Ατομική Διπλωματική Εργασία

ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΑΠΟ ΤΟ TWITTER ΓΙΑ TWEETS ΠΟΥ ΣΧΕΤΙΖΟΝΤΑΙ ΜΕ ΑΕΡΟΠΟΡΙΚΑ ΤΑΞΙΔΙΑ

Στέφανη Κασινοπούλου

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάιος 2019

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΑΠΟ ΤΟ TWITTER ΓΙΑ TWEETS ΠΟΥ ΣΧΕΤΙΖΟΝΤΑΙ ΜΕ ΑΕΡΟΠΟΡΙΚΑ ΤΑΞΙΔΙΑ

Στέφανη Κασινοπούλου

Επιβλέπων Καθηγητής

Μάριος Δικαιάκος

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

Μάιος 2019

Περίληψη

Ζούμε σε μια εποχή όπου τα μέσα κοινωνικής δικτύωσης έχουν γίνει κομμάτι του σύγχρονου ανθρώπου. Με το πάτημα ενός κουμπιού μπορούμε να ανταλλάξουμε σε πραγματικό χρόνο απόψεις, ιδέες, πληροφορίες και νέα με άτομα από κάθε γωνιά του κόσμου. Έτσι συλλέγονται ογκώδης δεδομένα από τις πλατφόρμες κοινωνικής δικτύωσης, ανοίγοντας νέους αναδυόμενους τομείς στην επιστήμη των δεδομένων. Στα πλαίσια της συγκεκριμένης ατομικής διπλωματική εργασίας θα εστιάσουμε στην πλατφόρμα κοινωνικής δικτύωσης του Twitter.

Η πλατφόρμα του Twitter μας δίνει την δυνατότητα να συλλέγουμε tweets από κανόνες φιλτραρίσματος που θέτουμε εμείς. Οπότε η συλλογή δεδομένων από το Twitter είναι ένας καλός τρόπος για συλλογή κειμένων γύρω από κάποιο θέμα. Το γεγονός ότι μπορούμε να συλλέξουμε μεγάλο όγκο tweets με συνδυασμό του εργαλείου ανάλυσης συναισθήματος, έχουμε την δυνατότητα να εξάγουμε χαρακτηριστικά αυτών των δεδομένων όπως την θετική ή αρνητική γνώμη των ανθρώπων γύρω από κάποιο θέμα.

Αυτή η εργασία θα επικεντρωθεί στην συλλογή των tweets που είναι σχετικά με το αεροπορικό ταξίδι. Μέσα από τα tweets που θα συλλεχθούν θα ανιχνευθεί ο βαθμός ικανοποίησης των ταξιδιωτών από τις υπηρεσίες που προσφέρουν οι αερογραμμές και τα αεροδρόμια. Θα παρουσιαστούν οι τρόποι εύρεσης των κατάλληλων φιλτραρισμάτων για την συλλογή των tweets γύρω από το συγκεκριμένο θέμα. Μετά θα ακολουθήσει η προετοιμασία και ο καθαρισμός των δεδομένων για την αργότερη ανάλυση και η αξιολόγηση της κάθε τεχνικής προ επεξεργασίας των δεδομένων. Τέλος γίνεται η ανάλυση συναισθήματος στα tweets και παρουσιάζεται εικονικά η υποκειμενική ικανοποίηση των ταξιδιωτών.

ΠΕΡΙΕΧΟΜΕΝΑ

Κεφάλαιο 1 - Εισαγωγή

1.1 Υποκίνηση της Εργασίας

Ζούμε σε μια εποχή όπου τα μέσα κοινωνικής δικτύωσης έχουν γίνει κομμάτι του σύγχρονου ανθρώπου. Με το πάτημα ενός κουμπιού μπορούμε να ανταλλάξουμε σε πραγματικό χρόνο απόψεις, ιδέες, πληροφορίες και νέα με άτομα από κάθε γωνιά του κόσμου. Έτσι συλλέγονται ογκώδης δεδομένα από τις πλατφόρμες κοινωνικής δικτύωσης, ανοίγοντας νέους αναδυόμενους τομείς στην επιστήμη των δεδομένων.

Οποιοσδήποτε που χρησιμοποιεί τα APIs (Application Programming Interfaces) που παρέχονται από τις πλατφόρμες κοινωνικής δικτύωσης μπορεί να ανιχνεύσει και να συλλέξει δεδομένα. Με την κατάλληλη επεξεργασία των δεδομένων αυτών μπορεί να παραχθεί πληροφορία η οποία να είναι χρήσιμη για διάφορους σκοπούς. Ένα εργαλείο για ανάλυση δεδομένων φυσικής γλώσσας είναι η ανάλυση συναισθήματος. Με το εργαλείο αυτό μπορούμε να εξάγουμε τις υποκειμενικές πληροφορίες ενός κειμένου. Άρα ο συνδυασμός μεγάλου όγκου δεδομένων και το εργαλείο που εξάγει το συναίσθημα δίνει την δυνατότητα εντοπισμού κάποιων χαρακτηριστικών των δεδομένων όπως θετική ή αρνητική γνώμη των ανθρώπων γύρω από κάποιο θέμα. Έτσι με την σωστή ανάλυση του συναισθήματος μπορούμε να εξάγουμε χρήσιμες πληροφορίες που θα βοηθήσουν στην καλύτερη κατανόηση των προβλημάτων των χρηστών και ακολούθως στην καλύτερη αντιμετώπιση τους.

Ένα από τα δημοφιλέστερα μέσα κοινωνικής δικτύωσης είναι το Twitter, όπου υπάρχουν πάνω από 336 εκατομμύρια ενεργούς χρήστες [1]. Τα χαρακτηριστικά που το κάνει να διαφέρει από τα υπόλοιπα μέσα κοινωνικής δικτύωσης είναι η χρήση διαφόρων ετικετών που διευκολύνει την αναζήτηση για συγκεκριμένα tweets και την αλληλεπίδραση των χρηστών με το Twitter. Επίσης, ακόμη ένα ξεχωριστό χαρακτηριστικό είναι ο περιορισμός συγγραφής στους 280 χαρακτήρες, αναγκάζοντας τους χρήστες να γίνονται δημιουργικοί με το γράψιμο τους .

Το Twitter χρησιμοποιείται τόσο από απλούς χρήστες όσο και από επιχειρήσεις. Οι χρήστες αναζητάνε συγκεκριμένες ετικέτες ή λέξεις στην μηχανή αναζήτησης του twitter για να μπορέσουν να παρακολουθήσουν συζητήσεις και απόψεις γύρω από κάποιο συγκεκριμένο θέμα που τους ενδιαφέρει και να εμπλακούν σε αυτό. Από την πλευρά των εταιριών, η χρήση των κατάλληλων ετικετών μπορεί να βοηθήσει στην απευθείας διαφήμιση και

εξυπηρέτηση σε ενδεχόμενους πελάτες. Έτσι το Twitter δίνει την δυνατότητα σε όλους τους χρήστες, χωρίς διακρίσεις, να ακουστούν και να εκφράσουν την άποψή τους.

Επίσης, το Twitter είναι γνωστό και στο χώρο των αεροπορικών εταιριών. Οι αεροπορικές εταιρίες δημιουργούν λογαριασμό στο Twitter για να έχουν την δυνατότητα να ενημερώνουν για διάφορες προσφορές και γεγονότα, να απαντάνε σε ερωτήσεις των ταξιδιωτών και να χειρίζονται τα παράπονα και σχόλια των πελατών τους. Αντιστοίχως οι πελάτες γράφουν αιτήσεις, ερωτήσεις, σχόλια, εμπειρίες, παράπονα κ.α. για την αεροπορική εταιρία που χρησιμοποιούν μέσω των tweets. Έτσι με την σωστή ανάλυση η κάθε αεροπορική εταιρία μπορεί να αξιολογήσει τις υπηρεσίες της μέσα από τα tweets των πελατών της.

Σε αυτή τη ατομική διπλωματική εργασία γίνεται η προσπάθεια να αναλυθούν τα tweets που σχετίζονται με αεροπορικά ταξίδια και στην συνέχεια να εξαχθούν συμπεράσματα για τις απόψεις και τα συναισθήματα των ταξιδιωτών.

1.2 Στόχοι της Εργασίας

Οι στόχοι που τέθηκαν για αυτή την εργασία αφορούν την ανάπτυξη ενός προγράμματος που θα αναλύει το συναίσθημα στα tweets που είναι σχετικά με αεροπορικά ταξίδια.

Το πρόγραμμα αυτό θα πρέπει να:

1) μαζεύει όσο το δυνατόν περισσότερα tweets που σχετίζονται με αεροδρόμια και αεροπορικά ταξίδια από επιβάτες και προσωπικό,
2) ταξινομεί το κείμενο των tweets με βάση το συναίσθημα τους, χρησιμοποιώντας τεχνικές εξόρυξης δεδομένων,
3) αναπαριστά εικονικά τα αποτελέσματα της ανάλυσης.

1.3 Περίγραμμα της Εργασίας

Στο παρόν κεφάλαιο παρουσιάστηκαν το κίνητρο, οι στόχοι και το περίγραμμα της διπλωματικής εργασίας.

Το δεύτερο κεφάλαιο περιέχει τις βασικές ορολογίες, τεχνολογίες και τεχνικές που συνέβαλαν στην ανάπτυξη του προγράμματος για καλύτερη κατανόηση της εργασίας. Σε

αυτό το κεφάλαιο ορίζεται το πρόβλημα της εργασίας, εξηγείται τι είναι το Twitter και το Twitter Streaming API και παρουσιάζονται τα εργαλεία Python, MongoDB και NLTK.

Το τρίτο κεφάλαιο περιέχει μια αναλυτική περιγραφή όλων των βημάτων της διαδικασίας, από την συλλογή των δεδομένων μέχρι την εξαγωγή των συμπερασμάτων. Περιγράφεται η διαδικασία για την εύρεση των μεθόδων που στοχεύουν στην μέγιστη δυνατή συλλογή των tweets που σχετίζονται με αεροπορικό ταξίδι. Στη συνέχεια παρουσιάζονται οι διάφορες τεχνικές για την προ επεξεργασία και προετοιμασία των δεδομένων για τις μετέπειτα αναλύσεις. Ακολούθως αξιολογούνται οι τεχνικές προ επεξεργασίας των δεδομένων για την προετοιμασία των κειμένων των tweets για την ανάλυση του συναισθήματος. Με την εύρεση της κατάλληλης τεχνικής προ επεξεργασίας, αναλύονται τα κείμενα των tweets και κατηγοριοποιούνται θετικά, αρνητικά και ουδέτερα αναλόγως της τιμής του συναισθήματος που υπολογίζει το εργαλείο ανάλυσης συναισθήματος.

Στο τέταρτο κεφάλαιο γίνεται σχολιασμός της ανάλυσης και εξάγεται το συμπέρασμα για την ικανοποίηση των ταξιδιωτών.

Κεφάλαιο 2 - Θεωρητικό Υπόβαθρο

2.1 Ορισμός του προβλήματος

Το ζητούμενο της παρούσας ατομικής διπλωματικής εργασίας είναι η περιγραφή και η υλοποίηση μιας μεθοδολογίας που θα ανιχνεύει μέσα από το Twitter τον βαθμό ικανοποίησης των ταξιδιωτών από τις υπηρεσίες που προσφέρουν οι αερογραμμές και τα αεροδρόμια. Η μεθοδολογία αυτή χωρίζεται σε μικρότερα μέρη, τα οποία θα παρουσιαστούν αναλυτικά στην συνέχεια.

Το πρώτο μέρος είναι η συλλογή των δεδομένων. Τα εισερχόμενα δεδομένα, δηλαδή τα tweets που μαζεύουμε με την βοήθεια του Twitter Streaming API, εξαρτώνται από την σύνταξη του κανόνα που γίνεται η αίτηση στην πλατφόρμα του Twitter. Η πρόκληση εδώ είναι η καλή σύνταξη των κανόνων για το Twitter Streaming. Καθημερινά δημοσιεύονται περίπου 500 εκατομμύρια tweets και το Twitter Streaming API μας επιτρέπει να εξάγουμε το 1% από αυτά. Για αυτό τον λόγο τα ερωτήματα που θα χρησιμοποιήσουμε πρέπει να είναι προσαρμοσμένα στο θέμα μας για να μπορέσουμε να πάρουμε όσο το δυνατό μεγαλύτερο ποσοστό χρήσιμων δεδομένων μέσα στο 1% που μας προσφέρει το Twitter Streaming API.

Το δεύτερο μέρος είναι η επεξεργασία των δεδομένων. Το κείμενο χρειάζεται να υποστεί κάποια προ επεξεργασία πριν δοθεί ως παράμετρο στην ανάλυση. Υπάρχουν διάφορες τεχνικές για την προ επεξεργασία του κειμένου, οι οποίες διαφέρουν στην κάθε ανάλυση.

Το τρίτο μέρος αφορά την ανάλυση των προ επεξεργασμένων δεδομένων. Σε αυτή τη φάση έχουμε μια ξεκάθαρη τεχνική για την επεξεργασία των κειμένων. Οπότε τώρα μπορούμε με την χρήση του εργαλείου VADER [1] της βιβλιοθήκης NLTK να περάσουμε ως είσοδο το επεξεργασμένο κείμενο και να πάρουμε το ποσοστό θετικότητας ή αρνητικότητας του κειμένου.

Το τέταρτο και τελευταίο μέρος είναι κάποιες επιπλέον επεξεργασίες στα δεδομένα για εξαγωγή πληροφορίας βάση το συναίσθημα των tweets. Η επεξεργασίες και αναλύσεις παρουσιάζονται σε γραφικές παραστάσεις έτσι ώστε να φαίνεται εύκολα η πληροφορία.

2.2 Twitter

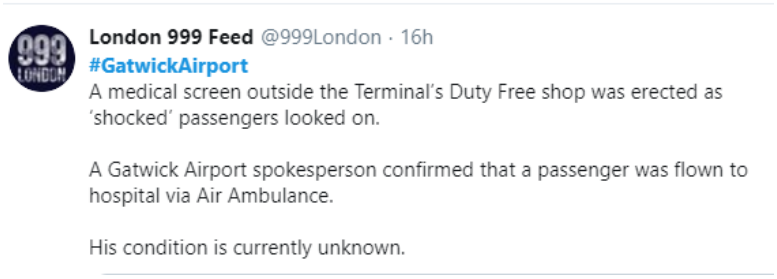
Το Twitter είναι μια αμερικάνικη πλατφόρμα κοινωνικής δικτύωσης που ιδρύθηκε πριν από 13 χρόνια. Είναι ένα από τα δημοφιλέστερα κοινωνικά δίκτυα ανά το παγκόσμιο. Οποιοσδήποτε μπορεί να δημιουργήσει λογαριασμό δωρεάν βάζοντας το μοναδικό ψευδώνυμο ο ίδιος επιθυμεί. Με την δημιουργία λογαριασμού μπορεί να δημοσιεύει tweets, να ακολουθεί - **following** χρήστες που τον ενδιαφέρουν και να ακολουθείτε από άλλους χρήστες – **follower.** Τα tweets των ατόμων που γίνονται following θα εμφανίζονται στην αρχική σελίδα και έτσι ο χρήστης θα μπορεί να ενημερώνεται για τα νέα των άλλων χρηστών.

Το χαρακτηριστικό του Twitter που το κάνει να ξεχωρίζει από τα υπόλοιπα κοινωνικά δίκτυα είναι ότι οι δημοσιεύσεις, τα λεγόμενα tweets, περιορίζονται στους 280 χαρακτήρες. Με αυτό τον τρόπο μπορεί κάποιος να παρακολουθήσει εκατοντάδες άλλους χρήστες του Twitter και να διαβάσει το περιεχόμενο με μια ματιά. Αυτές οι συνθήκες είναι ιδανικές για τον σύγχρονο πολυάσχολο άνθρωπο.

Στα tweets εκτός από το πραγματικό κείμενο μπορούν να χρησιμοποιηθούν επιπρόσθετα κάποιες ετικέτες ή εντολές οι οποίες συνεισφέρουν σε μια καλύτερη εμπειρία χρήσης του Twitter. Η πρώτη ετικέτα είναι το **mention "@"**. Η ετικέτα mention χρησιμοποιείται στις περιπτώσεις όπου κάποιος θέλει να απευθυνθεί σε ένα συγκεκριμένο άτομο. Αυτό μπορεί να το επιτευχθεί με την εισαγωγή του χαρακτήρα @ και στη συνέχεια το όνομα του ατόμου που θέλει να απευθυνθεί χωρίς κενό μεταξύ του χαρακτήρα και το όνομα του χρήστη. Έτσι το tweet που δημοσιεύεται θα εμφανιστεί στην αρχική σελίδα του χρήστη που αναφέρεται το @, ανεξαρτήτως αν τον ακολουθεί ή όχι. Ένα τέτοιο μήνυμα μπορεί να φαίνεται ως εξής:

Η δεύτερη ετικέτα είναι το **hashtag "#"**. Η χρήση της ετικέτα hashtag είναι για τον καθορισμό του θέματος του tweet. Αυτό βοηθά στην εξεύρεση tweets σχετικά με ένα

συγκεκριμένο θέμα και επιτρέπει στους χρήστες να παρακολουθούν όλα τα νέα tweets σε ένα δεδομένο θέμα σε πραγματικό χρόνο. Ένα τέτοιο μήνυμα μπορεί να φαίνεται ως εξής:



Αυτές είναι οι δύο κύριες ετικέτες που χρειάζεται να ξέρουμε στα πλαίσια της συγκεκριμένης ατομικής διπλωματικής εργασίας.

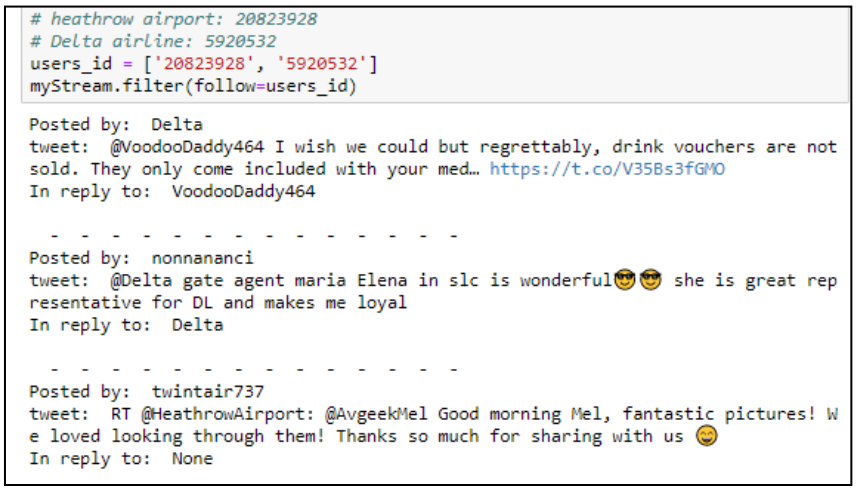
2.3 Twitter Streaming API

Το API ή αλλιώς Διεπαφή Προγραμματισμού Εφαρμογών αναφέρεται στο σετ των εντολών που παρέχει κάποιο πρόγραμμα ή βιβλιοθήκη στους προγραμματιστές για να κάνουν αιτήσεις προς αυτά για ανταλλαγή ή επεξεργασία δεδομένων. Σε αυτή την περίπτωση, το Twitter παρέχει το Twitter Streaming API για την παρακολούθηση μιας ροής από tweets σε πραγματικό χρόνο με βάση κάποιο φιλτράρισμα.

Το Twitter Streaming API παρέχει δύο επιλογές για την κατανάλωση tweets σε πραγματικό χρόνο. Η πρώτη επιλογή είναι το PowerTrack API. Είναι ένα επί πληρωμή API που προϋποθέτει την μακρόχρονη συνεργασία με την ομάδα του Twitter. Το PowerTrack API παρέχει στο πελάτη τη δυνατότητα να φιλτράρει μόνο δεδομένα που τον ενδιαφέρουν χωρίς ιδιαίτερους περιορισμούς. Δεύτερη επιλογή, η οποία θα χρησιμοποιηθεί στα πλαίσια της συγκεκριμένης εργασίας, είναι το statuses/filter API. Αυτή η επιλογή επιστρέφει tweets που ταιριάζουν σε έναν κανόνα φιλτραρίσματος ανά επιτρεπτή σύνδεση. Ένας κανόνας φιλτραρίσματος είναι μια ή περισσότερες παραμέτρους που όρισε ο πελάτης. Οι παράμετροι που μπορεί να ορίσει ο χρήστης είναι οι εξής τρείς:

Follow – Μια λίστα που διαχωρίζεται με κόμμα και περιέχει IDs χρηστών. Η λίστα αύτη υποδηλώνει ότι το API θα μας επιστρέψει τα tweets που δημιουργούνται από τους χρήστες της λίστας μας, απαντήσεις άλλων χρηστών σε κάθε tweet που δημιουργήθηκε από τους

χρήστες της λίστας και tweets των χρηστών της λίστας που αναδημοσιεύονται από άλλους χρήστες. Το μέγεθος της λίστας για το status/filter API περιορίζεται στα 5000 IDs χρηστών.

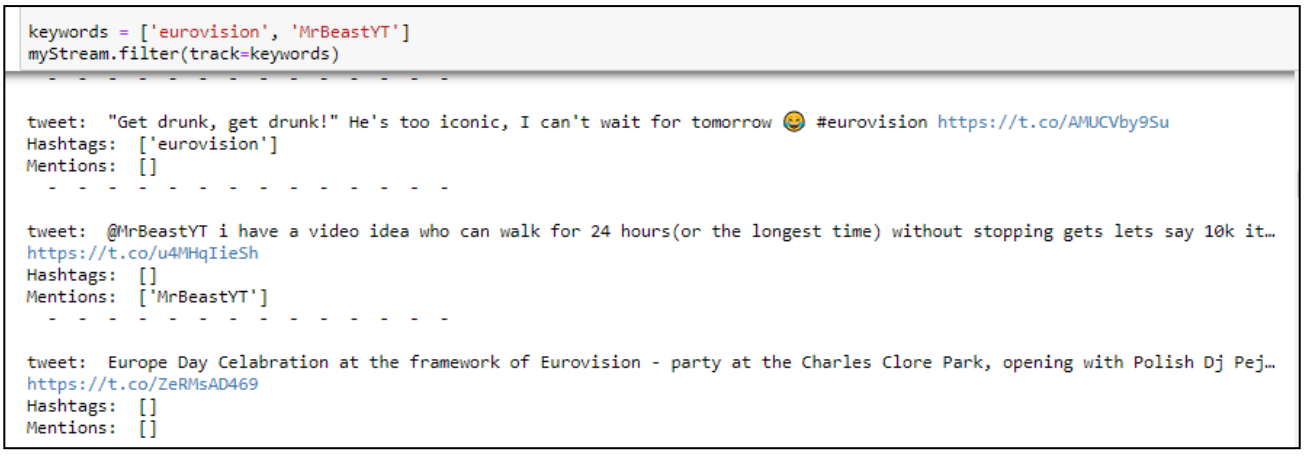


```
# heathrow airport: 20823928
# Delta airline: 5920532
users_id = ['20823928', '5920532']
myStream.filter(follow=users_id)
```

```
Posted by:  Delta
tweet:  @VoodooDaddy464 I wish we could but regrettably, drink vouchers are not
sold. They only come included with your med… https://t.co/V35Bs3fGMO
In reply to:  VoodooDaddy464

  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -
Posted by:  nonnananci
tweet:  @Delta gate agent maria Elena in slc is wonderful😎😎 she is great rep
resentative for DL and makes me loyal
In reply to:  Delta

  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -
Posted by:  twintair737
tweet:  RT @HeathrowAirport: @AvgeekMel Good morning Mel, fantastic pictures! W
e loved looking through them! Thanks so much for sharing with us 😊
In reply to:  None
```

Στην πιο πάνω εικόνα παρουσιάζεται ένα παράδειγμα συλλογής tweets με την παράμετρο follow με τα IDs των λογαριασμών της αερογραμμής Delta και του αεροδρομίου του Heathrow στο Λονδίνο. Στην πρώτη περίπτωση φαίνεται ότι το tweet έχει δημιουργηθεί από τη Delta προς απάντηση του χρήστη *VoodooDaddy464*. Στη δεύτερη περίπτωση ο χρήστης *nonnananci* απαντάει σε tweet που δημιουργήθηκε από τη Delta. Στη τελευταία περίπτωση ο χρήστης *twintair737* αναδημοσιεύει το tweet που δημοσίευσε το Heathrow αεροδρόμιο.

Track – Μια λίστα χωρισμένη με κόμμα που περιέχει φράσεις. Μια φράση μπορεί να είναι ένας ή περισσότεροι όροι που χωρίζονται από κενό και περιορίζεται στους 60 χαρακτήρες. Υποδηλώνει ότι το API θα μας επιστρέψει tweets που στο κείμενο τους ταιριάζουν όλοι οι όροι κάποιας φράσης από την λίστα, ανεξαρτήτως της σειράς που παρουσιάζονται οι όροι και ανεξαρτήτως των κεφαλαίων ή πεζών γραμμάτων. Εκτός από το κείμενο του tweet γίνονται αντιστοιχήσεις και σε άλλες οντότητες του tweets όπως στην οντότητα των hashtags, στην οντότητα των mentions και στην οντότητα των συνδέσμων. Το μέγεθος της λίστας για το status/filter API έχει ως όριο 400 φράσεις.

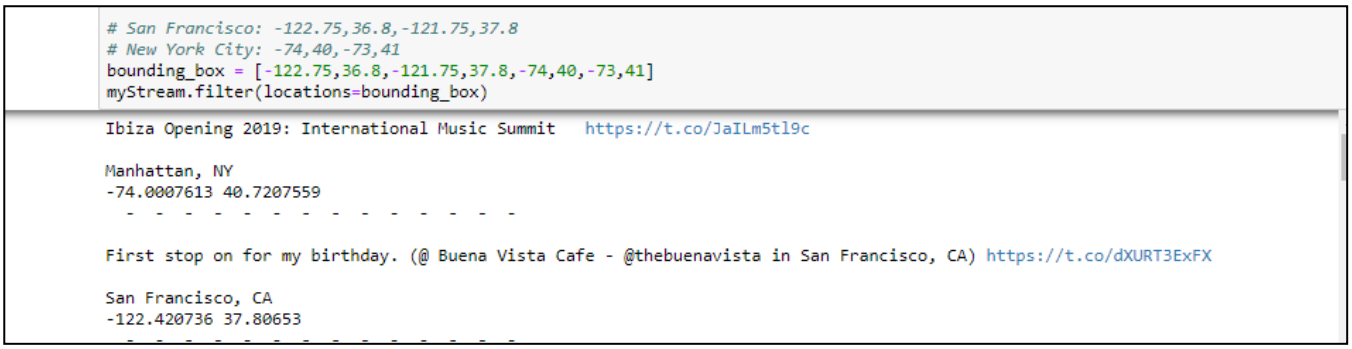


```
keywords = ['eurovision', 'MrBeastYT']
myStream.filter(track=keywords)
```

```
tweet:  "Get drunk, get drunk!" He's too iconic, I can't wait for tomorrow 😂 #eurovision https://t.co/AMUCVby9Su
Hashtags:  ['eurovision']
Mentions:  []
  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -

tweet:  @MrBeastYT i have a video idea who can walk for 24 hours(or the longest time) without stopping gets lets say 10k it…
https://t.co/u4MHqIieSh
Hashtags:  []
Mentions:  ['MrBeastYT']
  -  -  -  -  -  -  -  -  -  -  -  -  -  -  -

tweet:  Europe Day Celabration at the framework of Eurovision - party at the Charles Clore Park, opening with Polish Dj Pej…
https://t.co/ZeRMsAD469
Hashtags:  []
Mentions:  []
```

Στην εικόνα φαίνεται ο τρόπος συλλογής των tweets με την παράμετρο track για τις φράσεις «Eurovision» και «MrBeastYT». Το πρώτο tweet έχει αντιστοίχιση με την φράση «Eurovision» στην οντότητα hashtag. Το δεύτερο tweet έχει αντιστοίχιση με την φράση «MrBeastYT» στην οντότητα mentions. Το τρίτο tweet φαίνεται να έχει αντιστοίχιση με την φράση «Eurovision» μέσα στο κείμενο του.

Locations – Μια λίστα χωρισμένη με κόμμα που περιέχει ζεύγη γεωγραφικού μήκους και πλάτους που προσδιορίζουν ένα σύνολο πλαισίων οριοθέτησης. Το API θα επιστρέφει μόνο γεωγραφικά tweets που ανήκουν στα ζητούμενα πλαίσια οριοθέτησης. Το κάθε πλαίσιο οριοθέτησης πρέπει να ορίζεται ως ζεύγος από ζεύγη γεωγραφικού μήκους και πλάτους, όπου πρώτα θα ορίζεται το ζεύγος της νοτιοδυτικής γωνίας του πλαισίου οριοθέτησης και μετά το ζεύγος της βορειοανατολικής γωνίας. Το μέγεθος της λίστας έχει όριο μέχρι 25 πλαίσια οριοθέτησης.



```
# San Francisco: -122.75,36.8,-121.75,37.8
# New York City: -74,40,-73,41
bounding_box = [-122.75,36.8,-121.75,37.8,-74,40,-73,41]
myStream.filter(locations=bounding_box)
```

```
Ibiza Opening 2019: International Music Summit   https://t.co/JaILm5tl9c

Manhattan, NY
-74.0007613 40.7207559
  -  -  -  -  -  -  -  -  -  -  -  -  -  -

First stop on for my birthday. (@ Buena Vista Cafe - @thebuenavista in San Francisco, CA) https://t.co/dXURT3ExFX

San Francisco, CA
-122.420736 37.80653
```

Στο πιο πάνω παράδειγμα χρήσης της παραμέτρου locations φαίνεται η συλλογή δεδομένων από τις χώρες San Francisco και New York. Το πρώτο tweet βρίσκεται μέσα στα πλαίσια οριοθέτησης της New York, ενώ το δεύτερο tweet βρίσκεται εντός των πλαισίων οριοθέτησης του San Francisco.

Οι παράμετροι follow, track και locations συνδυάζονται με ένα χειριστή OR. Δηλαδή αν έχουμε τις παραμέτρους track=haloumi και location= -15.64, 33.86, 28.29, 59.61, το API επιστρέφει tweets που ταιριάζουν με “haloumi” ή βρίσκονται σε τοποθεσία εντός των -15.64, 33.86, 28.29, 59.61 συντεταγμένων.

2.4 Εργαλεία που χρησιμοποιήθηκαν

2.4.1 MongoDB

Το MongoDB είναι ένα πρόγραμμα μη-σχεσιακής βάσης δεδομένων σε μορφή πλατφόρμας. Έχει δυναμικό χαρακτήρα για μη-δομημένα δεδομένα και έτσι τα δεδομένα μπορούν να αποθηκεύονται με πολλούς τρόπους. Χρησιμοποιεί έγγραφα τύπου JSON και υποστηρίζει αναζητήσεις εύρους, κανονικής έκφρασης και γεωχωρική αναζήτηση.

2.4.2 Γλώσσα προγραμματισμού Python

Η Python είναι αντικειμενοστραφής, υψηλού-επιπέδου γλώσσα με ενσωματωμένες δομές δεδομένων που εύκολα μπορεί να την μάθει κάποιος. Η Python είναι ένα πολύτιμο εργαλείο ενός αναλυτή δεδομένων, αφού είναι ειδικά σχεδιασμένη για εκτέλεση επαναλαμβανόμενων εργασιών και επεξεργασίας μεγάλου όγκου δεδομένων. Έχει πολυάριθμες βιβλιοθήκες που βοηθούν στην διεκπεραίωση διαφόρων εργασιών στην ανάλυση δεδομένων.

Στη βάση δεδομένων μας έχουμε μεγάλο αριθμό tweets, τα οποία πρέπει να διαβαστούν για να γίνει περαιτέρω επεξεργασία σε κάποια πεδία των tweets. Η εκτέλεση εργασιών σε τέτοια ογκώδη δεδομένα μπορεί να είναι χρονοβόρα. Ως εκ τούτου, η Python είναι κατάλληλη διότι προσφέρει βιβλιοθήκες που μπορούν να επεξεργαστούν μεγάλο όγκο δεδομένων σε σύντομο χρονικό διάστημα με την χρήση παράλληλης επεξεργασίας όπως **NumPy** [4] και **Pandas** [5].

Το NumPy ή αλλιώς “Numerical Python” είναι μια βιβλιοθήκη ανοικτού κώδικα, η οποία παρέχει γρήγορους μαθηματικούς υπολογισμούς σε πίνακες. Παρέχει τις απαραίτητες πολυδιάστατες λειτουργίες της υπολογιστικής σχεδιασμένες για υψηλού επιπέδου μαθηματικές λειτουργίες. Η βιβλιοθήκη Pandas είναι παρόμοια με το NumPy και χρησιμοποιείται ευρέως στην επιστήμη των δεδομένων. Παρέχει δομές υψηλής ανάλυσης, εύκολες στην χρήση και εργαλεία ανάλυσης δεδομένων. Σε αντίθεση με την βιβλιοθήκη NumPy, η οποία παρέχει αντικείμενα για πολυδιάστατους πίνακες, το Pandas παρέχει το αντικείμενο DataFrame που είναι δισδιάστατος πίνακας. Το DataFrame είναι σαν ένα υπολογιστικό φύλλο με ονόματα στηλών και ετικέτες γραμμών. Ως εκ τούτου με τα DataFrame, το Pandas είναι ικανό να παρέχει πολλές πρόσθετες λειτουργίες για την

επεξεργασία μεγάλων όγκων δεδομένα όπως δημιουργία πινάκων, υπολογισμούς νέων στηλών βάση άλλων στηλών και παράσταση γραφημάτων.

Επίσης, η Python προσφέρει βιβλιοθήκες που βοηθούν στην εξαγωγή δεδομένων από το διαδίκτυο. Πιο συγκεκριμένα η βιβλιοθήκη **tweepy** παρέχει πρόσβαση στις μεθόδους του Twitter Streaming API, όπου η κάθε μέθοδος μπορεί να δεχτεί κάποιες παραμέτρους και να επιστρέψει μια σχετική απάντηση. Πιο συγκεκριμένα, χειρίζεται την πιστοποίηση ταυτότητας και υποστηρίζει τη παρακολούθηση των tweets.

Επιπρόσθετα, η Python προσφέρει την δυνατότητα εικονογραφικής απεικόνισης των δεδομένων. Βλέποντας τόσα πολλά δεδομένα είναι δύσκολο να βγάλουμε κάποιο συμπέρασμα. Ο μόνος τρόπος για να καταλάβουμε καλύτερα τα δεδομένα μας είναι με τη μορφή αριθμών όπως πίνακες, ιστογράμματα και άλλες γραφικές παραστάσεις. Οι βιβλιοθήκες που χρησιμοποιούνται στην παρόν εργασία είναι το **Matplotlib** και το **Seaborn**.

2.4.3 NLTK

Το NLTK – Natural Language Toolkit, είναι ένα πακέτο της Python που είναι υπεύθυνο για διάφορες λειτουργίες επεξεργασίας της φυσικής γλώσσας. Η χρήση του NLTK είναι δωρεάν και εύκολη με μεγάλη κοινότητα που το εμπλουτίζει και υποστηρίζει. Παρέχει συνηθισμένους αλγορίθμους όπως διάσπαση κειμένου, ανάλυση συναισθήματος και κατηγοριοποίηση θέματος. Έτσι το NLTK συμβάλει σημαντικά στην επεξεργασία, στην ανάλυση και στη κατανόηση του γραπτού κειμένου.

Το NLTK προσφέρει το εργαλείο VADER (Valence Aware Dictionary and sEntiment Reasoner) για ανάλυση συναισθήματος. Το VADER βασίζεται σε «λεξικά του συναισθήματος» και σε κανόνες για την ανάλυση συναισθήματος και είναι ειδικά προσαρμοσμένο στα συναισθήματα που εκφράζονται στα κοινωνικά μέσα. Έχει βρεθεί αρκετά επιτυχημένο όταν χρησιμοποιείται για κείμενα κοινωνικών μέσων διότι το VADER όχι μόνο επιστρέφει αν το κείμενο είναι θετικό και αρνητικό, αλλά επίσης εκτιμάει και το ποσοστό της θετικότητας και αρνητικότητας του κειμένου.

Κεφάλαιο 3 - Περιγραφή Μεθοδολογίας

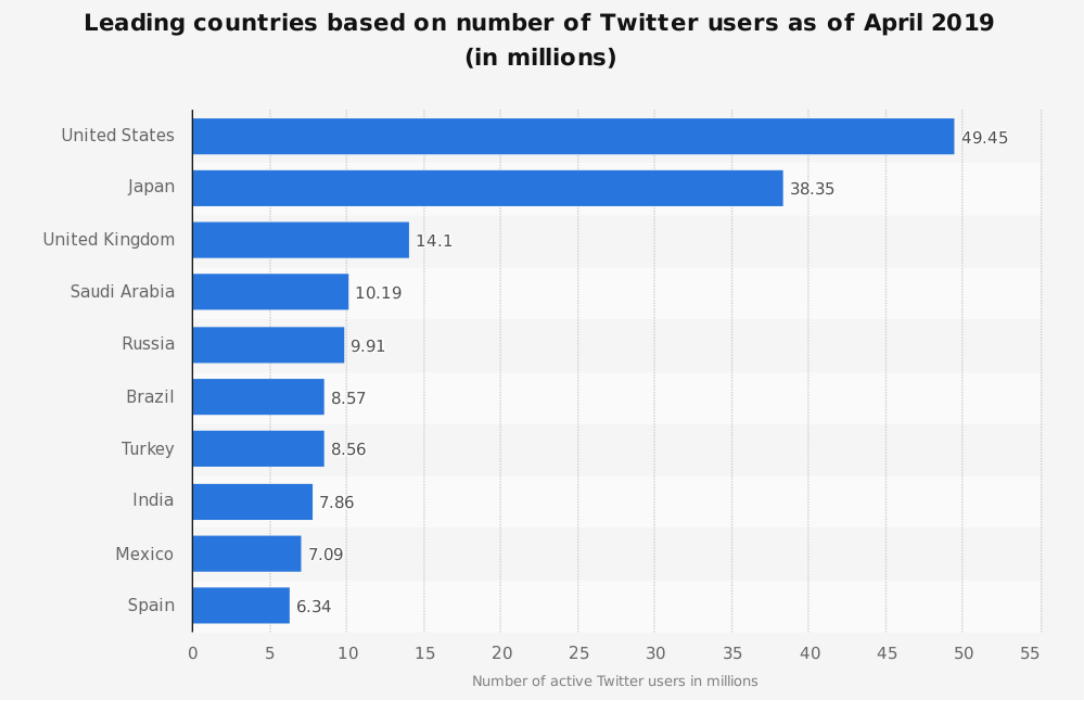
3.1Συλλογή Δεδομένων

Ένας από τους στόχους που τέθηκαν είναι η συλλογή όσο το δυνατόν περισσότερων tweets που σχετίζονται με αεροδρόμια και αεροπορικά ταξίδια από επιβάτες και προσωπικό. Οπότε το κύριο ζητούμενο στη συλλογή δεδομένων είναι η εύρεση των κατάλληλων κανόνων με τη χρήση των κατάλληλων λέξεων-κλειδιά και των γεωχωρικών πλαισίων οριοθέτησης.

3.1.1 Εντοπισμός δεδομένων για αναζήτηση

Γνωρίζουμε ότι διάφορες αεροπορικές εταιρίες και αεροδρόμια έχουν δημιουργήσει λογαριασμό στο Twitter ώστε να ενημερώνουν για διάφορα γεγονότα και προσφορές ή και να εξυπηρετούν πελάτες. Κάποιος πελάτης μπορεί να αναφερθεί σε μια εταιρία χρησιμοποιώντας το mention - @ και στην συνέχεια το όνομα της εταιρίας που θέλει να αναφερθεί. Ως εκ τούτου θα ήταν καλή πρακτική να μαζευτούν τα ονόματα των αεροπορικών εταιριών και αεροδρομίων όπως αυτά είναι στο Twitter.

Σε πρώτη φάση επιλέχθηκαν 2 αεροδρόμια. Εντόπισα πρώτα τις χώρες που έχουν τους περισσότερους χρήστες στο Twitter [3]. Από αυτές τις χώρες πήρα τις δύο με τους περισσότερους χρήστες και ομιλείται η αγγλική γλώσσα.



Στην περίπτωση μας είναι η Ηνωμένες Πολιτείες και το Ηνωμένο Βασίλειο. Από αυτές τις δύο χώρες έψαξα στο Wikipedia, μια πολύγλωσση ηλεκτρονική εγκυκλοπαίδεια, τα τρία πιο πολυσύχναστα αεροδρόμια της κάθε χώρας. Στη συνέχεια επέλεξα τα τρία πιο πολυσύχναστα αεροδρόμια της Ολλανδίας και τα τρία πιο πολυσύχναστα αεροδρόμια της Γερμανίας. Για το κάθε αεροδρόμιο καταγράφηκαν χειροκίνητα κάποιες βασικές πληροφορίες όπως όνομα αεροδρομίου, χώρα και πόλη που βρίσκεται, γεωγραφικές συντεταγμένες, το γεωχωρικό πλαίσιο οριοθέτησης, όνομα του λογαριασμού στο Twitter και τα hashtag που χρησιμοποιούνται για το συγκεκριμένο αεροδρόμιο.

Το επίσημο όνομα των αεροδρόμιων, η χώρα, η πόλη και οι γεωγραφικές συντεταγμένες πάρθηκαν από τη Wikipedia. Το γεωχωρικό πλαίσιο οριοθέτησης εντοπίστηκε με το μάτι βάση των συντεταγμένων που συλλέξαμε, με την βοήθεια του Online εργαλείου BoundingBox. Το όνομα λογαριασμού του κάθε αεροδρομίου στο Twitter βρέθηκε από την επίσημη ιστοσελίδα τους στο πεδίο επικοινωνία όπου συνήθως αναγράφετε. Τα hashtag που καταγράφηκαν είναι τα hashtag που παρουσιάστηκαν πρώτα στην αναζήτηση στην μηχανή αναζήτησης της Google, τα οποία πολλές φορές σύμπιπταν να είναι ίδια με το όνομα λογαριασμού του κάθε αεροδρομίου ενώ άλλες φορές ήταν συντομογραφία του ονόματος του αεροδρομίου.

Όλα αυτά τα έγγραφα αποθηκεύονται στην βάση δεδομένων MongoDB σε μορφή JSON στην συλλογή με όνομα airports. Τα ονόματα των αεροδρομίων που χρησιμοποιήθηκαν είναι τα εξής:

London Heathrow Airport, London Gatwick Airport, Edinburgh Airport, Hamburg Airport, Munich International Airport, Frankfurt am Main International Airport, Rotterdam The Hague Airport, Amsterdam Airport Schiphol, Eindhoven Airport, John F. Kennedy International Airport, Hartsfield–Jackson Atlanta International Airport, Los Angeles International Airport.

Αντιστοίχως επιλέχθηκαν 12 αερογραμμές οι οποίες έχουν στάση στα αεροδρόμια που έχουμε στην βάση δεδομένων. Καταγράφηκαν χειροκίνητα κάποιες βασικές πληροφορίες όπως όνομα αερογραμμής, όνομα του λογαριασμού στο Twitter, τα hashtag που χρησιμοποιούνται για τη συγκεκριμένη αερογραμμή, αριθμό χρηστών που ακολουθά – following η αερογραμμή και των αριθμό χρηστών που ακολουθούν – followers την αερογραμμή.

Το επίσημο όνομα της κάθε αερογραμμής αντιγράφηκε από την ηλεκτρονική εγκυκλοπαίδεια Wikipedia. Το όνομα λογαριασμού στο Twitter βρέθηκε από την επίσημη ιστοσελίδα της κάθε αερογραμμής. Ο αριθμός χρηστών που ακολουθά και ο αριθμός χρηστών που ακολουθούν την κάθε αερογραμμή αντιγράφηκε από το επίσημο λογαριασμό της κάθε αερογραμμής. Τα hashtag καταγράφηκαν με παρόμοιο τρόπο με την καταγραφή από τα αεροδρόμια, δηλαδή μέσω των πρώτων hashtag που παρουσιάστηκαν στη μηχανή αναζήτησης της Google.

Τα έγγραφα αυτά αποθηκεύονται στην συλλογή airlines. Τα ονόματα των αερογραμμών που χρησιμοποιήθηκαν είναι τα εξής:

EasyJet, British Airways, Virgin Atlantic, American Airlines, Delta, United Airlines, Southwest Airlines, Flybe, Jet2.com, Ryanair, Emirates Airline, Hawaiian Airlines

Πέρα από την καταγραφή αεροδρομίων και αερογραμμών, έγινε και αναζήτηση των πιο δημοφιλή hashtag που είναι σχετικά με αεροπορικό ταξίδι και καταγράφηκαν σε ένα αρχείο. Τα hashtag που καταγράφηκαν είναι τα εξής:

#airtravel #travel #airplane #aviation #boeing #flight #airport #aircraft #flying #clouds #avgeek #takeoff #plane #instatravel #airbus #cats #pilot #travelphotography #fly #planes #aviationgeek #sky #instaplane #travelling #aviationphotography #instaaviation #of #aeroplane #internationaltravel #bhfyp

Επειδή όμως κάποια από τα hashtag μπορούν να χρησιμοποιηθούν και σε άλλες περιπτώσεις που δεν σχετίζονται με αεροπορικό ταξίδι, προσθέτοντας πολύ θόρυβο στα δεδομένα μας, για τους σκοπούς της παρούσας εργασίας. Κράτησα μόνο τα hashtag που θεωρώ ότι σχετίζονται περισσότερο με αεροπορικό ταξίδι. Τα hashtag που παρέμειναν είναι τα εξής:

#airtravel #airplane #aviation #boeing #flight #airport #aircraft #flying #clouds #avgeek #plane #airbus #pilot #fly #planes #aviationgeek #sky #instaplane #aviationphotography #instaaviation #aeroplane

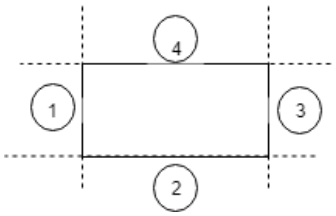
Μέχρι στιγμής παρουσιάστηκε η διαδικασία συλλογής κάποιων δεδομένων που θα βοηθήσουν στην δημιουργία κανόνων για αναζήτηση tweets που είναι σχετικά με αεροπορικό ταξίδι. Ο λόγος που επιλέχθηκαν λίγα αεροδρόμια και λίγες αερογραμμές είναι επειδή είναι πιο εύκολο να διαχειριστούν και επειδή η καταγραφή των πληροφοριών τους είναι χρονοβόρο. Εύκολα μπορούν να χρησιμοποιηθούν περισσότερα αεροδρόμια και αερογραμμές αν καταγραφούν οι πληροφορίες τους στην βάση δεδομένων. Στη συνέχεια θα παρουσιαστούν οι μέθοδοι συλλογής tweets με την βοήθεια των πληροφοριών που έχουν μαζευτεί.

3.1.2 Δημιουργία κανόνων για αναζήτηση

Όπως ανέφερα στο προηγούμενο κεφάλαιο θα χρησιμοποιηθεί το statuses/filter API για τη συλλογή των tweets που θα ταιριάζουν στους κανόνες που εισήγαγα. Οι κανόνες που εισήγαγα χωρίζονται σε 4 μεθόδους συλλογής tweets.

Η πρώτη μέθοδος είναι η συλλογή των tweets που δημοσιευτήκαν μέσα στα γεωχωρικά πλαίσια οριοθέτησης των αεροδρομίων που υπάρχουν στη βάση δεδομένων. Αυτό επιτυγχάνεται χρησιμοποιώντας την παράμετρο locations. Η παράμετρος location επιτρέπει να ορίσουμε μια γεωγραφική περιοχή τεσσάρων πλευρών και να συλλέξουμε τα tweets που εμπίπτουν σε αυτή την περιοχή. Μια γεωγραφική περιοχή πρέπει να οριστεί με την εξής σειρά:

[δυτικό γεωγραφικό μήκος, νότιο γεωγραφικό πλάτος, ανατολικό γεωγραφικό μήκος, βόρειο γεωγραφικό πλάτος]



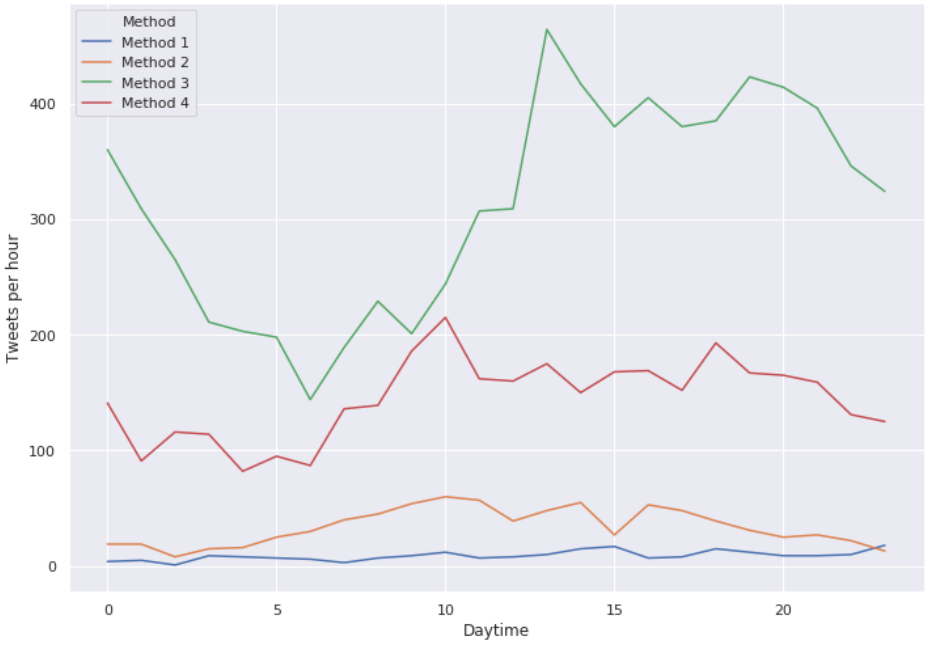
Για εισαγωγή γεωγραφικής περιοχής περισσότερο από ένα αεροδρόμιο πρέπει να συνεχιστεί η εισαγωγή στην ίδια λίστα οι 4 τιμές για το επόμενο αεροδρόμιο με την σειρά που έδωσα πιο πάνω.

Η δεύτερη μέθοδος είναι η συλλογή των tweets με βάση τις λέξεις-κλειδιά, δηλαδή hashtags και mention, που έχουμε αποθηκευμένες στη βάση δεδομένων για κάθε αεροδρόμιο. Με την χρήση της παραμέτρου track συλλέγουμε τα tweets που περιέχουν κάποια από αυτές τις λέξεις στο πεδίο hashtag ή mention στα μεταδεδομένα τους.

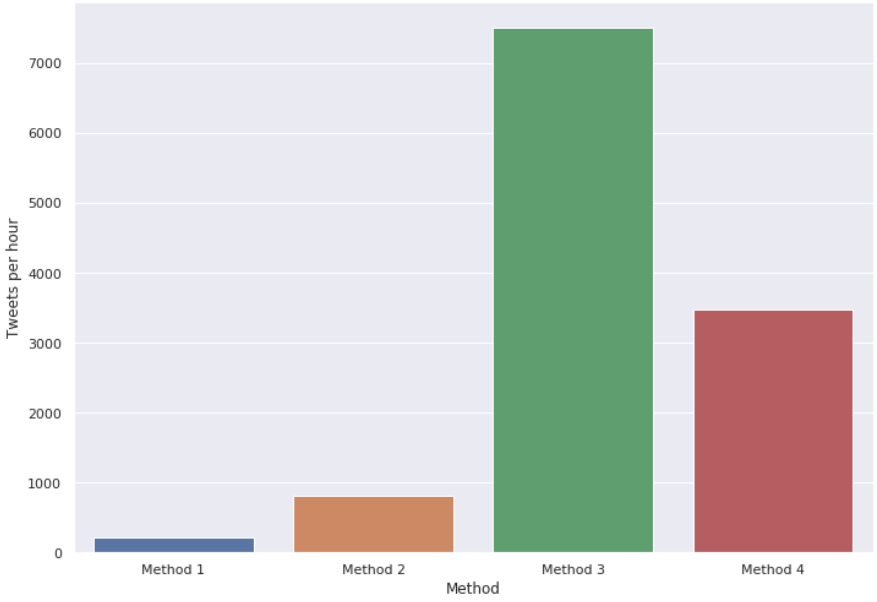
Η τρίτη μέθοδος είναι παρόμοια με την δεύτερη αλλά αντί για αεροδρόμια, υλοποιείται για τις αερογραμμές. Δηλαδή με την χρήση της παραμέτρου track ανιχνεύουμε τα tweets που περιέχουν στα μεταδεδομένα τους τις λέξεις των hashtags και mention των αερογραμμών που έχουμε αποθηκευμένες στη βάση δεδομένων.

Με τη τέταρτη μέθοδο γίνεται πάλι χρήση της παραμέτρου track με λέξεις-κλειδιά τα πιο δημοφιλή hashtags που σχετίζονται με αεροπορικό ταξίδι.

Έχοντας υπόψη τις τέσσερις μεθόδους συλλογής tweet κατέληξα σε 2 κανόνες συλλογής δεδομένων. Ο πρώτος κανόνας χρησιμοποιεί την πρώτη μέθοδο συλλογής tweets με τη χρήση της παραμέτρου location και ο δεύτερος κανόνας αποτελείται από τη δεύτερη, τρίτη και τέταρτη μέθοδο συλλογής tweet με τη χρήση της παραμέτρου track. Στις ακόλουθες γραφικές παραστάσεις φαίνεται η συχνότητα συλλογής των tweets καθόλη τη διάρκεια κάποιας ημέρας και το σύνολο των tweets που σύλλεξε η κάθε μέθοδος ξεχωριστά.



Εικόνα 1: Η συχνότητα συλλογής tweets καθόλη τη διάρκεια της 5ης Μαρτίου



Εικόνα 2: Το σύνολο των tweets που συλλέχθηκαν στις 5 του Μάρτη

Ο αριθμός των tweets που συλλέχθηκαν στης 5 του Μάρτη για τις τέσσερις μεθόδους είναι 216, 815, 7503 και 3478.

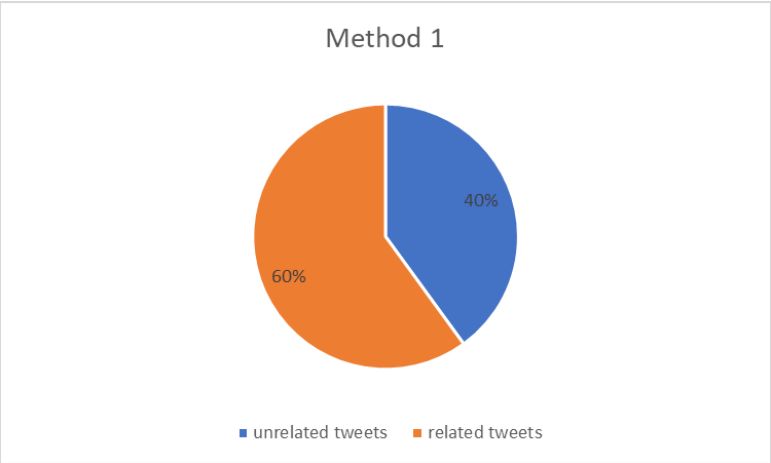
3.1.3 Εγκυρότητα της συλλογής των δεδομένων

Η εγκυρότητα είναι σημαντική διότι καθορίζει ποιες μεθόδους συλλογής δεδομένων πρέπει να χρησιμοποιηθούν για να διασφαλίσουμε την συλλογή όσο το δυνατόν περισσότερων tweets που σχετίζονται με αεροπορικά ταξίδια και να αποφύγουμε τη συλλογή ανεπιθύμητων tweets. Για την εξάλειψη των πηγών που οδηγούν σε σφάλματα θα κάνουμε κριτική στο κείμενο των tweets από τις μεθόδους που χρησιμοποιήθηκαν για την συλλογή των tweets. Για την εξάλειψη των πηγών που οδηγούν σε σφάλματα άτομα της ερευνητικής μας ομάδας θα κάνουν κριτική στο κείμενο των tweets που βρέθηκαν με τις τέσσερις μεθόδους συλλογής δεδομένων. Συγκεκριμένα, βάση του κειμένου, προσδιορίσουμε εκείνα τα tweets τα οποία σχετίζονται με αεροπορικά ταξίδια έτσι ώστε να υπολογίσουμε ακολούθως το ποσοστό ακριβείας της κάθε μεθόδου. Ένα παράδειγμα αξιολόγησης είναι το εξής:

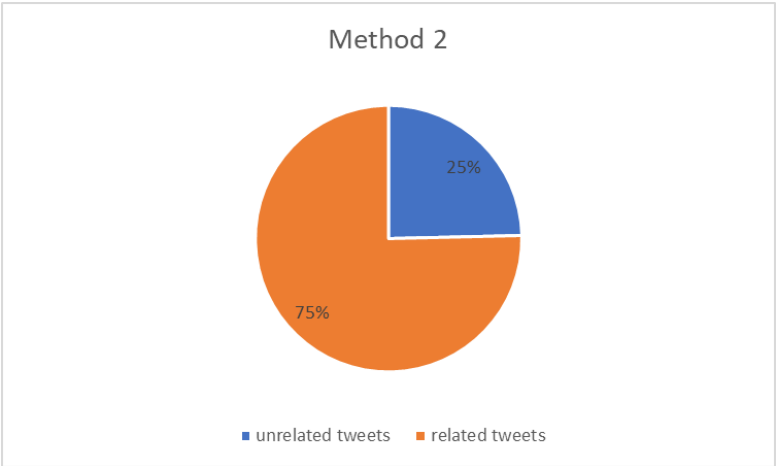
full_text	id	Related to Air Travel
@jonasbrothers @AmericanAir @Mastercard We got ours!	5cd19f99ead3c32b9646a540	FALSE
I appreciate @united for holding the plane for passengers with a tight connection. Apparently people are flying to Maui from Germany. They were even greeted with water. #goodservice	5cd0d75bead3c32b96468ebc	TRUE

Λόγω όμως του μεγάλου αριθμού tweets που συλλέγονται με τις διάφορες μεθόδους, η αξιολόγηση θα βασιστεί σε δείγμα των tweets που συλλέχθηκαν με την κάθε μέθοδο κατά τη διάρκεια μιας ημέρας. Το δείγμα των tweets θα παρθεί τυχαία ενώ το μέγεθος του δείγματος για την κάθε μέθοδο θα είναι διαφορετικό και θα εξαρτάται από το συνολικό αριθμό των tweets που συλλέχθηκε με την αντίστοιχη μέθοδο την συγκεκριμένη εκείνη μέρα. Ο καθορισμός του μεγέθους του δείγματος (sample size) θα γίνει με την βοήθεια του Online εργαλείου SurveySystem [7] με επίπεδο στατιστικής σημαντικότητας των ελέγχων α (confidence level) στο 95% και διαστήματα εμπιστοσύνης (confidence interval) στα 5%.

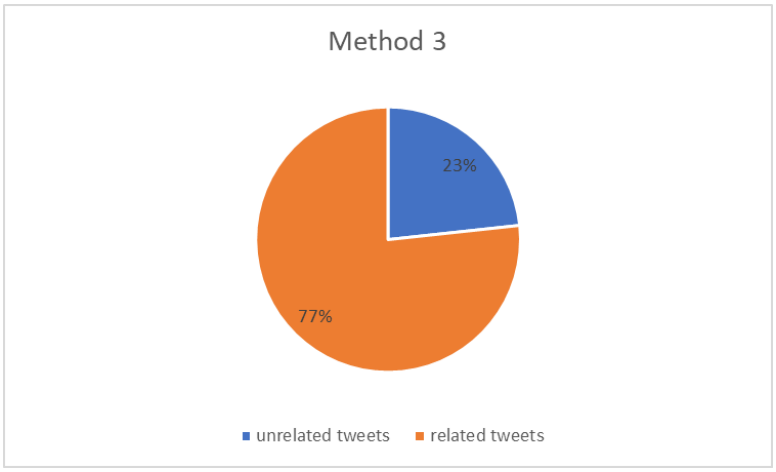
Η πρώτη μέθοδος είναι η συλλογή των tweets που δημοσιευτήκαν μέσα στα γεωχωρικά πλαίσια οριοθέτησης των αεροδρομίων. Με αυτή την μέθοδο έχουν συλλεχθεί 244 tweets στις 5 του Μάη του 2019. Το δείγμα που θα αξιολογηθεί, βάση του υπολογισμού του μεγέθος τους δείγματος, θα αποτελείται από 150 tweets. Η καταμέτρηση έδειξε ότι 60 tweets είναι άσχετα με το αεροπορικό ταξίδι ενώ τα υπόλοιπα 90 tweets είναι σχετικά. Έτσι καταλήγουμε στο ποσοστό των 60% σχετικών tweets. Αυτό το αποτέλεσμα που υπολογίστηκε από δείγμα του πραγματικού συνόλου μπορεί να ερμηνευθεί ως εξής: αν εξετάζαμε όλα τα 244 tweets που βρέθηκαν με αυτή την μέθοδο θα βρίσκαμε, κατά 95% πιθανότητα, ότι ένα υποσύνολο μεταξύ 55% (60-5) και 65% (60+5) των tweets θα ήταν σχετικό με αεροπορικά ταξίδια.



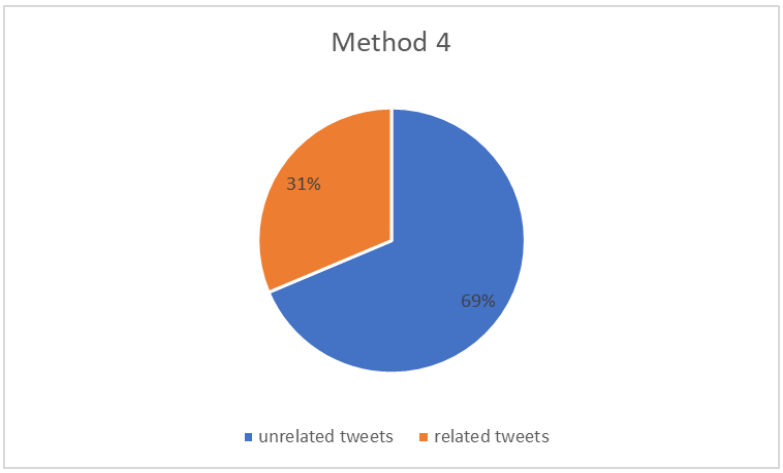
Η δεύτερη μέθοδος, που είναι η συλλογή των tweets βάση των hashtag και mention των αεροδρομίων, έχει συλλέξει 879 την ίδια μέρα. Έτσι το μέγεθος του δείγματος που θα αξιολογηθεί είναι 268 tweets. Με αυτή την μέθοδο βρέθηκαν 66 άσχετα tweets και 202 σχετικά, έχοντας έτσι 75.09% επιτυχία σχετικών tweets.



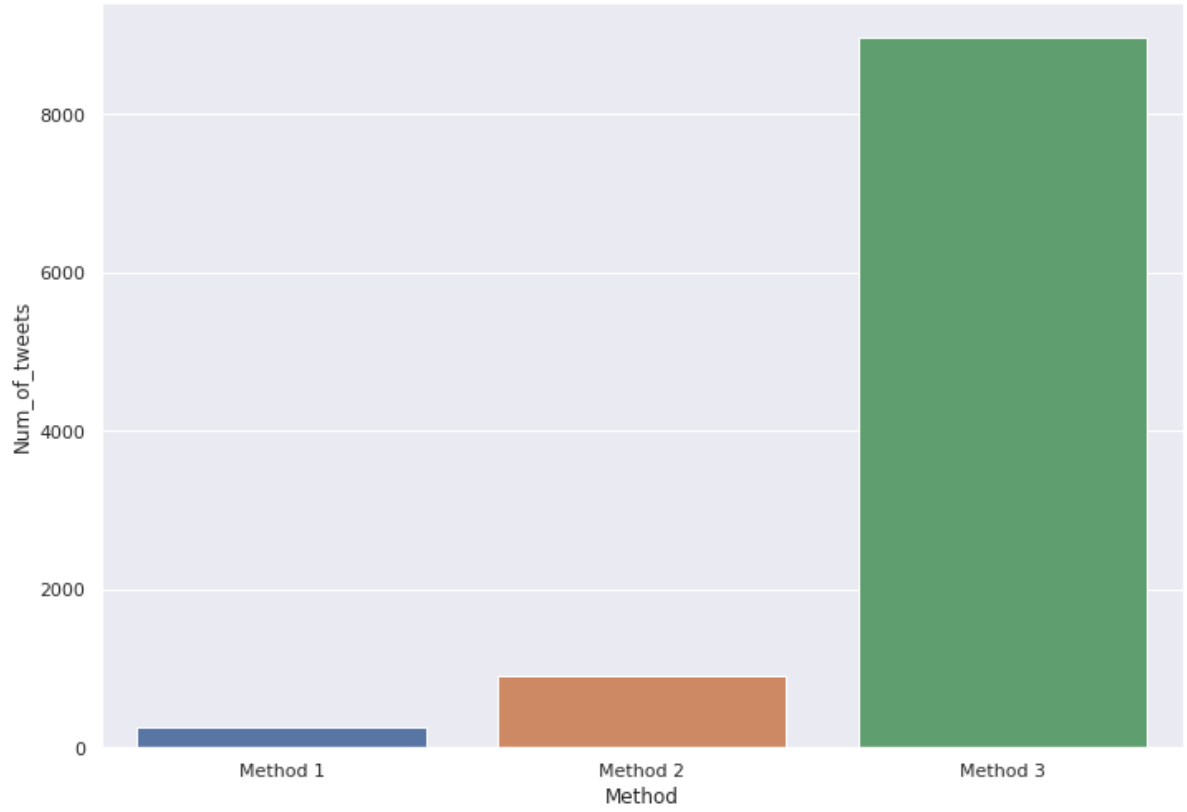
Η τρίτη μέθοδος, που είναι η συλλογή των tweets βάση των hashtag και mention των αερογραμμών, έχει συλλέξει 8343 tweets αυτή την μέρα. Το μέγεθος του δείγματος που θα αξιολογηθεί είναι 367 tweets. Με τη τρίτη μέθοδο παρατηρήθηκαν 85 άσχετα tweets και 282 σχετικά, έχοντας 76.51% επιτυχία στα σχετικά tweets.



Τελευταία μέθοδος, που είναι η συλλογή των tweets βάση των δημοφιλές hashtag που σχετίζονται με αεροπορικό ταξίδι, έχει συλλέξει 3017 tweets την συγκεκριμένη μέρα. Το μέγεθος του δείγματος είναι 341 του οποίου τα 234 είναι tweets άσχετα με αεροπορικό ταξίδι και 107 σχετικά tweets. Έτσι έχουμε επιτυχία μόνο 31.29% σχετικών tweets.



Σύμφωνα με τα πιο πάνω ποσοστά επιτυχίας σχετικών tweets, συμπεραίνουμε ότι η δεύτερη και η τρίτη μέθοδος έχουν τα καλύτερα ποσοστά επιτυχίας, ενώ στη τέταρτη μέθοδο μόνο το 31% των tweets που συλλέγει είναι σχετικά με αεροπορικό ταξίδι. Για τον λόγο ότι η τέταρτη μέθοδος προσθέτει πολύ θόρυβο στα δεδομένα μας και το γεγονός ότι χρησιμοποιεί πόρους από τις αιτήσεις των tweets από το Twitter Streaming API, θα χρειαστεί να την απορρίψουμε ως μέθοδο από το σύστημα.



Εικόνα 3: Μέσος όρος συλλογής tweets ανά ημέρα

Αφού καταλήξαμε ποιες μέθοδοι συλλογής δεδομένων θα χρησιμοποιηθούν βρέθηκε ο μέσος όρος συλλογής tweets ανά ημέρα εντός 40 ημερών. Ο μέσος όρος συλλογής tweets ανά ημέρα για τις τρεις μεθόδους είναι 248, 910 και 8965.

3.2 Καθαρισμός και Προετοιμασία των Δεδομένων

Σε αυτό το σημείο έχουμε συλλέξει και συνεχίζουμε να συλλέγουμε tweets με τους κανόνες που αναφέραμε προηγουμένως. Προτού όμως να κάνουμε ανάλυση των δεδομένων και να εξάγουμε κάποια πληροφορία χρειάζεται να κάνουμε επεξεργασία κειμένου των tweets και να το προετοιμάσουμε για την ανάλυσή του.

3.2.1 Επεξεργασία κειμένου

Υπάρχουν διάφορα βήματα για προ επεξεργασία κειμένου τα οποία θα εφαρμοστούν αναλόγως σε κάθε φάση ανάλυσης. Θα δοκιμαστούν διάφοροι συνδυασμοί προ επεξεργασίας κειμένου για κάθε ανάλυση και θα χρησιμοποιηθεί ο συνδυασμός που δημιουργεί τα καλύτερα αποτελέσματα στην κάθε ανάλυση.

Ένα είδος προ επεξεργασίας είναι να μετατρέψουμε όλους τους χαρακτήρες του κειμένου σε πεζούς για να υπάρχει μια συνεπή μορφή μεταξύ όλων των κειμένων. Έτσι διασφαλίζουμε την ασφάλεια για ανάλυση ιδίων λέξεων από διαφορετικά κείμενα να μην υποστούν ζητήματα ασυνέπειας μεταξύ των λέξεων λόγω κάποιου κεφαλαίου ή πεζού γράμματος.

Άλλο βήμα επεξεργασίας κειμένου είναι η αφαίρεση των συνδέσμων, των mentions και hashtags αν υπάρχουν στο κείμενο. Αφαιρούνται επειδή συνήθως επαναλαμβάνονται από πολλά tweets και δεν μας δίνουν χρήσιμη πληροφορία στην ανάλυση του συναισθήματος του tweet.

Επίσης ένα άλλο κομμάτι της προ επεξεργασία κειμένου είναι η αφαίρεση των ειδικών χαρακτήρων, σημεία στίξης και αριθμοί επειδή έχουν μεγάλη συχνότητα χρήσης τους και δεν δίνουν καμία πληροφορία. Οπότε αφαιρούνται για να μην επηρεάσουν την μετέπειτα ανάλυση.

Επίσης μπορούν να αφαιρεθούν οι stop words λέξεις, δηλαδή αφαιρούνται οι πιο συνηθισμένες λέξεις σε μια γλώσσα που στην περίπτωση μας είναι η αγγλική. Ένα δείγμα των stop words είναι το εξής:

```
In [8]: print(stop_words[:50])

['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be']
```

Αφαιρούνται επειδή παρουσιάζονται πολύ συχνά σε όλα τα tweets και συνήθως δεν δίνουν κάποια πληροφορία, οπότε δεν χρειάζεται να αναλυθούν.

Τέλος, στα tweets του κάθε αεροδρομίου ή αερογραμμής αφαιρείται από το κείμενο του tweet το όνομα της τρέχον εταιρίας αν υπάρχει. Αυτό γίνεται επειδή στην ανάλυση των πιο πολυχρησιμοποιημένων λέξεων αργότερα θα μας παρουσιαστούν πολλές φορές τα ονόματα δίνοντας την ίδια πληροφορία που ήδη γνωρίζουμε.

3.3 Ανάλυση Δεδομένων

Σε αυτό το στάδιο θα γίνει ανάλυση στα δεδομένα έτσι ώστε στη συνέχεια να εξάγουμε πληροφορία από τα δεδομένα που συλλέξαμε. Πιο συγκεκριμένα σε αυτό το κεφάλαιο θα γίνει η ανάλυση συναισθήματος του κειμένου των tweets που συλλέχθηκαν και η αξιολόγηση στις διάφορες παραμέτρους της προ επεξεργασίας του κειμένου που επηρεάζουν την κατηγοριοποίηση του κειμένου.

3.3.1 Ανάλυση συναισθήματος

Η ανάλυση συναισθήματος ή αλλιώς εξόρυξη γνώμης, είναι ένα χρήσιμο εργαλείο της επεξεργασίας φυσικής γλώσσας που μας βοηθά να εντοπίσουμε και να μελετήσουμε υποκειμενικές πληροφορίες. Το γεγονός ότι μπορούμε να συλλέξουμε μεγάλο όγκο tweets με συνδυασμό του εργαλείου ανάλυσης συναισθήματος, έχουμε την δυνατότητα να εξάγουμε χαρακτηριστικά αυτών των δεδομένων όπως την θετική ή αρνητική γνώμη των ανθρώπων γύρω από κάποιο θέμα.

Σε αυτή την εργασία η ανάλυση του συναισθήματος θα γίνει με την βοήθεια της βιβλιοθήκης VADER του πακέτου NLTK της Python. Με την χρήση της μεθόδου polarity_scores() παίρνουμε την τιμή της πολικότητας της δεδομένης πρότασης.

```
        sentiment_score = sia.polarity_scores(text)
        sentiment.append(sentiment_score['compound'])
        print(text,"\n")
        print(sentiment_score,"\n")
        print("  -  -  -  -  -  -  -  -  -  -")
```

```
Ready to fly! See you on the other side!   Manchester Airport

{'neg': 0.0, 'neu': 0.764, 'pos': 0.236, 'compound': 0.4738}
```

Μετρική συναισθήματος	**Σκορ**
Θετικό	0.236
Ουδέτερο	0.764
Αρνητικό	0.0
Σύνθεση	0.4738

Το θετικό, ουδέτερο και αρνητικό σκορ αντιπροσωπεύουν το ποσοστό του συναισθήματος του δεδομένου κειμένου που εμπίπτει σε αυτές τις κατηγορίες. Αυτό σημαίνει ότι το κείμενο του παραδείγματος της εικόνας εμπίπτει κατά 0.24% ως θετικό, 0.76% ως ουδέτερο και 0% ως αρνητικό. Επομένως όλες αυτές οι βαθμολογίες πρέπει να συνθέσουν την τελική τιμή του που θα χαρακτηρίζει το συναίσθημα του κειμένου. Έτσι η σύνθετη βαθμολογία είναι μια μέτρηση που υπολογίζεται αθροίζοντας τις βαθμολογίες σθένος κάθε λέξης του λεξικού και στη συνέχεια κανονικοποιείται μεταξύ την πιο ακραία αρνητική τιμή -1 και την πιο ακραία θετική τιμή 1. Αυτή η σύνθετη μέτρηση είναι χρήσιμη όταν θέλουμε να ορίσουμε μια δεδομένη πρόταση με ένα μονοδιάστατο μέτρο συναισθήματος.

3.3.2 Σύγκριση τεχνικών προ επεξεργασίας κειμένου για ανάλυση συναισθήματος

Για τη μέγιστη δυνατή απόδοση του εργαλείου VADER χρειάζεται να εφαρμοστεί η κατάλληλη προ επεξεργασία κειμένου. Για αυτό στα πλαίσια αυτής της διπλωματικής εργασίας θα περάσουμε από τη διαδικασία αξιολόγησης των διάφορων τεχνικών προ επεξεργασίας κειμένου και θα μελετήσουμε πώς αυτές οι τεχνικές επηρεάζουν τον υπολογισμό της σύνθετης συναισθηματικής μέτρησης του κειμένου.

Η σύγκριση των διάφορων τεχνικών προ επεξεργασίας κειμένου για συναισθηματική ανάλυση γίνεται με την χρήση του δείγματος των 149 tweets που συλλέχθηκαν βάση την γεωγραφική τους θέση. Η χρήση του συγκεκριμένου δείγματος γίνεται επειδή είναι σχετικά μικρό δείγμα και άρα θα παρατηρηθεί πιο εύκολα η επίδραση της προ επεξεργασίας κειμένου. Για κάθε τεχνική προ επεξεργασίας κειμένου αποθηκεύετε η σύνθετη μέτρηση συναισθήματος που υπολογίζεται για το κείμενο του κάθε tweet. Σε όλες τις μεθόδους προ επεξεργασίας κειμένου αφαιρούνται τα hashtags, τα mentions και οι σύνδεσμοι διότι δεν παρέχουν καμία πληροφορία στο συναίσθημα του κειμένου. Οι παράμετροι που θα ελέγξουμε είναι αν οι stop words, η κεφαλαιοποίηση (capitalization) των λέξεων και η χρήση ειδικών χαρακτήρων επηρεάζουν στον υπολογισμό του συναισθήματος του κειμένου.

Ο πρώτος έλεγχος της προ επεξεργασίας κειμένου είναι αν και πως η αφαίρεση των stop words επηρεάζουν στον υπολογισμό του συναισθήματος του κειμένου. Για να το αξιολογήσουμε χρειάζεται να χρησιμοποιήσουμε το εργαλείο VADER στο κείμενο των tweets που έχουμε, δύο φορές διατηρώντας τα χαρακτηριστικά του κειμένου σταθερά στα δύο περάσματα και να αλλάξουμε μόνο την χρήση των stop words. Το πρώτο πέρασμα θα είναι με τις stop words ως έχουν και το δεύτερο πέρασμα θα είναι με την αφαίρεση τους από το κείμενο. Έτσι το πρώτο πέρασμα γίνεται με παράμετρο το κείμενο του οποίου αφαιρούνται τα hashtags, τα mentions, οι σύνδεσμοι, οι ειδικοί χαρακτήρες και η μετατροπή όλων των γραμμάτων σε πεζά. Το δεύτερο πέρασμα γίνεται με παράμετρο το κείμενο του οποίου αφαιρούνται οι stop words, τα hashtags, τα mentions, οι σύνδεσμοι, οι ειδικοί χαρακτήρες και η μετατροπή όλων των γραμμάτων σε πεζά.

```
processing_df.tail()
```

	full_text	id	sentiment1	sentiment2
145	I'm at Los Angeles International Airport - @fl...	5cd0faecead3c34fc153a345	0.0000	0.0000
146	Ready to fly! See you on the other side! #hone...	5cd14bdbead3c34fc153a836	0.3612	0.3612
147	Japan Airlines 787 rotating out of RWY 27R JP✈...	5cd197d9ead3c34fc153af07	-0.1677	-0.2732
148	Cleared: Construction on #JohnFKennedyExpressw...	5cd1be63ead3c34fc153b4d8	0.1027	0.1027
149	Every time I walk past this, I giggle. Everyti...	5cd0f75dead3c34fc153a2ff	0.5023	0.4215

Στην εικόνα φαίνονται τα τελευταία 5 έγγραφα του DataFrame που στην πρώτη στήλη περιέχει τα κείμενα των tweets πριν την επεξεργασία τους, στην δεύτερη στήλη περιέχει τον

μοναδικό αριθμό των αντικειμένων tweets στην βάση δεδομένων, στην τρίτη στήλη έχει την τιμή της συναισθηματικής μέτρησης του κειμένου που περιείχε τις stop words και στην τελευταία στήλη έχει την τιμή της συναισθηματικής μέτρησης του κειμένου που δεν περιείχε τις stop words.

```
count = 0
for index, row in data.iterrows():
    if row['sentiment2'] != row['sentiment1']:
        count+=1
print("Different sentiment polarity on {} tweets.".format(count))
```

```
Different sentiment polarity on 21 tweets.
```

Γίνεται αξιολόγηση των δύο στηλών, sentiment1 και sentiment2, όπου συγκρίνεται η τιμή της συναισθηματικής μέτρησης ανάλογα με το κείμενο. Σύμφωνα με την πιο πάνω εικόνα φαίνεται ότι οι δύο τεχνικές επεξεργασίας κειμένου παίζει σημασία στον υπολογισμό της πολικότητας του συναισθήματος των κειμένων. Μετά από αναλυτική παρακολούθηση των κειμένων και των τιμών της συναισθηματικής μέτρησης για την κάθε τεχνική, καταλήγουμε πως η χρήση των stop words είναι σημαντική για τον υπολογισμό της τιμής της πολικότητας του συναισθήματος.

```
IST or PHILLY!?  sorry...but #alwayschooseadventurescolorado @ Edinburgh Airp
ort https://t.co/OMPsN02P6w
sentiment1: 0.0  sentiment2: -0.0772

I'm at Bus Stop N - Harlington Corner https://t.co/iTfa1FvzCK
sentiment1: -0.29600000000000004         sentiment2: -0.296

“Energy is neither created nor destroyed” expands to deeper and deeper levels
of understanding now.
.
I stand in virtually the same place, 365 days later, minus 82 minutes. This t
ime I… https://t.co/WBwN13uWdV
sentiment1: -0.2825      sentiment2: 0.4003

There’s no place like home🙏😛
#sunrise🤲 #untilnexttimecali❤️ @ Lax International Airport Los Angeles Cali
fornia https://t.co/0HiDfmZwP2
sentiment1: 0.0772       sentiment2: 0.3612
```

Ο επόμενος έλεγχος της προ επεξεργασίας κειμένου είναι κατά πόσο επηρεάζει η χρήση των ειδικών χαρακτήρων στον υπολογισμό του συναισθήματος του κειμένου. Οπότε για την πραγματοποίηση αυτού του ελέγχου θα εξεταστούν οι τιμές για την συναισθηματική κατηγοριοποίηση των κειμένων των tweets όπου θα αφαιρούνται οι ειδικοί χαρακτήρες και

των κειμένων των tweets που θα παραμείνουν οι ειδικοί χαρακτήρες. Έτσι το πρώτο πέρασμα για τον υπολογισμό της τιμής θα περνιέται ως παράμετρο το κείμενο του οποίου αφαιρούνται τα stop words, τα hashtags, τα mentions, οι σύνδεσμοι, οι ειδικοί χαρακτήρες και η μετατροπή όλων των γραμμάτων σε πεζά. Το δεύτερο πέρασμα γίνεται με παράμετρο το κείμενο του οποίου αφαιρούνται οι stop words, τα hashtags, τα mentions, οι σύνδεσμοι και η μετατροπή όλων των γραμμάτων σε πεζά, αλλά οι ειδικοί χαρακτήρες θα παραμείνουν.

	full_text	id	sentiment1	sentiment3
145	I'm at Los Angeles International Airport - @fl...	5cd0faecead3c34fc153a345	0.0000	0.0000
146	Ready to fly! See you on the other side! #hone...	5cd14bdbead3c34fc153a836	0.3612	0.4738
147	Japan Airlines 787 rotating out of RWY 27R JP ✈...	5cd197d9ead3c34fc153af07	-0.1677	-0.1677
148	Cleared: Construction on #JohnFKennedyExpressw...	5cd1be63ead3c34fc153b4d8	0.1027	0.1027
149	Every time I walk past this, I giggle. Everyti...	5cd0f75dead3c34fc153a2ff	0.5023	0.5023

Στην εικόνα φαίνονται τα τελευταία 5 έγγραφα του DataFrame με τις ίδιες στήλες όπως και στην προηγούμενη αξιολόγηση με την διαφορά στην τρίτη στήλη έχει την τιμή της συναισθηματικής μέτρησης του κειμένου που δεν περιείχε τους ειδικούς χαρακτήρες και στην τελευταία στήλη έχει την τιμή της συναισθηματικής μέτρησης του κειμένου που περιείχε τους ειδικούς χαρακτήρες. Όπως παρουσιάζετε από αυτά τα 5 έγγραφα φαίνεται πώς οι ειδικοί χαρακτήρες παίζουν ρόλο στην αξιολόγηση του κειμένου.

```
count = 0
for index, row in data.iterrows():
    if row['sentiment1'] != row['sentiment3']:
        count+=1
print("Different sentiment polarity on {} tweets".format(count))
```

```
Different sentiment polarity on 27 tweets
```

Πιο συγκεκριμένα υπάρχουν 27 διαφορετικά έγγραφα που αξιολογήθηκαν διαφορετικά στην κάθε τεχνική προ επεξεργασίας λόγω των ειδικών χαρακτήρων. Μετά από αναλυτική παρακολούθηση των κειμένων και των τιμών της συναισθηματικής μέτρησης για την κάθε τεχνική, καταλήγουμε πως η χρήση των ειδικών χαρακτήρων είναι σημαντική για τον υπολογισμό της τιμής της πολικότητας του συναισθήματος.

```
Toronto here I come!
Coffeeshops better be open!!!
CA
#coffee #travel #food #toronto #coffeeattendant @ Schiphol, Noord-Holland, Netherlands https://t.co/f0hWpOfsrx
sentiment1: 0.4404       sentiment3: 0.6209

Barely made my flight due to horrible traffic, (left my apartment at 6:15pm arrrived at LAX at 10:20)thank goodness the TSA w
as super quick today! But Talk about last minute lol..... but… https://t.co/Iy3RT16Kmy
sentiment1: 0.7684       sentiment3: 0.3595

Lovey to share an amazing day and night with novakski74 and wardie75 for their beautiful wedding! Thanks also to neilpodolans
ki and @Queentj for the photobombing and laughs!!… https://t.co/p677af6Ifd
sentiment1: 0.9432       sentiment3: 0.9284

Congratulations to our very special (and incredibly talented) Estee Lauder Counter Manager: Miguelina. She is celebrating her
Longevity Awards for her 18 years at #iSDutyFree… https://t.co/yTTaf3Qr1W
sentiment1: 0.953        sentiment3: 0.9273
```

Τελευταίος παράγοντας που χρειάζεται να ελέγξουμε είναι αν τα πεζά ή κεφαλαία γράμματα παίζουν ρόλο στον υπολογισμό της τιμής για το συναίσθημα του κειμένου. Οπότε για την πραγματοποίηση αυτού του ελέγχου θα εξεταστούν οι τιμές για την συναισθηματική κατηγοριοποίηση των κειμένων των tweets όπου θα μετατρέπονται πεζά όλα τα γράμματα του κειμένου και των κειμένων των tweets που θα παραμείνει η κεφαλαιοποίηση ως έχει.

Έτσι το πρώτο πέρασμα για τον υπολογισμό της τιμής θα περνιέται ως παράμετρο το κείμενο του οποίου αφαιρούνται τα stop words, τα hashtags, τα mentions και οι σύνδεσμοι, θα παραμένουν οι ειδικοί χαρακτήρες και η μετατροπή όλων των γραμμάτων θα γίνονται σε πεζά. Το δεύτερο πέρασμα γίνεται με παράμετρο το κείμενο του οποίου αφαιρούνται οι stop words, τα hashtags, τα mentions και οι σύνδεσμοι, θα παραμένουν οι ειδικοί χαρακτήρες και τα γράμματα δεν θα μετατρέπονται σε πεζά αλλά θα παραμένουν ως έχει.

```
processing_df.tail()
```

	full_text	id	sentiment4	sentiment5
145	I'm at Los Angeles International Airport - @fl...	5cd0faecead3c34fc153a345	0.0000	0.0000
146	Ready to fly! See you on the other side! #hone...	5cd14bdbead3c34fc153a836	0.4738	0.4738
147	Japan Airlines 787 rotating out of RWY 27R JP✈...	5cd197d9ead3c34fc153af07	-0.1677	-0.1677
148	Cleared: Construction on #JohnFKennedyExpressw...	5cd1be63ead3c34fc153b4d8	0.1027	0.1027
149	Every time I walk past this, I giggle. Everyti...	5cd0f75dead3c34fc153a2ff	0.5023	0.5023

Στην εικόνα φαίνονται τα τελευταία 5 έγγραφα του DataFrame με τις ίδιες στήλες όπως και στις προηγούμενες αξιολογήσεις με την διαφορά στην τρίτη στήλη έχει την τιμή της συναισθηματικής μέτρησης του κειμένου που όλοι οι χαρακτήρες μετατράπηκαν σε πεζούς χαρακτήρες και στην τελευταία στήλη έχει την τιμή της συναισθηματικής μέτρησης του κειμένου που οι χαρακτήρες παραμένουν ως έχει.

```
count = 0
for index, row in data.iterrows():
    if row['sentiment5'] != row['sentiment4']:
        count+=1

print("Different sentiment polarity on {} tweets".format(countu))
```

```
Different sentiment polarity on 6 tweets
```

Υπάρχουν 6 διαφορετικά έγγραφα που αξιολογήθηκαν διαφορετικά στην κάθε τεχνική προ επεξεργασίας κειμένου λόγω της κεφαλαιοποίησης των γραμμάτων. Μετά από αναλυτική παρακολούθηση των κειμένων και των τιμών της συναισθηματικής μέτρησης για την κάθε τεχνική, καταλήγουμε πως η χρήση κεφαλαίων ή πεζών γραμμάτων παίζει ρόλο στον υπολογισμό της πολικότητας του συναισθήματος του κειμένου. Μέσα από την ανάλυση των κειμένων σε συνδυασμό με των τιμών της κάθε τεχνικής φάνηκε πως αν τα γράμματα είναι όλα πεζά η τιμή της πολικότητας αντιπροσωπεύει καλύτερα την πραγματικότητα του συναισθήματος του κειμένου παρά όταν τα γράμματα παραμένουν ως έχουν.

```
#ATL almost feels like enemy territory – the land of a million @delta flight
s. But I'm loyal to my #Airmance @americanair ❤️US✈️ #SSU2019 (@ Hartsfield-
Jackson Atlanta International Airport - @atlairport in Atlanta, GA) https://
t.co/1KR8nWMtFC https://t.co/0XtV4xtAqC
sentiment4: 0.5636       sentiment5: 0.2698

I spend more time in planes than at home. But still love it. Ready for a big
 week. #sales #salesleadership #saleslife https://t.co/tlBahqW5tF
sentiment4: 0.8765       sentiment5: 0.7717

Barely made my flight due to horrible traffic, (left my apartment at 6:15pm a
rrrived at LAX at 10:20)thank goodness the TSA was super quick today! But Tal
k about last minute lol..... but… https://t.co/Iy3RT16Kmy
sentiment4: 0.3595       sentiment5: 0.5707
```

Οπότε η τελική τεχνική προ επεξεργασίας κειμένου που θα χρησιμοποιηθεί στην ανάλυση είναι η μετατροπή όλων των γραμμάτων σε πεζά, η χρήση των stop words και η αφαίρεση των hashtags, των mentions και των συνδέσμων.

3.3.3 Εύρεση κατωφλίου για κατηγοριοποίηση

Η εύρεση της τεχνικής προ επεξεργασίας κειμένου για την μέγιστη απόδοση του εργαλείου VADER μας παίρνει ένα βήμα πριν την τελική ανάλυση του συναισθήματος και την κατηγοριοποίηση του κειμένου σε θετικό, αρνητικό ή ουδέτερο. Για να γινεί η κατηγοριοποίηση του κειμένου χρειάζεται να θέσουμε ένα όριο για τα tweets που χαρακτηρίζονται θετικά , αρνητικά ή ουδέτερα.

Εφαρμόστηκε η προ επεξεργασία κειμένου σε 418 tweets και βρέθηκε η πολικότητα του συναισθήματος του κάθε κειμένου. Αρχικά φιλτραρίστηκαν τα tweets που η τιμή της πολικότητας τους ήταν μεγαλύτερη από 0.5, τα λεγόμενα θετικά και τα tweets που η τιμή της πολικότητας τους ήταν μικρότερη από -0.5, τα λεγόμενα αρνητικά. Όλα τα υπόλοιπα tweets θεωρούνται ουδέτερα. Από τα 418 tweets τα 107 χαρακτηρίστηκαν ως θετικά και τα 17 χαρακτηρίστηκαν αρνητικά. Στη συνέχεια μελετήθηκαν τα κείμενα των θετικών και αρνητικών κειμένων. Από τα 107 κείμενα που χαρακτηρίστηκαν ότι έχουν θετικό συναίσθημα, εντοπίστηκαν 14 που δεν τηρούσαν αυτό το χαρακτηριστικό, ενώ τα 17 που θεωρήθηκαν ότι είχαν αρνητικό συναίσθημα όντως και τα 17 διατηρούσαν αυτό το γνώρισμα. Με κατώφλι μεγαλύτερο από 0.5 για θετικά και μικρότερο από -0.5 για αρνητικά το εργαλείο για ανάλυση συναισθήματος έχει επιτυχία 86.92%. Με κατώφλι μεγαλύτερο από 0.6 για θετικά και μικρότερο από -0.6 για αρνητικά το εργαλείο για ανάλυση συναισθήματος έχει επιτυχία 78.68%. Με κατώφλι μεγαλύτερο από 0.4 για θετικά και μικρότερο από -0.4 για αρνητικά το εργαλείο για ανάλυση συναισθήματος έχει επιτυχία 81.29%.

Έτσι καταλήγουμε στα όρια για τιμές 0.5 και μεγαλύτερες να χαρακτηρίζονται ως θετικά και για τιμές -0.5 και μικρότερες να χαρακτηρίζονται ως αρνητικά.

Κεφάλαιο 4 - Εξαγωγή πληροφορίας από ανάλυση

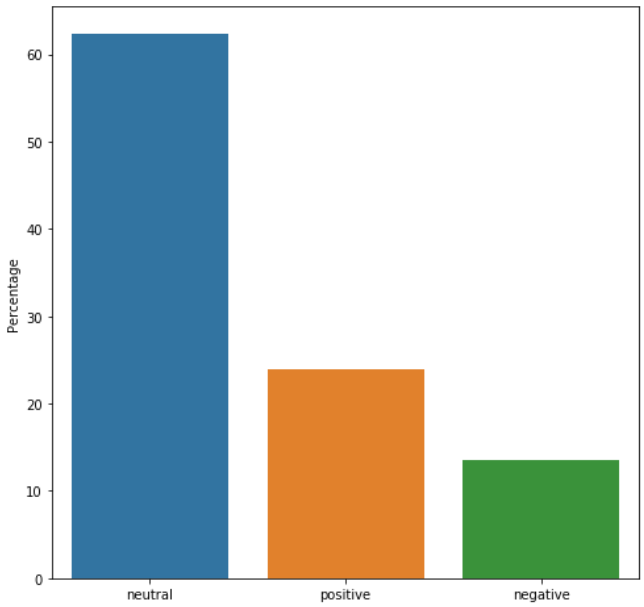
4.1 Εξαγωγή γνώσης

Πλέον έχει γινεί η ανάλυση του συναισθήματος στα tweets και τώρα θα δείξουμε με γραφικές παραστάσεις την πληροφορία που εξάγεται από αυτή την ανάλυση. Αρχικά πήραμε όλα τα tweets που μαζευτήκαν μέχρι στιγμής και τα βάλαμε σε ένα DataFrame όπως φαίνεται στην ακόλουθη εικόνα:

```
df.tail()
```

	Company	Date	compound	neg	neu	pos	text	label
3693657	Hawaiian Airlines	2019-05-21 11:00:00	0.0000	0.000	1.000	0.000	#southwestair i've been on hold for over 3 hou...	neutral
3693658	Hawaiian Airlines	2019-05-21 15:00:00	-0.7871	0.298	0.702	0.000	i ♥ @southwestair but they have the worst on-t...	negative
3693659	Hawaiian Airlines	2019-05-21 16:00:00	0.4588	0.000	0.700	0.300	@sgurman @lauraajarrett i will be viewing with...	neutral
3693660	Hawaiian Airlines	2019-05-22 05:00:00	0.4389	0.157	0.606	0.237	in spite of a three hour flight delay (den to ...	neutral
3693661	Hawaiian Airlines	2019-05-22 09:00:00	-0.4019	0.278	0.722	0.000	#southwestairlines mechanics dispute ends as #...	neutral

Στις στήλες του DataFrame έχουμε τις αεροπορικές εταιρίες που αφορά το κάθε tweet, την ημερομηνία δημιουργίας του tweets, τα αποτελέσματα της συναισθηματικής ανάλυσης, το κείμενο του κάθε tweet και την κατηγορία του συναισθήματος που εμπίπτει το κείμενο του tweet βάση της τιμής του compound. Οπότε έτσι έχουμε τρείς κατηγορίες των tweets βάση το συναίσθημα. Έχουν μαζευτεί 3,693,662 tweets μέχρι στιγμής των οποίων τα 886,105 έχουν κατηγοριοποιηθεί ως θετικά, τα 501,875 έχουν κατηγοριοποιηθεί ως αρνητικά και τα υπόλοιπα 2,305,682 έχουν κατηγοριοποιηθεί ως ουδέτερα. Στη συνέχεια υπολογίζεται το ποσοστό των θετικών, αρνητικών και ουδέτερων tweets. Παρατηρούμε ότι περισσότερα από τα μισά είναι ουδέτερα, ενώ τα θετικά tweets είναι σχεδόν διπλάσια από τα αρνητικά.



Στη συνέχεια χώρισα σε μία λίστα τα tweets που κατηγοριοποιήθηκαν ως θετικά και σε μια άλλη λίστα τα tweets που κατηγοριοποιήθηκαν ως αρνητικά. Από αυτές τις λίστες έκανα τρεις διαφορετικές επεξεργασίες του κειμένου για να εξάγω πληροφορία.

4.2 Εύρεση πιο πολυσύχναστων hashtag

Η πρώτη τεχνική είναι η αφαίρεση όλων των λέξεων από το κείμενο των tweets εκτός από τα hashtag και υπολόγισα την συχνότητα εμφάνισης του κάθε hashtag στα θετικά tweets και στα αρνητικά tweets.

```
Most used hashtags in positive tweets

[('#united', 22435), ('#delta', 12752), ('#avgeek', 10686), ('#unitedairlines', 8756), ('#americanairlines', 8180), ('#aviatio
n', 7178), ('#aateam', 7061), ('#travel', 6308), ('#britishairways', 4641), ('#easyjet', 4472), ('#planespotting', 3777), ('#bo
eing', 3754), ('#aviationphotography', 3314), ('#deltaairlines', 3107), ('#aviationlovers', 2957), ('#sfo', 2930), ('#sanfranci
sco', 2665), ('#sf', 2618), ('#cricketandtravelsuperfan', 2470), ('#sky', 2399), ('#lifeinthesky', 2320), ('#vegan', 2218), ('#
customerservice', 2110), ('#jfk', 2040), ('#insidethecockpit', 1644), ('#manchesterairport', 1634), ('#veganiseyourmenu', 155
0), ('#airlines', 1536), ('#manchester', 1532), ('#heathrow', 1523), ('#ba100', 1523), ('#kingdom', 1519), ('#heathrowairport',
1502), ('#london', 1499), ('#airport', 1488), ('#boeing777', 1398), ('#beingunited', 1353), ('#flight', 1234), ('#virginatlanti
c', 1222), ('#mufc', 1219), ('#boeing787', 1202), ('#ba', 1168), ('#emirates', 1162), ('#aateam.', 1107), ('#airbus', 1106),
('#dubnewyork', 1085), ('#championship', 1085), ('#', 1082), ('#southwestairlines', 1070), ('#dubunited', 1050)]


Most used hashtags in negative tweets

[('#americanairlines', 13152), ('#united', 7644), ('#delta', 5700), ('#unitedairlines', 4769), ('#aateam', 2904), ('#britishair
ways', 2460), ('#easyjet', 2438), ('#flybe', 1867), ('#deltaairlines', 1859), ('#jfk', 1532), ('#manchesterairport', 1486), ('#
ryanair', 1193), ('#americanair', 1192), ('#customerservice', 1180), ('#emirates', 1167), ('#airlines', 1150), ('#travel', 113
1), ('#neveragain', 1129), ('#fail', 1126), ('#unionstrong', 937), ('#', 846), ('#flybetterdontflyemirates', 712), ('#ba', 65
7), ('#flybescam', 641), ('#americanairlinessucks', 610), ('#aa', 594), ('#boycottdelta', 576), ('#southwestairlines', 558),
('#badcustomerservice', 545), ('#badservice', 517), ('#southwest', 512), ('#kingdom', 490), ('#dfw', 479), ('#flights', 473),
('#premier_league', 448), ('#avgeek', 437), ('#airline', 434), ('#unitedsucks', 434), ('#disappointed', 429), ('#aviation', 39
8), ('#heathrowairport', 389), ('#union', 389), ('#mufc', 378), ('#manchester', 373), ('#customerexperience', 370), ('#poorserv
ice', 366), ('#championship', 364), ('#unacceptable', 356), ('#poorcustomerservice', 343), ('#deltasucks', 336)]
```

Παρατηρούμε ότι υπάρχουν αρκετά hashtag που χρησιμοποιούνται με μεγάλη συχνότητα και δεν υπάρχουν στη βάση δεδομένων μας. Αυτά τα hashtag θα μπορούσαν να εξετασθούν και να ενσωματωθούν αναλόγως στο σύστημα αν τα tweets που αντιστοιχούνται είναι σχετικά με αεροπορικό ταξίδι. Επίσης παρατηρούμε ότι το hashtag με την μεγαλύτερη συχνότητα εμφάνισης στα αρνητικά tweets είναι το “#americanairlines”, ενώ το hashtag με την μεγαλύτερη συχνότητα εμφάνισης στα θετικά tweets είναι το “#united”.

4.3 Εύρεση πιο πολυσύχναστων mention

Η δεύτερη τεχνική είναι η αφαίρεση όλων των λέξεων από το κείμενο των tweets εκτός από τα mention και υπολόγισα την συχνότητα εμφάνισης του κάθε mention στα θετικά tweets και στα αρνητικά tweets.

```
Most used mentions in positive tweets

[('@delta', 162187), ('@americanair', 142280), ('@british_airways', 120329), ('@united', 99150), ('@heathrowairport', 75913),
('@southwestair', 74281), ('@easyjet', 55727), ('@virginatlantic', 39024), ('@manairport', 31511), ('@gatwick_airport', 23819),
('@flybe', 20657), ('@ryanair', 13763), ('@edi_airport', 8473), ('@jet2tweets', 8162), ('@emirates', 7750), ('@jfkairport', 700
3), ('@airbus', 6503), ('@', 5798), ('@boeingairplanes', 5414), ('@eoinhiggins_', 5176), ('@dublinairport', 4698), ('@davewalls
worth', 4599), ('@paytmtravel', 4550), ('@schiphol', 4253), ('@jetblue', 4207), ('@delta.', 4134), ('@alaskaair', 3584), ('@fly
sfo', 3504), ('@delta,', 3410), ('@britishairways', 2981), ('@klm', 2942), ('@baretrojets', 2842), ('@lufthansa', 2746), ('@boe
ing', 2592), ('@dfwairport', 2556), ('@airport_fra', 2546), ('@weareunited', 2501), ('@aflcio', 2496), ('@bt77w', 2412), ('@aer
lingus', 2358), ('@americanair.', 2243), ('@americanair,', 2219), ('@aircanada', 2085), ('@airbusintheuk', 2031), ('@pilotcharl
otte', 2016), ('@airfrance', 2000), ('@jumbo747pilot', 1947), ('@delta!', 1934), ('@spiritairlines', 1922), ('@flightradar24',
1918)]


Most used mentions in negative tweets

[('@americanair', 149213), ('@delta', 103740), ('@united', 59982), ('@british_airways', 45819), ('@southwestair', 36247), ('@ea
syjet', 27297), ('@flybe', 15787), ('@heathrowairport', 15285), ('@virginatlantic', 10226), ('@ryanair', 10217), ('@manairpor
t', 8964), ('@gatwick_airport', 7406), ('@eoinhiggins_', 5549), ('@delta.', 3239), ('@emirates', 3225), ('@americanair.', 313
3), ('@jfkairport', 3026), ('@edi_airport', 2917), ('@jet2tweets', 2412), ('@delta,', 2114), ('@britishairways', 2004), ('@jetb
lue', 1986), ('@aflcio', 1967), ('@americanair,', 1916), ('@', 1666), ('@cocacola', 1574), ('@united.', 1419), ('@spiritairline
s', 1302), ('@bbcsussex', 1287), ('@emiratessupport', 1099), ('@alaskaair', 1081), ('@dfwairport', 1033), ('@unitedairlines', 1
018), ('@basicallyidowrk', 1013), ('@machinistsunion', 963), ('@easyjet_press', 935), ('@dxb', 926), ('@aircanada', 914), ('@sc
hiphol', 891), ('@heathrownoise', 875), ('@united,', 868), ('@transportgovuk', 861), ('@richardbranson', 856), ('@travisakers',
840), ('@alex_cruz', 808), ('@kristyswansonxo', 796), ('@tsa', 779), ('@cnn', 761), ('@klm', 759), ('@homedepot', 754)]
```

Παρατηρούμε ότι και πάλι η American airline έχει τη μεγαλύτερη συχνότητα εμφάνισης στα αρνητικά tweets, ενώ η αερογραμμή Delta έχει τη μεγαλύτερη αναφορά στα θετικά tweets.

4.4 Εύρεση των πιο πολυσύχναστων αεροπορικών εταιριών

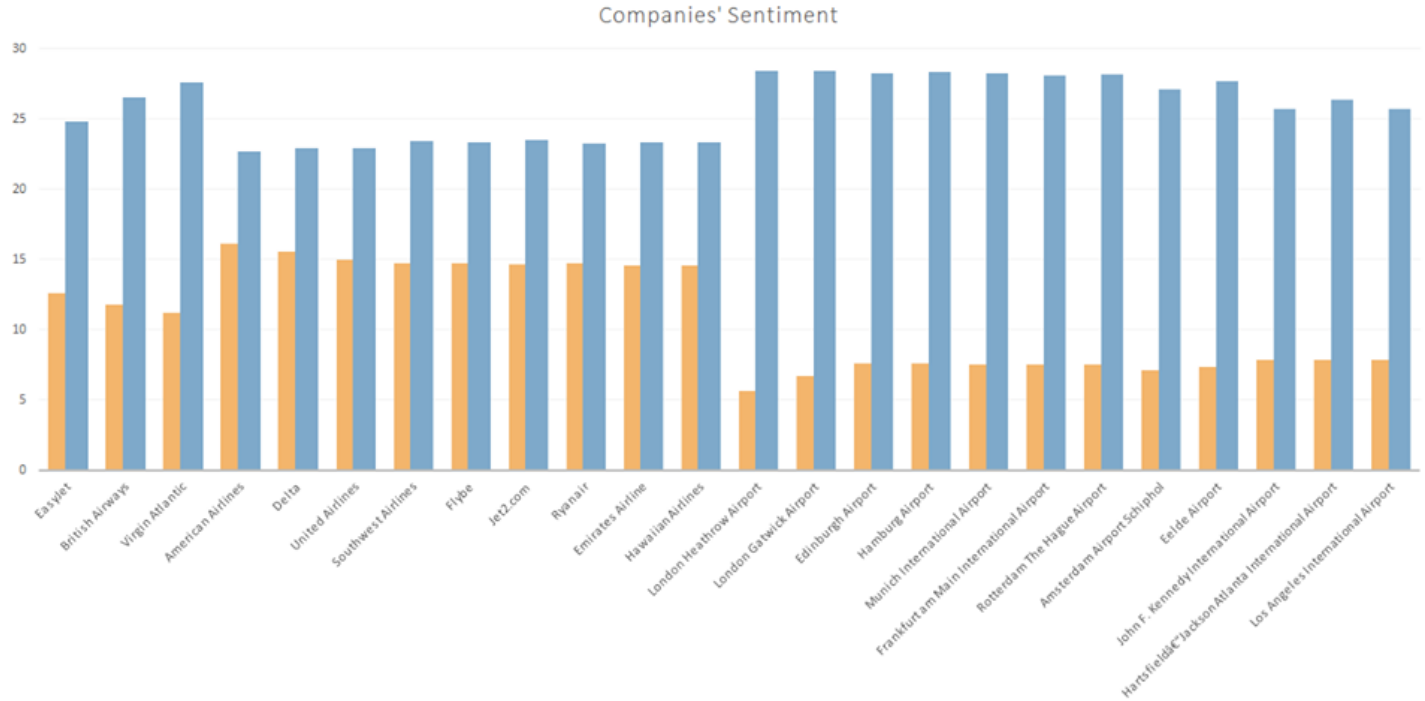
Η τρίτη τεχνική είναι η εύρεση των εταιριών που αφορά το κάθε tweets και ο υπολογισμός της συχνότητας εμφάνισης της κάθε αεροπορικής εταιρίας στα θετικά tweets και στα αρνητικά tweets.

```
Most used company in positive tweets

[('Hawaiian Airlines', 95436), ('Emirates Airline', 95036), ('Ryanair', 91311), ('Jet2.com', 87730), ('Flybe', 85914), ('Southw
est Airlines', 82004), ('United Airlines', 70445), ('Delta', 54199), ('American Airlines', 33695), ('Virgin Atlantic', 17774),
('Los Angeles International Airport', 15451), ('Hartsfield-Jackson Atlanta International Airport', 14358), ('British Airways',
14347), ('John F. Kennedy International Airport', 13734), ('Athens International Airport', 13442), ('Amsterdam Airport Schipho
l', 11535), ('Eelde Airport', 11320), ('Frankfurt am Main International Airport', 10691), ('Rotterdam The Hague Airport', 1063
5), ('Munich International Airport', 10373), ('Hamburg Airport', 10176), ('Edinburgh Airport', 10138), ('Manchester Airport', 9
451), ('London Gatwick Airport', 6909), ('London Heathrow Airport', 5529), ('EasyJet', 4472)]


Most used company in negative tweets

[('Hawaiian Airlines', 59425), ('Emirates Airline', 59293), ('Ryanair', 57770), ('Jet2.com', 54854), ('Flybe', 54354), ('Southw
est Airlines', 51467), ('United Airlines', 45891), ('Delta', 36679), ('American Airlines', 23965), ('Virgin Atlantic', 7234),
('British Airways', 6368), ('Los Angeles International Airport', 4694), ('Hartsfield-Jackson Atlanta International Airport', 42
57), ('John F. Kennedy International Airport', 4172), ('Athens International Airport', 4057), ('Amsterdam Airport Schiphol', 30
16), ('Eelde Airport', 2992), ('Frankfurt am Main International Airport', 2849), ('Rotterdam The Hague Airport', 2847), ('Munic
h International Airport', 2768), ('Hamburg Airport', 2736), ('Edinburgh Airport', 2717), ('Manchester Airport', 2457), ('EasyJe
t', 2274), ('London Gatwick Airport', 1636), ('London Heathrow Airport', 1103)]
```

Παρατηρούμε ότι η συχνότητα εμφάνισης σχεδόν κάθε εταιρίας έχει την ίδια σειρά εμφάνισης στα θετικά και στα αρνητικά tweets. Αυτό συμβαίνει επειδή μπορεί να υπάρχουν διαφορετικό σε πλήθος tweets που αναφέρονται σε κάθε εταιρία. Οπότε μπορούμε να

μελετήσουμε το πλήθος των θετικών και αρνητικών tweets σε σχέση με το συνολικό πλήθος των tweets της κάθε εταιρίας ξεχωριστά. Με αυτό τον τρόπο μπορούμε να εκτιμήσουμε συγκριτικά τον βαθμό ικανοποιήσεις των ταξιδιωτών της κάθε εταιρίας.

Μετά από υπολογισμούς βρέθηκε το ποσοστό των θετικών και αρνητικών tweets της κάθε εταιρίας και αναπαραστάθηκαν γραφικά όλα τα ποσοστά θετικότητας και αρνητικότητάς της κάθε εταιρίας μαζί. Έτσι εύκολα μπορούμε να συγκρίνουμε τον βαθμό ικανοποίησης των ταξιδιωτών της κάθε εταιρίας σε σχέση με άλλες εταιρίες.

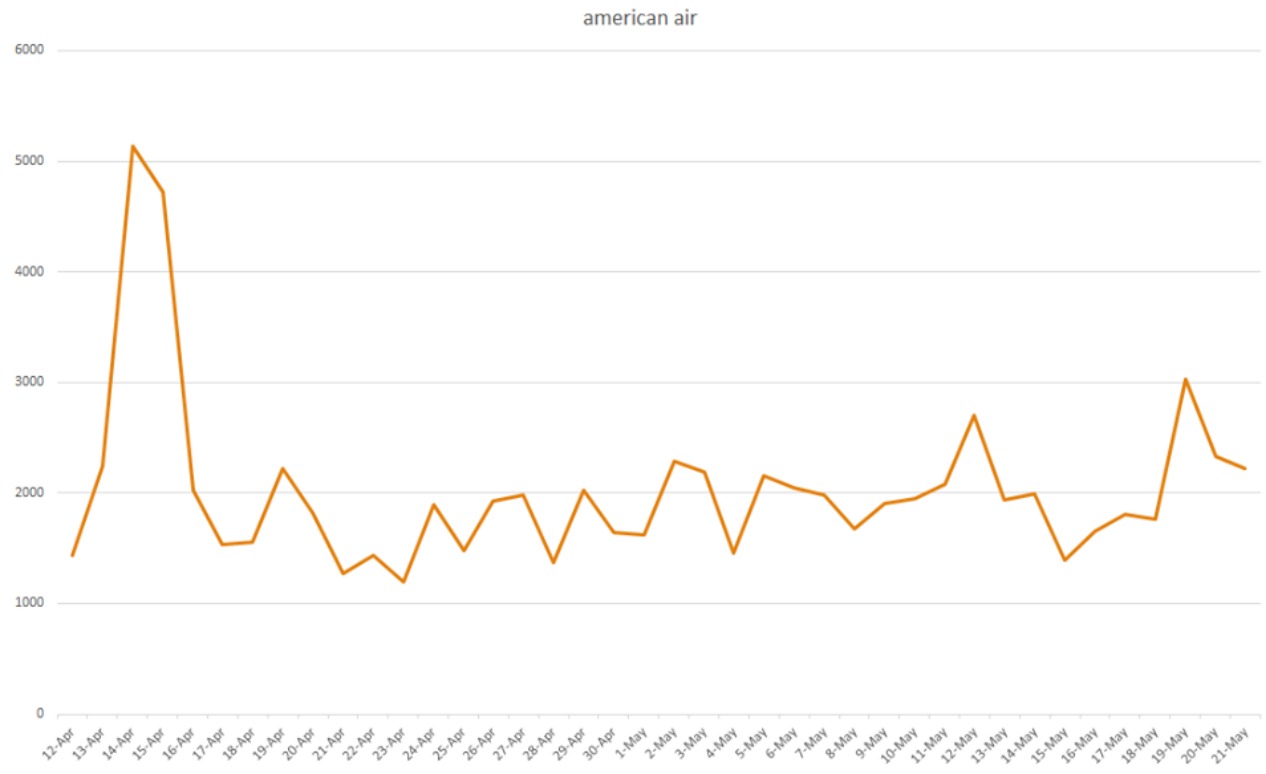


Βλέποντας την γραφική παράσταση μπορούμε να παρατηρήσουμε ότι τα αεροδρόμια έχουν σχετικά μεγαλύτερο ποσοστό θετικότητας και αρκετά μικρότερο ποσοστό αρνητικότητας σε σχέση με τις αερογραμμές. Αυτό ίσως συμβαίνει επειδή οι ταξιδιώτες δεν έχουν τόσες πολλές απαιτήσεις από τα αεροδρόμια σε σχέση με τις αερογραμμές.

Η εταιρία με το μεγαλύτερο ποσοστό θετικότητας είναι το London Gatwick Airport, ενώ η εταιρία με το μεγαλύτερο ποσοστό αρνητικών tweets έχει η American Airlines. Το γεγονός ότι τα hashtag και mention της American Airlines στις προηγούμενες δυο τεχνικές είχαν την μεγαλύτερη συχνότητα εμφάνισης στα αρνητικά tweets με πρότρεψε να αναλύσω περαιτέρω πληροφορίες για την συγκεκριμένη εταιρία.

4.5 American Airlines

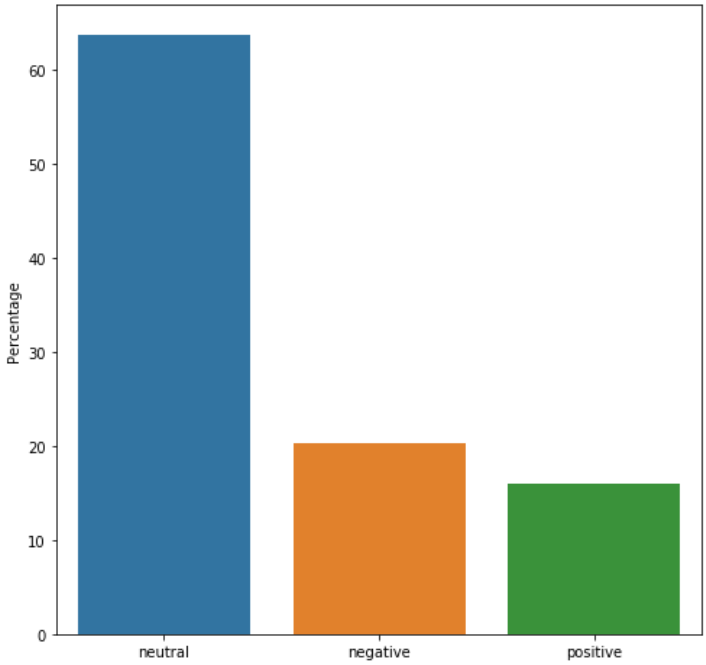
Η American Airlines είχε εμφανιστεί ως η πιο πολυσύχναστη φράση στα αρνητικά tweets στις τρείς προηγούμενες τεχνικές επεξεργασίας των tweets για εξαγωγή πληροφορίας. Για αυτό τον λόγο μάζεψα όλα τα tweets που σχετίζονται με την American Airlines. Τα tweets που σχετίζονται με την American Airlines είναι αυτά που συλλέχθηκαν βάση τας hashtag και mention της. Στη συνέχεια δημιούργησα ένα ιστόγραμμα που απεικονίζει το πλήθος των tweets που συλλέχθηκαν ανά ημέρα για περίοδο 40 ημερών.



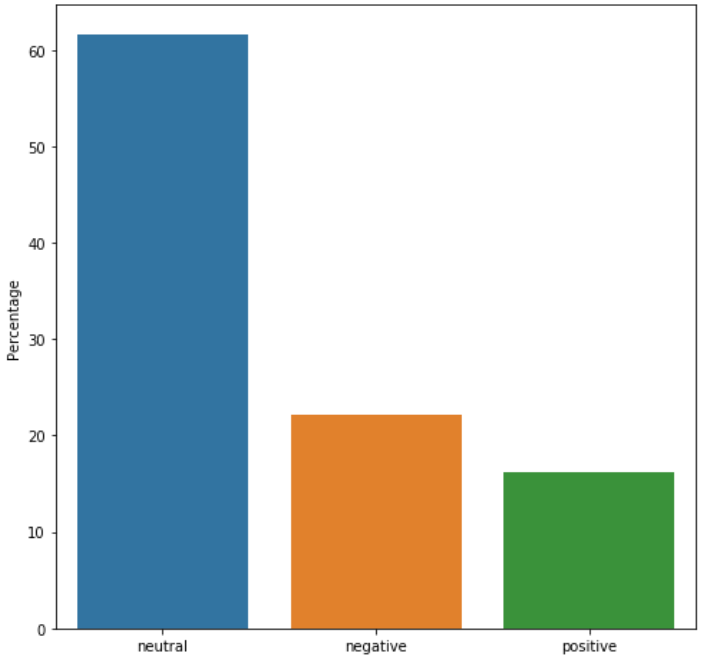
Εικόνα 4: Ιστόγραμμα που απεικονίζει το πλήθος των tweets που συλλέχθηκαν ανά ημέρα για περίοδο 40 ημερών

Στις 14 και 15 του Απρίλη παρατηρείται μία απότομη άνοδος στο ιστόγραμμα για την American Airlines. Εφάρμοσα συναισθηματική ανάλυση στα tweets της American Airlines τις ημέρες 14 και 15 του Απρίλη ξεχωριστά.

Για τις 14 του Απρίλη συλλέχθηκαν συνολικά 5139 tweets για την American Airlines των οποίων το 63% ήταν ουδέτερα, το 21% των tweets ήταν αρνητικά και 16% θετικά.



Για της 15 του Απρίλη συλλέχθηκαν συνολικά 4724 tweets των οποίων το 62% ήταν ουδέτερα, το 22% των tweets ήταν αρνητικά και 16% θετικά.



Στη συνέχεια αναζήτησα στο Google “14 April American Airlines” και εμφανίστηκαν δύο αποτελέσματα που δικαιολογούν αυτή την άνοδο στο πλήθος των tweets που συλλέχθηκαν στις 14 και 15 του Απρίλη. Το πρώτο αποτέλεσμα είναι ένα βίντεο που δείχνει μια τεράστια

ουρά στη γραμμή εξυπηρέτησης πελατών της American Airlines στην οποία υπήρχε χρόνος αναμονής 2 ωρών μέχρι να εξυπηρετηθεί κάποιος πελάτης. Το δεύτερο αποτέλεσμα ήταν μια ανακοίνωση από τις ειδήσεις του BBC που δημοσιεύτηκε στις 14 του Απρίλη. Αυτή ανακοίνωση αφορούσε την American Airlines και έλεγε πως θα χρειαστεί να ακυρωθούν όλες οι πτήσεις μέχρι τον Αύγουστο που γίνονται με το μοντέλο αεροπλάνου Boeing 737. Ο λόγος που ακυρώνονται οι πτήσεις με το μοντέλο Boeing 737 είναι επειδή στις 10 του Μάρτη είχε γινεί αεροπορικό δυστύχημα στην Αιθιοπία με αυτό το συγκεκριμένο μοντέλο αεροπλάνου, το οποίο χρησιμοποιείται αρκετά στην American Airlines. Για αυτό τον λόγο χρειάζονται να γινούν διάφορες εξετάσεις σε αυτό το μοντέλο προτού χρησιμοποιηθεί ξανά στα αεροπορικά ταξίδια.

Κεφάλαιο 5 – Συμπεράσματα – Μελλοντική Εργασία

5.1 Συμπεράσματα

Συνοψίζοντας, με την υφιστάμενη διπλωματική εργασία έχουμε υλοποιήσει μια μεθοδολογία η οποία μας παρουσιάζει γραφικά την ικανοποίηση των ταξιδιωτών από διάφορες αεροπορικές εταιρίες. Επίσης μπορούμε να εντοπίσουμε διάφορα μεγάλα γεγονότα της κάθε εταιρίας μέσα από τα ιστογράμματα της.

Αρχικά μετά την περιγραφή των διάφορων τεχνολογιών που χρησιμοποιούνται στα πλαίσια της συγκεκριμένης εργασίας, συλλέχθηκαν διάφορα δεδομένα όπως ονόματα λογαριασμών και hashtags για αεροδρόμια και αερογραμμές, πλαίσια οριοθέτησης των αεροδρομίων και δημοφιλή hashtags που σχετίζονται με αεροπορικό ταξίδι. Τα δεδομένα αυτά χρησιμοποιήθηκαν για το φιλτράρισμα των tweets που είναι σχετικά με αεροπορικό ταξίδι. Από τα δεδομένα αυτά φάνηκε από την αξιολόγηση που κάναμε ότι η χρήση των mentions και hashtags για αερογραμμές ήταν η πιο αποδοτική διότι το 77% των tweets που σύλλεγε ήταν σχετικά με το θέμα μας, ενώ η χρήση των δημοφιλή hashtag για αεροπορικά ταξίδια είχε μόνο 31% σχετικά tweets προσθέτοντας πολύ θόρυβο στα δεδομένα μας για αυτό στη συνέχεια αφαιρέθηκε από την μεθοδολογία. Επίσης η χρήση των mentions και hashtags για αεροδρόμια είχε αρκετά καλή απόδοση με 75% σχετικών tweets, ενώ η χρήση των πλαισίων οριοθέτησης των αεροδρομίων είχε καλή επίδοση με 60% σχετικών tweets. Η χρήση των πλαισίων οριοθέτησης των αεροδρομίων συλλέγει λίγα (248) tweets γεγονός το οποίο οφείλεται στο ότι η πλειοψηφία των χρηστών έχει απενεργοποιημένη την επιλογή εμφάνισης της τοποθεσίας.

Στην συνέχεια μαζεύτηκε μεγάλος όγκος από tweets ο οποίος χρησιμοποιήθηκε στην ανάλυση του συναισθήματος. Για την ανάλυση του συναισθήματος χρειάζεται να προ επεξεργαστούμε το κείμενο και να μετατρέψουμε όλους τους χαρακτήρες σε πεζούς και να αφαιρέσουμε τα hashtags, τα mentions και τους συνδέσμους αν τυχόν υπάρχουν. Μετά από την προ επεξεργασία του κειμένου εφαρμόστηκε η ανάλυση του συναισθήματος σε όλα τα tweets που μαζεύτηκαν μέχρι στιγμής και φάνηκε ότι το 62% όλων των tweets που συλλέχθηκαν ήταν ουδέτερα, ενώ το 24% θετικά και 14% αρνητικά. Παρατηρούμε ότι τα θετικά tweets είναι σχεδόν διπλάσια από τα αρνητικά. Στη συνέχεια είχαν εφαρμοστεί επιπλέον αναλύσεις όπως εύρεση πιο πολυσύχναστων λέξεων, mentions, hashtags και εταιρειών στα θετικά και αρνητικά tweets. Σε αυτό το σημείο είναι σημαντικό να αναφερθεί

ότι σε όλες τις αναλύσεις που έγιναν η American Airlines είχε την πιο συχνή εμφάνιση στην ανάλυση των πιο πολυσύχναστων mentions, hashtags και εταιρειών στα αρνητικά tweets.

Παρακάτω σε αυτή την εργασία αναλύθηκε το ποσοστό θετικών και αρνητικών tweets για την κάθε αεροπορική εταιρία. Έτσι μετά από σύγκριση των ποσοστών της κάθε εταιρίας φάνηκε πως τα αεροδρόμια έχουν σχετικά μεγαλύτερο ποσοστό θετικότητας και αρκετά μικρότερο ποσοστό αρνητικότητας σε σχέση με τις αερογραμμές. Ο λόγος που μπορεί να συμβαίνει αυτό είναι επειδή οι ταξιδιώτες δεν έχουν τόσες πολλές απαιτήσεις από τα αεροδρόμια σε σχέση με τις αερογραμμές. Το αεροδρόμιο με το μεγαλύτερο ποσοστό θετικών tweets (28.39%) είναι το London Gatwick Airport και το αεροδρόμιο με το μεγαλύτερο ποσοστό αρνητικών tweets (7.82%) είναι το Los Angeles International Airport. Η αερογραμμή με το μεγαλύτερο ποσοστό θετικών tweets (27.57%) είναι η Virgin Airlines και η αερογραμμή με το μεγαλύτερο ποσοστό αρνητικών tweets (16.15%) είναι η American Airlines.

Για το λόγο ότι η American Airlines είχε το μεγαλύτερο ποσοστό αρνητικών tweets σε σχέση με τις υπόλοιπες εταιρίες και την πιο συχνή εμφάνιση στην ανάλυση των πιο πολυσύχναστων mentions, hashtags και εταιρειών στα αρνητικά tweets, επιλέχθηκε για να γίνει πιο εξειδικευμένη ανάλυση. Έγινε το ιστόγραμμα που απεικονίζει το πλήθος των tweets που συλλέχθηκαν ανά ημέρα για περίοδο 40 ημερών, όπου φάνηκε μια απότομη άνοδος στις 14 και 15 του Απρίλη 2019. Στη συνέχεια έγινε ανάλυση του συναισθήματος για αυτές τις δύο μέρες και παρουσιάστηκε πως το 22% των tweets ήταν αρνητικά ενώ μόνο 16% ήταν θετικά. Μετά από ψάξιμο στο Google βρέθηκαν άρθρα και βίντεο που δικαιολογούν αυτή την αρνητική σε συναίσθημα άνοδο. Αυτή άνοδος πιθανόν να οφείλεται στο γεγονός ότι στις 14 του Απρίλη 2019 η American Airlines ανακοίνωσε ότι θα ακυρώσει σημαντικό αριθμό πτήσεων για τους επόμενους 5 μήνες ώστε να γίνουν έλεγχοι στα αεροπλάνα μοντέλου Boeing 737. Αυτές οι ακυρώσεις ήταν απαραίτητες καθώς το συγκεκριμένο μοντέλο ενεπλάκη σε δύο αεροπορικά δυστυχήματα τους προηγούμενους μήνες. Επίσης, το υψηλό ποσοστό αρνητικών tweets για την American Airlines επιβεβαιώθηκε με ένα βίντεο στο διαδίκτυο το οποίο δημοσιεύτηκε στις 14 του Απρίλη 2019 και το οποίο δείχνει μια τεράστια ουρά αναμονής στη γραμμή εξυπηρέτησης πελατών της συγκεκριμένης αερογραμμής. Από τα σχόλια των χρηστών στο βίντεο εκείνο είναι φανερό ότι οι πελάτες ήταν δυσαρεστημένοι με τις ακυρώσεις των πτήσεων καθώς και με την εξυπηρέτηση της αερογραμμής

Η μεθοδολογία μας φαίνεται να δουλεύει καθώς μπορούμε να ανιχνεύσουμε μέσω των tweets κάποιες σημαντικές αλλαγές στον αριθμό των tweets και στα ποσοστά θετικών και

αρνητικών συναισθημάτων οι οποίες συμβαδίζουνε με συγκεκριμένα γεγονότα όπως ακύρωση πολλαπλών πτήσεων ή μεγάλη καθυστέρηση εξυπηρέτησης των ταξιδιωτών.

Τέλος, αυτή η μεθοδολογία μπορεί να δώσει σε αερογραμμές και αεροδρόμια διάφορες ιδέες για να βελτιώσουν πιο εξειδικευμένες υπηρεσίες τους. Για παράδειγμα, βάση των tweets θα μπορούσε κάποια αερογραμμή να μάθει ποια γεύματα προτιμούν οι πελάτες κατά την διάρκεια της πτήσης. Για να γίνει αυτό χρειάζεται κάποια ενθάρρυνση από τις αερογραμμές ή τα αεροδρόμια για μεγαλύτερη συμμετοχή των πελατών τους στο Twitter και αυτό μπορεί να επιτευχθεί όταν οι πελάτες βλέπουν κάποιο αντίκτυπο στις υπηρεσίες των αεροδρομίων ή αερογραμμών όταν αυτοί εκφράζουν τις απόψεις τους. Έτσι με την ενθάρρυνση για μεγαλύτερη συμμετοχή στο Twitter των πελατών, οι αερογραμμές και τα αεροδρόμια θα έχουν αρκετό υλικό να επεξεργαστούν με απώτερο σκοπό την βελτίωση διαφόρων εξειδικευμένων υπηρεσιών.

5.2 Μελλοντική Εργασία

Η μεθοδολογία αυτής της εργασίας έχει αρκετά σημεία που θα μπορούσε να βελτιωθεί. Για παράδειγμα στη φάση της ανάλυσης των πιο πολυσύχναστων hashtag εντοπιστήκαν διάφορα hashtag που χρησιμοποιούνται σε μεγάλη συχνότητα και δεν υπάρχουν στη βάση δεδομένων. Θα μπορούσαν τα εξετασθούν αυτά hashtag και να ενσωματωθούν στο σύστημα μας εάν η συλλογή των tweets με αυτά είναι σχετικά. Έτσι θα έχουμε μεγαλύτερο αριθμό από tweets και άρα περισσότερη πληροφορία για να επεξεργαστούμε.

Αυτή εργασία εφάρμοζε ανάλυση συναισθήματος και τα tweets κατηγοριοποιούταν σε θετικά ή αρνητικά. Η εργασία θα μπορούσε να επεκταθεί με διάφορες επιπλέον πιο εξειδικευμένες επεξεργασίας φυσικής γλώσσας και να χώριζε σε διάφορες κατηγορίες τα θετικά και αρνητικά tweets, όπως για παράδειγμα σε κατηγορία «φαγητό». Έτσι η εταιρίες θα γνωρίζουν πιο καλά σε ποια σημεία αναφέρονται οι πελάτες τους.

Βιβλιογραφία

[1] C. Hutto και E. Gilbert, «VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text,» 2014.

[2] «Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2019 (in millions): Statista,» Statista, April 2019. [Ηλεκτρονικό]. Available: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.

[3] «Number of active twitter users in selected countries: Statista,» Statista, 2019. [Ηλεκτρονικό]. Available: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/.

[4] «NumPy,» Scipy.org, 2019. [Ηλεκτρονικό]. Available: https://www.numpy.org/.

[5] «pandas.pydata.org,» 12 March 2019. [Ηλεκτρονικό]. Available: https://pandas.pydata.org/pandas-docs/stable/.

[6] «hashtags: Best-hashtags,» 2017. [Ηλεκτρονικό]. Available: http://best-hashtags.com/hashtag/airtravel/.

[7] «SurveySystem,» Creative Research Systems , 2012. [Ηλεκτρονικό]. Available: https://www.surveysystem.com/sscalc.htm.

[8] Klokan Technologies, «boundingbox.klokantech,» Klokan Technologies, 2017. [Ηλεκτρονικό]. Available: https://boundingbox.klokantech.com/.

[9] H. Efstathiades, D. Antoniades, G. Pallis και M. Dikaiakos, «Identification of Key Locations based on Online Social Network Activity».