτράρισμα ή συνάθροιση – ανάλογα με τις απαιτήσεις και αποθηκεύει το αποτέλεσμα σε μία αποθήκη δεδομένων, απ’ όπου η διαδικασία χειρισμού HTTP αναζητεί αποτελέσματα σε απάντηση των αιτημάτων του χρήστη.

Συνεπώς, αντί της προηγούμενης προσέγγισης, κατα την οποία στοχεύαμε απεθείας σε όλους τους Έλληνες χρήστες του Twitter, η καινούργια προσέγγιση στόχευε απευθείας σε λέξεις που ήταν γραμμένες στα ελληνικά.

Αυτό επιτεύχθηκε με τη χρήση ελληνικών Stopwords ('ο', 'ή', 'είναι' κλπ) αλλά και συντομογραφιών αυτών, έτσι ώστε να "τραβήξουμε" σχεδόν οποιοδήποτε Tweet αφορούσε το ελληνικό κοινό του Twitter.

Η χρήση του Streaming API έγινε μέσω της βιβλιοθήκης Tweepy. Μια απλή σε χρήση βιβλιοθήκη της Python για να την εύκολη διαχείρηση του API του Twitter.

3.2.5. Καταγραφή Δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν για έρευνα στη παρούσα εργασία, αφορούν τον Νοέμβριο του 2016. Και περιέχουν Tweets, από την 1η Νοεμβρίου του 2016 στις 00.00 μέχρι και την 30η του ίδιου μήνα στις 23.59. Ο συνολικός αριθμός τους είναι 1.258.400 tweets.

3.3. Keyword Extraction από Tweet

3.3.1. Αιτίες μη αποτελεσματικότητας των εξελιγμένων Extractors

Το Keyword Extraction από Tweets διαφέρει σε μεγάλο βαθμό από αυτό σε News [11]. Και εργαλεία όπως το KEA δεν είναι κατάλληλα, γιατί βασίζονται κατά ένα μέρος στην επανάληψη των φράσεων κλειδιών. Τα άρθρα συνήθως έχουν μέγεθος

μεγαλύτερο από μία παράγραφο, αφού δεν υπάρχει περιορισμός στον αριθμό των χαρακτήρων που θα χρησιμοποιηθεί για να γραφτούνε, όπως συμβαίνει στα tweets. Τα datasets των tweets, περιέχουν μερικό μέρος του περιεχομένου του κάθε tweet(λόγω των #hastags, των @mentions, και των urls), συχνή χρήση λεξιλογικών παραλλαγών για απλές λέξεις (που αποτελεί πρόβλημα για τη χρήση λεξικού που κάνουν τα ανεπτυγμένα εργαλεία εξαγωγής) και παρατηρείτε υψηλή απόκλιση του πραγματικού πλήθους των keywords που χρησιμοποιούνται σε κάθε tweet.

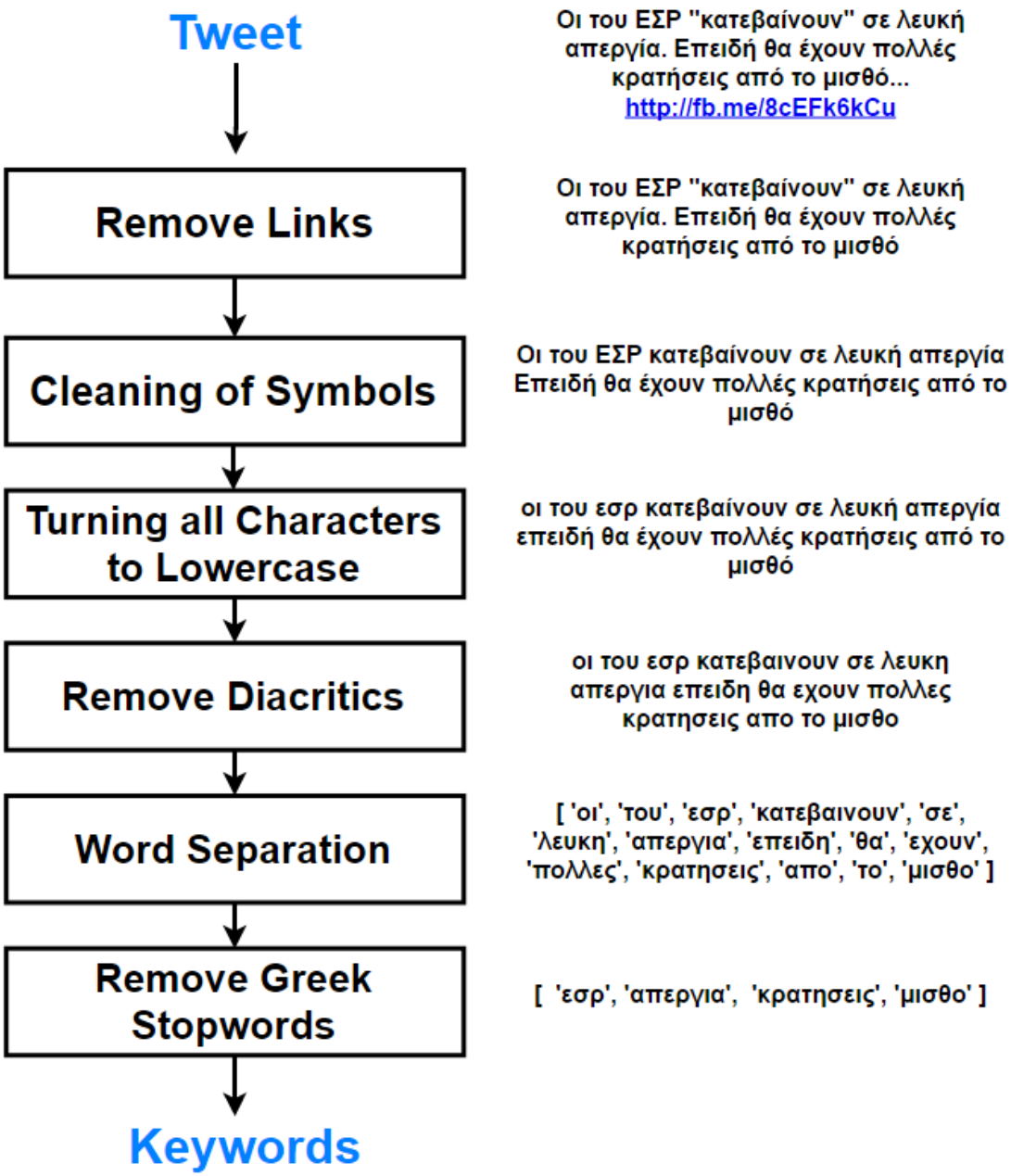
Με άλλα λόγια τα microbloggs τείνουν να είναι πολύ μικρότερα από ιστοσελίδες με άρθρα, ειδικά στο Twitter, όπου το περιεχόμενο πρέπει να περιορίζεται σε 140 χαρακτήρες. Η γλώσσα είναι, επίσης, πιο casual με πολλά μηνύματα που περιέχουν ορθογραφικά λάθη, αργκό, και συντομογραφίες (πχ "σμρ", αντί για "σήμερα"). Επίσης, δεν υπάρχει αυστηρή τήρηση σαφήνειας.

Πιο συγκεκριμένα, εξελιγμένα Keyword Extraction Tools, όπως το KEA, συνήθως χρησιμοποιούν TF-IDF (συχνότητα όρου - αντίστροφη συχνότητα εγγράφων). Η συχνότητα όρου στο δεδομένο έγγραφο δίνει ένα μέτρο σπουδαιότητας για τον όρο μέσα στο έγγραφο. Η συχνότητα εγγράφων είναι ένα μέτρο γενικής σπουδαιότητας του όρου (είναι ο λόγος του αριθμού όλων των εγγράφων διαιρεμένος με τον αριθμό εγγράφων που περιέχουν τον όρο). Κατά τη λειτουργία τους, k λέξεις με την υψηλότερη αξία TF-IDF επιλέγονται ως λέξεις-κλειδιά. Ο αριθμός k είναι προσαρμοσμένος ανάλογα με το κείμενο προς επεξεργασία. Αυτό εφαρμόζεται αρκετά αποτελεσματικά σε κείμενα , όπως άρθρα ειδήσεων, και ηλεκτρονικά βιβλία, επειδή οι όροι που αναζητούνται ως λέξεις κλειδιά τείνουν να εμφανίζονται συχνά μέσα σε αυτά. Αλλά κάτι τέτοιο δε συμβαίνει σε κείμενα με μέγεθος τόσο μικρό όσο αυτό που επιτρέπει το Twitter. Στα tweets οι προτάσεις τείνουν να είναι σύντομες και γενικά οι όροι τους να εμφανίζονται μόνο μία φορά, συμπεριλαμβανομένων άρα και των όρων που αντιπροσωπεύουν τις λέξεις-κλειδιά τους. Ως εκ τούτου το χαρακτηριστικό της συχνότητας-όρου, δεν είναι πολύ αποτελεσματικό καθώς με το TF-IDF θα επωφεληθούν απλά οι λέξεις που εμφανίζονται σπάνια, αφού αυτές έχουν πολύ χαμηλή αντίστροφη συχνότητα εγγράφου(IDF).

3.3.2. Υπαρκτά Εργαλεία και Προσέγγιση(Brute Force)

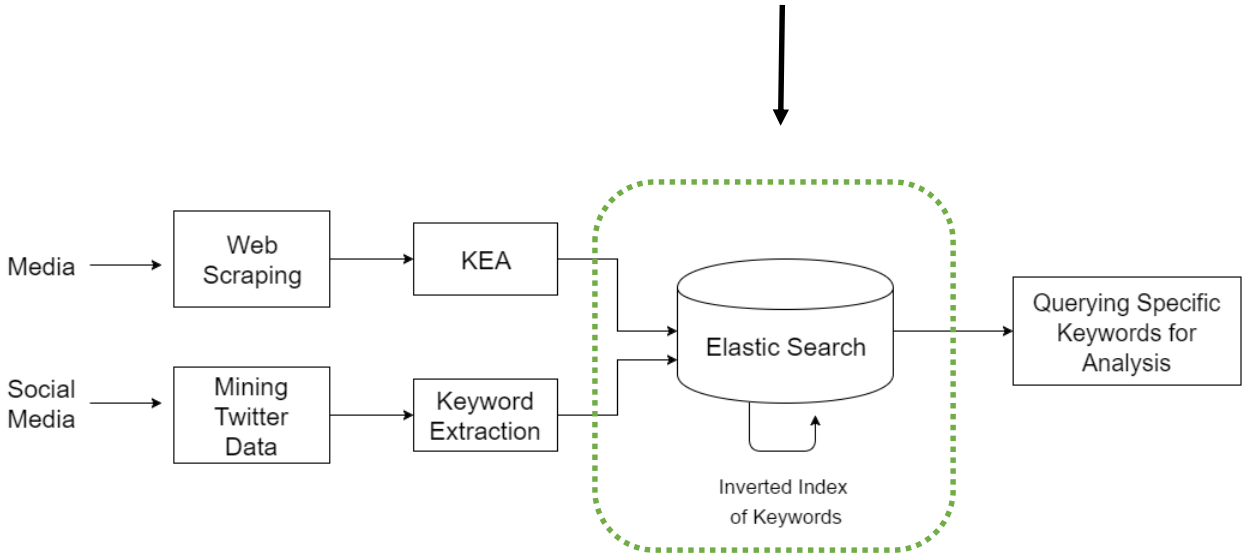
Έχουν επιτευχθεί προσπάθειες δημιουργίας έξυπνων εργαλείων για την εξαγωγή λέξεων κλειδιών αποκλειστικά για το Twitter(όπως αυτή του Luis Marujo του Wang Ling, ή το Twitter Keyword Graph). Παρόλα αυτά τα εργαλεία αυτά είτε δεν είναι διαθέσιμα, αυτή τη στιγμή, είτε δεν έχουν υλοποιηθεί για ελληνικά. Πραγματοποιήσαμε απόπειρα δημιουργίας εφαρμογής σύμφωνα με το συνδυασμό των αλγορίθμων που χρησιμοποιούν τα παραπάνω εργαλεία(Word Vectores και Brown Clusters), αλλά δεν αποτελούσε εφικτή επιλογή η υλοποίηση εκτελέσιμου που να βασιζόταν στους αλγορίθμους τους, καθώς είναι ιδιαίτερα εξειδικευμένη, και κάτι τέτοιο θα αποτελούσε άλλη εργασία. Επομένως, αρκεστήκαμε σε μία απλή προσέγγιση λύσης(Brute Force). Τη δημιουργία εφαρμογής που καθαρίζει ένα Tweet από το "θόρυβο", δηλαδή τις λέξεις που δεν είναι λέξεις κλειδιά. Αυτό πραγματοποιήθηκε με την συνεχή ενίσχυση των stopwords με "άχρηστες" λέξεις, βάσει των επανειλημμένων αποτελεσμάτων μετά από κάθε τρέξιμο.
Συνεπώς, χρησιμοποιήθηκε ένας πίνακας με μεγάλη γκάμα συνηθισμένων και ασυνήθιστων λέξεων (όπως 'τν', 'ααα', 'επιτέλοοους' κλπ) που περιείχαν ακόμη και την επανεγγραφή των κλασικών stopwords ανορθόγραφα.

3.3.3. Υλοποίηση



Εικόνα 8 - Στάδια Υλοποίησης keyword-extractor για tweets

Το κείμενο ενός tweet εισέρχεται ως είσοδος στον αλγόριθμο. Στη συνέχεια αφαιρούνται από αυτό τα links που ενδέχεται να εμπεριέχονται στο κείμενο. Μετά από αυτό, επιλέγεται η απομάκρυνση όλων των συμβόλων από την πρότασή μας, καθώς υπάρχει αυξημένη πιθανότητα αυτά να είναι "κολλητά" με κάποιες από τις λέξεις. Πράγμα που θα είχε ως συνέπεια την αδύνατη σύγκρισή τους με τις stopwords. Το κείμενο χωρίς τα links και τα σύμβολα, ακολουθεί την μετατροπή όλων των χαρακτήρων του σε Lowercase(), δηλαδή όλα τα κεφαλαία γράμματα γίνονται πεζά("μικρά"). Στην επόμενη επεξεργασία, αφαιρούνται όλα τα τονικά σύμβολα πάνω από τους χαρακτήρες. Τα παραπάνω δύο στάδια στοχεύουν στην απλοποίηση της αναπαράστασης των ελληνικών χαρακτήρων ώστε να ελαχιστοποιηθεί όσο γίνεται ο αριθμός των λέξεων στον πίνακα τον stopwords με τις οποίο θα συγκριθούνε. Τέλος, οι λέξεις που περιέχονται μέσα στη πρόταση χωρίζονται σε ξεχωριστά στοιχεία, τα οποία τοποθετούνται σε πίνακα, από τον οποίο συγκρίνονται μία προς μία με τον πίνακα των stopwords. Το αποτέλεσμα που προκύπτει αποτελεί είναι τα Keywords του Tweet.



Εικόνα 9 - Απεικόνιση Συστήματος (Τρίτο στάδιο)

4. Βάση Δεδομένων

4.1. ElasticSearch

Το ElasticSearch χρησιμοποιείται ως εργαλείο για αναζήτηση λέξεων μέσα από δεδομένα κειμένου. Το εύρος των δυνατοτήτων του είναι αρκετά μεγάλο και μερικές από αυτές τις δυνατότητες θα αναλυθούν στη συνέχεια, αλλά κατά κόρον είναι φτιαγμένο, για να αποτελεί την αναγκαία υποδομή, πάνω στην οποία δημιουργούνται μηχανές αναζήτησης, αλλά και για να παρέχει στατιστικές αναλύσεις πάνω σε

σώματα κειμένων. Χαρακτηριστικά και δομή του ElasticSearch. Πιο συγκεκριμένα, το ElasticSearch είναι ένας ανεξάρτητος εξυπηρετητής βάσης δεδομένων (standalone database server), γραμμένο σε Java, που δέχεται δεδομένα και τα αποθηκεύει με τέτοιο τρόπο, ώστε να διευκολύνει αναζητήσεις πάνω σ' αυτά. Είναι ιδιαίτερα εύχρηστο προγραμματιστικά, καθώς το κυρίως πρωτόκολλο που χρησιμοποιεί βασίζεται σε HTTP/JSON. Ο ίδιος ο μηχανισμός του HTTP είναι εξ ολοκλήρου ασύγχρονος από μόνος του, επομένως κανένα νήμα πληροφορίας που αποστέλλεται δεν χρειάζεται να περιμένει απόκριση.

Το ElasticSearch προσφέρει μεγάλες δυνατότητες επεκτασιμότητας, αφήνοντας περιθώρια ποικίλων ομαδοποιήσεων, πράγμα το οποίο αντικατοπτρίζει και το ίδιο το όνομά του. Επίσης, η χρήση του μπορεί να εφαρμοστεί σε διάφορα πεδία αναζητήσεων, όπως παραδείγματος χάρη, στα προϊόντα που διαθέτει μια βάση δεδομένων ή στα άρθρα που διαθέτει ένα ιστολόγιο. Το ElasticSearch διαθέτει τις δυνατότητες να ξεπερνά τα εμπόδια που δημιουργεί στις αναζητήσεις η φυσική γλώσσα[14].

4.1.1. Apache Lucene

Ο πυρήνας του ElasticSearch στην ουσία είναι ένα άλλο λογισμικό, ιδιαίτερα γνωστό και με ισχυρές δυνατότητες, το Apache Lucene. Το ElasticSearch μπορεί να μελετηθεί περισσότερο σε βάθος, αν μελετηθεί παράλληλα με το Lucene και τις δικές του υποδομές. Οι αλγόριθμοι του ElasticSearch που σχετίζονται είτε με ταίριασμα κειμένου (text matching) είτε με βελτιστοποιημένα ευρετήρια όρων προς αναζήτηση προέρχονται από τις υποδομές του Lucene.

Το νέο που κομίζει το ElasticSearch στην ήδη υπάρχουσα τεχνολογία του Lucene είναι το πιο εύχρηστο και ακριβές API, η επεκτασιμότητα, η διαλειτουργικότητα με γλώσσες προγραμματισμού πέραν της Java, η ευχρηστία προγραμματιστικής διαχείρισής του, η ομαδοποίηση (clustering) και η αντικατάσταση (replication), γνωρίσματα που ήδη αναφέρθηκαν στις NoSQL τεχνολογίες και καινούργιες, πιο

εύχρηστες εφαρμογές, οι οποίες αποτελούν επεκτάσεις των Java κλάσεων του Lucene.

Βασικό συστατικό του ElasticSearch είναι το ευρετήριο (index), όπου και αποθηκεύονται τα δεδομένα. Όπως είχε προαναφερθεί και στην περιγραφή των NoSQL βάσεων δεδομένων, τα ευρετήρια για το ElasticSearch έχουν ακριβώς την ίδια λειτουργία με τους πίνακες των σχεσιακών βάσεων. Αλλά αντίθετα από αυτές, οι τιμές που είναι αποθηκευμένες σ' ένα ευρετήριο προορίζονται, ώστε να συνδράμουν στην επιτάχυνση μιας πιθανής αναζήτησης σε δεδομένα κειμένου[14].

4.1.2. Δομή

Η κυρίως αποθηκευτική οντότητα του ElasticSearch είναι το έγγραφο. Σε αναλογία προς τις σχεσιακές βάσεις, ένα έγγραφο είναι μια γραμμή από δεδομένα σ' έναν πίνακα. . Τα έγγραφα αποτελούνται από πεδία (γραμμές), αλλά κάθε πεδίο μπορεί να εμφανίζεται περισσότερες από μια φορές, τότε σ' αυτήν την περίπτωση ονομάζεται πεδίο πολλαπλών τιμών (multivalued). Κάθε πεδίο διαθέτει ένα τύπο (κείμενο, νούμερο, ημερομηνία κοκ), όπως επίσης μπορεί να εμπεριέχει ένα μέρος από κάποιο έγγραφο (subdocument) ή και πίνακες. Οι τύποι των πεδίων μπορούν επίσης να είναι σύνθετοι. Ο τύπος των πεδίων έχει ιδιαίτερη σημασία στο ElasticSearch, γιατί παρέχει στην εκάστοτε μηχανή αναζήτησης πληροφορίες για το πόσο συχνές διαδικασίες, παραδείγματος χάρη σύγκρισης ή ταξινόμησης, πρέπει να λάβουν χώρα. Στην περίπτωση του ElasticSearch βέβαια ο καθορισμός αυτών των διαδικασιών επιτυγχάνεται με αυτόματο τρόπο. Από την άλλη, στις σχεσιακές βάσεις τα έγγραφα δεν απαιτείται να έχουν προκαθορισμένη δομή. Το κάθε έγγραφο μπορεί να διαθέτει διαφορετικό σύνολο πεδίων και επίσης τα πεδία δεν είναι απαραίτητο να είναι γνωστά πριν από την ανάπτυξη της εφαρμογής, παρά μόνο αν από πριν καθοριστεί κάποιου είδους σχήματος.

Εξίσου σημαντικό στο ElasticSearch είναι και ο τύπος του εγγράφου, όπου σ' ένα ευρετήριο μπορούν να αποθηκευτούν πολλές οντότητες με διαφορετικούς σκοπούς. Για παράδειγμα, σε ένα ιστολόγιο που διαθέτει ταυτόχρονα άρθρα και σχόλια, ο

τύπος του εγγράφου διαφοροποιεί εύκολα αυτά τα δύο διαφορετικά είδη κειμενικών δεδομένων.

Είναι επίσης πολύ σημαντικό να επισημανθεί πως το κάθε έγγραφο μπορεί να έχει τη δική του διαφορετική δομή και πως αυτός ο διαχωρισμός βοηθάει σημαντικά στη διαχείριση των δεδομένων. Όμως δεν πρέπει να αγνοούνται και οι περιορισμοί που υφίστανται, καθώς οι διαφορετικοί τύποι ενός εγγράφου δεν μπορούν να θέσουν διαφορετικούς τύπους για την ίδια οντότητα[14].

4.1.3. Ανεξάρτητός εξυπηρετητής

Επίσης, το ElasticSearch έχει την ιδιότητα να λειτουργεί και ως ανεξάρτητος εξυπηρετητής, αλλά και να λειτουργεί μέσω πολλαπλών συνεργαζόμενων εξυπηρετητών, ώστε να μπορεί να επεξεργάζεται ογκώδη σύνολα δεδομένων και να διαθέτει ανοχή σε τυχών σφάλματα. Το σύνολο αυτών των συνεργαζόμενων εξυπηρετητών ονομάζεται σύμπλεγμα (cluster) και ο καθένας από αυτούς ξεχωριστά κόμβος (node).

4.1.4. Διαδικασία ευρετηρίασης(index)

Στη συνέχεια, μετά την περιγραφή της δομής και των χαρακτηριστικών του ElasticSeach, θα ήταν χρήσιμο να γίνει αναφορά για τη διαδικασία που ακολουθείται από την στιγμή που ένα έγγραφο αποθηκεύεται εντός του elasticSearch μέχρι τη στιγμή που ανακτάται μέσω κάποιου ερωτήματος.

Η πρόσθεση δεδομένων μέσα σε ένα ευρετήριο μπορεί να γίνει είτε απευθείας, απλώς με την αποθήκευση μερικών εγγράφων χωρίς προηγούμενο καθορισμό κάποιου ειδικού ευρετηρίου είτε μπορεί ο χρήστης να επιλέξει τον ευρετηριασμό(indexing) των δεδομένων με συγκεκριμένο τρόπο. Το ElasticSearch διαθέτει αρκετές επιλογές για τον τρόπο, με τον οποίο μπορεί να γίνει ευρετηρίαση δεδομένων. Οι δύο βασικοί μέθοδοι που ακολουθούνται για τον

εξειδικευμένο

καθορισμό ενός ευρετηρίου είναι η ανάλυση (analysis) και η χαρτογράφηση (mapping), οι οποίες και θα αναλυθούν παρακάτω[14].

4.1.5. Ανάλυση

Στο ElasticSearch, με τη δημιουργία ενός ευρετηρίου ταυτόχρονα καθορίζεται και ο αριθμός των θραυσμάτων και των αντιγράφων που θα χρησιμοποιηθούν. Όπως προαναφέρθηκε, ένα θραύσμα είναι ένα κομμάτι από τα δεδομένα και το αντίγραφο είναι μια επανάληψη των ίδιων δεδομένων με τη μορφή ενός αντιγράφου ασφαλείας (backup).

Όταν λοιπόν υπάρχει ένας κόμβος, τότε όλα τα θραύσματα και όλα τα αντίγραφα βρίσκονται μέσα σε αυτόν τον κόμβο. Ενώ στην περίπτωση που οι κόμβοι είναι περισσότεροι, τότε τα δεδομένα κατανέμονται ανάμεσα σ' αυτούς τους κόμβους. Τα δεδομένα στο ElasticSearch αναλύονται σε δύο φάσεις. Η πρώτη φορά είναι όταν ένα έγγραφο αποθηκεύεται και η πληροφορία της ανάλυσής του καταγράφεται και αυτή στο ευρετήριο και η δεύτερη φορά είναι όταν διενεργείται κάποιο ερώτημα. Το ElasticSearch αναλύει το ερώτημα της αναζήτησης και ψάχνει στις αποθηκευμένες πληροφορίες του ευρετηρίου.Η διαδικασία της ανάλυσης ευρετηρίου (index analysis) λειτουργεί ως ένα ρυθμιζόμενο μητρώο αναλυτών (analyzers), οι οποίοι μπορούν τόσο να αναλύσουν σε πεδία ένα έγγραφο κατά την είσοδό του όσο και ένα ερώτημα με τη μορφή σύμβολοακολουθίας (string).

Οι αναλυτές γενικότερα αποτελούνται από έναν και μόνο διαχωριστή λεκτικών μονάδων (tokenizer) και από κανένα μέχρι περισσότερα φίλτρα λεκτικών μονάδων. Ως διαχωριστής λεκτικών μονάδων ορίζεται το εργαλείο εκείνο που λαμβάνει ως είσοδο μια συμβολοακολουθία (string) και επιστρέφει το σύνολο των λεκτικών μονάδων που εμπεριέχονται στην ακολουθία αυτή. Παραδείγματος χάρη, ο προκαθορισμένος αναλυτής (standard analyzer) αποτελείται από τον προκαθορισμένο διαχωριστή λεκτικών μονάδων, ο οποίος εφαρμόζει το προκαθορισμένο φίλτρο λεκτικών μονάδων, όπου κανονικοποιεί όλες τις λεκτικές

μονάδες, το φίλτρο μικρογράμματων λέξεων, όπου μετατρέπει όλες τις λεκτικές μονάδες με μικρά γράμματα και το φίλτρο των stop-words λέξεων, όπου απομακρύνει όλες τις λέξεις που γραμματικά είναι άρθρα, αντωνυμίες, σύνδεσμοι κτλ, δηλαδή μικρές σε μέγεθος λέξεις με περισσότερο λειτουργικό χαρακτήρα και λιγότερο σημασιολογικό[14].

4.1.6. Χαρτογράφηση

Στη συνέχεια ακολουθεί η χαρτογράφηση, κατά την οποία καθορίζεται το είδος των δεδομένων που θα τοποθετηθεί σε κάθε πεδίο. Αν δεν υπάρξει καθορισμός χαρτογράφησης, τότε το ElasticSearch θα προσπαθήσει να μαντέψει το είδος το δεδομένων και θα τα χαρτογραφήσει αυτόματα, χωρίς αυτό να συνιστά αλάνθαστη διαδικασία. Στη χαρτογράφηση, πέρα από το είδος των δεδομένων σε κάθε πεδίο, δίνεται η δυνατότητα να προσδοθεί βαρύτητα (boost) σε ορισμένα πεδία, ώστε σ' αυτά να δίνεται προτεραιότητα κατά τη διάρκεια των αναζητήσεων.

4.1.7. Αναζήτηση

Το επόμενο βήμα που ακολουθεί, εφόσον έχουν ολοκληρωθεί η παραπάνω διαδικασίες και έχουν προστεθεί κάποια δεδομένα στο ευρετήριο, είναι η αναζήτηση. Μια βασική αναζήτηση εκτελεί ένα ερώτημα διαμέσου του ElasticSearch. Το επιστρεφόμενο αποτέλεσμα του ερωτήματος μπορεί και αυτό να φιλτραριστεί είτε με τον καθορισμό συγκεκριμένων φίλτρων, όπως παραδείγματος χάρη λέξεων κλειδιών ή εύρος ημερομηνιών είτε με τον καθορισμό συγκεκριμένων όψεων (facets), ώστε να επιστρέφεται μόνο ένα συγκεκριμένο κομμάτι των δεδομένων, όπως για παράδειγμα, να επιστρέφονται αποτελέσματα μόνο από ένα συγκεκριμένο πεδίο, μέχρι ένα ορισμένο μέγεθος και σε φθίνουσα σειρά.

4.1.8. Πλεονεκτήματα ElasticSearch

Συνοψίζοντας για το ElasticSearch, θα αναφερθούν μερικές από τις περιπτώσεις, όπου η χρήση του στην κατασκευή μηχανών αναζήτησης είναι ιδιαίτερη χρήσιμή. Τα παρακάτω παραδείγματα είναι χαρακτηριστικές περιπτώσεις των προβλημάτων που μπορεί να επιλύσει[14].

• Έχει την ικανότητα να επιστρέφει τα καλύτερα αποτελέσματα μέσα από έναν μεγάλο αριθμό περιγραφών προϊόντων, ιδιαίτερα όταν αναζητούνται φράσεις, όπως “chef’sknife” για παράδειγμα. Αυτό καθίσταται δυνατό με την χρήση των διαχωριστών λεκτικών μονάδων, που περιγράφηκαν παραπάνω.

• Δοθέντος του προηγούμενου παραδείγματος, υπάρχει η δυνατότητα να αναζητηθούν σε ποια συγκεκριμένα υποκαταστήματα υπάρχει το συγκεκριμένο προϊόν και από ποιες ποσότητες και πάνω. Σ’ αυτό μπορεί να βοηθήσει ο καθορισμός συγκεκριμένων όψεων στα επιστρεφόμενα αποτελέσματα.

• Υπάρχει και η δυνατότητα αυτοσυμπλήρωσης (autocompleting) στο κουτί αναζήτησης σε μερικώς γραμμένες λέξεις βασιζόμενη σε προηγούμενες αναζητήσεις.

Σε γενικές γραμμές, το ElasticSearch αποτελεί την τέλεια λύση στο να επιστρέφει αποτελέσματα κατά προσέγγιση από δεδομένα, καθώς βαθμολογεί τα αποτελέσματα βάσει της ποιότητας. Ενώ το ElasticSearch μπορεί να φέρει εις πέρας ακριβές ταίριασμα σε κειμενικές αναζητήσεις και στατιστικούς υπολογισμούς, στην πραγματικότητα ο τρόπος που το πετυχαίνει είναι μια εγγενώς προσεγγιστική διαδικασία. Αυτή του η ιδιότητα το ξεχωρίζει και από τις περισσότερες παραδοσιακές βάσεις δεδομένων.

4.2. Elasticsearch για την αποθήκευση άρθρων και tweets

Στην παρούσα εργασία χρησιμοποιήσαμε τη βάση δεδομένων elasticsearch καθώς λειτουργεί πολύ γρήγορα στα text searches μέσω ενός πλήθους queries δυνατοτήτων σε σχέση με άλλες βάσεις. Χρησιμοποιήσαμε το java api του για να κάνουμε indexing και quering στα δεδομένα, αλλά υπήρχε δυσκολία στο να μάθουμε να το χειριζόμαστε λόγω του ελλιπούς documentation που προσφέρει αλλά και τις λιγοστές πληροφορίες που βρήκαμε σε κοινότητες όπως το stackoverflow κτλ [15].

Εγκαταστήσαμε τη βάση στον χώρο εργασίας μας καθώς επίσης και τα plugins: head και kibana για την καλύτερη οπτικοποίηση των δεδομένων.

Στη συνέχεια συνδέσαμε το elasticsearch με το project μας μέσω του TransportClient

```
Client client = TransportClient.builder().build()
                            .addTransportAddress(new
InetSocketTransportAddress(InetAddress.getByName("localhost"), 9300));
```

Εικόνα 10 - Κώδικας σύνδεσης του elasticsearch transport client με την εφαρμογή

Indexing ενός άρθρου στο elasticsearch

```
response = client.prepareIndex("index", "type", id)
       .setSource(jsonBuilder()
                    .startObject()
                    .field("Tile:",entry.getTitle())
                    .field("Publish Date: ", entry.getPublishedDate())
                    .field("URL",entry.getLink())
                    .field("Keywords",keywords)
                    .field("Content", Boilerpipe.content)
                    .endObject()
       .get();
```

Εικόνα 11 - Παράδειγμα κώδικα indexing ενός άρθρου στο elasticsearch

Με τη βοήθεια της κλάσης SearchResponse μπορούμε να κάνουμε queries που αναφέρονται σε τιμές τον πεδίων τιμών του πίνακα με τα άρθρα σελίδων και με βάσει αυτών η συνάρτηση να επιστρέψει καταχωρήσει που ικανοποιούν τα κριτήρια του query.

Παράδειγμα κώδικα που μας επιστρέφει το πεδίο “Content” όλων τον καταχωρήσεων του index = “index” και type = “type” με ημερομηνία “Publish Date” από
«2016-11-09T11:01:00Z» μέχρι «2016-11-11T11:01:00Z» το οποίο είναι ένα συγκεκριμένο date format που αναγνωρίζει και το elasticsearch.

```
SearchResponse response = client.prepareSearch("index")
            .setTypes("type")

.setSearchType(SearchType.DFS_QUERY_THEN_FETCH)
.setFetchSource(new String[]{"Content"}, null)
.setQuery(QueryBuilders.rangeQuery("Publish Date:
").from("2016-11-09T11:01:00Z").to("2016-11-
11T11:01:00Z"))
.setFrom(0).setSize(60).setExplain(true)

.get();
SearchHit[] results = response.getHits().getHits();
```

Εικόνα 12 - Παράδειγμα ενός query στο elasticsearch

_index	_type	_id	_score	Title:	Publish Date: ▼	URL:
innews	all	17713		Ο Τσίπρας συναντά την Νοτοπούλου - Σήμερα εγκαινιάζει τα γραφεία στη Θεσσαλονίκη	2016-11-26T09:35:00Z	http://www.protothema.gr/politics/article/631714/o-tsipras-sunada
innews	all	17709		Προσποιήθηκε τον σεΐχη και έκλεψε πανάκριβη Rolls Royce από γκαράζ	2016-11-26T09:32:00Z	http://www.newsbomb.gr/kosmos/news/story/749083/prospoiithike
innews	all	17711		Χωρίς λεωφορεία και πάλι η Θεσσαλονίκη	2016-11-26T09:30:00Z	http://www.lifo.gr/now/greece/122956
innews	all	17708		Μαριέττο Χρουσαλά - Λέων Πατίτσας: Κόνουν μοϊ πεζοπορίο! | Lifestyle - Ειδήσεις NewsIt.gr	2016-11-26T09:29:00Z	http://www.newsit.gr/lifestyle/marietta-xrysala-leon-patitsas-kar
innews	all	17710		Happy Birthday Rita Ora! Αυτές είναι οι δέκα πιο εντυπωσιακές της εμφανίσεις	2016-11-26T09:27:00Z	http://www.zougla.gr/lifestyle/article/hapy-birthday-rita-ora-aftes-i
innews	all	17706		Κουβανοί βγήκαν στους δρόμους του Μαϊάμι -Πανηγυρίζουν για τον θάνατο του Κάστρο [εικόνες & βίντεο]	2016-11-26T09:25:00Z	http://www.iefimerida.gr/news/303645/koyvanoi-vgikan-stoys-dror
innews	all	17712		Η Μαριάν Φέιθφουλ ανεβαίνει στη σκηνή του Μπατακλάν και αποτίνει φόρο τιμής στον Λέοναρντ Κοέν	2016-11-26T09:24:00Z	http://www.lifo.gr/now/people/122955
innews	all	17701		Ποια επιδόματα και ποιες φοροαπαλλαγές καταργούνται	2016-11-26T09:23:00Z	http://www.protothema.gr/economy/article/631710/poia-epidomat
innews	all	17705		Οπερμον (SPD): Δεν πιστεύω πως η Τουρκία θα καταγγείλει τη συμφωνία με την Ε.Ε	2016-11-26T09:19:00Z	http://www.iefimerida.gr/news/303644/operman-spd-den-pisteyo-p
innews	all	17697		Μην τους ξεχάσετε! Ποιοι γιορτάζουν σήμερα;	2016-11-26T09:13:00Z	http://www.newsbomb.gr/bombplus/fun/story/749082/min-toys-xe
innews	all	17698		Στη Θεσσαλονίκη σήμερα ο Τσίπρας -Σύσκεψη με φορείς της πόλης	2016-11-26T09:12:00Z	http://www.iefimerida.gr/news/303643/stl-thessaloniki-simera-o-tsi
innews	all	17695		Καιρός: Έκτακτο δελτίο επιδείνωσης με βροχές και καταιγίδες! | Καιρός - Ειδήσεις NewsIt.gr	2016-11-26T09:12:00Z	http://www.newsit.gr/kairos/kairos-ektakto-deltio-epideinosis-me-
innews	all	17694		Βίντεο: Ο Ραούλ Κάστρο ανακοινώνει τον θάνατο του Φιντέλ	2016-11-26T09:12:00Z	http://www.protothema.gr/world/article/631709/video-o-raoul-kast
innews	all	17707		Το εξώφυλλο του Spiegel για τον Τράμπ: Σε θρόνο, με την Ιβάνκα δίπλα του και την Μελάνια μακριά [εικόνες]	2016-11-26T09:08:00Z	http://www.iefimerida.gr/news/303642/exofyllo-toy-spiegel-gia-ton
innews	all	17696		Η στιγμή της ανακοίνωσης του θανάτου του Φιντέλ από την κουβανική τηλεόραση	2016-11-26T09:07:00Z	http://www.lifo.gr/now/world/122954
innews	all	17686		ΟΗΕ: Προς επιβολή νέων κυρώσεων στη Β. Κορέα	2016-11-26T09:06:00Z	http://www.protothema.gr/world/article/631708/oie-pros-epivoli-ne
innews	all	17688		Σε μπλόφες και κόλπα αναζητά διέξοδο η κυβέρνηση	2016-11-26T09:05:00Z	http://www.newsbomb.gr/politikh/news/story/749081/se-mplofes-k
innews	all	17699		Τα κόλπα της μαφίας των διαρρηκτών - Η λεία της ξεπέρασε το 1 εκατ. ευρώ!	2016-11-26T09:04:59Z	http://www.newsbomb.gr/ellada/astynomiko-reportaz/story/74908
innews	all	17680		Τα πρωτοσέλιδα των εφημερίδων του Σαββάτου 26 Νοεμβρίου με μία ματιά	2016-11-26T09:03:00Z	http://www.iefimerida.gr/news/303641/ta-protoselida-ton-efimerid
innews	all	17682		Φιντέλ Κάστρο: Ποιος ήταν ο ηγέτης της κουβανικής επανάστασης (pics+vid)	2016-11-26T09:02:00Z	http://www.newsbomb.gr/kosmos/news/story/749079/fintel-kastro
innews	all	17687		GP Άμπου Ντάμπι: Αυτό είναι το ελαστικό του 2017	2016-11-26T09:02:00Z	http://www.protothema.gr/car-and-speed/article/631707/gp-abou-e
innews	all	17702		Ηλεκτροτεχνίτης στην ΔΕΗ Αγρινίου	2016-11-26T09:00:00Z	http://www.altsantiri.gr/ergasia/ilektrotechnitis-stin-dei-agriniou/
innews	all	17685		Θάνατος Φιντέλ Κάστρο: "Hasta la victoria siempre"! Το ιστορικό διάγγελμα [vid] | Κόσμος - Ειδήσεις NewsIt.gr	2016-11-26T08:58:00Z	http://www.newsit.gr/kosmos/THanatos-Fintel-Kastro-Hasta-la-vict
innews	all	17679		Γκέραλντ Κνάους: Τουρκία και ΕΕ παίζουν ρώσικη ρουλέτα με το προσφυγικό -Στο τέλος θα την πληρώσει η Ελλάδα	2016-11-26T08:53:00Z	http://www.iefimerida.gr/news/303640/gkeralnt-knaoys-toyrkia-kai
innews	all	17683		Θα αποτεφρωθεί η σορός του Φιντέλ Κάστρο	2016-11-26T08:53:00Z	http://www.lifo.gr/now/world/122953
innews	all	17678		Η Joanna Krupa... ξανοχτύπα με μικροσκοπικό μπικίνι	2016-11-26T08:51:00Z	http://www.protothema.gr/life-style/article/631644/h-joanna-krupa
innews	all	17677		Κίνα: Τουλάχιστον ένας νεκρός σε σεισμική δόνηση 6,5 βαθμών	2016-11-26T08:49:40Z	http://www.protothema.gr/world/article/631705/kina-toulahiston-e
innews	all	17676		Καιρός: Πάρτε ομπρέλα σήμερα θα τη χρειαστείτε - Έρχονται βροχές και ισχυρές καταιγίδες	2016-11-26T08:48:00Z	http://www.newsbomb.gr/kairos/prognwsh-emy/story/749077/kair
innews	all	17684		Βροχές και ισχυρές καταιγίδες σήμερα	2016-11-26T08:45:00Z	http://www.lifo.gr/now/greece/122952
innews	all	17704		Προσλήψεις στον Δήμο Ζαγοράς – Μουρεσίου	2016-11-26T08:45:00Z	http://www.altsantiri.gr/ergasia/proslipsis-ston-dimo-zagoras-mour
innews	all	17675		Φιντέλ Κάστρο: Ιστορική στιγμή - Ο Ραούλ Κάστρο ανακοινώνει το θάνατο του αδερφού του (vid)	2016-11-26T08:40:00Z	http://www.newsbomb.gr/kosmos/news/story/749078/fintel-kastro
innews	all	17673		Απώλεια έως κι ενός μισθού το χρόνο από τις αυξήσεις – φωτιά στους έμμεσους φόρους	2016-11-26T08:40:00Z	http://www.protothema.gr/economy/article/631704/-apoleia-eos-ki
innews	all	17681		Ινδονησία: Συνελήφθη ισλαμιστής που σχεδίαζε επίθεση εναντίον της πρεσβείας της Μιανμάρ	2016-11-26T08:38:00Z	http://www.iefimerida.gr/news/303639/indonisia-synelifthi-islamisti
innews	all	17669		Путин с Эрдоганом обсудил вопросы урегулирования сирийского кризиса	2016-11-26T08:33:00Z	http://www.newsbomb.gr/ru/story/749040/putin-s-erdoganom-obs
innews	all	17671		Επτέλους χειμώνας! Χιόνια και πτώση της θερμοκρασίας έως 10 βαθμούς	2016-11-26T08:33:00Z	http://www.protothema.gr/greece/article/631703/epitelous-heimon
innews	all	17670		Η κρίση του 1931 και η ξενοφοβία που τροφοδότησε	2016-11-26T08:30:00Z	http://www.protothema.gr/stories/article/631560/i-krisi-tou-1931-k
innews	all	17666		Μία ιστορική στιγμή: Ο Ραούλ Κάστρο ανακοινώνει σε διάγγελμα τον θάνατο του Φιντέλ Κάστρο [βίντεο]	2016-11-26T08:29:00Z	http://www.iefimerida.gr/news/303638/mia-istoriki-stigmi-o-raoyl-k
innews	all	17674		Πέθανε ο Φιντέλ Κάστρο	2016-11-26T08:27:00Z	http://www.altsantiri.gr/kosmos/pethane-o-fintel-kastro
innews	all	17672		Εφημερίδες: Διαβάστε το σημερινό (26/11/2016) πρωτοσέλιδο	2016-11-26T08:25:00Z	http://www.newsbomb.gr/media-agb/typos/story/749076/efimerid
innews	all	17667		Στο Bataclan η τραγουδίστρια-θρύλος των 60's Μαρίν Φέιθφουλ	2016-11-26T08:24:00Z	http://www.iefimerida.gr/news/303637/sto-bataclan-i-tragoydistria

Εικόνα 13 - Οπτικοποίηση του index στο plugin heard του

Result Source

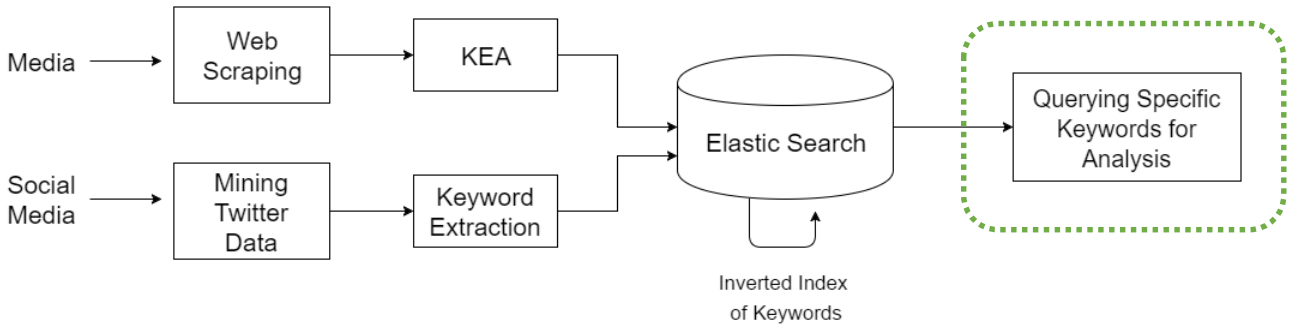

```
{
    "_index": "news",
    "_type": "rss",
    "_id": "0",
    "_version": 1,
    "_score": 1,
    "_source": {
        "Title": "Σουηδία: Οικιακή θέρμανση μέσω αναζητήσεων στο διαδίκτυο",
        "Publish Date": "2016-11-01T02:38:00Z",
        "Keywords": "περιβάλλον, ηλεκτρικής ενέργειας, Οικιακή
        θέρμανση,Σουηδία,περιβάλλον,διαδίκτυο",
        "Content": "Τα κέντρα δεδομένων που είναι απαραίτητα για τη διαδικτυακή
        δραστηριότητα παράγουν υπερβολική θερμότητα και απορροφούν σημαντική ποσότητα
        ηλεκτρικής ενέργειας για λόγους ψύξης.Ωστόσο, αξιωματούχοι στη Στοκχόλμη της
        Σουηδίας σκέπτονται να αξιοποιήσουν όλη αυτήν την απορριπτόμενη θερμότητα για το
        σκοπό της οικιακής θέρμανσης.Τα κέντρα δεδομένων δεν είναι φιλικά προς το
        περιβάλλον, καταναλώνοντας παγκοσμίως περίπου την ίδια ποσότητα ενέργειας με την
        αεροπορική βιομηχανία.Μάλιστα, η ποσότητα ηλεκτρικής ενέργειας που χρησιμοποιείται
        στα κέντρα δεδομένων μπορεί έως και να τριπλασιαστεί την προσεχή δεκαετία. Ωστόσο,
        η θερμότητα από τα κέντρα δεδομένων μπορεί να βοηθήσει στην απεξάρτηση από τα
        ορυκτά καύσιμα.Μόνο ένα κέντρο δεδομένων των δέκα μεγαβάτ μπορεί να παρέχει
        αρκετή θερμότητα για 20.000 διαμερίσματα και να μειώσει τις εκπομπές αερίων του
        θερμοκηπίου κατά 8.000 τόνους.το πλαίσιο της πρωτοβουλίας «Πάρκα Δεδομένων
        Στοκχόλμης», τα κέντρα δεδομένων θα τροφοδοτούνται από ανανεώσιμες πηγές
        ενέργειας, και η θερμότητα που θα παράγεται θα πωλείται στη δημοτική εταιρεία
        θέρμανσης, η οποία ήδη σκόπευε να στραφεί στη βιομάζα ή άλλες πηγές αντί των
        ορυκτών καυσίμων. Η εταιρεία θέρμανσης έχει ήδη αρχίσει να εργάζεται με μικρά κέντρα
        δεδομένων, και η πρωτοβουλία ανακοίνωσε στην ιστοσελίδα της ότι θα φέρει σε επαφή,
        προετοιμάσει και προσφέρει «όλα τα απαραίτητα στοιχεία υποδομής κατάλληλα για
        δραστηριότητα κέντρων δεδομένων»Δεδομένου ότι ο στόχος της κυβέρνησης της
        Σουηδίας είναι να απομακρυνθεί πλήρως από τα ορυκτά καύσιμα ως το 2040, η
        συγκεκριμένη πρωτοβουλία αναμένεται να ωθήσει την σκανδιναβική χώρα πιο κοντά σε
        αυτό το στόχο. "
    }
}
```


Εικόνα 14 - Παράδειγμα καταχώρησης ενός άρθρου στο elasticsearch

Στη συνέχεια χρησιμοποιούμε την κλάση TermVectors για να υπολογίσουμε την inverted index λίστα με τα counts των keywords ανά μέρα.

```
TermVectorsResponse resp = client.prepareTermVectors().setIndex("in-
dexName")
                .setType("docType").setId("docId").execute().actionGet()
```

5. Δημιουργία μοντέλου συσχέτισης των δύο μέσων με σκοπό την πρόβλεψη



Εικόνα 15 - Οπτικοποίηση Συστήματος (Στάδιο τέταρτο)

5.1 Trend analysis

Για την εξαγωγή συμπερασμάτων για το πώς τα trends των media επηρεάζουν αυτά των social media(twitter) και αντίστροφα, αρχικά πήραμε μερικά παραδείγματα λέξεων κλειδιών που όπως παρατηρήσαμε είχαν υψηλά ποσοστά εμφάνισης και στα δύο μέσα. Και σχηματίσαμε στατιστικούς πίνακες με σκοπό την απεικόνισή τους σε γραφήματα, ώστε να διαπιστωθεί διαισθητικά αν υπάρχει μεταξύ τους κάποιο κοινό pattern.

Για να συγκρίνουμε τις εμφανίσεις των λέξεων αυτών ανά μέρα χρησιμοποιήσαμε το λόγο, των εμφανίσεων μιας λέξης-κλειδί προς το συνολικό άθροισμα εμφανίσεων όλων των keywords στα άρθρα μια ημέρας. Αντίστοιχα για τα tweets, χρησιμοποιήσαμε το λόγο των εμφανίσεων μιας λέξης-κλειδί, προς των αριθμό των tweets(αυτό έγινε εξαιτίας του «θορύβου» που προέκυπτε από λέξεις-κλειδιά στα αποτελέσματα του brute force αλγόριθμου για τα tweets) .

Έτσι φτιάξαμε έναν πίνακα με 2 στήλες και 30 γραμμές όπου η πρώτη στήλη αναφερόταν στα media και δεύτερη στα social media για κάθε μια από τις 30 μέρες του Νοεμβρίου 2016.

5.1.1 Ανάλυση Λέξης

Παράδειγμα η λέξη "Trump"

Ο πίνακας στα αριστερά δείχνει τις εμφανίσεις των keywords στα άρθρα των media και των social media

Ο πίνακας στα δεξιά είναι μετά τη διαίρεση προς το σύνολο των keywords για τα media και των tweets για το twitter

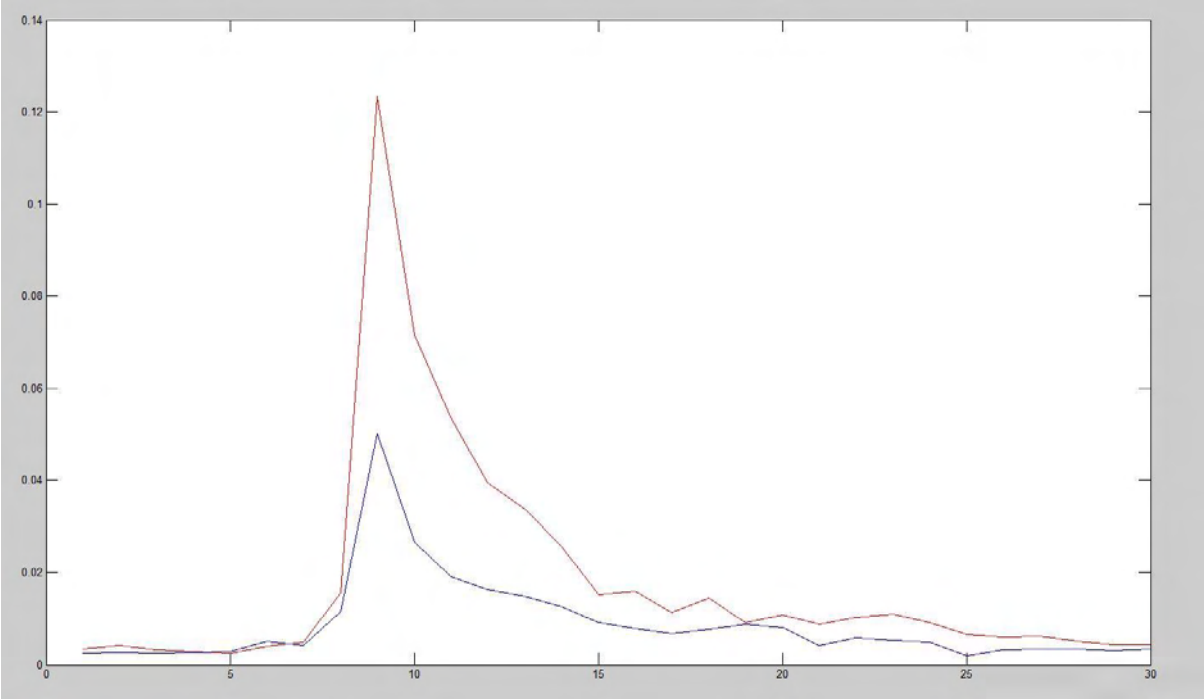
KeyWord Counts

	Media	social
1	640	197
2	701	192
3	639	73
4	674	56
5	480	43
6	854	70
7	1064	95
8	2952	299
9	13635	2420
10	6670	1359
11	4624	998
12	2678	681
13	2281	560
14	3144	461
15	2278	289
16	1921	302
17	1638	217
18	1844	272
19	1434	155
20	1281	176
21	1060	169
22	1548	194
23	1395	219
24	1307	183
25	496	129
26	590	106
27	604	106
28	882	101
29	815	89
30	911	88

Counts/keywords-tweets

	Media	Social
1	0.0025258405	0.00345038
2	0.0027565867	0.00413971
3	0.0024189793	0.00321997
4	0.0026411175	0.00293685
5	0.0027663314	0.00244443
6	0.0051106509	0.00396196
7	0.0041744971	0.00493455
8	0.0114322892	0.01555914
9	0.0501725774	0.12329953
10	0.0265732817	0.07154514
11	0.0190940248	0.05347192
12	0.0163875239	0.03952637
13	0.0147790592	0.03378786
14	0.0126788508	0.02563532
15	0.0090770352	0.01518894
16	0.0077376996	0.01599830
17	0.0066403162	0.01138091
18	0.0076791418	0.01443813
19	0.0086646002	0.00917702
20	0.0079799162	0.01080815
21	0.0041192403	0.00872168
22	0.0058815260	0.01016185
23	0.0052694004	0.01098680
24	0.0048930602	0.00909407
25	0.0019048789	0.00653528
26	0.0033035084	0.00593405
27	0.0034965237	0.00616816
28	0.0033839000	0.00504924
29	0.0030287939	0.00432689
30	0.0033654236	0.00429121

Στη συνέχεια κάναμε plot τον δεύτερο πίνακα



Εικόνα 16 - Γραφική παράσταση της συχνότητας εμφάνισης της λέξης "Trump" στα δύο μέσα

Ο **x** άξονας αντιπροσωπεύει τις μέρες

Ο **y** άξονας αντιπροσωπεύει το ποσοστό εμφάνισης

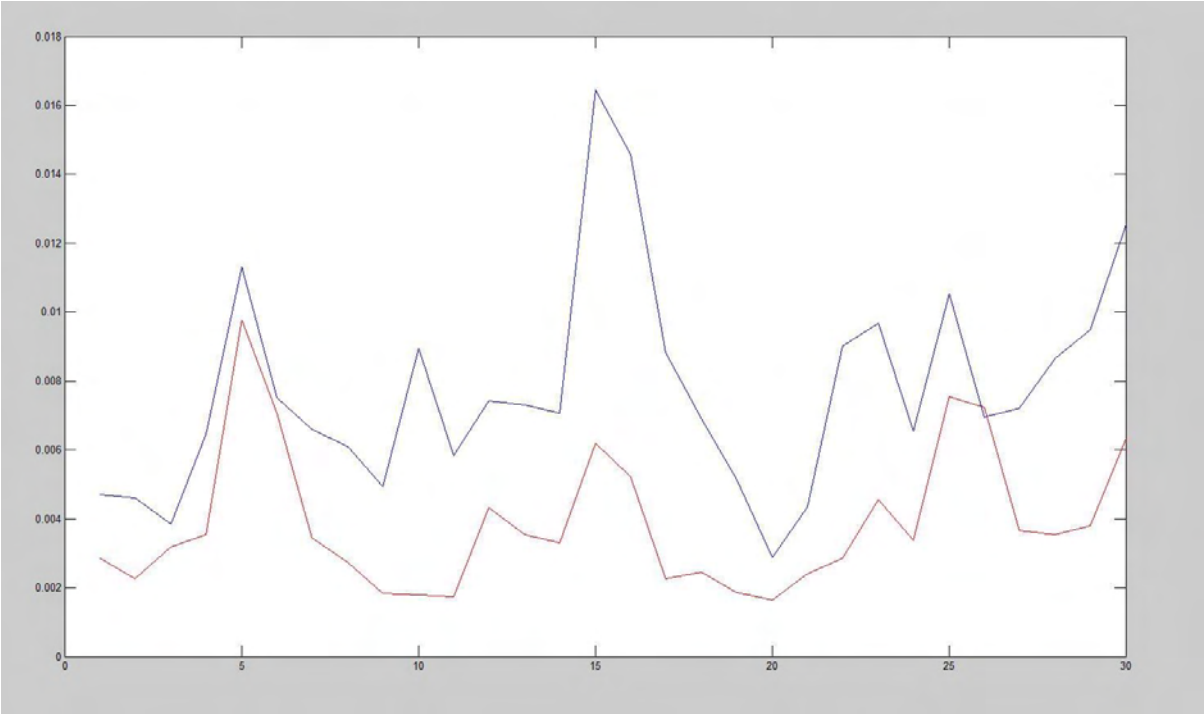
Η συνάρτηση με το κόκκινο χρώμα αναπαριστά το ποσοστό εμφάνισης ανά μέρα στα media, ενώ το μπλε χρώμα στο Twitter.

Παρατηρούμε λοιπόν μια εμφανή συσχέτιση των δυο συναρτήσεων η οποία αποδεικνύει την επιρροή του ενός μέσου στο άλλο.

Αναφορικά το peak της συνάρτησης συμπίπτει με εκλογή του Trump ως νέου προέδρου των ΗΠΑ.

Με παρόμοιο τρόπο ελέγξαμε και άλλες λέξεις με παρόμοια δημοφιλία(ποσοστό εμφάνισης) στα δύο μέσα, αλλά και ομάδες λέξεων με κοινή θεματολογία.

Παράδειγμα για τη λέξη “Τσίπρας”



Εικόνα 17 - Γραφική παράσταση της συχνότητας εμφάνισης της λέξης “Τσίπρας” στα δύο μέσα

Ο **x** άξονας αντιπροσωπεύει τις μέρες

Ο **y** άξονας αντιπροσωπεύει το ποσοστό εμφάνισης

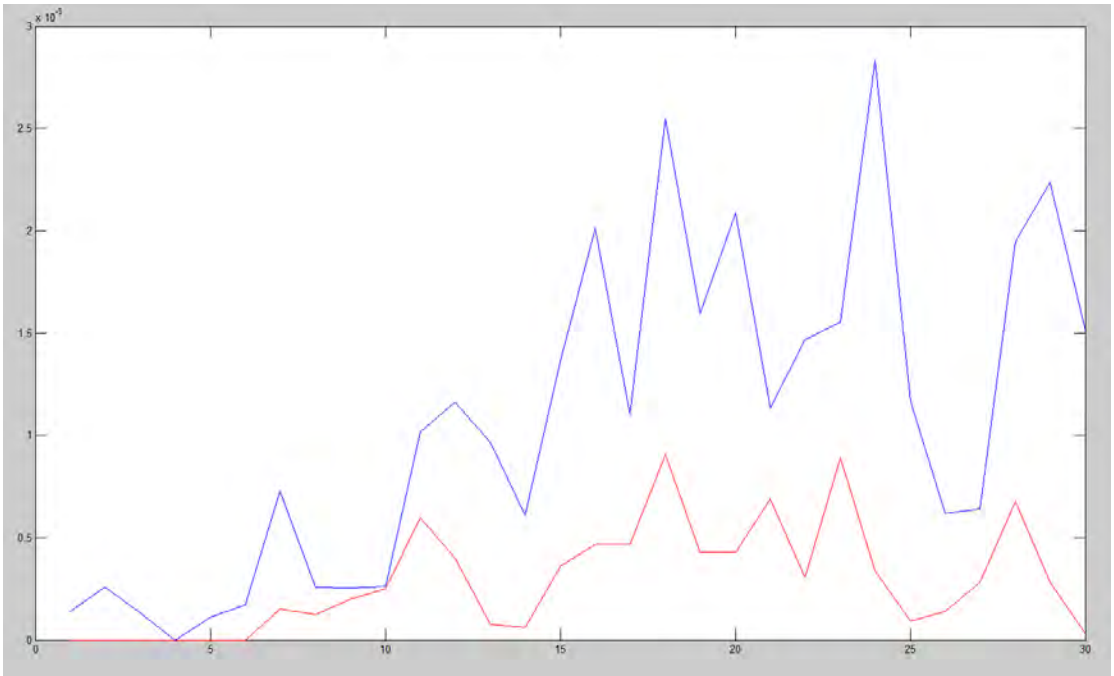
Η συνάρτηση με το κόκκινο χρώμα αναπαριστά το ποσοστό εμφάνισης ανά μέρα στα media, ενώ το μπλε χρώμα στο Twitter.

Και εδώ παρατηρείται εμφανή συσχέτιση μεταξύ των δυο συναρτήσεων η οποία αποδεικνύει την επιρροή του ενός μέσου με το άλλο. Αν και σε αυτό το διάγραμμα η συνάρτηση που αντικατοπτρίζει το Twitter δείχνει πιο έντονη εμφάνιση της λέξης αυτής στο συγκεκριμένο μέσο και όχι στα media.

Το αντίθετο δηλαδή με αυτό που συνέβαινε με το προηγούμενο γράφημα της λέξης “Trump”.

Αυτό δείχνει τη μεγαλύτερη βαρύτητα της λέξης “Τσίπρας” σε σχέση με τη λέξη “Trump”.

Παράδειγμα για τη λέξη “Mannequin”(έγινε πρώτα “viral” στο Twitter)



Εικόνα 18 - Γραφική παράσταση της συχνότητας εμφάνισης της λέξης “Mannequin” στα δύο μέσα

Ο **x** άξονας αντιπροσωπεύει τις μέρες

Ο **y** άξονας αντιπροσωπεύει το ποσοστό εμφάνισης

Η συνάρτηση με το κόκκινο χρώμα αναπαριστά το ποσοστό εμφάνισης ανά μέρα στα media, ενώ το μπλε χρώμα στο Twitter.

Το παράδειγμα της συγκεκριμένης λέξης επιλέχτηκε εξαιτίας του γεγονότος ότι αυτή η λέξη έγινε πρώτα “viral” στο Twitter μέσα στο Νοέμβριο του 2016 και στη συνέχεια ασχολήθηκαν με αυτή τα παραδοσιακά ΜΜΕ. Αυτό φαίνεται και στο παραπάνω γράφημα στο οποίο η συνάρτηση που απεικονίζει το Twitter έχει τιμές στον άξονα y από την πρώτη μέρα του μήνα, ενώ η συνάρτηση των media ξεκινάει να έχει τιμές στον άξονα y και να «ακολουθεί» τη συνάρτηση του Twitter μετά την $6^{η}$ μέρα.

Τα media αντιλαμβάνονται την έκταση του φαινομένου και αρχίζουν την δημοσίευση άρθρων που αναφέρονται σε αυτή τη λέξη.

5.1.2. Ανάλυση ομάδων λέξεων

Στη συνέχεια παρουσιάζονται τα αποτελέσματα για τα group λέξεων που εξετάστηκαν.

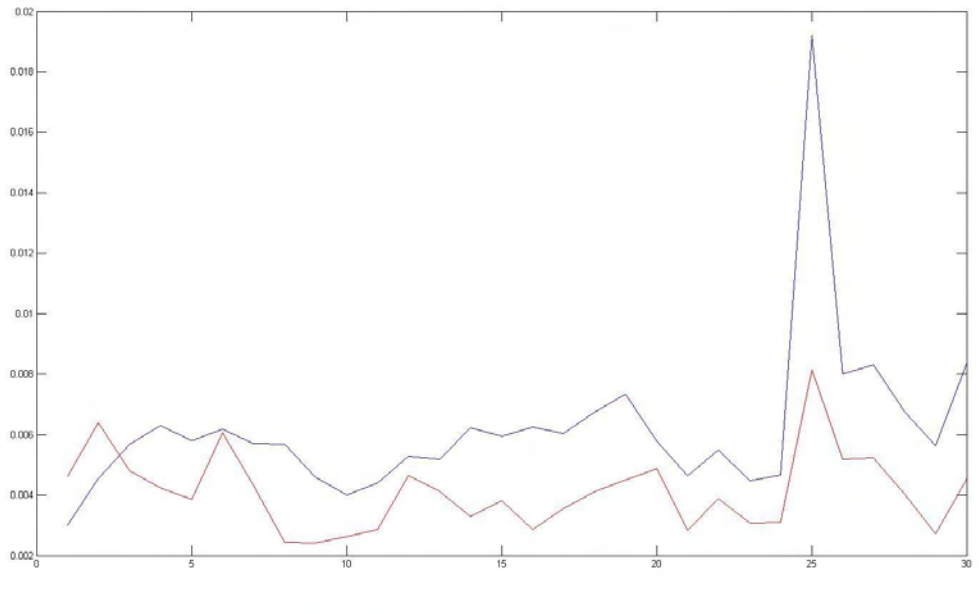
Αυτά αφορούσαν τις δημοφιλέστερες λέξεις κλειδιά από τους εξής θεματικούς άξονες:

Ευρωπαϊκή Ένωση

Αθλητικά

Προσφυγικό Ζήτημα

Προσφυγικό Ζήτημα



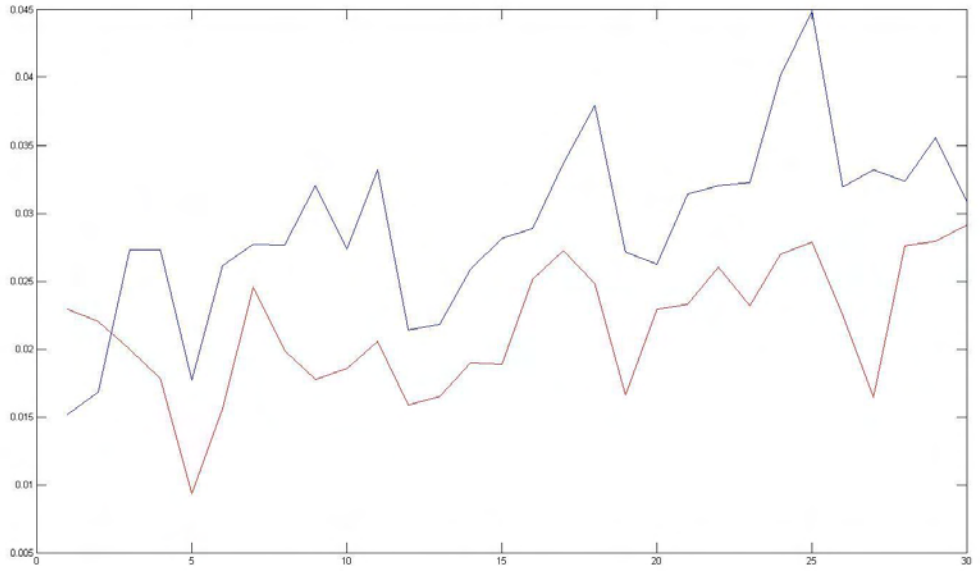
Εικόνα 19 - Γραφική παράσταση της συχνότητας εμφάνισης συνόλου λέξεων με θεματολογία το "προσφυγικό ζήτημα" στα δύο μέσα

Ο **x** άξονας αντιπροσωπεύει τις μέρες

Ο **y** άξονας αντιπροσωπεύει το ποσοστό εμφάνισης

Η συνάρτηση με το κόκκινο χρώμα αναπαριστά το ποσοστό εμφάνισης ανά μέρα στα media, ενώ το μπλε χρώμα στο Twitter.

Ευρωπαϊκή Ένωση



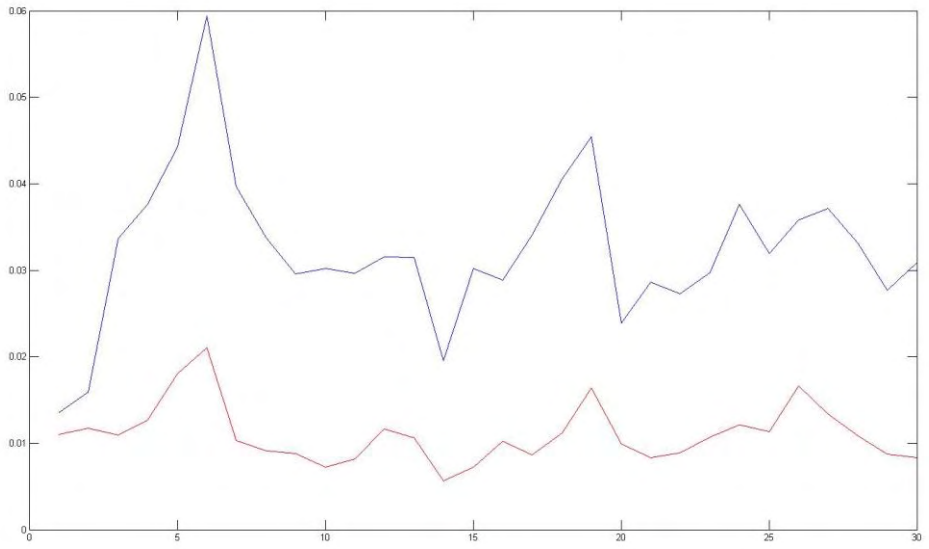
Εικόνα 20 - Γραφική παράσταση της συχνότητας εμφάνισης συνόλου λέξεων με θεματολογία την " Ευρωπαϊκή Ένωση" στα δύο μέσα

Ο **x** άξονας αντιπροσωπεύει τις μέρες

Ο **y** άξονας αντιπροσωπεύει το ποσοστό εμφάνισης

Η συνάρτηση με το κόκκινο χρώμα αναπαριστά το ποσοστό εμφάνισης ανά μέρα στα media, ενώ το μπλε χρώμα στο Twitter.

Αθλητικά



Εικόνα 21 - Γραφική παράσταση της συχνότητας εμφάνισης συνόλου λέξεων με θεματολογία "Αθλητικά" στα δύο μέσα

Ο **x** άξονας αντιπροσωπεύει τις μέρες

Ο **y** άξονας αντιπροσωπεύει το ποσοστό εμφάνισης

Η συνάρτηση με το κόκκινο χρώμα αναπαριστά το ποσοστό εμφάνισης ανά μέρα στα media, ενώ το μπλε χρώμα στο Twitter.

Εφόσον παρατηρήσαμε ότι η συσχέτιση είναι τόσο προφανής, δοκιμάσαμε να τρέξουμε κάποιους αλγορίθμους Μηχανικής Μάθησης στο Dataset μας.
Αυτό έγινε με τη χρήση του εργαλείου Weka.

5.2. Μάθηση Μηχανών και Εφαρμογές

Η Μηχανική Μάθηση ή Γνωσιακές Μηχανές (Machine Learning), ένας κλάδος της Τεχνητής Νοημοσύνης, είναι ένα επιστημονικό πεδίο που αναφέρεται στο σχεδιασμό και ανάπτυξη αλγορίθμων που δέχονται ως είσοδο (input) εμπειρικά δεδομένα, όπως εκείνα που προέρχονται από αισθητήρες (sensor) ή βάσεις

δεδομένων, και δίνει σχέδια η σχετικές προβλέψεις για τα χαρακτηριστικά των εμπλεκόμενων μηχανισμών που δημιούργησαν τα δεδομένα. Τα κύρια χαρακτηριστικά των αγνώστων βασικών κατανομών πιθανοτήτων μπορούν τα γίνουν γνωστά έτσι ώστε τα δεδομένα να χρησιμοποιηθούν με αποδοτικό τρόπο από έναν εκπαιδευόμενο. Τέτοια δεδομένα μπορούν να θεωρηθούν ως περιπτώσεις πιθανών σχέσεων μεταξύ των παρατηρούμενων μεταβλητών[16].

Ένα από τα κύρια αντικείμενα της έρευνας μάθησης μηχανών είναι ο σχεδιασμός αλγορίθμων που αναγνωρίζουν πολυσύνθετα σχέδια και λάβουν νοήμονες αποφάσεις βασισμένες στα δεδομένα εισόδου. Μια βασική δυσκολία είναι ότι η ομάδα όλων των δυνατών συμπεριφορών με ολα τα πιθανά δεδομένα εισόδου είναι πολύ μεγάλη για να συμπεριληφθεί σε ένα σύνολο δεδομένων που έχουν παρατηρηθεί (επιλεγμένα δεδομένα , training data). Με βάση τα προηγούμενα, ο εκπαιδευόμενος (learner) πρέπει να γενικεύσει από τα δεδομένα παραδείγματα έτσι ώστε να μπορεί να παράγει χρήσιμα συμπεράσματα για νέα προβλήματα.
Η Μάθηση Μηχανών μπορεί εναλλακτικά να ορισθεί ως ένα επιστημονικό πεδίο που προσδίδει στους υπολογιστές την ικανότητα να μάθουν χωρίς να έχουν άμεσα προγραμματισθεί για τέτοιο σκοπό.
Αναφέρεται ότι μία μηχανή μαθαίνει κάθε φορά που αλλάζει την δομή του, το πρόγραμμα ή δεδομένα, που βασίζονται στα δεδομένα εισόδου ή σε ανταπόκριση εξωτερικών πληροφοριών, έτσι ώστε η αναμενόμενη απόδοση να βελτιωθεί. Τέτοιες αλλαγές όπως η πρόσθεση εγγραφών σε μια βάση δεδομένων, ανήκουν σε δικαιοδοσίες άλλων γνωστικών αντικειμένων και γενικά είναι γνωστές ως μάθηση. Για παράδειγμα, όταν η απόδοση μίας μηχανής αναγνώρισης ομιλίας βελτιώνεται μετά από το άκουσμα
διαφόρων δειγμάτων της ομιλίας ενός ατόμου κατανοούμε και μπορούμε να πούμε ότι η μηχανή έχει μάθει.
Η Μάθηση Μηχανών συνήθως αναφέρεται σε αλλαγές σε συστήματα
που εκτελούν διαδικασίες (αναγνώριση, διάγνωση, σχεδιασμός, έλεγχος ρομπότ, προβλέψεις, κλπ.), που σχετίζονται με την Τεχνητή Νοημοσύνη (AI) [16].

5.2.1. Weka

Το WEKA (Waikato Environment for Knowledge Analysis) είναι μια σουίτα λογισμικού για μηχανική μάθηση και Εξόρυξη Δεδομένων. Αναπτύχθηκε στο Πανεπιστήμιο του Waikato της Ν. Ζηλανδίας και πήρε το όνομα του από το Weka, ένα μικρό και υπό εξαφάνιση πουλί της Ν. Ζηλανδίας. Το WEKA ανήκει στην κατηγορία του λεγόμενου "ελεύθερου λογισμικού" (freeware) και διατίθεται δημοσίως σύμφωνα με τους όρους της άδειας GNU General Public License, η οποία επιτρέπει στους χρήστες να χρησιμοποιούν, αλλά και να τροποποιούν ελεύθερα το λογισμικό [17][18].

Το WEKA είναι ένα από τα πιο διαδεδομένα λογισμικά Εξόρυξης Δεδομένων. Έχει χρησιμοποιηθεί σε μεγάλο αριθμό επιστημονικών εργασιών, και αρκετά βιβλία Εξόρυξης Δεδομένων αναφέρονται σε αυτό. Η μεγάλη δημοφιλία του οφείλεται στα ειδικά χαρακτηριστικά του και στις δυνατότητες που προσφέρει.

Αναλυτικότερα το WEKA:

● Περιέχει αρκετά μεγάλη ποικιλία μεθόδων για κατηγοριοποίηση, παλινδρόμηση, ανάλυση συστάδων, και κανόνες συσχέτισης. Επίσης, παρέχει δυνατότητες για προεπεξεργασία των δεδομένων, καθώς και εργαλεία οπτικοποίησης.

● Είναι λογισμικό ανοικτού κώδικα. Αυτό σημαίνει ότι ο πηγαίος κώδικας είναι δημοσίως διαθέσιμος. Χρήστες με γνώσεις προγραμματισμού μπορούν να τροποποιούν και να εξελίσσουν τους αλγορίθμους.

● Είναι γραμμένο σε γλώσσα Java, γεγονός που το καθιστά ικανό να εγκαθίσταται σε διαφορετικές πλατφόρμες υλικού και λογισμικού.

● Διαθέτει γραφικό περιβάλλον εργασίας. Στο διαδίκτυο υπάρχει διαθέσιμη μεγάλη ποικιλία βιβλιοθηκών για μηχανική μάθηση και εξόρυξη δεδομένων. Ωστόσο, η χρήση τους απαιτεί τη συγγραφή κώδικα. Αντιθέτως, το γραφικό περιβάλλον του WEKA επιτρέπει τη χρήση του λογισμικού από τελικούς χρήστες, οι οποίοι δεν διαθέτουν γνώσεις προγραμματισμού.

5.2.2. Linear Regression

Αρχικά, για την μοντελοποίηση του προβλήματος και κατ' επέκταση την δημιουργία μοντέλου πρόβλεψης χρησιμοποιήσαμε τον απλό αλγόριθμο Linear Regression.

Στην απλή γραμμική παλινδρόμηση(Linear Regression) έχουμε ένα σύνολο με δείγματα τιμών $\{x_i \ , y_i\}$. Σκοπός είναι να βρούμε ένα απλό μαθηματικό μοντέλο, το οποίο να περιγράφει την σχέση αυτών των δύο μεταβλητών την x και την y. Το απλό μαθηματικό μοντέλο που είναι μια ευθεία γραμμή της μορφής

$f(x) = y = ax + \beta$ η οποία "ταιριάζει" καλύτερα στο σύνολο των δειγμάτων. Έχοντας αυτό το μοντέλο μπορούμε να "προβλέψουμε" τις τιμές του y για νέες τιμές του x. Η μεθοδολογία αυτή χρησιμοποιείται στην μηχανική μάθηση (machine learning).

Παρατήρηση για τη συγκεκριμένη περίπτωση:

Η μεταβλητή x, δηλαδή αυτή προς πρόβλεψη αφορά τα media.
Η ανεξάρτητη μεταβλητή y, αφορά τα social media.

Αρχικά δοκιμάστηκε η λέξη "Trump" και το αποτέλεσμα του αλγορίθμου έδειξε πολύ μικρό σφάλμα (17.8933%).

```
=== Run information ===

Scheme:       weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4
Relation:     Trump-weka
Instances:    30
Attributes:   2
              media
              social
Test mode:    evaluate on training data

=== Classifier model (full training set) ===


Linear Regression Model

social =

      0.3763 * media +
      0.0019

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.02 seconds

=== Summary ===

Correlation coefficient                  0.9883
Mean absolute error                      0.0011
Root mean squared error                  0.0015
Relative absolute error                 17.8933 %
Root relative squared error             15.2749 %
Total Number of Instances               30
```

Εικόνα 22 - Output του Weka για τη λέξη "Trump" (Linear Regression)

Στη συνέχεια δοκιμάστηκε το dataset της λέξης “Τσίπρας”, στην οποία όμως το σφάλμα ήταν αρκετά μεγάλο (78.9238%)

```
=== Run information ===

Scheme:       weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4
Relation:     Tsipras-weka
Instances:    30
Attributes:   2
              media
              social
Test mode:    evaluate on training data

=== Classifier model (full training set) ===


Linear Regression Model

social =

      0.925  * media +
      0.0042

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correlation coefficient                  0.6114
Mean absolute error                      0.0018
Root mean squared error                  0.0024
Relative absolute error                 78.9238 %
Root relative squared error             79.1334 %
Total Number of Instances               30
```

Εικόνα 23 - Output του Weka για τη λέξη "Τσίπρας" (Linear Regression)

Αυτό οφείλεται στο γεγονός ότι η απόσταση μεταξύ των δύο συναρτήσεων δεν είναι σταθερή στο μεγαλύτερο μέρος της όπως ισχύει με την προηγούμενη περίπτωση, αλλά υπάρχει μία φανερή τμηματική ομοιότητα που ο αλγόριθμος δεν μπορούσε να αντιληφθεί.

Για το λόγο αυτό έπρεπε να βρούμε ένα νέο μοντέλο που να λαμβάνει υπόψη την σχέση των κοντινών χρονικά τιμών μεταξύ τους, έτσι ώστε να γίνει ασφαλέστερη πρόβλεψη.

Το καταλληλότερο μοντέλο για την περίπτωση μας ήταν η προσέγγιση των Forecasting Time Series.

5.2.3. Time Series Forecasting και SV Regression

Το μοντέλο Forecasting Time Series χρησιμοποιείται για να προβλέψει μελλοντικές τιμές που βασίζονται σε προηγούμενες παρατηρήσεις τιμών. Ενώ το regression analysis είναι συχνά εξοπλισμένο με τέτοιο τρόπο ώστε να ελέγχει θεωρίες σχετικά με το εάν οι τρέχουσες τιμές από μία ή περισσότερες ανεξάρτητες χρονοσειρές επηρεάζουν την την τρέχουσα τιμή μιας άλλης χρονοσειράς, αυτός ο τύπος της ανάλυσης χρονοσειρών δεν επικεντρώνεται στο να συγκρίνει τιμές ενός time series η πολλαπλών εξαρτημένων time series σε διαφορετικά σημεία στο χρόνο.

Παρόλα αυτά το Weka δεν υποστηρίζει την εφαρμογή του συγκεκριμένου μοντέλου(ARIMA, ARMA) αλλά αντί για αυτό προτείνει τη χρήση του Support Vector Regression[19].
Ενός Machine Learning μοντέλου που συνήθως χρησιμοποιείται για classification, αλλά έχει πολύ καλά αποτελέσματα και για continues τιμές.

Επομένως, εφαρμόσαμε το πακέτο time series analysis forecasting and control του weka σε συνδυασμό με τον αλγόριθμο support vector machine το οποίο χρησιμοποιείται για να προβλέψει μοντέλα που εξελίσσονται με το χρόνο.

Σε αυτή την περίπτωση οι δύο περιπτώσεις(media και Twitter για την ίδια λέξη) εξετάζονται σαν δύο διαφορετικές συναρτήσεις στο ίδιο σύστημα αξόνων όπου η μία επηρεάζει την περιοδικότητα της άλλης.

Παρατήρηση για τη συγκεκριμένη περίπτωση:

Το y είναι το ποσοστό εμφάνισης της λέξης και x οι μέρες του μήνα. Για να γίνει η πρόβλεψη ο αλγόριθμος παρατηρεί την ομοιότητα των γραφημάτων.

Εφαρμογή στη λέξη “Τσίπρας”

Εικόνα 24 - Γραφική παράσταση πρόβλεψης χρονικών σειρών της λέξης “Τσίπρας” για τα δύο μέσα

Εξαιτίας του Training που κάνει ο αλγόριθμος με βάση τα προηγούμενα δεδομένα το γράφημα μας ξεκινάει τις προβλέψεις από την 8η μέρα και μετά.

Με πράσινο χρόνο απεικονίζονται οι πραγματικές τιμές για τα ποσοστά εμφάνισης της λέξης στο Twitter ενώ με κίτρινο χρώμα το τι προέβλεψε ο αλγόριθμος για τις μελλοντικές τιμές του Twitter ένα βήμα κάθε φορά πριν σχεδιάσει το επόμενο σημείο λαμβάνοντας υπόψη το γράφημα των media(κόκκινο χρώμα).

Αντίστοιχα, με κόκκινο χρώμα απεικονίζονται οι πραγματικές τιμές και τα ποσοστά εμφάνισης της λέξης στα media, ενώ με μπλε χρώμα το τι προέβλεψε ο αλγόριθμος για τις μελλοντικές τιμές των media ένα βήμα κάθε φορά πριν σχεδιάσει το επόμενο σημείο λαμβάνοντας υπόψη το γράφημα του Twitter(πράσινο χρώμα).

Τα ποσοστά λάθους που προέκυψαν κατά το output του αλγορίθμου είναι πολύ μικρά.

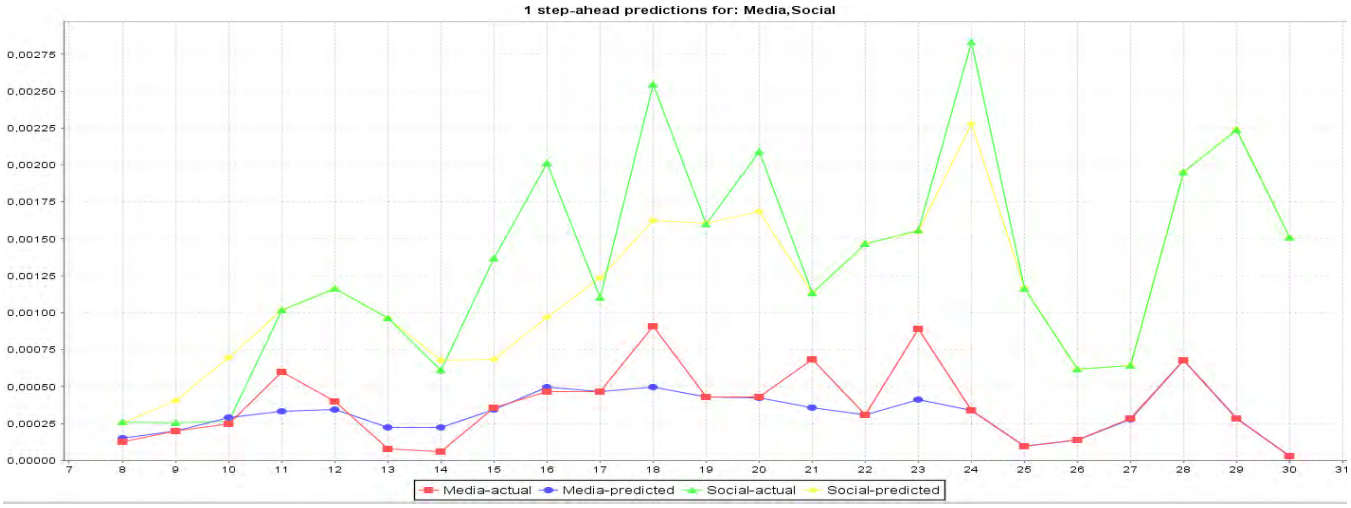
```
Number of kernel evaluations: 465 (98.445% cached)

=== Evaluation on training data ===
Target                       1-step-ahead
=======================================
media
  N                                  23
  Mean absolute error            0.0004
  Root mean squared error         0.001
social
  N                                  23
  Mean absolute error            0.0011
  Root mean squared error        0.0026

Total number of instances: 30
```

Εικόνα 25 - Output του Weka για το σφάλμα πρόβλεψης, της λέξης "Τσίπρας" (support vector regression)

Εφαρμογή στη λέξη "Mannequin"



Εικόνα 26 - Γραφική παράσταση πρόβλεψης χρονικών σειρών της λέξης "Mannequin" για τα δύο μέσα

Εξαιτίας του Training που κάνει ο αλγόριθμος με βάση τα προηγούμενα δεδομένα το γράφημα μας ξεκινάει τις προβλέψεις από την 7η μέρα και μετά.

Με πράσινο χρόνο απεικονίζονται οι πραγματικές τιμές για τα ποσοστά εμφάνισης της λέξης στο Twitter ενώ με κίτρινο χρώμα το τι προέβλεψε ο αλγόριθμος για τις μελλοντικές τιμές του Twitter ένα βήμα κάθε φορά πριν σχεδιάσει το επόμενο σημείο λαμβάνοντας υπόψη το γράφημα των media(κόκκινο χρώμα).

Αντίστοιχα, με κόκκινο χρώμα απεικονίζονται οι πραγματικές τιμές και τα ποσοστά εμφάνισης της λέξης στα media, ενώ με μπλε χρώμα το τι προέβλεψε ο αλγόριθμος για τις μελλοντικές τιμές των media ένα βήμα κάθε φορά πριν σχεδιάσει το επόμενο σημείο λαμβάνοντας υπόψη το γράφημα του Twitter(πράσινο χρώμα).

Τα ποσοστά λάθους που προέκυψαν κατά το output του αλγορίθμου είναι στις μισές τιμές μικρά και στις άλλες μισές μεγάλα.

```
Number of kernel evaluations: 465 (99.388% cached)

=== Evaluation on training data ===
Target                      1-step-ahead
=======================================
Media
  N                                23
  Mean absolute error            0.0001
  Root mean squared error        0.0002
Social
  N                                23
  Mean absolute error            0.0002
  Root mean squared error        0.0004

Total number of instances: 30
```

Εικόνα 27 - Output του Weka για το σφάλμα πρόβλεψης, της λέξης "Mannequin" (support vector regression)

Από τα πειράματα προκύπτει ότι μπορεί να γίνει πρόβλεψη από το ένα μέσο στο άλλο(είτε Twitter προς media, είτε media προς Twitter). Παρόλα αυτά η πρόβλεψη των μελλοντικών τιμών του Twitter σύμφωνα με τις τιμές των media είναι αρκετά πιο ακριβής, σε σχέση με την πρόβλεψη των μελλοντικών τιμών των media, σύμφωνα με τις τιμές του Twitter. Άρα είναι πιο εύκολο να προβλέψουμε τις αντιδράσεις του Twitter με βάση τα media και δυσκολότερο αυτές του Twitter με βάση τα media.

6. Συμπεράσματα

Παρατηρούμε λοιπόν πως υπάρχει συσχέτιση μεταξύ των media και των κοινωνικών δικτύων καθώς τόσο τα ποσοστά μεμονομένων λέξεων όσο και τα ποσοστά από τα σύνολα λέξεων μιας θεματολογίας εμφανίζουν παρόμοια κατανομή στο χρόνο.

Αυτό γίνεται ιδιαίτερα εμφανές στα γραφήματα ειδικά σε χρονικές στιγμές όπου μια είδηση έρχεται στην επικαιρότητα και οι λέξεις που τη συνοδεύουν γίνονται trends. Στα σημεία αυτά παρατηρούμε τοπικά μέγιστα στις συναρτήσεις, ενώ τόσο η άνοδος μιας τάση όσο και η πτώση φαίνεται να συμπίπτουν χρονικά.

Όσον αφορά την πρόβλεψη μελοντικών τάσεων στα social media φαίνεται πως αυτό είναι δυνατόν καθώς σε πολλά σημεία η άνοδος των ποσοστών για μια λέξη ακολουθει μια αντίστοιχη άνοδα στα media, ενώ στις περισσότερα περιπτώσεις που εξετάσαμε παρατηρούμε μια μικρή χρονική καθυστέρηση στην απόκριση ενός trend από sites στο Twitter.

Αυτό το αποδείξαμε επίσης με τεχνικές μηχανικής μάθηση όπου τα ποσοστά πρόβλεψης είναι ιδιαίτερα αποτελεσματικά ειδικά όταν η επόμενη βρίσκεται χρονικά πολύ κοντά από τα δεδομένα που έχουμε μέχρι εκείνη τη στιγμή.

7. References

[1] Μαίδου Σεβαστή,"Η χρήση του Twitter ως μέσο συμμετοχής στην πολιτική",(2016),Available:https://dspace.lib.uom.gr/bitstream/2159/19039/6/MaidouSevastiMsc2016.pdf-

[2] "Web Scraping", Wikipedia, Available: https://en.wikipedia.org/wiki/Web_scraping

[3] Μπαρζόκα Βασίλειου,"Εξόρυξη στοιχείων σχετικών με την αγορά εργασίας πληροφορικής",(2012) ,Πτυχιακή εργασία,Τεχνολογικό Εκπαιδευτικό Ιδρυμα Σερρών Σχολή Τεχνολογικών Εφαρμογών Τμήμα Πληροφορικής & Επικοινωνιών, Available:http://informatics.teicm.gr/attachments/links/barzokas2012.pdf

[4] Μαθιουλάκης Μανώλης και Χαιρέτης Διονύσης,"Ανάπτυξη εφαρμογής ιστού για άντληση και επεξεργασία rss feeds",(2011), Available: http://nefeli.lib.teicrete.gr/browse/stef/epp/2011/MathioulakisManolis,ChairetisDionysios/attached-document-1320150866-361739-32090/Chairetis_Mathioulakis2011.pdf

[5] ”Τι είναι τα meta tags και τρόποι αξιοποίησής τους”, Hiremycode.com , blog, available: https://www.hiremycode.com/blog/ti-ine-ta-meta-tags-ke-tropi-axiopiisis-tous/

[6] KEA keyphrase extraction algorithm, OpenSource Tool, Available: http://www.nzdl.org/Kea/

[7] Aytuğ Onan, Hasan Bulut and Serdar Korukoğlu, “Ensemble of keyword extraction methods and classifiers in text classification”,(2016), Celal Bayar University, Department of Computer Engineering, Muradiye, Manisa, Turkey, Ege University, Department of Computer Engineering, Bornova, Izmir, Turkey

[8] Α. Εμμανουηλίδης, “Σχεδιασμός και υλοποίηση αλγορίθμων ταιριάσματος με χρήση λέξεων κλειδιά και εφαρμογή τους στο ταίριασμα βιογραφικών”, (2011), Διπλωματική Εργασία, Τμήμα Πληροφορικής, Οικονομικό Πανεπιστήμιο Αθηνών

[9] Αντώνιος Α. Απειρανθίτης και Βασίλειος Α. Τσαχανσάχης , ”Ανάλυση Δεδομένων από Κοινωνικά Δίκτυα για τη Μελέτη των Συναισθημάτων των Χρηστών”, (2016), Διπλωματική Εργασία, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ, Εθνικό Μετσόβιο Πολυτεχνείο

[10] “Twitter Developer Documentation“, Available: https://dev.twitter.com/overview/api/twitter-ids-json-and-snowflake

[11] L. Marujo, W. Ling, I. Trancoso, C. Dyer, A. W. Black, A. Gershman, D. Martins de Matos, J. P. Neto, and J. Carbonell, “Automatic Keyword Extraction on Twitter”, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA,

Instituto Superior Tecnico, Universidade de Lisboa, Lisbon, Portugal, INESC-ID, Lisbon, Portugal

[12] Boilerpipe algorithm, Java Api, Available:http://boilerpipe-web.appspot.com/

[13] Robert J. Moore, ”Twitter Data Analysis: An Investor's Perspective”, (2009), Ηλεκτρονικό Άρθρο, Available: https://techcrunch.com/2009/10/05/twitter-data-analysis-an-investors-perspective-2/

[14] Δημήτρης Τσαρούχας,“Ανάπτυξη διαδικτυακής εφαρμογής για την αναζήτηση κειμενικών δεδομένων με τη συνεργασία NoSql βάσης δεδομένων και του Elasticsearch”,(2013),Μεταπτυχιακή Διατριβή,Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής, Πρόγραμμα Μεταπτυχιακών Σπουδών “Πληροφορική”,Available:http://dione.lib.unipi.gr/xmlui/bitstream/handle/unipi/6029/Tsarouchas.pdf?sequence=2

[15] Elasticsearch,Elastic Stack and Product Documentation,Java API 2.3,Available: https://www.elastic.co/guide/en/elasticsearch/client/java-api/2.3/index.html

[16] Λυπιτάκη Αναστασία-Δήμητρα,”Μηχανική μάθηση σε ανομοιογενή δεδομένα”,(2014),Μεταπτυχιακή Διατριβή,Πανεπιστήμιο Πατρών

[17] Weka, Data Mining Software in Java,University of Waikato,Available:http://www.cs.waikato.ac.nz/ml/weka/

[18] Κύρκος, Ε. 2015. Οδηγός WEKA. (Κεφάλαιο Συγγράμματος). Στο Κύρκος, Ε. 2015. *Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων*. (ηλεκτρ. βιβλ.)

Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, κεφ 13. Available: http://hdl.handle.net/11419/1239

[19] Pentaho Data Mining Community Documentation, “Time Series Analysis and Forecasting with Weka”,Available: http://wiki.pentaho.com/display/DATAMINING/Time+Series+Analysis+and+Forecasting+with+Weka

[20] Charisios Zafeiris, “Forecasting trends in users’ impact on shaping the internet media agenda”, (2017), Διπλωματική Εργασία, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας