ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΠΡΩΗΝ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΥ ΣΧΕΔΙΑΣΜΟΥ ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ
ΤΜΗΜΑ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ (ΠΑΤΡΑ)

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΕΞΑΓΩΓΗ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΤΟ TWITTER

ΣΠΟΥΔΑΣΤΗΣ : ΓΕΩΡΓΙΟΣ ΤΖΑΝΗ

ΕΠΟΠΤΕΥΩΝ ΚΑΘΗΓΗΤΗΣ : ΣΤΑΜΟΣ ΚΩΝΣΤΑΝΤΙΝΟΣ

ΠΑΤΡΑ - 4/12/2016

Ευχαριστώ τους γονείς μου και τον κ. Στάμο για την υποστήριξη και τη βοήθεια που μου προσέφεραν.

Περίληψη

Η παρούσα πτυχιακή ασχολείται με την εξόρυξη δεδομένων από το Twitter, δηλαδή εξαγωγή συνόλων δεδομένων από το Twitter με τη βοήθεια προγράμματος που θα δημιουργήσουμε, και κατάταξη των δεδομένων για καλύτερη ανάγνωση και εύρεση αυτού που αναζητάμε, με στόχο την ανακάλυψη νέας γνώσης, έπειτα κάνουμε ανάλυση των αποτελεσμάτων και τέλος βγάζουμε τα τελικά συμπεράσματα.

Abstract

This Thesis is about mining Twitter data, namely extraction of data sets from Twitter with the help of a program we will create, and classify the data for better reading and finding what we search for, aiming to acquire new knowledge, then we analyze the results and in the end we make our final conclusions.

Περιεχόμενα

Κεφάλαιο 1 Εισαγωγή.....Σελίδα 5
Κεφάλαιο 2 Big Data.....Σελίδα 6
Κεφάλαιο 3 Επιστήμη των δεδομένων.....Σελίδα 7
Κεφάλαιο 4 Machine learning.....Σελίδα 8
Κεφάλαιο 5 Ανωνυμοποίηση των δεδομένων.....Σελίδα 9
Κεφάλαιο 6 Παραβίαση των δεδομένων.....Σελίδα 11
Κεφάλαιο 7 Cloud computing.....Σελίδα 12
Κεφάλαιο 8 Αποθήκευση δεδομένων.....Σελίδα 13
Κεφάλαιο 9 Ζητήματα εξόρυξης δεδομένων.....Σελίδα 16
Κεφάλαιο 10 Ανακάλυψη της γνώσης.....Σελίδα 18
Κεφάλαιο 11 Εξόρυξη δεδομένων.....Σελίδα 19
Κεφάλαιο 12 Εξόρυξη του Παγκόσμιου ιστού.....Σελίδα 23
Κεφάλαιο 13 Twitter APIs.....Σελίδα 25
Κεφάλαιο 14 Εξόρυξη δεδομένων με χρήση Python 3.4.....Σελίδα 28
Κεφάλαιο 15 Αποτελέσματα εξόρυξης δεδομένων.....Σελίδα 42
Κεφάλαιο 16 Συμπεράσματα.....Σελίδα 45
Βιβλιογραφία.....Σελίδα 48

Κεφάλαιο 1. Εισαγωγή

Ο στόχος της παρούσας πτυχιακής είναι η δημιουργία ενός προγράμματος, με τη βοήθεια του οποίου θα μπορέσουμε να εξάγουμε δεδομένα από το Twitter. Η εξαγωγή δεδομένων είναι μία ολοκληρωμένη διαδικασία που αποσκοπεί στην ανακάλυψη γνώσης σε ότι αφορά την συμπεριφορά των χρηστών. Τα αποτελέσματα αυτής της διαδικασίας είναι ένα σύνολο από κατάλληλα μοντελοποιημένα δεδομένα. Αυτά τα μοτίβα αξιολογούνται και στη συνέχεια εκμεταλλεύονται για σκοπούς όπως είναι η βελτίωση της δομής του διαδικτυακού τόπου και η βελτίωση της εμπειρίας χρήστη.

Στα επόμενα κεφάλαια αναλύω διάφορες θεωρίες για τα δεδομένα, πως τα διαχειρίζονται, πως εξάγονται, και πως διαχειρίζονται τη γνώση από την εξαγωγή διάφορων δεδομένων. Στη συνέχεια παρουσιάζω το πρόγραμμα που δημιούργησα, διάφορα βήματα που ακολούθησα, τη χρησιμοποίησα, τα αποτελέσματα που έλαβα από την εξαγωγή δεδομένων, και τέλος τα συμπεράσματα που κατέληξα.

Κεφάλαιο 2. Big Data

Big Data είναι ένας όρος για σύνολα δεδομένων που είναι τόσο μεγάλα ή σύνθετα που οι παραδοσιακές εφαρμογές επεξεργασίας δεδομένων είναι ανεπαρκείς για την αντιμετώπισή τους. Στις προκλήσεις περιλαμβάνονται η ανάλυση, η σύλληψη, η επιμέλεια δεδομένων, η αναζήτηση, η κοινή χρήση, η αποθήκευση, η μεταφορά, η απεικόνιση, οι επερωτήσεις, η ενημέρωση και η προστασία ιδιωτικών πληροφοριών. Ο όρος Big Data συχνά αναφέρεται στη χρήση των predictive analytics, user behavior analytics, ή σε ορισμένες άλλες προηγμένες μεθόδους των data analytics που εξάγουν αξία από τα δεδομένα, και σπάνια σε ένα συγκεκριμένο μέγεθος του συνόλου δεδομένων. Δεν υπάρχει αμφιβολία ότι οι ποσότητες των τώρα διαθέσιμων δεδομένων είναι πράγματι μεγάλες, αλλά αυτό δεν είναι το σημαντικότερο χαρακτηριστικό του νέου αυτού οικοσυστήματος δεδομένων.

Η ανάλυση των συνόλων δεδομένων μπορεί να βρει νέες συσχετίσεις στον εντοπισμό επιχειρηματικών τάσεων, στην πρόληψη των ασθενειών, στην καταπολέμηση της εγκληματικότητας και ούτω καθεξής. Οι επιστήμονες, στελέχη επιχειρήσεων, επαγγελματίες της ιατρικής, της διαφήμισης και των κυβερνήσεων συναντούν τακτικά δυσκολίες με μεγάλα σύνολα δεδομένων σε τομείς όπως η αναζήτηση στο Διαδίκτυο, τη χρηματοδότηση, την αστική πληροφορικής και των επιχειρήσεων πληροφορικής. Οι επιστήμονες αντιμετωπίζουν περιορισμούς στο έργο e-Science, συμπεριλαμβανομένων τη μετεωρολογία, τη γονιδιωματική, τη connectomics, τις πολύπλοκες προσομοιώσεις φυσικής, τη βιολογία και τη περιβαλλοντική έρευνα.

Η αύξηση στην ποσότητα των διαθέσιμων στοιχείων παρουσιάζει ευκαιρίες αλλά και προβλήματα. Σε γενικές γραμμές, έχοντας περισσότερα δεδομένα σχετικά με τους πελάτες κάποιου (και πιθανούς πελάτες) θα πρέπει να επιτρέπει στις εταιρείες να προσαρμόσουν καλύτερα τα προϊόντα τους και τις προσπάθειες μάρκετινγκ, προκειμένου να δημιουργήσουν ένα υψηλότερο επίπεδο ικανοποίησης και συνέχεια συνεργασιών. Οι εταιρείες που μπορούν να εισπράξουν μεγάλο μέρος των δεδομένων τους παρέχετε η ευκαιρία να διεξαχθεί βαθύτερη και πλουσιότερη ανάλυση.

Αν και η καλύτερη ανάλυση είναι κάτι θετικό, τα Big Data μπορούν επίσης να δημιουργήσουν υπερφόρτωση και θόρυβο. Οι εταιρείες πρέπει να είναι σε θέση να χειριστούν μεγαλύτερο όγκο δεδομένων, καθώς προσδιορίζουν ποια δεδομένα αντιπροσωπεύουν τα σήματα σε σύγκριση με το θόρυβο. Ο προσδιορισμός του τι κάνει σχετικά τα δεδομένα γίνεται ένας βασικός παράγοντας. Δομημένα δεδομένα, που αποτελούνται από αριθμητικές τιμές, μπορούν εύκολα να αποθηκευτούν και να ταξινομηθούν. Αδόμητα δεδομένα, όπως ηλεκτρονικά μηνύματα, βίντεο και έγγραφα κειμένου, μπορεί να απαιτούν πιο εξελιγμένες τεχνικές που πρέπει να εφαρμοστούν πριν τα δεδομένα γίνουν χρήσιμα.

Τα Big Data πιο συχνά αποθηκεύονται σε βάσεις δεδομένων ηλεκτρονικών υπολογιστών, και αναλύονται χρησιμοποιώντας το λογισμικό που έχει σχεδιαστεί ειδικά για να χειριστεί μεγάλα, σύνθετα σύνολα δεδομένων. Πολλές Software-as-a-Service (SaaS) εταιρείες ειδικεύονται στη διαχείριση αυτού του είδους πολύπλοκων δεδομένων. Οι αναλυτές δεδομένων εξετάζουν τη σχέση μεταξύ των διαφόρων τύπων δεδομένων, όπως δημογραφικά δεδομένα και ιστορικά αγορών, για να καθοριστεί αν υπάρχει συσχέτιση.

Σχεδόν κάθε τμήμα σε μια εταιρεία μπορεί να αξιοποιήσει τα ευρήματα από την ανάλυση των δεδομένων: από το ανθρώπινο δυναμικό και τη τεχνολογία, έως την εμπορία και τις πωλήσεις.

Κεφάλαιο 3.Επιστήμη των δεδομένων

Ένα πεδίο των Big Data, που επιδιώκει να προμηθεύσει σημαντικές πληροφορίες από μεγάλες ποσότητες σύνθετων δεδομένων. Η επιστήμη των δεδομένων συνδυάζει διαφορετικούς τομείς εργασίας των στατιστικών στοιχείων και του υπολογισμού, προκειμένου να ερμηνεύσουν τα δεδομένα για το σκοπό της λήψης αποφάσεων.

Η ενσωμάτωση της τεχνολογίας στην καθημερινή μας ζωή έχει καταστεί δυνατή από τη διαθεσιμότητα των δεδομένων σε τεράστιες ποσότητες. Τα δεδομένα προέρχονται από διάφορους τομείς και πλατφόρμες συμπεριλαμβανομένων των κινητών τηλεφώνων, τα κοινωνικά μέσα, διαφημιστικές ιστοσελίδες, έρευνες της υγειονομικής περίθαλψης, αναζητήσεις στο διαδίκτυο, κ.λπ. Η αύξηση στην ποσότητα των διαθέσιμων στοιχείων άνοιξε τη πόρτα σε ένα νέο πεδίο μελέτης που ονομάζεται Big Data το οποίο αναφέρεται στο τεράστιο όγκο των διαθέσιμων πληροφοριών που μπορεί να αξιοποιηθεί για να παραχθούν ακόμα καλύτερα εργαλεία για τις επιχειρήσεις σε όλους τους τομείς, συμπεριλαμβανομένων των μεταφορών, τη χρηματοδότηση, την κατασκευή, και τη ρύθμιση. Τα συνεχώς αυξανόμενα σύνολα δεδομένων και η εύκολη πρόσβαση στα δεδομένα γίνεται δυνατή από μια συνεργασία των εταιρειών γνωστή ως fintech που χρησιμοποιούν την τεχνολογία για να καινοτομούν και να ενισχύσουν τα παραδοσιακά χρηματοοικονομικά προϊόντα και υπηρεσίες. Τα δεδομένα που παράγονται χρησιμοποιούνται για να δημιουργήσει ακόμα περισσότερα δεδομένα που μοιράζονται εύκολα ανάμεσα σε όλες τις οντότητες χάρη στα αναδυόμενα προϊόντα fintech όπως το “cloud computing and storage”. Ωστόσο, η ερμηνεία των τεράστιων ποσοτήτων των αδόμητων δεδομένων για την αποτελεσματική λήψη αποφάσεων μπορεί να αποδειχθεί υπερβολικά πολύπλοκη και χρονοβόρα για τις επιχειρήσεις, ως εκ τούτου, η εμφάνιση της επιστήμης των δεδομένων.

Η επιστήμη των δεδομένων ενσωματώνει εργαλεία από πολλούς κλάδους για τη συγκέντρωση ενός συνόλου δεδομένων, την επεξεργασία και άντληση ιδεών από το σύνολο των δεδομένων, την εξαγωγή χρήσιμων δεδομένων από το σύνολο και την ερμηνεία τους για σκοπούς λήψης αποφάσεων. Οι πειθαρχικοί τομείς που συνθέτουν το πεδίο της επιστήμης των δεδομένων περιλαμβάνουν την εξόρυξη, τη στατιστική, machine learning, analytics, και το προγραμματισμό. Η εξόρυξη δεδομένων εφαρμόζει αλγορίθμους στα σύνθετα δεδομένα που αποκαλύπτουν μοτίβα τα οποία στη συνέχεια χρησιμοποιούνται για την εξαγωγή αξιοποιήσιμων και των σχετικών δεδομένων από το σύνολο. Τα στατιστικά μέτρα όπως τα predictive analytics χρησιμοποιούν τα εξαγόμενα δεδομένα για να αξιολογήσουν γεγονότα που είναι πιθανό να συμβούν στο μέλλον, με βάση αυτά τα στοιχεία που δείχνουν τη συνέβη στο παρελθόν. Machine learning είναι ένα εργαλείο τεχνητής νοημοσύνης που επεξεργάζεται μαζικές ποσότητες στοιχείων που ένας άνθρωπος δε μπορεί να επεξεργαστεί στη διάρκεια μιας ζωής. Το Machine learning τελειοποιεί το μοντέλο απόφασης που έχει υποβληθεί στο πλαίσιο των predictive analytics που ταιριάζει με την πιθανότητα ενός γεγονότος να συμβεί σε ό, τι πραγματικά συνέβη κατά την προβλεπόμενη ώρα. Σύμφωνα με το Analytics, ο αναλυτής δεδομένων συλλέγει και επεξεργάζεται τα δομημένα δεδομένα από το στάδιο του machine learning με τη χρήση αλγορίθμων. Ερμηνεύει, μετατρέπει, και συνοψίζει τα δεδομένα σε μια συνεκτική γλώσσα που η ομάδα λήψης αποφάσεων μπορεί να καταλάβει. Αυτές οι περιοχές που αναφέρονται δεν είναι μια πλήρη λίστα με το τι περιλαμβάνει η επιστήμη των δεδομένων. Καθώς ο ρόλος του επιστήμονα δεδομένων γίνεται καλύτερα κατανοητός, περισσότερα σύνολα ικανότητας θα προστεθούν στο πεδίο που θα καλύπτουν τομείς όπως η αρχιτεκτονική των δεδομένων, η μηχανική των δεδομένων και διαχειριστής των δεδομένων.

Κεφάλαιο 4.Machine Learning

Η ιδέα ότι ένα πρόγραμμα υπολογιστή μπορεί να μάθει και να προσαρμοστεί σε νέα δεδομένα, χωρίς ανθρώπινη παρέμβαση. Το Machine learning είναι ένα πεδίο της τεχνητής νοημοσύνης που κρατά τους ενσωματωμένους αλγορίθμους ενός υπολογιστή ενημερωμένους, ανεξάρτητα από τις αλλαγές στην παγκόσμια οικονομία.

Οι διάφορες εφαρμογές δεδομένων του machine learning που σχηματίζονται μέσω ενός πολύπλοκου αλγορίθμου ή του πηγαίου κώδικα που είναι ενσωματωμένος στο μηχάνημα ή τον υπολογιστή. Αυτός ο κώδικας προγραμματισμού δημιουργεί ένα μοντέλο το οποίο προσδιορίζει τα δεδομένα και δημιουργεί προβλέψεις γύρω από τα δεδομένα που προσδιορίζει. Το μοντέλο χρησιμοποιεί παραμέτρους που χτίστηκαν μέσα στον αλγόριθμο για να σχηματίσουν μοτίβα για τη διαδικασία λήψης αποφάσεων. Όταν νέα ή πρόσθετα δεδομένα γίνουν διαθέσιμα, ο αλγόριθμος ρυθμίζει αυτόματα τις παραμέτρους για να ελέγξει για μια αλλαγή μοτίβου, εάν υπάρχουν. Ωστόσο, το μοντέλο δεν πρέπει να αλλάξει.

Το πως λειτουργεί το machine learning μπορεί να εξηγηθεί καλύτερα με ένα παράδειγμα στον οικονομικό κόσμο. Παραδοσιακά, οι χαρακτήρες επενδύσεων στην αγορά κινητών αξιών, όπως οι οικονομικοί ερευνητές, αναλυτές, διαχειριστές κεφαλαίων, ιδιώτες επενδυτές αναζητούν μέσα από πολλές πληροφορίες από διαφορετικές εταιρείες σε όλο τον κόσμο για να κάνουν κερδοφόρες επενδυτικές αποφάσεις. Ωστόσο, κάποιες σχετικές πληροφορίες δεν μπορούν να δημοσιοποιηθούν ευρέως από τα μέσα ενημέρωσης και μπορεί να είναι ιδιωτικές σε επίλεκτο κοινό που έχουν το πλεονέκτημα ότι είναι υπάλληλοι της εταιρείας ή κάτοικοι της χώρας από όπου η πληροφορία προέρχεται. Επιπλέον, υπάρχει κάποιο όριο στο πόσες πληροφορίες μπορούν οι άνθρωποι να συλλέξουν και να επεξεργαστούν μέσα σε συγκεκριμένο χρονικό πλαίσιο. Εδώ είναι που μπαίνει το machine learning.

Μια εταιρεία διαχείρισης περιουσιακών στοιχείων μπορεί να χρησιμοποιεί το machine learning στην επενδυτική της ανάλυση και την περιοχή έρευνας. Πείτε πως ο διαχειριστής περιουσιακών στοιχείων επενδύει μόνο σε μετοχές ορυχείων. Το μοντέλο ενσωματωμένο στο σύστημα σαρώνει το World Wide Web και συλλέγει όλα τα είδη των νέων ειδήσεων από τις επιχειρήσεις, βιομηχανίες, πόλεις και χώρες, το σύνολο των δεδομένων που συγκεντρώθηκε περιέχει τις πληροφορίες αυτές. Όλες οι πληροφορίες που εισάγονται στο σύνολο δεδομένων είναι οι πληροφορίες που οι διαχειριστές κεφαλαίων και οι ερευνητές της εταιρείας δεν θα ήταν σε θέση να αποκτήσουν χρησιμοποιώντας όλο το ανθρώπινο δυναμικό και τις διασυνδέσεις τους. Οι παράμετροι που χτίστηκαν παράλληλα με το μοντέλο εξάγουν μόνο τα δεδομένα σχετικά με τις εταιρείες εξόρυξης, ρυθμιστικές πολιτικές στον τομέα της εξερεύνησης, και πολιτικά γεγονότα σε επιλεγμένες χώρες από το σύνολο των δεδομένων. Ας πούμε, μια εταιρεία εξόρυξης XYZ μόλις ανακάλυψε ένα ορυχείο διαμαντιών σε μια μικρή πόλη στη Νότια Αφρική, το machine learning app θα το σημειώσει αυτό, ως σχετικό στοιχείο. Το μοντέλο θα μπορούσε στη συνέχεια να χρησιμοποιήσει ένα εργαλείο ανάλυσης που ονομάζεται predictive analytics για να γίνουν προβλέψεις σχετικά με το εάν η εξορυκτική βιομηχανία θα είναι επικερδής για ένα χρονικό διάστημα, ή ποιες μετοχές εξόρυξης είναι πιθανό να αυξηθούν σε αξία για ένα ορισμένο χρονικό διάστημα. Αυτές οι πληροφορίες αναμεταδίδονται στο διαχειριστή περιουσιακών στοιχείων για να αναλύσει και να λάβει μια απόφαση για το χαρτοφυλάκιό του. Ο διαχειριστής περιουσιακών στοιχείων μπορεί να λάβει μια απόφαση να επενδύσει εκατομμύρια δολάρια σε μετοχές XYZ.

Κεφάλαιο 5.Ανωνυμοποίηση των δεδομένων

Μια τεχνική προστασίας των προσωπικών δεδομένων που αποσκοπεί στην προστασία των ιδιωτικών ή ευαίσθητων δεδομένων με τη διαγραφή ή την κρυπτογράφηση προσωπικών αναγνωρίσιμων πληροφοριών από μια βάση δεδομένων. Ανωνυμοποίηση των δεδομένων γίνεται με σκοπό την προστασία ενός ατόμου ή των ιδιωτικών δραστηριοτήτων μιας εταιρείας, διατηρώντας παράλληλα την ακεραιότητα των δεδομένων που συλλέγονται και μοιράζονται. Επίσης γνωστό ως Συσκότιση δεδομένων, Data Masking και Μη-αναγνώριση δεδομένων.

Ανωνυμοποίηση των δεδομένων πραγματοποιείται από τις περισσότερες βιομηχανίες που ασχολούνται με ευαίσθητες πληροφορίες, όπως η υγειονομική περίθαλψη, οι οικονομικές βιομηχανίες και βιομηχανίες ψηφιακών μέσων, προωθώντας παράλληλα την ακεραιότητα της κοινής χρήσης των δεδομένων. Η ανωνυμοποίηση των δεδομένων μειώνει τον κίνδυνο ακούσιας αποκάλυψης κατά την ανταλλαγή δεδομένων μεταξύ των χωρών, των βιομηχανιών, και ακόμη και τμημάτων εντός της ίδιας εταιρείας. Για παράδειγμα, μια ανταλλαγή εμπιστευτικών δεδομένων για τους ασθενείς ενός νοσοκομείου με ένα ιατρικό ερευνητικό εργαστήριο ή μια φαρμακευτική εταιρεία θεωρείτε ηθικό μόνο αν κρατά τους ασθενείς τους ανώνυμους. Αυτό μπορεί να γίνει με την αφαίρεση των ονομάτων, των αριθμών κοινωνικής ασφάλισης, των ημερομηνιών γέννησης και των διευθύνσεων των ασθενών τους από τη λίστα που θα μοιράσει, αφήνοντας τα σημαντικά συστατικά που απαιτούνται για την ιατρική έρευνα, όπως η ηλικία, ασθένειες, το ύψος, το βάρος, το φύλο, τη φυλή, κ.λπ.

Η ανωνυμοποίηση των δεδομένων σύμφωνα με την οποία διαβαθμισμένες πληροφορίες εξυγιαίνονται και μεταμφιέζονται θα πρέπει να γίνετε με τέτοιο τρόπο ώστε αν συμβεί μια παραβίαση, τα δεδομένα που αποκτούνται να είναι άχρηστα για τους ενόχους. Η ανάγκη για την προστασία των δεδομένων θα πρέπει να διατηρείτε σε υψηλή προτεραιότητα σε κάθε οργάνωση, καθώς διαβαθμισμένες πληροφορίες που πέφτουν σε λάθος χέρια μπορεί να χρησιμοποιηθούν καταχρηστικά, ηθελημένα ή αθέλητα. Έλλειψη ευαισθησίας, στο χειρισμό ευαίσθητων πληροφοριών πελάτη μπορεί να έρθει σε μεγάλο κόστος στις επιχειρήσεις λόγω ρυθμιστικών αρχών που πατάζουν τη βαριά αμέλεια. Νομικές απαιτήσεις και απαιτήσεις συμμόρφωσης, όπως το PCI DSS (Payment Card Industry Data Security Standard) επιβάλουν βαρύ πρόστιμα στα χρηματοπιστωτικά ιδρύματα σε περίπτωση παραβίασης της πιστωτικής κάρτας. PIPEDA, ένας νόμος του Καναδά, διέπει την αποκάλυψη και χρήση των προσωπικών πληροφοριών από εταιρείες. Υπάρχουν και άλλοι πολλαπλοί ρυθμιστικοί φορείς που έχουν συσταθεί για να παρακολουθούν τη χρήση ή κακή χρήση των προσωπικών δεδομένων από τους οργανισμούς.
Η αποκωδικοποίηση ανώνυμων δεδομένων είναι δυνατή μέσω μιας διαδικασίας που είναι γνωστή ως Αντι-Ανωνυμοποίηση (ή Επαναγνώριση). Λόγω του γεγονότος ότι τα ανώνυμα δεδομένα μπορούν να αποκωδικοποιηθούν και να ξεμπλεχτούν, οι κριτικοί πιστεύουν πως η ανωνυμοποίηση προσφέρει μια ψευδή αίσθηση ασφάλειας. Ερευνητές στο MIT και το Πανεπιστήμιο Catholique de Louvain στο Βέλγιο, ανέλυσαν στοιχεία για 1,5 εκατομμύρια χρήστες κινητών τηλεφώνων σε μια μικρή ευρωπαϊκή χώρα για μια χρονική περίοδο 15 μηνών και διαπίστωσαν ότι μόνο τέσσερα σημεία αναφοράς, με αρκετά χαμηλή χωρική και χρονική ανάλυση, ήταν αρκετά για να προσδιορίσουν επακριβώς το 95 τοις εκατό από αυτούς. Με άλλα λόγια, για να εξαγάγετε ολοκληρωμένες πληροφορίες τοποθεσίας για ένα άτομο από "ανώνυμα" σύνολα δεδομένων άνω του ενός εκατομμυρίου ανθρώπων, το μόνο που θα χρειαστεί να κάνετε είναι να τοποθετήσετε το άτομο σε απόσταση περίπου 2 μέτρων

ενός πομπού κινητού τηλεφώνου, κάποια στιγμή κατά τη διάρκεια μιας ώρας, τέσσερις φορές μέσα σε ένα χρόνο. Λίγες δημοσιεύσεις από το Twitter θα παράσχουν κατά πάσα πιθανότητα όλες τις πληροφορίες που χρειάζονται, εάν περιέχουν συγκεκριμένες πληροφορίες για την τύχη του ατόμου.

Η επαναγνώριση αντιστρέφει τη διαδικασία της ανωνυμοποίησης ταιριάζοντας κοινόχρηστα αλλά περιορισμένα σύνολα δεδομένων με σύνολα δεδομένων που είναι εύκολα προσβάσιμα στο διαδίκτυο. Οι εξορύκτες δεδομένων μπορούν στη συνέχεια να ανακτήσουν κάποιες πληροφορίες από κάθε διαθέσιμο σύνολο δεδομένων και να συναρμολογήσουν την ταυτότητα ή τη συναλλαγή ενός ατόμου. Για παράδειγμα, ένας εξορύκτης δεδομένων θα μπορούσε να ανακτήσει ένα σύνολο δεδομένων που μοιράζετε μια εταιρεία τηλεπικοινωνιών, μια ιστοσελίδα κοινωνικών μέσων ενημέρωσης, μια πλατφόρμα ηλεκτρονικού εμπορίου, και ένα δημόσιο αποτέλεσμα απογραφής για να προσδιορίσει το όνομα και συχνές δραστηριότητες ενός χρήστη.

Η επαναγνώριση μπορεί να είναι επιτυχής, όταν νέες πληροφορίες απελευθερώνεται ή όταν η στρατηγική ανωνυμοποίησης που έχει εφαρμοστεί δεν έχει γίνει σωστά. Με μια μεγάλη προμήθεια δεδομένων και τον περιορισμένο διαθέσιμο χρόνο ανά ημέρα, οι αναλυτές δεδομένων και εξορύκτες εφαρμόζουν συντομεύσεις γνωστές ως «heuristics» στη λήψη αποφάσεων. Ενώ τα heuristics εξοικονομούν πολύτιμο χρόνο και πόρους στην έρευνα ενός συνόλου δεδομένων, θα μπορούσε επίσης να δημιουργήσει κενά που θα μπορούσαν να εκμεταλλευτούν εάν είχε εφαρμοστεί το λάθος ευρετικό εργαλείο. Αυτά τα κενά θα μπορούσαν να εντοπιστούν από τους εξορύκτες δεδομένων που επιδιώκουν να επαναγνωρίσουν ένα σύνολο δεδομένων είτε για νομικούς ή παράνομους σκοπούς.
Τα Heuristic προέρχονται από ελληνική λέξη που σημαίνει ανακαλύπτω.

Προσωπικά αναγνωρίσιμες πληροφορίες που αποκτήθηκαν παράνομα από τις τεχνικές αντι-ανωνυμοποίησης μπορούν να πωληθούν σε μαύρες αγορές, οι οποίες είναι επίσης μια μορφή πλατφορμών ανωνυμοποίησης. Πληροφορίες που πέφτουν σε λάθος χέρια μπορεί να χρησιμοποιηθούν για εξαναγκασμό, εκβιασμό και εκφοβισμό που οδηγούν σε ανησυχίες για τη προστασία της ιδιωτικής ζωής και τεράστιο κόστος για τις επιχειρήσεις που πέφτουν θύματα.

Η αντι-ανωνυμοποίηση μπορεί επίσης να χρησιμοποιηθεί νόμιμα. Για παράδειγμα, η ιστοσελίδα Silk Road, μια μαύρη αγορά παράνομων ναρκωτικών, φιλοξενήθηκε από ανώνυμο δίκτυο που ονομάζεται Tor το οποίο χρησιμοποιεί μια στρατηγική κρεμμύδι που θολώνει τις διευθύνσεις IP των χρηστών της. Το δίκτυο Tor φιλοξενεί επίσης και άλλες παράνομες αγορές όπως εμπόριο όπλων, εμπόριο κλεμμένων πιστωτικών καρτών, και ευαίσθητες εταιρικές πληροφορίες. Με τη χρήση πολύπλοκων εργαλείων αντι-ανωνυμοποίησης, το FBI ήταν επιτυχής στο σπάσιμο και κλείσιμο του Silk Road και ιστοσελίδων που συμμετείχαν σε παιδική πορνογραφία.

Επιτυχία στις διαδικασίες επαναγνώρισης έχουν αποδείξει ότι η ανωνυμία δεν είναι εγγυημένη. Ακόμα κι αν πρωτοποριακά εργαλεία ανωνυμίας εφαρμόζονταν σήμερα για να συγκαλύψουν τα δεδομένα, τα δεδομένα θα μπορούσαν να επαναπροσδιοριστούν σε ένα-δύο χρόνια, καθώς νέες τεχνολογίες και νέα σύνολα δεδομένων θα γίνουν διαθέσιμα.

Κεφάλαιο 6.Παραβίαση των δεδομένων

Μια μη εξουσιοδοτημένη πρόσβαση και ανάκτηση ευαίσθητων πληροφοριών από ένα άτομο, μια ομάδα ή σύστημα λογισμικού. Η παραβίαση δεδομένων είναι ένα ατύχημα ασφαλείας στον κυβερνοχώρο το οποίο συμβαίνει όταν τα δεδομένα, ηθελημένα ή αθέλητα, πέφτουν σε λάθος χέρια, χωρίς τη γνώση του χρήστη ή του ιδιοκτήτη. Επίσης γνωστό ως διαρροή δεδομένων.

Μια παραβίαση των δεδομένων μπορεί να πραγματοποιηθεί ακούσια ή εκούσια. Μια ακούσια παραβίαση των δεδομένων συμβαίνει όταν ένας νόμιμος φύλακας των πληροφοριών, όπως ένας υπάλληλος που χάνει ή χρησιμοποιεί εταιρικά εργαλεία εξ αμελείας. Ένας εργαζόμενος που έχει πρόσβαση σε ανασφάλιστες ιστοσελίδες, που κατεβάζει ένα επικίνδυνο πρόγραμμα λογισμικού σε ένα φορητό υπολογιστή εργασίας, που συνδέεται σε ένα μη ασφαλές δίκτυο WiFi, που χάνει το φορητό υπολογιστή ή το κινητό του σε μια δημόσια θέση, κλπ, διατρέχει τον κίνδυνο παραβίασης των στοιχείων της εταιρείας του. Το 2015, η Nutmeg, μια εταιρεία διαχείρισης επενδύσεων online, είχε τα δεδομένα τις εκτεθειμένα, όταν ένας εσφαλμένος κώδικας στο σύστημα οδήγησε σε αποστολή μέσω του ηλεκτρονικού ταχυδρομείου τις προσωπικά αναγνωρίσιμες πληροφορίες (PII) 32 λογαριασμών σε λάθος παραλήπτες. Οι πληροφορίες που στάλθηκαν περιλάμβαναν τα ονόματα, τις διευθύνσεις και λεπτομέρειες των επενδύσεων και έβαλαν τους κατόχους λογαριασμών σε κίνδυνο κλοπής ταυτότητας.

Μια σκόπιμη παραβίαση των δεδομένων συμβαίνει όταν ένας εισβολέας στο κυβερνοχώρο χακάρει το σύστημα ενός ατόμου ή μιας εταιρείας με σκοπό την πρόσβαση σε αποκλειστικές και προσωπικές πληροφορίες. Οι χάκερς στο κυβερνοχώρο χρησιμοποιούν μια ποικιλία τρόπων για να μπουν σε ένα σύστημα. Μερικοί ενθέτουν κακόβουλο λογισμικό σε δικτυακούς τόπους ή σε συνημμένα ηλεκτρονικού ταχυδρομείου που, όταν γίνει πρόσβαση, κάνει το σύστημα του υπολογιστή ευάλωτο καταλήγοντας σε εύκολη είσοδο και προσβασιμότητα των δεδομένων από χάκερ. Μερικοί χάκερς χρησιμοποιούν botnets, τα οποία είναι μολυσμένοι υπολογιστές, για πρόσβαση σε αρχεία άλλων υπολογιστών. Τα Botnets επιτρέπουν στους δράστες να αποκτήσουν πρόσβαση σε πολλούς υπολογιστές ταυτόχρονα, χρησιμοποιώντας το ίδιο κακόβουλο λογισμικό εργαλείο. Οι χάκερς μπορούν επίσης να χρησιμοποιήσουν μια επίθεση της αλυσίδας εφοδιασμού για να έχουν πρόσβαση σε πληροφορίες. Όταν μια εταιρεία έχει ένα σταθερό και αδιαπέραστο μέτρο ασφαλείας στη θέση του, ένας χάκερ μπορεί να περάσει μέσα από ένα μέλος του δικτύου της εφοδιαστικής αλυσίδας της εταιρείας που έχει ένα ευάλωτο σύστημα ασφαλείας. Μόλις ο χάκερ μπει στο σύστημα του υπολογιστή του μέλους, μπορεί να αποκτήσει πρόσβαση στο δίκτυο της εταιρείας που είχε θέσει ως στόχο.

Οι ιδιοκτήτες και οι χρήστες ενός παραβιασμένου συστήματος ή δικτύου δεν γνωρίζουν πάντα αμέσως, όταν συνέβη η παραβίαση. Το 2016, το Yahoo ανακοίνωσε τι θα μπορούσε να είναι η μεγαλύτερη παραβίαση ασφαλείας στον κυβερνοχώρο όταν ισχυρίστηκε ότι περίπου 500 εκατομμύρια λογαριασμοί παραβιάστηκαν. Περαιτέρω έρευνα αποκάλυψε ότι η παραβίαση των δεδομένων είχε πραγματικά συνέβη δύο χρόνια πριν από το 2014.

Κεφάλαιο 7.Cloud Computing

Το cloud computing είναι ένα μοντέλο για την παροχή υπηρεσιών πληροφορικής στο οποίο οι πόροι ανακτώνται από το διαδίκτυο μέσω εργαλείων και εφαρμογών που βασίζονται στο διαδίκτυο παρά από μια άμεση σύνδεση με ένα διακομιστή. Τα πακέτα δεδομένων και λογισμικού αποθηκεύονται σε διακομιστές, ωστόσο, μια δομή cloud computing επιτρέπει την πρόσβαση σε πληροφορίες εφόσον μια ηλεκτρονική συσκευή έχει πρόσβαση στο διαδίκτυο. Αυτός ο τύπος συστήματος επιτρέπει στους εργαζόμενους να εργάζονται εξ’ αποστάσεως.

Το cloud computing ονομάστηκε έτσι, επειδή οι πληροφορίες στις οποίες έχουμε πρόσβαση βρίσκονται στο "σύννεφο" και δεν απαιτούν από τον χρήστη να είναι σε ένα συγκεκριμένο μέρος για να έχουν πρόσβαση σε αυτές. Οι εταιρείες έχουν διαπιστώσει ότι το cloud computing τους επιτρέπει τη μείωση του κόστους της διαχείρισης των πληροφοριών, δεδομένου ότι δεν είναι υποχρεωμένοι να έχουν δικούς τους διακομιστές και μπορούν να χρησιμοποιήσουν χώρο μισθωμένα από τρίτους. Επιπλέον, η cloud-like δομή επιτρέπει στις εταιρείες να αναβαθμίσουν το λογισμικό πιο γρήγορα.

Πριν το σύννεφο γίνει μια βιώσιμη εναλλακτική λύση, οι εταιρείες έπρεπε να αγοράσουν, να κατασκευάσουν και να συντηρήσουν δαπανηρές υποδομές για τη τεχνολογία των πληροφοριών (IT). Η πρόσβαση λογισμικού μέσα από το σύννεφο εξαλείφει τα προβλήματα διοικητικής μέριμνας και παρέχει άμεσα διαθέσιμες πλατφόρμες για τους χρήστες σε όλο το επεκτατικό γεωγραφικό φάσμα. Έτσι, ο ρυθμός με τον οποίο οι επιχειρήσεις υιοθετούν και χρησιμοποιούν τα συστήματα που βασίζονται στο Διαδίκτυο έχει επιταχυνθεί. Η Oracle Corporation απέκτησε 3.600 πελάτες και 690 εκατομμύρια $ το τέταρτο τρίμηνο του 2015 από επιχείρηση που διαχειρίζονταν "σύννεφο".

Ένα από τα κύρια πλεονεκτήματα του cloud computing εκτείνεται σε εταιρείες λογισμικού που μπορούν να προσφέρουν τα προϊόντα τους μέσω του διαδικτύου παρά μέσω πιο παραδοσιακών μεθόδων που αφορούν δίσκους ή άλλα υλικά μέσα. Το 2013, η Adobe Systems ανακοίνωσε πως όλες οι επόμενες εκδόσεις του Photoshop, καθώς και άλλα στοιχεία του Creative Suite της, θα είναι διαθέσιμες μόνο μέσω μιας συνδρομής στο Διαδίκτυο. Το Photoshop χρησιμοποιεί το "σύννεφο" σαν αποθήκη, αλλά η επεξεργασία δεν ολοκληρώνεται μέσω του διαδικτύου.

Οι cloud computing πλατφόρμες επιτρέπουν τις λογισμικό ως υπηρεσία (SaaS) εταιρείες να κατέχουν πολλά πλεονεκτήματα σε σχέση με το λογισμικό εφαρμογών που αναπτύσσεται μέσα από φυσικούς τρόπους. Μία από τις πιο ευρέως διαδεδομένες εφαρμογές είναι η τηλεδιάσκεψη ή η συγκέντρωση εξ’ αποστάσεως συμμετεχόντων στη σύσκεψη που μοιράζονται δυνατότητες ήχου, βίντεο και παρουσίαση μέσω του διαδικτύου. Η εφαρμογή GoToMeeting της Citrix Systems επιτρέπει στους χρήστες να κατεβάσουν γρήγορα μια εφαρμογή που δημιουργεί ένα απομακρυσμένο περιβάλλον σύσκεψης με την οποία οι συμμετέχοντες συνεργάζονται ανεξάρτητα από την τοποθεσία. Οι μόνες απαιτήσεις υλικού και λογισμικού είναι ένα desktop ή κινητή υπολογιστική συσκευή και μια σύνδεση στο internet.

Κεφάλαιο 8.Αποθήκευση Δεδομένων

Η αποθήκευση των δεδομένων είναι η ηλεκτρονική αποθήκευση ενός μεγάλου όγκου πληροφοριών από μια επιχείρηση. Τα αποθηκευμένα δεδομένα πρέπει να αποθηκεύονται κατά τρόπο που να είναι ασφαλή, αξιόπιστα, εύκολα ανακτήσιμα και εύκολα διαχειρίσιμα. Η έννοια της αποθήκευσης δεδομένων προέρχεται από το 1988 με το έργο των ερευνητών της IBM Barry Devlin και Paul Murphy. Η ανάγκη για αποθήκη δεδομένων εξελίχθηκε καθώς τα συστήματα ηλεκτρονικών υπολογιστών έγιναν πιο περίπλοκα και χειρίζονταν αυξανόμενες ποσότητες δεδομένων.

Οι επιχειρήσεις μπορεί να αποθηκεύουν δεδομένα για χρήση στην εξερεύνηση και στην εξόρυξη δεδομένων, αναζητώντας μοτίβα πληροφοριών που θα τους βοηθήσουν να βελτιώσουν τις επιχειρήσεις τους. Ένα καλό σύστημα αποθήκευσης δεδομένων μπορεί επίσης να κάνει ευκολότερη για τα διαφορετικά τμήματα μέσα σε μια εταιρεία τη πρόσβαση σε δεδομένα άλλου τμήματος. Για παράδειγμα, μια αποθήκη δεδομένων θα μπορούσε να επιτρέψει στο CEO μιας εταιρείας να εξετάσει εύκολα τα δεδομένα της ομάδας πωλήσεων και να τον βοηθήσει να πάρει αποφάσεις για το πώς να βελτιώσουν τις πωλήσεις ή για τον εξορθολογισμό του τμήματος. Η αποτελεσματική αποθήκευση και διαχείριση των δεδομένων είναι επίσης αυτό που κάνει τα πράγματα όπως τις κρατήσεις ταξιδιών και τη χρήση αυτόματων ταμειολογιστικών μηχανών δυνατόν.

Η αποθήκευση δεδομένων περιλαμβάνει τον καθαρισμό των δεδομένων, την ενσωμάτωση των δεδομένων και την ενοποίηση των δεδομένων. Για την ενσωμάτωση ετερογενών βάσεων δεδομένων, έχουμε τις εξής δύο προσεγγίσεις:

- **Προσέγγιση με γνώμονα την ερώτηση**
- **Προσέγγιση με γνώμονα την ενημέρωση**

Προσέγγιση με γνώμονα την ερώτηση

Αυτή είναι η παραδοσιακή προσέγγιση για την ενσωμάτωση ετερογενών βάσεων δεδομένων. Αυτή η προσέγγιση χρησιμοποιείται για την κατασκευή καλυμμάτων και ολοκληρωτών στην κορυφή των πολλαπλών ετερογενών βάσεων δεδομένων. Αυτοί οι ολοκληρωτές είναι επίσης γνωστοί ως μεσολαβητές.

1. Όταν ένα ερώτημα εκδίδεται από τη πλευρά του πελάτη, ένα λεξικό μεταδεδομένων μεταφράζει το ερώτημα σε ερωτήματα, κατάλληλα για τη προσωπική ετερογενή ιστοσελίδα που εμπλέκεται.
2. Τώρα αυτά τα ερωτήματα χαρτογραφούνται και αποστέλλονται στον τοπικό επεξεργαστή ερωτημάτων.
3. Τα αποτελέσματα από ετερογενείς ιστοσελίδες ενσωματώνονται σε ένα παγκόσμιο σύνολο απαντήσεων.

Αυτή η προσέγγιση έχει τα ακόλουθα μειονεκτήματα:

- Η προσέγγιση με γνώμονα την ερώτηση χρειάζεται πολύπλοκη ενσωμάτωση και διεργασίες φιλτραρίσματος.
- Είναι πολύ αναποτελεσματικό και πολύ ακριβό για συχνές ερωτήσεις.
- Αυτή η προσέγγιση είναι ακριβή για ερωτήματα που απαιτούν συσσωμάτωση.

Προσέγγιση με γνώμονα την ενημέρωση

Τα σημερινά συστήματα αποθήκευσης δεδομένων ακολουθούν την προσέγγιση με γνώμονα την ενημέρωση παρά την παραδοσιακή προσέγγιση που συζητήθηκε νωρίτερα. Στην προσέγγιση με γνώμονα την ενημέρωση, οι πληροφορίες από πολλαπλές ετερογενείς πηγές ενσωματώνεται εκ των προτέρων και αποθηκεύεται σε μια αποθήκη. Οι πληροφορίες αυτές είναι διαθέσιμες για άμεση επερώτηση και ανάλυση.

Η προσέγγιση αυτή έχει τα εξής πλεονεκτήματα:

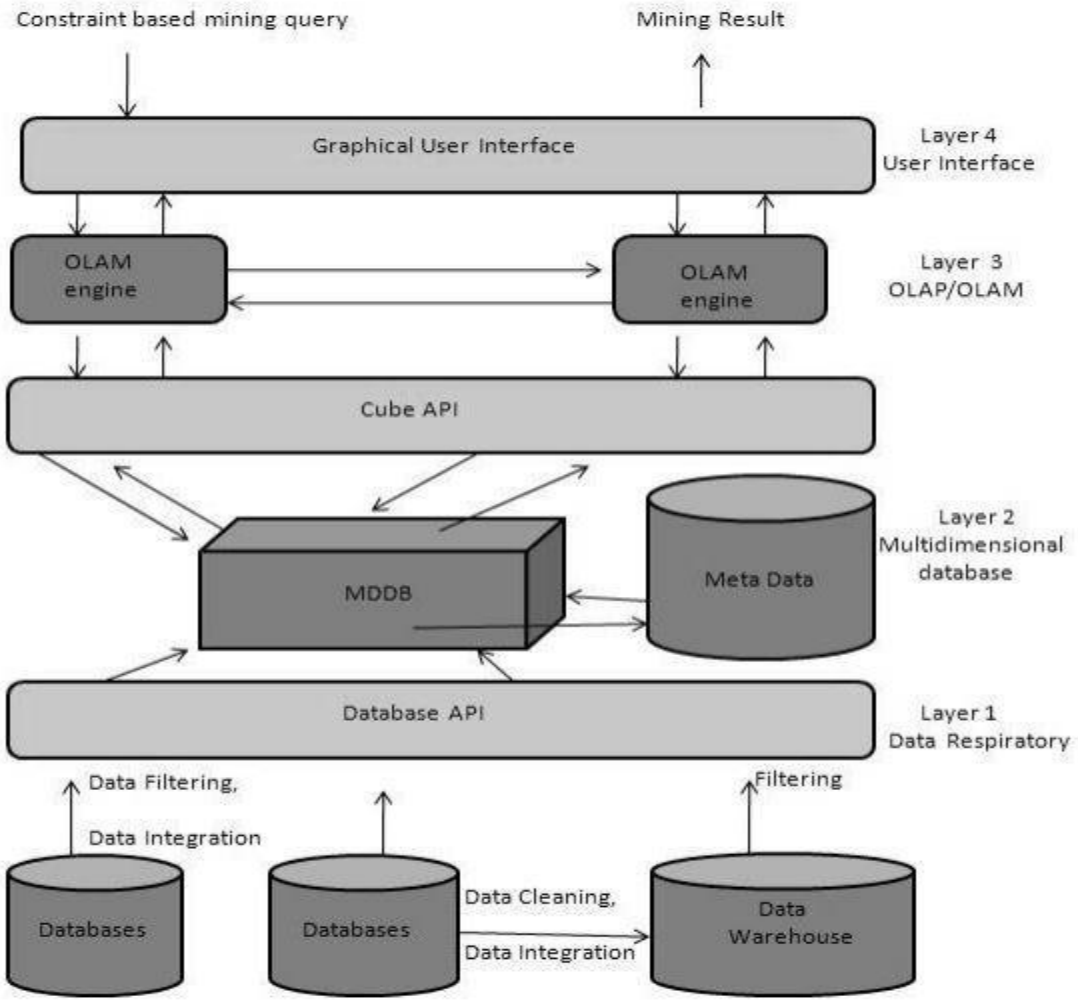
- Η προσέγγιση αυτή παρέχει υψηλή απόδοση.
- Τα δεδομένα μπορούν να αντιγραφούν, επεξεργαστούν, ενσωματωθούν, σχολιαστούν, συνοψισθούν και αναδιαρθρωθούν στη σημασιολογική μνήμη των δεδομένων εκ των προτέρων.

Η επεξεργασία ερωτημάτων δεν απαιτεί διασύνδεση με την επεξεργασία σε τοπικές πηγές.

Από Αποθήκευση Δεδομένων (OLAP) σε Εξόρυξη Δεδομένων (OLAM)

Η ηλεκτρονική αναλυτική εξόρυξη ενσωματώνεται με την ηλεκτρονική ανάλυση επεξεργασίας, την εξόρυξη δεδομένων και την εξόρυξη γνώσης σε πολυδιάστατες βάσεις δεδομένων. Εδώ είναι το διάγραμμα που δείχνει την ένταξη των δύο OLAP και OLAM:

https://www.tutorialspoint.com/data_mining/data_mining_tutorial.pdf



Το OLAM είναι σημαντικό για τους ακόλουθους λόγους:

- **Υψηλή ποιότητα δεδομένων σε αποθήκες δεδομένων** - Τα εργαλεία εξόρυξης δεδομένων για να εργαστούν απαιτούνται ολοκληρωμένα, συνεκτικά, και καθαρά δεδομένα. Αυτά τα βήματα είναι πολύ δαπανηρά στην προεπεξεργασία των δεδομένων. Οι αποθήκες δεδομένων κατασκευασμένες με τέτοια προεπεξεργασία είναι πολύτιμες πηγές υψηλής ποιότητας δεδομένων για την OLAP καθώς και την εξόρυξη δεδομένων.
- **Διαθέσιμες υποδομές επεξεργασίας πληροφοριών που αφορούν τις αποθήκες δεδομένων** – Οι υποδομές επεξεργασίας πληροφοριών αναφέρονται στην πρόσβαση, ενσωμάτωση, στην ενοποίηση και το μετασχηματισμό των πολλαπλών ετερογενών βάσεων δεδομένων, στην πρόσβαση στο διαδίκτυο και στις εγκαταστάσεις εξυπηρέτησης, στην υποβολή εκθέσεων και στα εργαλεία ανάλυσης OLAP.
- **Διερευνητική ανάλυση των δεδομένων που βασίζεται σε OLAP**- Η διερευνητική ανάλυση των δεδομένων απαιτείται για την αποτελεσματική εξόρυξη δεδομένων. Το OLAM παρέχει τη δυνατότητα για εξόρυξη δεδομένων σε διάφορα υποσύνολα των δεδομένων και σε διαφορετικά επίπεδα αφαίρεσης.
- **Ηλεκτρονική επιλογή των λειτουργιών για εξόρυξη δεδομένων** – Η ενσωμάτωση του OLAP με πολλαπλές λειτουργίες εξόρυξης δεδομένων και ηλεκτρονική αναλυτική εξόρυξη παρέχει στους χρήστες την ευελιξία να επιλέξουν τις επιθυμητές λειτουργίες εξόρυξης δεδομένων και να αλλάζουν εργασίες εξόρυξης δεδομένων δυναμικά.

Μια αποθήκη δεδομένων παρουσιάζει τα ακόλουθα χαρακτηριστικά για την υποστήριξη της διαδικασίας λήψης αποφάσεων της διοίκησης:

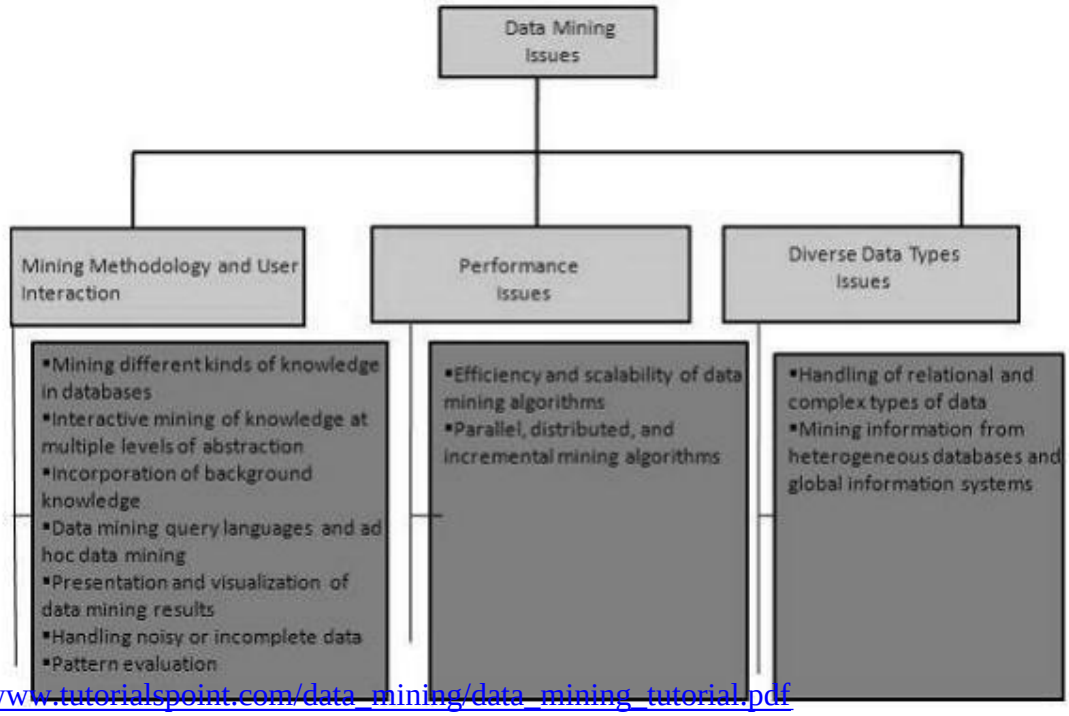
- **Προσανατολισμένη στο θέμα** – Η αποθήκη δεδομένων είναι προσανατολισμένη στο θέμα, διότι μας παρέχει τις πληροφορίες γύρω από ένα θέμα και όχι τις συνεχιζόμενες δραστηριότητες του οργανισμού. Αυτά τα θέματα μπορεί να είναι προϊόν, πελάτες, προμηθευτές, πωλήσεις, έσοδα, κ.λπ. Η αποθήκη δεδομένων δεν επικεντρώνεται στις συνεχιζόμενες δραστηριότητες, αλλά εστιάζει στην μοντελοποίηση και ανάλυση των δεδομένων για τη λήψη αποφάσεων.
- **Ενσωματωμένη** – Η αποθήκη δεδομένων κατασκευάζεται από την ενσωμάτωση δεδομένων από ετερογενείς πηγές, όπως σχεσιακές βάσεις δεδομένων, επίπεδα, αρχεία κ.λπ. Αυτή η ενσωμάτωση ενισχύει την αποτελεσματική ανάλυση των δεδομένων.
- **Παραλλαγή του χρόνου** - Τα δεδομένα που συλλέγονται σε μια αποθήκη δεδομένων ταυτίζονται με μια συγκεκριμένη χρονική περίοδο. Τα δεδομένα σε μια αποθήκη δεδομένων παρέχουν πληροφορίες από ιστορική άποψη.
- **Μη πτητική** - Μη πτητική σημαίνει ότι τα προηγούμενα δεδομένα δεν αφαιρούνται όταν νέα δεδομένα προστίθενται σε αυτή. Η αποθήκη δεδομένων διατηρείται χωριστά από το επιχειρησιακή βάση δεδομένων, επομένως, οι συχνές αλλαγές στην επιχειρησιακή βάση δεδομένων δεν αντικατοπτρίζεται στην αποθήκη δεδομένων.

Κεφάλαιο 9.Ζητήματα εξόρυξης δεδομένων

Η εξόρυξη δεδομένων δεν είναι ένα εύκολο έργο, καθώς οι αλγόριθμοι που χρησιμοποιούνται μπορεί να είναι πολύ περίπλοκοι και τα δεδομένα δεν είναι πάντα διαθέσιμα σε ένα μέρος. Θα πρέπει να ενσωματωθούν από διάφορες ετερογενείς πηγές δεδομένων. Αυτοί οι παράγοντες δημιουργούν επίσης ορισμένα ζητήματα. Θα συζητήσουμε τα σημαντικά θέματα που αφορούν:

- Μεθοδολογία εξόρυξης και αλληλεπίδραση χρήστη
- Ζητήματα απόδοσης
- Ποικίλα θέματα Τύπων Δεδομένων

Το παρακάτω διάγραμμα περιγράφει τα σημαντικά θέματα.



https://www.tutorialspoint.com/data_mining/data_mining_tutorial.pdf

Μεθοδολογία εξόρυξης και αλληλεπίδραση χρήστη

Αναφέρεται στα ακόλουθα είδη θεμάτων:

- **Εξόρυξη διαφόρων ειδών γνώσης από βάσεις δεδομένων** - Οι διάφοροι χρήστες μπορεί να ενδιαφέρονται για διαφορετικά είδη γνώσης. Συνεπώς, είναι απαραίτητο για την εξόρυξη δεδομένων να καλύψει ένα ευρύ φάσμα για το έργο της ανακάλυψης της γνώσης.
- **Διαδραστική εξόρυξη γνώσης σε πολλαπλά επίπεδα αφαίρεσης** - Η διαδικασία εξόρυξης δεδομένων πρέπει να είναι διαδραστική, διότι επιτρέπει στους χρήστες να εστιάσουν στην αναζήτηση για μοτίβα, παρέχοντας και διυλίζοντας αιτήματα εξόρυξης δεδομένων με βάση τα αποτελέσματα που επιστράφηκαν.
- **Ενσωμάτωση της γνώσης υποβάθρου** - Για την καθοδήγηση της διαδικασίας ανακάλυψης και για να εκφράσουν τα πρότυπα που ανακαλύπτονται, το γνωστικό υπόβαθρο μπορεί να χρησιμοποιηθεί. Το γνωστικό υπόβαθρο μπορεί να χρησιμοποιηθεί για να εκφράσει τα μοτίβα που βρεθήκαν όχι μόνο σε συνοπτική άποψη, αλλά σε πολλαπλά επίπεδα αφαίρεσης.
- **Γλώσσες επερωτήσεων για εξόρυξη δεδομένων και ad hoc εξόρυξη δεδομένων** – Η γλώσσα επερωτήσεων για εξόρυξη δεδομένων που επιτρέπει στο χρήστη να

περιγράψει ad hoc εργασίες εξόρυξης, θα πρέπει να ενσωματωθεί με μια γλώσσα επερωτήσεων για αποθήκη δεδομένων και να βελτιστοποιηθεί για αποτελεσματική και ευέλικτη εξόρυξη δεδομένων.

- **Παρουσίαση και οπτικοποίηση των αποτελεσμάτων της εξόρυξης δεδομένων** - Μόλις τα μοτίβα ανακαλυφθούν πρέπει να εκφραστούν σε γλώσσα υψηλού επιπέδου, και οπτικές αναπαραστάσεις. Οι παραστάσεις αυτές θα πρέπει να είναι εύκολα κατανοητές.
- **Χειρισμός θορυβώδη ή ελλιπή δεδομένων** - Οι μέθοδοι καθαρισμού των δεδομένων απαιτούνται για τον χειρισμό του θορύβου και των ελλιπή αντικειμένων, ενώ γίνεται εξόρυξη των κανονικοτήτων των δεδομένων. Εάν οι μέθοδοι καθαρισμού των δεδομένων δεν είναι εκεί, τότε η ακρίβεια των μοτίβων που θα βρεθούν θα είναι κακή.
- **Αξιολόγηση μοτίβων** - Τα μοτίβα που βρέθηκαν πρέπει να είναι ενδιαφέροντα επειδή είτε παρουσιάζουν κοινή γνώση ή έχουν έλλειψη καινοτομίας.

Ζητήματα απόδοσης

Μπορεί να υπάρχουν ζητήματα που σχετίζονται με την απόδοση, όπως τα εξής:

- **Αποτελεσματικότητα και επεκτασιμότητα των αλγορίθμων εξόρυξης δεδομένων** - Για να εξάγει αποτελεσματικά τις πληροφορίες από τεράστιο ποσό δεδομένων σε βάσεις δεδομένων, ο αλγόριθμος εξόρυξης δεδομένων θα πρέπει να είναι αποτελεσματικός και επεκτάσιμος.
- **Παράλληλοι, κατανεμημένοι και αυξητικοί αλγόριθμοι εξόρυξης**– Οι παράγοντες, όπως το τεράστιο μέγεθος των βάσεων δεδομένων, η ευρεία διανομή των δεδομένων, και η πολυπλοκότητα των μεθόδων εξόρυξης δεδομένων παρακινούν την ανάπτυξη των παράλληλων και κατανεμημένων αλγορίθμων εξόρυξης των δεδομένων. Αυτοί οι αλγόριθμοι χωρίζουν τα δεδομένα σε κατατμήσεις που υφίσταται περαιτέρω επεξεργασία με παράλληλο τρόπο. Μετά το αποτελέσματα από τις κατατμήσεις συγχωνεύεται. Οι αυξητικοί αλγόριθμοι, ενημερώνουν τις βάσεις δεδομένων, χωρίς να γίνει εξόρυξη των δεδομένων πάλι από την αρχή.

Ποικίλα θέματα τύπων δεδομένων

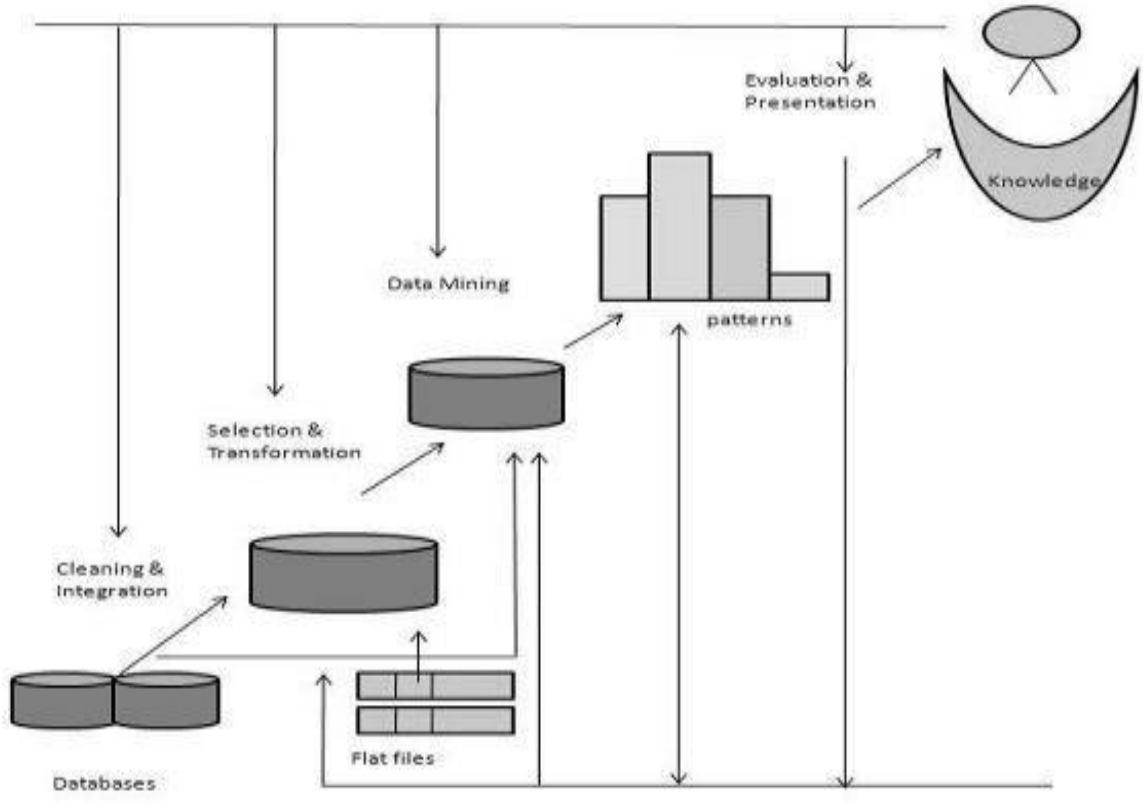
- **Χειρισμός σχεσιακών και σύνθετων τύπων δεδομένων** - Η βάση δεδομένων μπορεί να περιέχει πολύπλοκα αντικείμενα δεδομένων, αντικείμενα δεδομένων πολυμέσων, χωρικά δεδομένα, χρονικά δεδομένα κλπ. Δεν είναι δυνατόν για ένα σύστημα να εξορύξει όλα αυτά τα είδη των δεδομένων.
- **Εξόρυξη πληροφοριών από ετερογενείς βάσεις δεδομένων και παγκόσμια πληροφοριακά συστήματα** - Τα δεδομένα είναι διαθέσιμα σε διάφορες πηγές δεδομένων σε LAN ή WAN. Αυτές οι πηγές δεδομένων μπορεί να είναι δομημένες, ημι-δομημένες ή αδόμητες. Ως εκ τούτου, η εξόρυξη γνώσης από αυτές προσθέτει προκλήσεις στην εξόρυξη των δεδομένων.

Κεφάλαιο 10.Ανακάλυψη της γνώσης

Μερικοί άνθρωποι δεν διαφοροποιούν την εξόρυξη δεδομένων από την ανακάλυψη της γνώσης, ενώ άλλοι βλέπουν την εξόρυξη δεδομένων ως ένα ουσιαστικό βήμα στη διαδικασία της ανακάλυψης της γνώσης. Αυτή είναι η λίστα των βημάτων που εμπλέκονται στη διαδικασία ανακάλυψης της γνώσης:

- **Καθαρισμός δεδομένων** - Σε αυτό το βήμα, ο θόρυβος και τα ασυνεπής δεδομένα αφαιρούνται.
- **Ενσωμάτωση Δεδομένων** - Σε αυτό το στάδιο, οι πολλαπλές πηγές δεδομένων συνδυάζονται.
- **Επιλογή δεδομένων** - Σε αυτό το βήμα, τα δεδομένα που σχετίζονται με την εργασία ανάλυσης ανακτώνται από τη βάση δεδομένων.
- **Μετασχηματισμός δεδομένων** - Σε αυτό το βήμα, τα δεδομένα μετασχηματίζονται ή ενοποιούνται σε μορφές κατάλληλες για την εξόρυξη εκτελώντας περιληπτικές ή συναθροιστικές λειτουργίες.
- Εξόρυξη των δεδομένων - Σε αυτό το βήμα, ευφυείς μέθοδοι εφαρμόζονται για την εξαγωγή μοτίβων των δεδομένων.
- Αξιολόγηση μοτίβων - Σε αυτό το βήμα, τα μοτίβα των δεδομένα αξιολογούνται.
- Παρουσίαση της γνώσης - Σε αυτό το βήμα, η γνώση παρουσιάζεται.

Το παρακάτω διάγραμμα δείχνει τη διαδικασία της ανακάλυψης της γνώσης:
https://www.tutorialspoint.com/data_mining/data_mining_tutorial.pdf



Κεφάλαιο 11.Εξόρυξη δεδομένων

Η εξόρυξη δεδομένων ορίζεται ως η εξαγωγή των πληροφοριών από ένα τεράστιο σύνολο δεδομένων. Με άλλα λόγια μπορούμε να πούμε ότι η εξόρυξη δεδομένων είναι η εξόρυξη της γνώσης από τα δεδομένα. Αυτές οι πληροφορίες μπορούν να χρησιμοποιηθούν για οποιαδήποτε από τις ακόλουθες εφαρμογές:

- Ανάλυση της αγοράς
- Ανίχνευση απάτης
- Διατήρηση των πελατών
- Έλεγχος παραγωγής
- Εξερεύνηση της επιστήμης
- Αθλήματα
- Αστρολογία
- Internet Web Surf-Aid

Τα παντοπωλεία είναι γνωστοί χρήστες τεχνικών εξόρυξης δεδομένων. Πολλά σούπερ-μάρκετ προσφέρουν δωρεάν κάρτες μελών για τους πελάτες που τους δίνουν πρόσβαση σε μειωμένες τιμές που δεν είναι διαθέσιμες για τα μη μέλη. Οι κάρτες καθιστούν εύκολο για τα καταστήματα να παρακολουθούν ποιος αγοράζει τι, πότε και σε ποια τιμή. Τα καταστήματα μπορούν στη συνέχεια να χρησιμοποιήσουν αυτά τα δεδομένα, αφού τα αναλύσουν, για πολλούς λόγους, όπως προσφέροντας στους πελάτες κουπόνια που απευθύνονται στις αγοραστικές τους συνήθειες και να αποφασίσουν πότε να θέσουν τα προϊόντα σε προσφορά ή πότε να τα πουλήσουν σε πλήρη τιμή. Η εξόρυξη δεδομένων μπορεί να είναι μια αιτία ανησυχίας όταν μόνο επιλεγμένες πληροφορίες, οι οποίες δεν είναι αντιπροσωπευτικές της συνολικής ομάδας του δείγματος, χρησιμοποιείται για να αποδείξει κάποια υπόθεση.

Εφαρμογές εξόρυξης δεδομένων

Η εξόρυξη δεδομένων είναι πολύ χρήσιμη στους ακόλουθους τομείς:

- Ανάλυση και διαχείριση της αγοράς
- Εταιρική ανάλυση και διαχείριση κινδύνων
- Ανίχνευση απάτης

Ανάλυση και διαχείριση της αγοράς

Παρακάτω αναφέρονται τα διάφορα πεδία της αγοράς όπου χρησιμοποιείται η εξόρυξη των δεδομένων:

- **Προφίλ Πελατών** – Η εξόρυξη δεδομένων βοηθά στο να καθοριστεί τι είδους άνθρωποι αγοράζουν τη είδος προϊόντων.
- **Προσδιορισμός των απαιτήσεων των πελατών** – Η εξόρυξη δεδομένων βοηθά στον εντοπισμό των καλύτερων προϊόντων για διαφορετικούς πελάτες. Χρησιμοποιεί την πρόβλεψη για να βρει τους παράγοντες που μπορεί να προσελκύσουν νέους πελάτες.
- **Ανάλυση συνυφασμένων αγορών** – Η εξόρυξη δεδομένων εκτελεί ενώσεις / συσχετίσεις μεταξύ των πωλήσεων των προϊόντων.

- **Μάρκετινγκ στόχου** - Η εξόρυξη δεδομένων βοηθά στον εντοπισμό ομάδων μοντέλων πελατών που μοιράζονται τα ίδια χαρακτηριστικά, όπως τα συμφέροντα, τις καταναλωτικές συνήθειες, το εισόδημα κ.λπ.
- **Καθορισμός πρότυπου για την αγορά των πελατών** – Η εξόρυξη δεδομένων βοηθά στην καθορισμό μοτίβου αγοράς του πελάτη.
- **Παροχή συνοπτικών πληροφοριών** – Η εξόρυξη δεδομένων μας παρέχει διάφορες πολυδιάστατες συνοπτικές εκθέσεις.

Εταιρική ανάλυση και διαχείριση κινδύνων

Η εξόρυξη δεδομένων χρησιμοποιείτε στα ακόλουθα πεδία του εταιρικού τομέα:

- **Οικονομικό Προγραμματισμό και Αξιολόγηση Κεφαλαίου** - Αφορά την ανάλυση και πρόβλεψη των ταμειακών ροών, έκτακτη απαίτηση ανάλυσης για την αξιολόγηση των περιουσιακών στοιχείων.
- **Προγραμματισμός των Πόρων** - Επιτυγχάνεται συνοψίζοντας και συγκρίνοντας τους πόρους και τις δαπάνες.
- **Ανταγωνισμός** - Περιλαμβάνει τη παρακολούθηση των ανταγωνιστών και των κατευθύνσεων της αγοράς.

Ανίχνευση απάτης

Η εξόρυξη δεδομένων χρησιμοποιείται επίσης στους τομείς των υπηρεσιών πιστωτικών καρτών και τηλεπικοινωνιών για την ανίχνευση της απάτης. Σε τηλεφωνικές κλήσεις που αφορούν απάτη, βοηθά στον εντοπισμό του προορισμού της κλήσης, τη διάρκεια της κλήσης, την ώρα της ημέρας ή της εβδομάδας, κλπ. Επίσης αναλύει τα μοτίβα που αποκλίνουν από τις αναμενόμενες νόρμες.

Εξόρυξη δεδομένων κειμένου

Οι βάσεις δεδομένων κειμένου αποτελούνται από τεράστιες συλλογές εγγράφων. Συγκεντρώνουν αυτές τις πληροφορίες από διάφορες πηγές, όπως άρθρα ειδήσεων, βιβλία, ψηφιακές βιβλιοθήκες, ηλεκτρονικά μηνύματα, ιστοσελίδες, κλπ. Λόγω της αύξησης στην ποσότητα των πληροφοριών, οι βάσεις δεδομένων κειμένου αυξάνονται με ταχείς ρυθμούς. Σε πολλές από τις βάσεις δεδομένων κειμένου, τα δεδομένα είναι ημι-δομημένα.

Για παράδειγμα, ένα έγγραφο μπορεί να περιέχει μερικά δομημένα πεδία, όπως ο τίτλος, ο συγγραφέας, ημερομηνία δημοσίευσης, κ.λπ. Αλλά μαζί με τα δομημένα δεδομένα, το έγγραφο περιέχει επίσης αδόμητα συστατικά κειμένου, όπως το απόσπασμα και το περιεχόμενο. Χωρίς να γνωρίζει κάποιος τι θα μπορούσε να υπάρχει στα έγγραφα, είναι δύσκολο να διατυπώσει κανείς αποτελεσματικά ερωτήματα για την ανάλυση και την εξαγωγή χρήσιμων πληροφοριών από τα δεδομένα. Οι χρήστες απαιτούν εργαλεία για να συγκρίνουν τα έγγραφα και να κατατάξουν τη σημασία και συνάφεια τους. Ως εκ τούτου, η εξόρυξη κειμένου έχει γίνει δημοφιλές και ουσιαστικό θέμα στην εξόρυξη των δεδομένων.

Ανάκτηση πληροφοριών

Η ανάκτηση πληροφοριών ασχολείται με την ανάκτηση πληροφοριών από ένα μεγάλο αριθμό εγγράφων που βασίζονται σε κείμενο. Μερικά από τα συστήματα βάσεων δεδομένων δεν είναι συνήθως παρόντα σε συστήματα ανάκτησης πληροφοριών, διότι και τα δύο χειρίζονται διαφορετικά είδη δεδομένων. Παραδείγματα συστημάτων ανάκτησης πληροφοριών περιλαμβάνουν:

- Σύστημα καταλόγου ηλεκτρονικής βιβλιοθήκης
- Ηλεκτρονικά συστήματα διαχείρισης εγγράφων
- Συστήματα αναζήτησης ιστού κ.λπ.

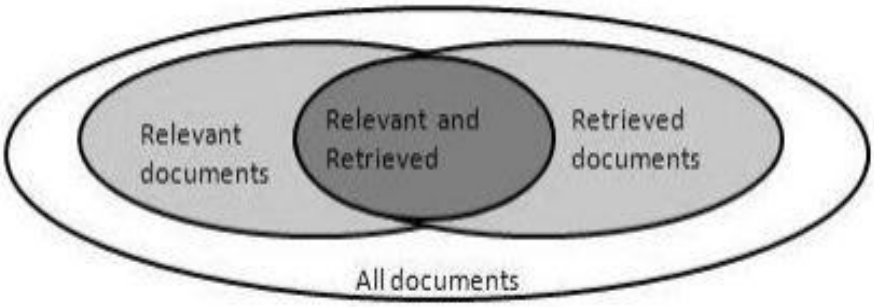
Το κύριο πρόβλημα σε ένα σύστημα ανάκτησης πληροφοριών είναι ο εντοπισμός των σχετικών εγγράφων σε μια συλλογή εγγράφων που βασίζονται στο ερώτημα του χρήστη. Αυτό το ερώτημα του χρήστη αποτελείται από ορισμένες λέξεις-κλειδιά που περιγράφουν τη πληροφορία ή πληροφορίες που αναζητά.

Σε τέτοια προβλήματα αναζήτησης, ο χρήστης λαμβάνει μια πρωτοβουλία για να τραβήξει τις σχετικές πληροφορίες έξω από μια συλλογή. Αυτό είναι κατάλληλο, όταν ο χρήστης έχει ανάγκη για ad-hoc πληροφορίες, δηλαδή, μια βραχυπρόθεσμη ανάγκη. Αλλά αν ο χρήστης έχει μια μακροχρόνια ανάγκη για πληροφορίες, τότε το σύστημα ανάκτησης μπορεί επίσης να αναλάβει πρωτοβουλία να ωθήσει οποιοδήποτε νεοαφιχθέν πληροφοριακό στοιχείο στο χρήστη.

Αυτό το είδος πρόσβασης στις πληροφορίες ονομάζεται φιλτράρισμα πληροφοριών. Και τα αντίστοιχα συστήματα είναι γνωστά ως Συστήματα Φιλτραρίσματος ή Συνιστά Συστήματα.

Βασικά μέτρα για ανάκτηση κειμένου

Πρέπει να ελέγξουμε την ακρίβεια του συστήματος, όταν ανακτά μια σειρά από έγγραφα βάσει της εισόδου του χρήστη. Ας θέσουμε το σύνολο των εγγράφων που σχετίζονται με το ερώτημα συμβολικά ως {Σχετικά} και το σύνολο των ανακτημένων εγγράφων ως {Ανακτήθηκαν}. Το σύνολο των εγγράφων που είναι σχετικά και ανακτώνται μπορεί να είναι συμβολίζεται ως {Σχετικά} ∩ {Ανακτήθηκαν}. Αυτό μπορεί να αποδειχθεί με τη μορφή ενός διαγράμματος Venn ως εξής:



https://www.tutorialspoint.com/data_mining/data_mining_tutorial.pdf

Υπάρχουν τρία θεμελιώδη μέτρα για την αξιολόγηση της ποιότητας της ανάκτησης κειμένου:

- Ακρίβεια
- Ανάκληση
- F-score

Ακρίβεια

Ακρίβεια είναι το ποσοστό των ανακτημένων εγγράφων που είναι όντως σχετικά με το ερώτημα. Η ακρίβεια μπορεί να οριστεί ως:

```
Precision= |{Relevant} ∩ {Retrieved}| /  |{Retrieved}|
```

Ανάκληση

Ανάκληση είναι το ποσοστό των εγγράφων που σχετίζονται με το ερώτημα και όντως ανακτήθηκαν. Η ανάκληση ορίζεται ως εξής:

```
Recall = |{Relevant} ∩ {Retrieved}| /  |{Relevant}|
```

F-score

F-score είναι το συχνά χρησιμοποιούμενο trade-off. Το σύστημα ανάκτησης πληροφοριών συχνά πρέπει να κάνει trade-off την ακρίβεια ή το αντίστροφο. Σαν F-score ορίζεται ο αρμονικός μέσος της ανάκλησης ή της ακρίβειας ως εξής:

```
F-score = recall x precision / (recall + precision) / 2
```

Επιπτώσεις εξόρυξης κειμένου

Μέχρι πρόσφατα, οι ιστοσελίδες χρησιμοποιούσαν πιο συχνά αναζητήσεις που βασίζονται σε κείμενο, το οποίο έβρισκε μόνο τα έγγραφα που περιείχαν συγκεκριμένες λέξεις ή φράσεις ορισμένες από το χρήστη. Τώρα, μέσω της χρήσης ενός σημασιολογικού ιστού, η εξόρυξη κειμένου μπορεί να βρει περιεχόμενο με βάση την έννοια και το πλαίσιο (και όχι μόνο από μια συγκεκριμένη λέξη). Επιπλέον, το λογισμικό εξόρυξης κειμένου μπορεί να χρησιμοποιηθεί για την κατασκευή μεγάλων φακέλων πληροφοριών για συγκεκριμένα πρόσωπα και γεγονότα. Για παράδειγμα, μεγάλα σύνολα δεδομένων με βάση τα δεδομένα που εξήχθησαν από εκθέσεις ειδήσεων μπορούν να κατασκευαστούν για να διευκολυνθεί η ανάλυση των κοινωνικών δικτύων ή της αντικατασκοπείας. Στην πραγματικότητα, το λογισμικό εξόρυξης κειμένου μπορεί να ενεργεί υπό ιδιότητα παρόμοια με ενός αναλυτή πληροφοριών ή ενός ερευνητή βιβλιοθηκάριο, αν και με πιο περιορισμένο πεδίο εφαρμογής της ανάλυσης. Η εξόρυξη κειμένου χρησιμοποιείται επίσης σε ορισμένα φίλτρα ανεπιθύμητης ηλεκτρονικής αλληλογραφίας ως τρόπος καθορισμού των χαρακτηριστικών των μηνυμάτων που είναι πιθανό να είναι διαφημίσεις ή άλλο ανεπιθύμητο υλικό. Η εξόρυξη κειμένου παίζει σημαντικό ρόλο στον καθορισμό κλίματος στις χρηματοπιστωτικές αγορές.

Κεφάλαιο 12.Εξόρυξη του παγκόσμιου ιστού

Ο παγκόσμιος ιστός περιέχει τεράστιες ποσότητες πληροφοριών που παρέχει μια πλούσια πηγή για την εξόρυξη δεδομένων.

Το διαδίκτυο δημιουργεί μεγάλες προκλήσεις για ανακάλυψη πόρων και γνώσης με βάση τις ακόλουθες παρατηρήσεις:

- **Το διαδίκτυο είναι υπερβολικά μεγάλο.** - Το μέγεθος του ιστού είναι τεράστιο και αυξάνεται ταχέως. Αυτό φανερώνει ότι το διαδίκτυο είναι πολύ μεγάλο για αποθήκευση δεδομένων και εξόρυξη δεδομένων.
- **Πολυπλοκότητα των ιστοσελίδων.** - Οι ιστοσελίδες δεν έχουν ενοποιητική δομή. Είναι πολύ περίπλοκες, σε σύγκριση με τα παραδοσιακά έγγραφα κειμένου. Υπάρχει τεράστιο ποσό εγγράφων σε ψηφιακές βιβλιοθήκες του ιστού. Αυτές οι βιβλιοθήκες δεν είναι διατεταγμένες σύμφωνα με οποιαδήποτε συγκεκριμένη ταξινομημένη σειρά.
- **Ο ιστός είναι δυναμική πηγή πληροφοριών**. - Οι πληροφορίες στο διαδίκτυο ενημερώνονται γρήγορα. Τα δεδομένα όπως ειδήσεις, χρηματιστηριακές αγορές, καιρικές συνθήκες, αθλητισμός, ψώνια, κλπ, ενημερώνονται τακτικά.
- **Ποικιλομορφία κοινοτήτων χρηστών.** - Η κοινότητα χρηστών στο διαδίκτυο είναι ταχέως αναπτυσσόμενη. Αυτοί οι χρήστες έχουν διαφορετικό υπόβαθρο, ενδιαφέροντα, και σκοπούς χρήσης. Υπάρχουν περισσότερες από 100 εκατομμύρια θέσεις εργασίας που είναι συνδεδεμένες στο Internet και εξακολουθούν να αυξάνονται ραγδαία.
- **Σχετικότητα των πληροφοριών.** - Θεωρείται ότι ένα συγκεκριμένο άτομο ενδιαφέρεται γενικά μόνο σε μικρό τμήμα του ιστού, ενώ το υπόλοιπο τμήμα του ιστού περιέχει τις πληροφορίες που δεν είναι σχετικές με τον χρήστη και μπορεί να κατακλύζουν τα επιθυμητά αποτελέσματα.

Εξόρυξη της δομής διάταξης ιστοσελίδας

Η βασική δομή της ιστοσελίδας βασίζεται στο Μοντέλο Αντικειμένου Εγγράφου (DOM – Document Object Model). Η δομή DOM αναφέρεται σε μια δομή που μοιάζει με δέντρο, όπου η ετικέτα HTML στη σελίδα αντιστοιχεί σε έναν κόμβο στο δέντρο DOM. Μπορούμε να τμηματοποιήσουμε την ιστοσελίδα χρησιμοποιώντας προκαθορισμένες ετικέτες HTML. Η σύνταξη HTML είναι ευέλικτη, ως εκ τούτου, οι ιστοσελίδες δεν ακολουθούν τις προδιαγραφές του W3C. Αν δεν ακολουθεί τις προδιαγραφές του W3C μπορεί να προκαλέσει σφάλμα στη δομή δέντρου DOM.

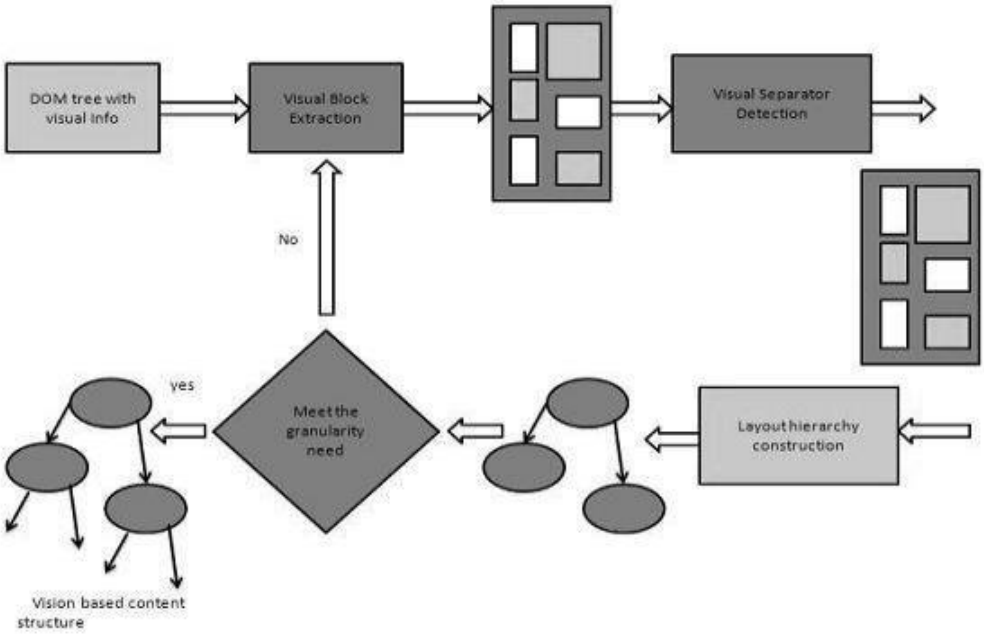
Η δομή DOM εισήχθη αρχικά για παρουσίαση στο πρόγραμμα περιήγησης και όχι για την περιγραφή της σημασιολογικής δομής της ιστοσελίδας. Η δομή DOM δεν μπορεί να αναγνωρίσει σωστά τη σημασιολογική σχέση μεταξύ των διαφόρων τμημάτων σε μια ιστοσελίδα.

Τμηματοποίηση σελίδων με βάση την όραση(VIPS)

- Ο σκοπός των VIPS είναι να εξάγει τη σημασιολογική δομή μιας ιστοσελίδας με βάση την οπτική παρουσίαση της.
- Μια τέτοια σημασιολογική δομή αντιστοιχεί σε μια δομή δέντρου. Σε αυτό το δέντρο κάθε κόμβος αντιστοιχεί σε ένα μπλοκ.
- Μια τιμή έχει εκχωρηθεί σε κάθε κόμβο. Η τιμή αυτή ονομάζεται Βαθμός Συνοχής. Αυτή η τιμή έχει εκχωρηθεί για να δείξει το συνεκτικό περιεχόμενο στο μπλοκ βασισμένο στην οπτική αντίληψη.
- Ο αλγόριθμος VIPS εξάγει πρώτα όλα τα κατάλληλα μπλοκ από το δέντρο HTML DOM. Μετά από αυτό βρίσκει τα διαχωριστικά μεταξύ αυτών των μπλοκ.
- Οι διαχωριστές αναφέρονται στις οριζόντιες ή κάθετες γραμμές σε μια ιστοσελίδα που οπτικά διασχίζουν χωρίς μπλοκ.
- Η σημασιολογία της ιστοσελίδας είναι κατασκευασμένη με βάση αυτά τα τμήματα.

Η παρακάτω εικόνα δείχνει τη διαδικασία του αλγορίθμου VIPS:

https://www.tutorialspoint.com/data_mining/data_mining_tutorial.pdf



Κεφάλαιο 13.Twitter APIs

Το Twitter παρέχει δύο τύπους API (API από το Application Programming Interface, που σημαίνει Διασύνδεση Προγραμματισμού Εφαρμογών) για πρόσβαση στα δεδομένα τους:

- RESTful APIs: Χρησιμοποιείται για τη παροχή δεδομένων σχετικά με τα υπάρχοντα αντικείμενα δεδομένων, όπως καταστάσεις "tweets", χρήστης, κλπ.
- Streaming APIs: Χρησιμοποιείται για ζωντανές καταστάσεις "tweets", καθώς αποστέλλονται.

REST APIs

Τα REST APIs παρέχουν πρόσβαση μέσω προγραμματισμού για διάβασμα και γράψιμο δεδομένων του Twitter. Δημιουργία ενός νέου Tweet, διάβασμα του προφίλ ενός χρήστη και τα δεδομένα των ακόλουθων, και πολλά άλλα. Το REST API προσδιορίζει εφαρμογές του Twitter και χρήστες χρησιμοποιώντας OAuth, οι απαντήσεις είναι σε μορφή JSON (JavaScript Object Notation), το οποίο είναι μια μορφή ανοιχτού πρότυπου που χρησιμοποιεί αναγνώσιμο από τον άνθρωπο κείμενο, για να μεταδώσει αντικείμενα δεδομένων που αποτελούνται από ζεύγη χαρακτηριστικών-τιμών.
Οι εφαρμογές πρέπει να πιστοποιούν όλα τα αιτήματα με OAuth 1.0a ή Application-only Authentication. Αυτό επιτρέπει στην αποτροπή καταχρηστικής συμπεριφοράς, και βοηθά επίσης σε περαιτέρω κατανόηση στο πώς κατηγορίες εφαρμογών χρησιμοποιούν το API. Εφαρμόζεται αυτή η κατανόηση για καλύτερη ανταπόκριση στις ανάγκες των προγραμματιστών, καθώς συνεχίζετε η εξέλιξη της πλατφόρμας.

The Search API

Το Twitter Search API είναι μέρος του REST API του Twitter. Επιτρέπει ερωτήματα εναντίον των δεικτών των τελευταίων ή δημοφιλών Tweets και συμπεριφέρεται παρόμοια με, αλλά όχι ακριβώς όπως και η αναζήτηση στη Twitter εφαρμογή για κινητά ή στην ιστοσελίδα του Twitter. Το Twitter Search API αναζητά έναντι δειγματοληψίας των πρόσφατων Tweets που δημοσιεύθηκαν τις τελευταίες 7 ημέρες. Είναι σημαντικό να γνωρίζουμε ότι το Search API επικεντρώνεται στη συνάφεια και όχι στη πληρότητα.

Επίσης η υπηρεσία αναζήτησης του Twitter και, κατ 'επέκταση, το Search API δεν είναι μια εξαντλητική πηγή για Tweets. Δεν αναπροσαρμόζονται ή διατίθενται μέσω της διεπαφής αναζήτησης όλα τα Tweets.

API.Search Παράμετροι

Q – Ένα UTF-8, URL-κωδικοποιημένο ερώτημα αναζήτησης με 500 χαρακτήρες κατ 'ανώτατο όριο, συμπεριλαμβανομένων των χειριστών. Ερωτήματα μπορεί επιπλέον να περιορίζονται από την πολυπλοκότητα.

Geocode - Επιστρέφει tweets από τους χρήστες που βρίσκονται μέσα σε μια δεδομένη ακτίνα του δεδομένου γεωγραφικού πλάτους / μήκους. Η τοποθεσία κατά προτίμηση λαμβάνετε από το Geotagging API, αλλά θα επιστρέψει στο

προφίλ του Twitter τους. Η τιμή της παραμέτρου καθορίζεται από "γεωγραφικό πλάτος, μήκος, ακτίνα", όπου οι μονάδες ακτίνας πρέπει να προσδιορίζονται ως είτε "mi" (μίλια) ή "km" (χιλιόμετρα). Δεν γίνετε να χρησιμοποιήσουμε το κοντινό χειριστή μέσω του API για να γεωκωδικοποιήσει αυθαίρετες θέσεις. Ωστόσο, μπορείτε να χρησιμοποιήσετε αυτήν την παράμετρο γεωκωδικοποίησης για να αναζητήσετε άμεσα κοντά γεωκωδικών. Ένα μέγιστο 1.000 διακριτών υπο-περιοχών θα εξεταστεί κατά τη χρήση του τροποποιητή ακτίνας.

Lang - Περιορίζει τα tweets για μια συγκεκριμένη γλώσσα, δίνεται από έναν 639-1 κωδικό ISO.

Locale - Καθορισμός της γλώσσας της ερώτησης που θέλετε να στείλετε (μόνο η ja είναι σήμερα σε ισχύ). Αυτό προορίζεται για καταναλωτές συγκεκριμένης γλώσσας και η προεπιλογή θα πρέπει να δουλεύει στην πλειοψηφία των περιπτώσεων.

Result_type – Προαιρετική. Καθορίζει τι είδους αποτελέσματα αναζήτησης θα προτιμούσες να λάβεις. Η τρέχουσα προεπιλογή είναι "Mixed". Έγκυρες τιμές περιλαμβάνουν:

* Mixed: Συμπεριλαμβάνονται δημοφιλές και σε πραγματικό χρόνο αποτελέσματα στην απόκριση.

* Recent: Επιστρέφει μόνο τα πιο πρόσφατα αποτελέσματα στην απόκριση

* Popular: Επιστρέφει μόνο τα πιο δημοφιλή αποτελέσματα στην απόκριση.

Count - Ο αριθμός των tweets που επιστρέφει ανά σελίδα, με ανώτατο όριο των 100. Προεπιλογές έως 15. Αυτό ήταν στο παρελθόν η παράμετρος "RPP" στο παλιό Search API.

Until - Επιστρέφει tweets που δημιουργήθηκαν πριν από την καθορισμένη ημερομηνία. Η ημερομηνία θα πρέπει να διαμορφωθεί ως EEEE-MM-HH. Έχοντας υπόψη ότι το ευρετήριο αναζήτησης έχει ένα όριο 7 ημερών. Με άλλα λόγια, δεν θα βρεθούν tweets για μια ημερομηνία μεγαλύτερη από μία εβδομάδα.

Since_id - Επιστρέφει τα αποτελέσματα με ένα αναγνωριστικό μεγαλύτερο από (δηλαδή, πιο πρόσφατο από) το καθορισμένο αναγνωριστικό. Υπάρχουν όρια στον αριθμό των tweets που μπορούν να αποκτηθούν μέσω του API. Αν το όριο των Tweets είναι πλήρες από την since_id, η since_id θα προβεί σε παλαιότερο διαθέσιμο αναγνωριστικό.

Max_id - Επιστρέφει αποτελέσματα με ένα αναγνωριστικό μικρότερο από (δηλαδή, παλαιότερο από) ή ίσο με την καθορισμένη ταυτότητα.

Include_entities - Ο κόμβος οντότητες δεν θα περιλαμβάνεται, όταν οριστεί σε false.

Streaming API

Το streaming API του Twitter μπορεί να παρέχει δεδομένα μέσω μιας απόκρισης ροής HTTP. Αυτό είναι πολύ παρόμοιο με τη λήψη ενός αρχείου, όπου μπορείτε να διαβάσετε μια σειρά από bytes και να το αποθηκεύσετε στο δίσκο και να επαναλάβετε μέχρι το τέλος του αρχείου. Η μόνη διαφορά είναι η ροή αυτή είναι ατελείωτη. Τα μόνα πράγματα που θα μπορούσαν να σταματήσουν αυτή τη ροή είναι:

- Αν κλείσετε τη σύνδεσή σας με την ανταπόκριση της ροής.
- Εάν η ταχύτητα της σύνδεσής σας δεν είναι σε θέση να λαμβάνει δεδομένα και το ρυθμιστικό των διακομιστών γεμίζει.

Αυτό σημαίνει ότι η διαδικασία αυτή θα χρησιμοποιεί το νήμα από όπου ξεκίνησε μέχρι να σταματήσει. Στην παραγωγή, θα πρέπει πάντα να ξεκινά από ένα διαφορετικό νήμα ή διαδικασία για να βεβαιωθούμε ότι το λογισμικό δεν θα παγώσει μέχρι να σταματήσουμε τη ροή.

Οι λόγοι που προτιμείται το Streaming API:

- Σύλληψη μεγάλης ποσότητας δεδομένων, επειδή το REST API έχει περιορισμένη πρόσβαση σε παλαιότερα δεδομένα.
- Ανάλυση σε πραγματικό χρόνο, όπως η παρακολούθηση κοινωνικών συζητήσεων για μια ζωντανή εκδήλωση.
- Αρχειοθέτηση μέσα σε μια οργάνωση σαν η αρχειοθέτηση μιας κοινωνικής συζήτησης σχετικά με το εμπορικό σήμα της οργάνωσης.
- Σύστημα ανταπόκρισης τεχνητής νοημοσύνης για λογαριασμό twitter, όπως αυτοματοποιημένες απαντήσεις και ερωτήσεις αρχειοθέτησης ή παροχή απαντήσεων.

Σύνδεση στο Streaming API

Υπάρχουν δύο τρόποι για να συνδεθεί κανείς σε μια ροή:

- Filter(follow, track, async, locations, stall_warning, languages, encoding, filter_level)
- Firehose(count, async)

Το Firehose συλλαμβάνει τα πάντα. Θα πρέπει να βεβαιωθούμε ότι έχουμε ταχύτητα σύνδεσης που μπορεί να χειριστεί τη ροή και πως έχουμε την ικανότητα αποθήκευσης που μπορεί να αποθηκεύσει αυτά τα tweets με τον ίδιο ρυθμό. Θα πρέπει να ορίσουμε μία από δύο παραμέτρους του Filter:

- Follow: Μια λίστα ταυτοτήτων χρηστών που θα ακολουθήσει. Αυτό θα μας επιστρέψει όλα τα tweets, retweets, και άλλους που έκαναν retweet τα tweets τους. Αυτό δεν περιλαμβάνει αναφορές και retweets όπου ο χρήστης δεν πατήσετε το κουμπί retweet.
- Track: Μια συμβολοσειρά ή λίστα συμβολοσειρών που θα χρησιμοποιηθεί για το φιλτράρισμα. Εάν χρησιμοποιηθούν πολλές λέξεις χωρισμένες με κενά, το κενό θεωρείτε ως τελεστής AND. Εάν χρησιμοποιηθούν πολλές λέξεις σε μια σειρά διαχωρισμένες με κόμμα ή μια λίστα με λέξεις αυτό θα θεωρηθεί ως τελεστής OR. Η Track κάνει διάκριση πεζών-κεφαλαίων.

Κεφάλαιο 14.Εξόρυξη δεδομένων με χρήση Python 3.4

Υπάρχουν πολλές γλώσσες προγραμματισμού που μπορεί να χρησιμοποιήσει κανείς και διάφοροι τρόποι για να συνδεθεί σε Twitter APIs για εξόρυξη των δεδομένων, όπως η java, η SQL, η C++ και άλλες πολλές. Αποφάσισα να χρησιμοποιήσω τη Python 3.4 μέσω του Jupyter Notebook, για καλύτερη παρουσίαση του κώδικα και επειδή υπάρχει η βιβλιοθήκη tweepy που επιτρέπει εύκολη πρόσβαση σε Twitter API μέσω OAuth. Οπότε έφτιαξα δύο προγράμματα, ένα χρησιμοποιώντας το Search API από το REST API του Twitter και το άλλο χρησιμοποιώντας το Streaming API του Twitter.

Εξόρυξη με REST API

Ξεκινώντας δημιούργησα ένα λογαριασμό Twitter τον οποίο θα χρησιμοποιήσουμε για πρόσβαση στο Twitter API, ο λογαριασμός μας παρέχει κάποια secrets και tokens που κάνουν τη πρόσβαση μοναδική ανά χρήστη. Στο επόμενο βήμα εισάγω τις βιβλιοθήκες που θα μου χρειαστούν, όπως tweepy, pandas, matplotlib.pyplot και numpy.

- **Tweepy -** Είναι βιβλιοθήκη της Python που μας παρέχει εύκολη πρόσβαση σε Twitter APIs.
- **Pandas** - Είναι μια εργαλειοθήκη ανάλυσης δεδομένων της Python.
- **Matplotlib.pyplot** - Είναι μια βιβλιοθήκη της Python για κατασκευή διαγραμμάτων.
- **Numpy** – Είναι θεμελιώδη πακέτο της Python που χρησιμοποιείτε για την επιστημονική υπολογιστική. Το Numpy μπορεί επίσης να χρησιμοποιηθεί ως αποτελεσματικό πολυδιάστατο δοχείο γενικών δεδομένων. Αυθαίρετοι τύποι δεδομένων μπορούν να οριστούν. Αυτό επιτρέπει στο Numpy απρόσκοπτα και γρήγορα να ενσωματωθεί με μια ευρεία ποικιλία βάσεων δεδομένων. Επίσης

ρυθμίζω το pandas με βάση την οθόνη μου.

```
import tweepy
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

pd.options.display.max_columns = 50
pd.options.display.max_rows= 50
pd.options.display.width= 120
```

Στη συνέχεια συνδέομαι στο API του Twitter χρησιμοποιώντας το tweepy και τα κλειδιά που πήρα από το Twitter μέσω το λογαριασμού μου.

```
consumer_key = "HfuLbUswsmo14w9fwVzmk2ANw"
consumer_secret = "VDVFCjyBuSeCmvFHAddSYp1dhqzFYnCdvOhxPnuBluc5hvODie"

auth = tweepy.OAuthHandler(consumer_key=consumer_key, consumer_secret=consumer_se

api = tweepy.API(auth)
```

Χρησιμοποιώ τη παράμετρο Q για να αναζητήσω tweets με συγκεκριμένη ονομασία, και αποθηκεύω τα δεδομένα μου σε μια μεταβλητή που ονομάζω results.

```
results = api.search(q="#FinalFantasyXV")
```

Έπειτα εξετάζω τα αποτελέσματα, βλέπω πόσα tweets αποκτά η μεταβλητή μου, από προεπιλογή θα τράβα μόνο 15 tweets αλλά αυτό μπορούμε να το αλλάξουμε, επίσης φτιάχνω μια απλή συνάρτηση που θα μου εμφανίζει μερικά στοιχεία για ένα tweet, όπως το όνομα, πότε δημιουργήθηκε και τη περιείχε.

Στη συνέχεια εξετάζω τη κατάσταση αντικειμένου και το αντικείμενο χρήστη θέτοντας σε μια νέα μεταβλητή tweet ένα από τα tweets που εξόρυξα και το καλώ με τη συνάρτηση dir, η οποία καλεί ότι υπάρχει μέσα στη μεταβλητή, αφαιρώ αντικείμενα τα οποία ξεκινάνε με «_» διότι είναι συνήθως περιττές πληροφορίες, το ίδιο κάνω και για το αντικείμενο χρήστη απλά θέτω νέα μεταβλητή user όπου εισάγω το συγγραφέα του tweet, αυτο που μας επιστρέφει είναι διάφορες πληροφορίες για το tweet και το χρήστη σε μορφή JSON.

```
In [25]: tweet=results[2]

for param in dir(tweet):
    if not param.startswith("_"):
        print ("%s : %s" % (param, eval("tweet." + param)))
```

```
author : User(utc_offset=None, profile_sidebar_fill_color='DDEEF6', translator_type='none',
profile_background_color='F5F8FA', default_profile_image=True, listed_count=0, id=735861127838617600,
favourites_count=1637, screen_name='anthonybranstel', notifications=None, url=None,
profile_sidebar_border_color='C0DEED', following=False, name='Anthony Branstett', profile_text_color='333333',
is_translator=False, time_zone=None, has_extended_profile=False, created_at=datetime.datetime(2016, 5, 26, 15, 52, 19),
contributors_enabled=False, profile_image_url='http://abs.twimg.com/sticky/default_profile_images
/default_profile_1_normal.png', profile_link_color='1DA1F2', id_str='735861127838617600', protected=False,
statuses_count=44, geo_enabled=False, is_translation_enabled=False, lang='fr', description='',
profile_image_url_https='https://abs.twimg.com/sticky/default_profile_images/default_profile_1_normal.png',
_api=<tweepy.api.API object at 0x05A3DA50>, follow_request_sent=None, location='', friends_count=29,
default_profile=True, verified=False, _json={'utc_offset': None, 'geo_enabled': False, 'profile_sidebar_fill_color':
'DDEEF6', 'is_translation_enabled': False, 'profile_background_color': 'F5F8FA', 'screen_name': 'anthonybranstel',
'lang': 'fr', 'description': '', 'profile_image_url_https': 'https://abs.twimg.com/sticky/default_profile_images
/default_profile_1_normal.png', 'default_profile_image': True, 'listed_count': 0, 'id': 735861127838617600, 'time_zone':
None, 'following': None, 'translator_type': 'none', 'url': None, 'notifications': None, 'has_extended_profile': False,
'default_profile': True, 'follow_request_sent': None, 'location': '', 'friends_count': 29,
'profile_sidebar_border_color': 'C0DEED', 'profile_background_image_url_https': None, 'followers_count': 9, 'name':
'Anthony Branstett', 'profile_text_color': '333333', 'verified': False, 'favourites_count': 1637, 'is_translator': False,
'profile_use_background_image': True, 'created_at': 'Thu May 26 15:52:19 +0000 2016', 'contributors_enabled': False,
'entities': {'description': {'urls': []}}, 'profile_image_url': 'http://abs.twimg.com/sticky/default_profile_images
/default_profile_1_normal.png', 'profile_background_tile': False, 'profile_link_color': '1DA1F2', 'statuses_count': 44,
'profile_background_image_url': None, 'id_str': '735861127838617600', 'protected': False},
profile_use_background_image=True, profile_background_image_url_https=None, entities={'description': {'urls': []}},
profile_background_tile=False, followers_count=9, profile_background_image_url=None)
```

```
In [26]: user=tweet.author

for param in dir(user):
    if not param.startswith("_"):
        print ("%s : %s" % (param, eval("user." + param)))
```

```
contributors_enabled : False
created_at : 2016-05-26 15:52:19
default_profile : True
default_profile_image : True
description :
entities : {'description': {'urls': []}}
favourites_count : 1637
follow : <bound method User.follow of User(utc_offset=None, profile_sidebar_fill_color='DDEEF6', translator_type='none',
profile_background_color='F5F8FA', default_profile_image=True, listed_count=0, id=735861127838617600,
favourites_count=1637, screen_name='anthonybranstel', notifications=None, url=None,
profile_sidebar_border_color='C0DEED', following=False, name='Anthony Branstett', profile_text_color='333333',
is_translator=False, time_zone=None, has_extended_profile=False, created_at=datetime.datetime(2016, 5, 26, 15, 52, 19),
contributors_enabled=False, profile_image_url='http://abs.twimg.com/sticky/default_profile_images
/default_profile_1_normal.png', profile_link_color='1DA1F2', id_str='735861127838617600', protected=False,
statuses_count=44, geo_enabled=False, is_translation_enabled=False, lang='fr', description='',
profile_image_url_https='https://abs.twimg.com/sticky/default_profile_images/default_profile_1_normal.png',
_api=<tweepy.api.API object at 0x05A3DA50>, follow_request_sent=None, location='', friends_count=29,
default_profile=True, verified=False, _json={'utc_offset': None, 'geo_enabled': False, 'profile_sidebar_fill_color':
'DDEEF6', 'is_translation_enabled': False, 'profile_background_color': 'F5F8FA', 'screen_name': 'anthonybranstel',
'lang': 'fr', 'description': '', 'profile_image_url_https': 'https://abs.twimg.com/sticky/default_profile_images
/default_profile_1_normal.png', 'default_profile_image': True, 'listed_count': 0, 'id': 735861127838617600, 'time_zone':
None, 'following': None, 'translator_type': 'none', 'url': None, 'notifications': None, 'has_extended_profile': False,
'default_profile': True, 'follow_request_sent': None, 'location': '', 'friends_count': 29,
'profile_sidebar_border_color': 'C0DEED', 'profile_background_image_url_https': None, 'followers_count': 9, 'name':
'Anthony Branstett', 'profile_text_color': '333333', 'verified': False, 'favourites_count': 1637, 'is_translator': False,
'profile_use_background_image': True, 'created_at': 'Thu May 26 15:52:19 +0000 2016', 'contributors_enabled': False,
'entities': {'description': {'urls': []}}, 'profile_image_url': 'http://abs.twimg.com/sticky/default_profile_images
/default_profile_1_normal.png', 'profile_background_tile': False, 'profile_link_color': '1DA1F2', 'statuses_count': 44,
'profile_background_image_url': None, 'id_str': '735861127838617600', 'protected': False},
profile_use_background_image=True, profile_background_image_url_https=None, entities={'description': {'urls': []}},
```

Χρησιμοποιώ τη κλάση cursor για σελιδοποίηση των αποτελεσμάτων και για να πάρω περισσότερα από 15 αποτελέσματα, η cursor χρησιμοποιείτε για καλύτερη διαχείριση βάσεων δεδομένων.

```
In [27]: results = []
for tweet in tweepy.Cursor(api.search, q="#FinalFantasyXV").items(100):
    results.append(tweet)

print (len(results))
```

```
100
```

Στη συνέχεια δημιουργώ μια συνάρτηση όπου φτιάχνω ένα πλαίσιο δεδομένων και αποθηκεύω συγκεκριμένα δεδομένα σε σχέση με το tweet και με το χρήστη από τα δεδομένα που έχω ήδη εξορύξει, για τη δημιουργία του πλαισίου χρησιμοποιώ pandas. Έπειτα βλέπω τα πρώτα πέντε στοιχεία του πίνακα και τα πέντε τελευταία.

In [28]:
```
def process_results(results):
    id_list = [tweet.id for tweet in results]
    data_set = pd.DataFrame(id_list, columns=["id"])

    # Processing Tweet Data

    data_set["text"] = [tweet.text for tweet in results]
    data_set["created_at"] = [tweet.created_at for tweet in results]
    data_set["retweet_count"] = [tweet.retweet_count for tweet in results]
    data_set["favorite_count"] = [tweet.favorite_count for tweet in results]
    data_set["source"] = [tweet.source for tweet in results]

    # Processing User Data
    data_set["user_id"] = [tweet.author.id for tweet in results]
    data_set["user_screen_name"] = [tweet.author.screen_name for tweet in results]
    data_set["user_name"] = [tweet.author.name for tweet in results]
    data_set["user_created_at"] = [tweet.author.created_at for tweet in results]
    data_set["user_description"] = [tweet.author.description for tweet in results]
    data_set["user_followers_count"] = [tweet.author.followers_count for tweet in results]
    data_set["user_friends_count"] = [tweet.author.friends_count for tweet in results]
    data_set["user_location"] = [tweet.author.location for tweet in results]

    return data_set
data_set = process_results(results)
```

In [29]:
```
data_set.head(5)
```

Out[29]:

	id	text	created_at	retweet_count	favorite_count	source	user_id	user_screen_name	user_name	user_cre
0	803938927002296320	RT @Kei_Katagiri_k: 海外のシドニーレイヤーさんが本家を超絶する勢いです。...	2016-11-30 12:29:11	375	0	Twitter for Android	3076092607	seven_Ernst	砂地ダチョウ	2015-03-1 05:26:32
1	803938921025409024	RT @Kei_Katagiri_k: 海外のシドニーレイヤーさんが本家を超絶する勢いです。...	2016-11-30 12:29:09	375	0	ツイタマ+ for Android	1957901478	hane0621	双葉	2013-10-1 03:50:18
2	803938914234839040	RT @Kei_Katagiri_k: 海外のシドニーレイヤーさんが本家を超絶する勢いです。...	2016-11-30 12:29:08	375	0	Twitter for iPhone	1328364037	NevsHi303	シオン・シュバルツ・フォン・グランツ	2013-04-(04:11:00
3	803938907595280384	RT @Kei_Katagiri_k: 海外のシドニーレイヤーさんが本家を超絶する勢いです。...	2016-11-30 12:29:06	375	0	ついっぷる	51492976	amagata	あまがた	2009-06-2 17:09:43
4	803938899852541953	RT @Kei_Katagiri_k: 海外のシドニーレイヤーさんが本家を超絶する勢いです。...	2016-11-30 12:29:04	375	0	fuyutiger	27561399	shima_zu	しまづ	2009-03-3 02:11:34

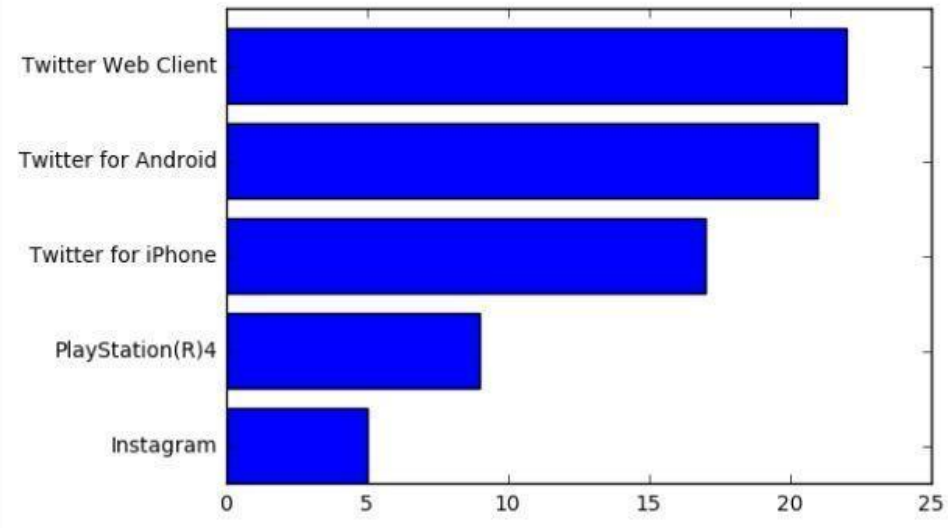
```
In [30]: data_set.tail(5)
```

Out[30]:

source	user_id	user_screen_name	user_name	user_created_at	user_description	user_followers_count	user_friends_count	user_locati
Mobile Web (M5)	2351706733	viralvinnie	viral vinnie	2014-02-19 13:35:01	Aspiring game developer, host of crappy youtub...	39	249	Not where y are
Twitter Web Client	179907944	mizukisu	水貴:PSO プチオンリーやるってよ	2010-08-18 11:27:49	基本萌えばっかり。すごくうるさい。フォローもブロックもお好きにどうぞ。アルトネリコで妄想激し...	732	736	甲子園のあ市
Twitter for iPhone	827577294	reverse_ll	えるつー	2012-09-16 19:03:12	がちむち受け好きしやわせいちゃらぶ好きのふじよしの人/DQは40鯖、FFはIfrit鯖をうろ...	61	102	
ついっぷる	79726870	nekokotatu	OYAMA	2009-10-04 13:10:20	ゲームばっかりしているかのように見えてゲームばっかりしているよ! アニメばっかり見ているよう...	92	92	福岡県
Instagram	202131472	Honest_JW	ㅈㅇ	2010-10-13 11:14:04	Since 10/10/13 #YNWA #Novak	94	181	디지털로

Στο τέλος δημιουργώ ένα διάγραμμα χρησιμοποιώντας τη matplotlib, όπου βλέπω ποιες πηγές χρησιμοποίησαν περισσότερο οι χρήστες για να δημιουργήσουν tweets.

```
In [18]: sources = data_set["source"].value_counts()[:5][::-1]

         plt.barh(range(len(sources)), sources.values)
         plt.yticks(np.arange(len(sources)) + 0.4, sources.index)
         plt.show()
```



Εξόρυξη με Streaming API

Ξεκινώντας εγκαθιστώ το MongoDB στον υπολογιστή μου το οποίο στήνω και συνδέομαι, έτσι έχω μία βάση δεδομένων έτοιμη για αποθήκευση αρχείων, έτσι θα αποθηκεύσουμε ότι εξορύξουμε από το Twitter.

Όπως για το REST API έτσι και για το streaming API πρέπει να εισάγω κάποιες βιβλιοθήκες στην αρχή, όπως και πριν εισάγω τις numpy, pandas ,tweepy και matplotlib.pyplot, επιπλέον νέες βιβλιοθήκες είναι οι pymongo, ipywidgets, display, CountVectorizer, re, και datetime.

- **PyMongo** – Προσφέρει εργαλεία τα οποία λειτουργούν με το MongoDB μέσω της Python.
- **IpyWidgets** – Τα Widgets είναι αντικείμενα της Python τα οποία εμφανίζονται στο πρόγραμμα περιήγησης συνήθως σαν ολισθητής, κουτί κειμένου, κτλ. Τα Widgets χρησιμοποιούνται για τη κατασκευή διαδραστικών GUI(Graphical User Interface) και επίσης για το συγχρονισμό πληροφοριών μεταξύ Python και Javascript.
- **Display** – Δημιουργεί ένα αντικείμενο display παρέχοντας του ανεπεξέργαστα δεδομένα.
- **CountVectorizer** – Μετατρέπει μια συλλογή από έγγραφα κειμένου σε ένα πίνακα όπου μετρά ενδείξεις ανάλογα με το τη θα ορίσουμε.
- **Re** – Η Regural expressions είναι ουσιαστικά μια μικρή, πολύ εξειδικευμένη γλώσσα προγραμματισμού ενσωματωμένη μέσα στη Python και διατίθενται χρησιμοποιώντας τη βιβλιοθήκη RE. Χρησιμοποιώντας αυτή τη γλώσσα, καθορίζουμε τους κανόνες για το πιθανό σύνολο συμβολοσειρών που θέλουμε να ταιριάζουν, αυτά τα σύνολα μπορεί να περιέχουν αγγλικές προτάσεις ή διευθύνσεις ηλεκτρονικού ταχυδρομείου ή οτιδήποτε θέλουμε. Έπειτα κάνουμε ερωτήσεις όπως «Ταιριάζει αυτή η συμβολοσειρά στο μοτίβο» ή «Ταιριάζει κάτι από το μοτίβο σε οτιδήποτε σε αυτή τη συμβολοσειρά». Γίνετε επίσης να χρησιμοποιηθεί η RE για να τροποποιήσουμε μια συμβολοσειρά ή να τη χωρίσουμε χώρια με διάφορους τρόπους.
- **Datetime** – Η βιβλιοθήκη datetime παρέχει classes για το χειρισμό ημερομηνιών και ωρών, με απλούς και σύνθετους τρόπους.

```
In [1]: import numpy as np
        import pandas as pd
        import tweepy
        import matplotlib.pyplot as plt
        import pymongo
        import ipywidgets as wgt
        from IPython.display import display
        from sklearn.feature_extraction.text import CountVectorizer
        import re
        from datetime import datetime

        %matplotlib inline
```

Στη συνέχεια συνδέομαι στο streaming API του twitter χρησιμοποιώντας το tweepy και τα κλειδιά που πήρα από το λογαριασμό μου, η μόνη διαφορά από τη REST API είναι πως χρειαζόμαστε τέσσερα κλειδιά, τα παίρνουμε όλα από την ιστοσελίδα https://apps.twitter.com/ έχοντας φτιάξει πρώτα λογαριασμό twitter.

```
In [2]: api_key = "HfuLbUswsmo14w9fwVzmk2ANw"
        api_secret = "VDVFCjyBuSeCmvFHAddSYp1dhqzFYnCdvOhxPnuBluc5hvODie"
        access_token = "794481828585467904-tZhnggL7U3zZKoDltJKnjMH4ITqGMV2"
        access_token_secret = "Mf79ERzfx9uDqy5t1VtVW5yemxAsRX8S1k1uFe2jEOvIN"


        auth = tweepy.OAuthHandler(api_key, api_secret)
        auth.set_access_token(access_token, access_token_secret)

        api = tweepy.API(auth)
```

Στο επόμενο βήμα συνδέομαι σε μια βάση δεδομένων χρησιμοποιώντας το MongoDB, συνδέομαι σε μια βάση με το όνομα tweets και σε μια συλλογή με το όνομα StreamingTutorial, και μετά μετράω πόσα αντικείμενα υπάρχουν μέσα στη συλλογή μου.

```
In [8]: col = pymongo.MongoClient()["tweets"]["StreamingTutorial"]
        col.count()

Out[8]: 100
```

Θα χρειαστούμε έναν StreamListener ο οποίος θα επεκτείνεται στη κλάση tweepy.StreamListener. Υπάρχουν αρκετές μέθοδοι που μπορούμε να χρησιμοποιήσουμε για να εκτελούν λειτουργίες, κάποιες από τις ποιο σημαντικές είναι:

- **On_status(self, status):**Αυτή θα μεταφέρει ένα αντικείμενο “tweet” όταν λαμβάνεται ένα tweet.
- **On_data(self, raw_data):**Καλείτε όταν οποιοδήποτε δεδομένο λαμβάνεται και μεταφέρονται τα ακατέργαστα δεδομένα.
- **On_error(self, status_code):**Καλείτε όταν παίρνουμε απάντηση μέσω κώδικα.

Εδώ χρησιμοποίησα τη μέθοδο on_status, αυτή δηλαδή θα καλείτε όταν θα τρέχουμε τον Listener. Τα υπόλοιπα μέσα στη κλάση είναι μετρητές που βοηθούν κυρίως στην οπτικοποίηση για την παρακολούθηση της διαδικασίας Streaming, και έχουμε καλή οπτικοποίηση με τη βοήθεια των widgets. Αξιοσημείωτο είναι πως η μέθοδος on_status όταν θα λαμβάνει ένα νέο “status” θα μεταφέρει το JSON κομμάτι αυτού στη συλλογή μας στο MongoDB. Στο τέλος δημιουργώ ένα νέο Listener και θέτω ως μέγιστο εκατό tweets, και ξεκινώ το Stream με αυτόν το Listener χρησιμοποιώντας τη ταυτότητα μου.

In [4]:
```
class MyStreamListener(tweepy.StreamListener):

    counter = 0

    def __init__(self, max_tweets=1000, *args, **kwargs):
        self.max_tweets = max_tweets
        self.counter = 0
        super().__init__(*args, **kwargs)

    def on_connect(self):
        self.counter = 0
        self.start_time = datetime.now()

    def on_status(self, status):
        # Increment counter
        self.counter += 1

        # Store tweet to MongoDB
        col.insert_one(status._json)


        if self.counter % 1 == 0:
            value = int(100.00 * self.counter / self.max_tweets)
            mining_time = datetime.now() - self.start_time
            progress_bar.value = value
            html_value = """<span class="label label-primary">Tweets/Sec: %.1f</span>""" % (self.counter / max(
[1,mining_time.seconds]))
            html_value += """ <span class="label label-success">Progress: %.1f%%</span>""" % (self.counter / se
lf.max_tweets * 100.0)
            html_value += """ <span class="label label-info">ETA: %.1f Sec</span>""" % ((self.max_tweets - self
.counter) / (self.counter / max([1,mining_time.seconds])))
            wgt_status.value = html_value
            #print("%s/%s" % (self.counter, self.max_tweets))
            if self.counter >= self.max_tweets:
                myStream.disconnect()
                print("Finished")
                print("Total Mining Time: %s" % (mining_time))
                print("Tweets/Sec: %.1f" % (self.max_tweets / mining_time.seconds))
                progress_bar.value = 0


myStreamListener = MyStreamListener(max_tweets=100)
myStream = tweepy.Stream(auth = api.auth, listener=myStreamListener)
```

Όπως είχα προαναφέρει υπάρχουν δύο τρόποι για να συνδεθούμε στο Stream με Filter ή Firehose, εδώ συνδέομαι με Filter και χρησιμοποιώ τη παράμετρο Track για να εξορύξω μια λίστα, με τη λίστα ο κώδικας είναι ποιο σαφής για διάβασμα. Με τη βοήθεια τον widgets δημιουργώ μια γραμμή προόδου όπου εμφανίζονται διάφορες πληροφορίες για το Stream. Ξεκινάω λοιπόν τη Filter με ένα μετρητή λάθους για να συλλέξω tweets που θα περιέχουν λέξεις από τη λίστα που έχω ορίσει.

```
In [5]: keywords = ["Twitch",
                    "Youtube",
                    "Netflix",
                   ]

        # Visualize a progress bar to track progress
        progress_bar = wgt.IntProgress(value=0)
        display(progress_bar)
        wgt_status = wgt.HTML(value="""<span class="label label-primary">Tweets/Sec: 0.0</span>""")
        display(wgt_status)

        # Start a filter with an error counter of 20
        for error_counter in range(20):
            try:
                myStream.filter(track=keywords)
                print("Tweets collected: %s" % myStream.listener.counter)
                print("Total tweets in collection: %s" % col.count())
                break
            except:
                print("ERROR# %s" % (error_counter + 1))

        Finished
        Total Mining Time: 0:00:03.009174
        Tweets/Sec: 33.3
        Tweets collected: 100
        Total tweets in collection: 100
```

Αφού συλλέξαμε τα tweets μπορούμε να ρίξουμε μια ματιά σε τι αποκτήσαμε, οπότε καλώ τυχαία ένα αντικείμενο.

```
In [6]: col.find_one()

Out[6]: {'_id': ObjectId('5829a973e85e110b9c332c27'),
         'contributors': None,
         'coordinates': None,
         'created_at': 'Mon Nov 14 12:09:23 +0000 2016',
         'entities': {'hashtags': [],
          'symbols': [],
          'urls': [{'display_url': 'youtube.com/watch?feature=…',
            'expanded_url': 'http://www.youtube.com/watch?feature=player_embedded&v=hlm
        gjUD1F3g',
            'indices': [16, 39],
            'url': 'https://t.co/RJjV9WjoAo'},
           {'display_url': 'akhbarnet.ml/%d9%85%d8%a2%d…',
            'expanded_url': 'http://akhbarnet.ml/%d9%85%d8%a2%d8%b3%d9%8a-%d8%ad%d9%84%
        d8%a8-%d8%aa%d8%aa%d9%88%d8%a7%d8%b5%d9%84/',
            'indices': [40, 63],
            'url': 'https://t.co/VIXf1wXD3U'}],
          'user_mentions': []},
         'favorite_count': 0,
         'favorited': False,
         'filter_level': 'low',
         'geo': None,
         'id': 798135739229896704,
         'id_str': '798135739229896704',
         'in_reply_to_screen_name': None,
         'in_reply_to_status_id': None,
         'in_reply_to_status_id_str': None,
         'in_reply_to_user_id': None,
         'in_reply_to_user_id_str': None,
         'is_quote_status': False,
         'lang': 'ar',
         'place': None,
         'possibly_sensitive': False,
         'retweet_count': 0,
         'retweeted': False,
         'source': '<a href="http://publicize.wp.com/" rel="nofollow">WordPress.com</a
        >',
         'text': 'مَسي حلب تتواصل https://t.co/RJjV9WjoAo https://t.co/VIXf1wXD3U',
         'timestamp_ms': '1479125363534',
         'truncated': False,
```

```
'timestamp_ms': '1479125363534',
'truncated': False,
'user': {'contributors_enabled': False,
 'created_at': 'Thu Oct 22 12:21:08 +0000 2015',
 'default_profile': True,
 'default_profile_image': False,
 'description': None,
 'favourites_count': 6,
 'follow_request_sent': None,
 'followers_count': 5388,
 'following': None,
 'friends_count': 4007,
 'geo_enabled': False,
 'id': 4016789897,
 'id_str': '4016789897',
 'is_translator': False,
 'lang': 'en',
 'listed_count': 4,
 'location': 'Qatar',
 'name': '300K صفاء',
 'notifications': None,
 'profile_background_color': 'C0DEED',
 'profile_background_image_url': 'http://abs.twimg.com/images/themes/theme1/bg.png',
 'profile_background_image_url_https': 'https://abs.twimg.com/images/themes/theme1/bg.png',
 'profile_background_tile': False,
 'profile_banner_url': 'https://pbs.twimg.com/profile_banners/4016789897/1446349928',
 'profile_image_url': 'http://pbs.twimg.com/profile_images/794031258137403392/rXfngTd7_normal.jpg',
 'profile_image_url_https': 'https://pbs.twimg.com/profile_images/794031258137403392/rXfngTd7_normal.jpg',
 'profile_link_color': '1DA1F2',
 'profile_sidebar_border_color': 'C0DEED',
 'profile_sidebar_fill_color': 'DDEEF6',
 'profile_text_color': '333333',
 'profile_use_background_image': True,
 'protected': False,
 'screen_name': 'safaenihal',
 'statuses_count': 5437,
 'time_zone': 'Pacific Time (US & Canada)',
 'url': 'http://akhbarnet.ml',
 'utc_offset': -28800,
 'verified': False}}
```


Στη συνέχεια βάζω τα αποτελέσματα της συλλογής μου σε ένα πλαίσιο δεδομένων, το οποίο περιέχει τέσσερις στήλες, πότε δημιουργήθηκε το tweet, από ποια πηγή προήρθε, τι περιείχε το tweet, και τον χρήστη.

In [7]:
```
dataset = [{"created_at": item["created_at"],
            "text": item["text"],
            "user": "@%s" % item["user"]["screen_name"],
            "source": item["source"],
           } for item in col.find()]

dataset = pd.DataFrame(dataset)
dataset
```

Out[7]:

	created_at	source	text	user
0	Mon Nov 14 12:09:23 +0000 2016	<a href="http://publicize.wp.com/" rel="nofol...	ماسي حلب نار اصل https://t.co/RJjV9WfjoAo https:...	@safaenihal
1	Mon Nov 14 12:09:23 +0000 2016	<a href="http://www.google.com/" rel="nofollow...	We Use to have an Elephant! https://t.co/eAN1...	@jeffgrantmedia
2	Mon Nov 14 12:09:23 +0000 2016	<a href="http://www.facebook.com/twitter" rel=...	Eripe eiohnhttps://t.co/yzGqYKZtzw https://t...	@ignaslak_micha
3	Mon Nov 14 12:09:23 +0000 2016	<a href="http://www.google.com/" rel="nofollow...	I liked a @YouTube video https://t.co/BShVUlHN2...	@Loppyyy
4	Mon Nov 14 12:09:23 +0000 2016	<a href="http://twitter.com/download /android" ...	RT @Millicent_Mazur: Buy Bitcoin With Paypal h...	@marguerileade
5	Mon Nov 14 12:09:23 +0000 2016	<a href="http://linkis.com" rel="nofollow">Put...	Видео "110" (https://t.co/qCgK7ZXsVf) на @YouT...	@cegrey2013
6	Mon Nov 14 12:09:23 +0000 2016	<a href="http://www.google.com/" rel="nofollow...	ฉันชอบวิดีโอ @YouTube https://t.co/p8H9ZN3WG0 ...	@TThetossil
7	Mon Nov 14 12:09:23 +0000 2016	<a href="http://twitter.com/download /android" ...	RT @Millicent_Mazur: Buy Bitcoin With Paypal h...	@MechemLibby
8	Mon Nov 14 12:09:23 +0000 2016	<a href="http://twitter.com/download /iphone" r...	2016	(احكم حبيبي ما (حصرياً بالكلمات	(A...	@etnajlaa
9	Mon Nov 14 12:09:23 +0000 2016	<a href="http://twitter.com/download /android" ...	RT @AntraxCLASH: Me gustó un video de @YouTube...	@johenny_padill
10	Mon Nov 14 12:09:22 +0000 2016	<a href="http://twitter.com" rel="nofollow">Tw...	[MY SMT] EXO-CBX 등장!https://t.co/n1Sn6G395E v...	@QueensPaoz
11	Mon Nov 14 12:09:23 +0000 2016	<a href="http://twitter.com/download /android" ...	https://t.co/mNTPSH0Xnq	@psalcha_bkly
12	Mon Nov 14 12:09:23 +0000 2016	<a href="http://twitter.com" rel="nofollow">Tw...	@ArianaGranPL https://t.co/NDakDkFlqQ Zajrzysz...	@luvmyhazzas

Χρησιμοποιώ το CountVectorizer για να μετρήσω τη συχνότητα των λέξεων που χρησιμοποιήθηκαν στα tweets, πόσες φορές δηλαδή χρησιμοποιήθηκαν κάποιες λέξεις, και τις εμφανίζω σε ένα πίνακα.

In [9]:
```
cv = CountVectorizer()
count_matrix = cv.fit_transform(dataset.text)

word_count = pd.DataFrame(cv.get_feature_names(), columns=["word"])
word_count["count"] = count_matrix.sum(axis=0).tolist()[0]
word_count = word_count.sort_values("count", ascending=False).reset_index(drop=Tr
word_count[:50]
```

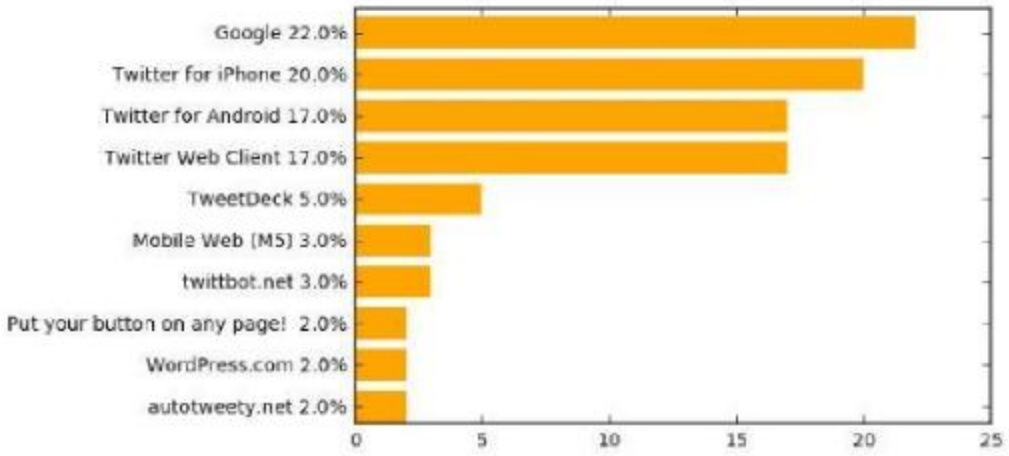
Out[9]:

	word	count
0	https	108
1	co	106
2	youtube	44
3	rt	31
4	de	15
5	video	13
6	to	10
7	the	10
8	my	9
9	via	8
10	liked	6
11	on	6
12	smt	6
13	vía	5
14	vídeo	5
15	2016	5
16	um	5
17	exo	5
18	of	5
19	with	4
20	gostei	4
21	등장	4
22	cbx	4
23	you	4
24	홈쇼핑	4

Στο τέλος θέλω να διακρίνω τις πηγές από τις οποίες προήρθαν τα tweets στο πλαίσιο δεδομένων μπορούμε να δούμε από που προήρθαν, αλλά υπάρχουν και περιττές πληροφορίες, οπότε χρησιμοποιώ τη βιβλιοθήκη Regural Expressions για να αποκτήσω μόνο το όνομα της πηγής από το πλαίσιο δεδομένων, έπειτα τις καταρτίζω για να δω τις πρώτες δέκα πηγές και τις συχνότητες τους, και φτιάχνω ένα διάγραμμα που δείχνει τη συχνότητα τους σε ποσοστά.

In [10]:
```
def get_source_name(x):
    value = re.findall(pattern="<[^>]+>([^<]+)</a>", string=x)
    if len(value) > 0:
        return value[0]
    else:
        return ""

dataset.source_name = dataset.source.apply(get_source_name)

source_counts = dataset.source_name.value_counts().sort_values()[-10:]

bottom = [index for index, item in enumerate(source_counts.index)]
plt.barh(bottom, width=source_counts, color="orange", linewidth=0)

y_labels = ["%s %.1f%%" % (item, 100.0*source_counts[item]/len(dataset)) for inde
plt.yticks(np.array(bottom)+0.4, y_labels)

source_counts
```

Out[10]:
```
autotweety.net                2
WordPress.com                 2
Put your button on any page!  2
twittbot.net                  3
Mobile Web (M5)               3
TweetDeck                     5
Twitter Web Client           17
Twitter for Android          17
Twitter for iPhone           20
Google                       22
Name: source, dtype: int64
```



Το Twitter παρέχει API και για τη Java, στο συγκεκριμένο θέμα είναι ποιο εύκολο να γράψεις και να διαβάσεις στη Python, και παρέχει περισσότερα εργαλεία για την ανάλυση των δεδομένων, αλλά αυτό δε σημαίνει πως δε μπορεί κανείς να κάνει την ίδια δουλειά χρησιμοποιώντας τη Java. Χρησιμοποιώντας το Jupyter Notebook έχω καλύτερο έλεγχο πάνω στο κώδικα που γράφω, και μπορώ να τμηματοποιήσω το κάθε μέρος του, επίσης μου επιτρέπει να κρατώ σημειώσεις πάνω σε κάθε τμήμα του κώδικα, το οποίο είναι χρήσιμο για επιστημονικές εργασίες.

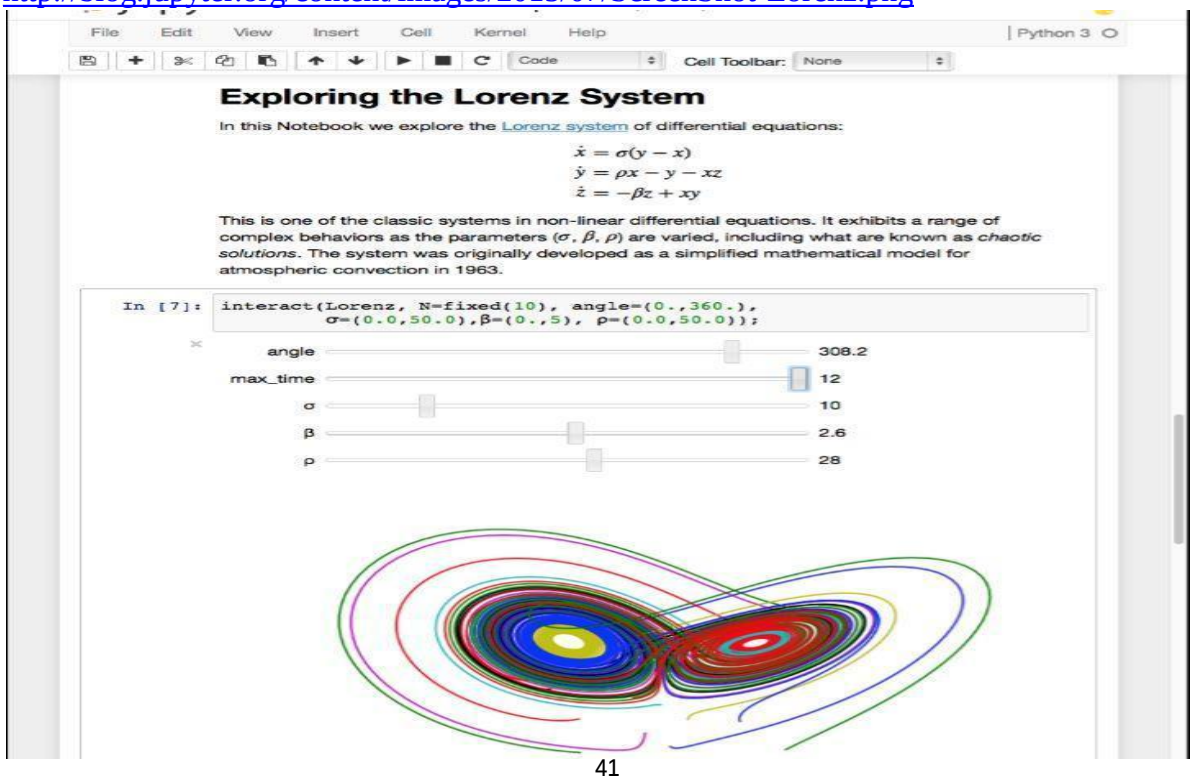
Η γενική μέθοδος της εξόρυξης δεδομένων που εμφανίζετε και στους δύο παραπάνω κώδικες είναι:

- **Συλλογή δεδομένων** – Πρώτα συλλέγουμε τα δεδομένα, χρησιμοποιώντας κάποιες λέξεις κλειδιά μπορούμε να συλλέξουμε δεδομένα για οποιοδήποτε αντικείμενο.
- **Δόμηση δεδομένων** – Δομούμε τα δεδομένα έτσι ώστε να είναι ποιο εύκολα στην ανάγνωση, και για να ξεχωρίσουμε τις σημαντικές πληροφορίες από τις περιττές.
- **Ανάλυση δεδομένων** – Αναλύουμε τα δεδομένα που έχουμε αποκτήσει για ανακάλυψη γνώσης, όπως συχνότητα των λέξεων που χρησιμοποιήθηκαν, πήγες από όπου προήλθαν τα δεδομένα, κτλ.

Έχοντας τα τρία πάνω βήματα στο μυαλό μπορεί κανείς να δημιουργήσει το δικό του σχέδιο για εξόρυξη δεδομένων, και όχι μόνο στο προγραμματιστικό τομέα, αλλά και στο τομέα των επιχειρήσεων όταν ζητάνε τη γνώμη του καταναλωτή είναι ένας τρόπος εξόρυξης δεδομένων, όπου αποκτούν γνώση πάνω στις προτιμήσεις των καταναλωτών.

Το θεωρητικό κομμάτι λοιπόν είναι εύκολο, η μεγαλύτερη δυσκολία που αντιμετώπισα γράφοντας τους παραπάνω κώδικες ήταν να συνδέσω όλα τα κομμάτια που χρειάστηκα μαζί, δηλαδή τις βιβλιοθήκες, τη βάση δεδομένων και να τα κάνω όλα να δουλέψουν στο πρόγραμμα περιήγησης χάρης το Jupyter Notebook.

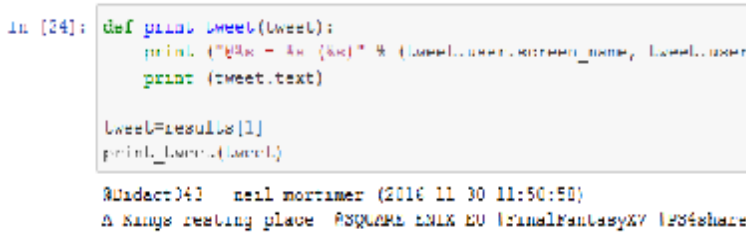
http://blog.jupyter.org/content/images/2015/07/ScreenShot-Lorenz.png



Κεφάλαιο 15.Αποτελέσματα εξόρυξης δεδομένων

Από τη στιγμή που ο κώδικας είναι έτοιμος για να τον τρέξουμε, τα αποτελέσματα είναι ποικίλα, ανάλογα με το τι δεδομένα αναζητούμε. Θέτοντας στο κώδικα τη να αναζητήσει έχουμε διάφορα αποτελέσματα σε σχέση με το κείμενο, τους χρήστες και από πού προήρθε το tweet. Στο REST API αναζήτησα πράγματα όπως FinalFantasyXV, BusinessAdministration, και άλλα που θα παρουσιάσω.

```
In [3]: results = api.search(q="#FinalFanta
```



Αλλάζοντας το τη ψάχνω θα μου εμφανίσει ποικίλα αποτελέσματα. Και καθώς καταρτίζω τα δεδομένα μου σε πίνακες, γίνονται πιο εύκολα στην ανάγνωση.

```
In [33]: results = []
         for tweet in tweepy.Cursor(api.search, q="Business
             results.append(tweet)

         print (len(results))

         100
```

```
In [35]: data_set.head(5)
```

Out[35]:

	id	text	created_at	retweet_count	favorite_count	source
0	834386721160650752	RT @wtcep: Learn new skills in #BusinessAdmini...	2017-02-22 12:57:51	1	0	Twitter Androi
1	834376192698548224	#Apprenticeship Vacancies with @dalsderby #Bus...	2017-02-22 12:16:00	0	0	TweetD
2	834368530535612416	Great seeing learners progressing so well. Wit...	2017-02-22 11:45:34	0	0	Twitter iPhone
3	834311826317910016	The Best Master's Degrees to Consider* in 2017...	2017-02-22 08:00:14	0	0	Hootsu
4	834311823092506625	The Best Master's Degrees to Consider* in 2017...	2017-02-22 08:00:14	0	0	Hootsu

Όπως βλέπουμε παίρνουμε πολλές πληροφορίες, για το πότε δημιουργήθηκε το tweet,ποιος είναι ο δημιουργός του, από ποια πηγή το δημιούργησε, τη τοποθεσία του χρήστη, και άλλες πολλές πληροφορίες. Διάφορα άλλα αποτελέσματα που βρήκα είναι:

```
In [36]: results = []
         for tweet in tweepy.Cursor(api.search, q="Cri
             results.append(tweet)

         print (len(results))
```

That's the challenge. I think I'm jus...	2017-02-22 13:37:00	0	1	for iPhone	617748826568134656	mage
RT @TessFowler: Glorious Gilmore. \n\n#Critica...	2017-02-22 13:36:57	62	0	Twitter for Android	30042733	emmi
RT @TessFowler: Glorious Gilmore. \n\n#Critica...	2017-02-22 13:35:50	62	0	Twitter for iPhone	389416809	Valer
RT @ThatNerdGurl_: Blessing yo timeline with t...	2017-02-22 13:32:54	14	0	Twitter Web Client	336442881	LyrWa
RT @Gamer_Artist: #criticalrole and so the cri...	2017-02-22 13:32:45	49	0	Twitter Web Client	336442881	LyrWa
RT @TessFowler: Glorious Gilmore, in progress....	2017-02-22 13:32:27	15	0	Twitter for iPhone	50839655	Dasa
Hey #Critters! Check out the #etched "Critical...	2017-02-22 13:32:15	0	0	Twitter Web Client	244176821	MC_E

```
In [46]: results = []
         for tweet in tweepy.Cursor(api.search, q
             results.append(tweet)

         print (len(results))
```

4	Spyros Chatzigiannis	2010-02-08 16:36:36	Freelance Journalist. Worked for 10 years (200...	1300
ns	ncctanos Ioannidis	2009-07-22 01:23:31	Brooklyn Filmmaker & Video Editor	887
;b	George Roussos	2010-09-04 20:13:14	Ερασιτέχνης της 7ης Τέχνης - Κινηγώντας @chimeres ...	4461
:ccas	Konstantino Roccas	2011-05-24 17:20:19	Grad of UBC J-School. Usually right, rarely wr	195
;b	George Roussos	2010-09-04 20:13:14	Ερασιτέχνης της 7ης Τέχνης - Κινηγώντας @chimeres	4461
4ich	SarantisMichal	2012-06-12 09:07:46	Journalist at TurkAct.v Media Network	610
;b	George Roussos	2010-09-04 20:13:14	Ερασιτέχνης της 7ης Τέχνης Κινηγώντας @chimeres ...	4461

Όπως βλέπουμε έχουμε διάφορα αποτελέσματα, με το REST API μπορούμε να ψάχνουμε για μια λέξη κλειδί ή μία σειρά χαρακτήρων κάθε φορά, το Streaming API μας επιτρέπει την αναζήτηση πολλαπλών λέξεων ή σειρών χαρακτήρων κάθε φορά, στο συγκεκριμένο έκανα αναζήτηση για τρεις λέξεις Twitch, Youtube, και Netflix.

Tracking keywords

```
keywords = ["Twitch",
            "Youtube",
            "Netflix",
           ]

# Visualize a progress bar to track progress
progress_bar = wgt.IntProgress(value=0)
display(progress_bar)
wgt_status = wgt.HTML(value="""<span class="label label-prim
display(wgt_status)

# Start a filter with an error counter of 20
for error_counter in range(20):
    try:
        myStream.filter(track=keywords)
        print("Tweets collected: %s" % myStream.listener.cou
        print("Total tweets in collection: %s" % col.count()
        break
    except:
        print("ERROR# %s" % (error_counter + 1))
```

```
Finished
Total Mining Time: 0:00:03.009174
Tweets/Sec: 33.3
Tweets collected: 100
Total tweets in collection: 100
```

Έπειτα αποθηκεύω τα αποτελέσματα σε μια βάση δεδομένων και φτιάχνω ένα πλαίσιο που παρουσιάζω συγκεκριμένες πληροφορίες σχετικά με τα δεδομένα που απέκτησα.

In [7]:
```
dataset = [{"created_at": item["created_at"],
            "text": item["text"],
            "user": "@%s" % item["user"]["screen_name"],
            "source": item["source"],
           } for item in col.find()]

dataset = pd.DataFrame(dataset)
dataset
```

Out[7]:

	created_at	source	text
0	Mon Nov 14 12:09:23 +0000 2016	<a href="http://publicize.wp.com/" rel="nofoll...	ملتي حلب تراسل h
1	Mon Nov 14 12:09:23 +0000 2016	<a href="http://www.google.com/" rel="nofollow...	We Use to have
2	Mon Nov 14 12:09:23 +0000 2016	<a href="http://www.facebook.com/twitter" rel=...	Enpe elol'nhttp
3	Mon Nov 14 12:09:23 +0000 2016	<a href="http://www.google.com/" rel="nofollow...	I liked a @YouT
4	Mon Nov 14 12:09:23 +0000 2016	<a href="http://twitter.com/download /android" ...	RT @Millicent_
5	Mon Nov 14 12:09:23 +0000 2016	<a href="http://links.com" rel="nofollow">Put...	Видео "110" (h
6	Mon Nov 14 12:09:23 +0000 2016	<a href="http://www.google.com/" rel="nofollow...	ดนตรีที่ใจ @
7	Mon Nov 14 12:09:23 +0000 2016	<a href="http://twitter.com/download /android" ...	RT @Millicent_
8	Mon Nov 14 12:09:23 +0000 2016	<a href="http://twitter.com/download /iphone" r...	2016 | (

70	Mon Nov 14 12:09:25 +0000 2016	<a href="http://twitter.com/download /android" ...	RT @PGYHomeT
71	Mon Nov 14 12:09:25 +0000 2016	<a href="http://twitter.com" rel="nofollow">Tw...	Celine Dion &amp
72	Mon Nov 14 12:09:25 +0000 2016	<a href="http://twitter.com/download /iphone" r...	パ党芸人 座談会 2 https://t.co/VNsE
73	Mon Nov 14 12:09:25 +0000 2016	<a href="https://about.twitter.com /products/tw...	RT @sacanagam
74	Mon Nov 14 12:09:25 +0000 2016	<a href="http://twitter.com/download /iphone" r...	Bella e Vale - Sca
75	Mon Nov 14 12:09:25 +0000 2016	<a href="http://twitter.com" rel="nofollow">Tw...	The Desert: Jour
76	Mon Nov 14 12:09:25 +0000 2016	<a href="http://twitter.com" rel="nofollow">Tw...	https://t.co/dZwq1
77	Mon Nov 14 12:09:25 +0000 2016	<a href="http://twitter.com/download /iphone" r...	RT @LastWeekT
78	Mon Nov 14 12:09:25 +0000 2016	<a href="http://www.google.com/" rel="nofollow...	Gostei de um vide
79	Mon Nov 14 12:09:25 +0000 2016	<a href="https://mobile.twitter.com" rel="nofo...	RT @DrashtiSupp
80	Mon Nov 14 12:09:25 +0000 2016	<a href="https://about.twitter.com /products/tw...	RT @mimyo_: L-H /T98rXo89q...
81	Mon Nov 14 12:09:25 +0000 2016	<a href="https://mobile.twitter.com" rel="nofo...	RT @SolidNaddic
82	Mon Nov 14 12:09:25 +0000 2016	<a href="http://twitter.com/download /iphone" r...	ผลิตผล - wawa (

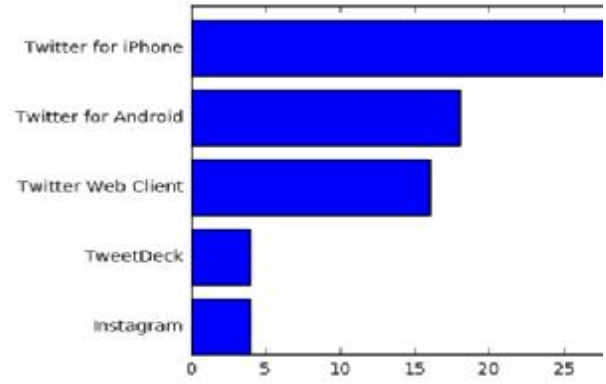
Τα αποτελέσματα είναι πάρα πολλά, αυτά που αντλώ από τα δεδομένα είναι το κείμενο του tweet, πότε δημιουργήθηκε, το όνομα του χρήστη που το δημιούργησε και τη πηγή από την οποία προήρθε. Έτσι μπορούμε να διακρίνουμε καλύτερα τη περιέχει κάθε tweet.

Κεφάλαιο 16. Συμπεράσματα

Έπειτα από όλες αυτές τις εξαγωγές δεδομένων, τι γνώσεις προκύπτουν από τα αποτελέσματα και ποια είναι τα τελικά συμπεράσματα για την εξαγωγή δεδομένων; Και με το REST API και το streaming API προσπάθησα να απομονώσω τις πήγες από τις οποίες προήρθαν τα tweet, με αποτέλεσμα να διακρίνω ποιες πηγές χρησιμοποιήθηκαν περισσότερο από κάποιες άλλες, επίσης κατατάσσω τις λέξεις που χρησιμοποιήθηκαν στα tweets ανάλογα με το ποιες είχαν περισσότερη χρήση.

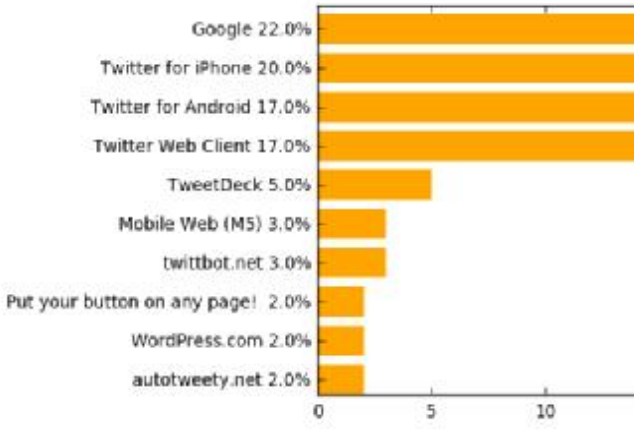
Visualization of the results

```
In [34]: sources = data_set["source"].value_counts()[:5][::-1]

         plt.barh(range(len(sources)), sources.values)
         plt.yticks(np.arange(len(sources)) + 0.4, sources.index)
         plt.show()
```



Όπως βλέπουμε σε αυτή τη περίπτωση του REST API τα περισσότερα tweets προήρθαν από χρήστες iPhone.

```
Out[10]: autotweety.net                 2
         WordPress.com                  2
         Put your button on any page!   2
         twittbot.net                   3
         Mobile Web (M5)                3
         TweetDeck                      5
         Twitter Web Client            17
         Twitter for Android           17
         Twitter for iPhone            20
         Google                        22
         Name: source, dtype: int64
```



Στο streaming API όπου αντλήσαμε ζωντανά δεδομένα βλέπουμε πως η πηγή που χρησιμοποιήθηκε περισσότερο είναι η Google.

Highest used words

In [9]:
```
cv = CountVectorizer()
count_matrix = cv.fit_transform(dataset.text)

word_count = pd.DataFrame(cv.get_feature_names(), columm
word_count["count"] = count_matrix.sum(axis=0).tolist()[
word_count = word_count.sort_values("count", ascending=F
word_count[:50]
```

Out[9]:

	word	count
0	https	108
1	co	106
2	youtube	44
3	rt	31
4	de	15
5	video	13
6	to	10
7	the	10
8	my	9

Στο streaming API χρησμοποίησα το CountVectorizer για να μετρήσω τη συχνότητα των λέξεων που υπήρχαν μέσα στα tweets. Το https σημαίνει ότι τη μεγαλύτερη συχνότητα την έχουν διάφορα Links, δηλαδή στα περισσότερα tweets εμφανίζονται σύνδεσμοι που παραπέμπουν σε άλλες ιστοσελίδες ή και στην ίδια ιστοσελίδα.

Όπως είδαμε λοιπόν είναι εύκολο να αποκτήσει κανείς μεγάλα σύνολα δεδομένων πάνω σε ένα θέμα, και να αντλήσει από αυτά διάφορες πληροφορίες για να αποκτήσει τη γνώση που χρειάζεται. Το twitter είναι μία πολύ καλή πηγή εξαγωγής δεδομένων, καθώς χρησιμοποιείτε από ποικίλες κατηγορίες ανθρώπων όπως πολιτικούς, καλλιτέχνες, αθλητές και πολλούς άλλους. Επίσης το tweeter παραχωρεί ελεύθερα τη χρήση των API του για ανάκτηση και διάβασμα κειμένου, απλά φτιάχνοντας ένα λογαριασμό. Μπορούμε να αντλήσουμε ζωντανά δεδομένα όταν για παράδειγμα εκδίδεται ένα νέο βιβλίο ή όταν ένα νέο παιχνίδι γίνεται διαθέσιμο, για να δούμε τη πιστεύουν για το νέο προιόν και πως το

δέχονται. Υπάρχουν πολλοί τομείς που μπορούν να επωφεληθούν από καλά προγράμματα εξαγωγής δεδομένων, θα ήταν πολύ χρήσιμο για τη Διεύθυνση Δίωξης Ηλεκτρονικού Εγκλήματος, για τη πρόβλεψη εγκλημάτων και τρομοκρατικών Χτυπημάτων, αλλά κυρίως χρησιμοποιείται για εταιρική κατασκοπεία και απόκτηση γνώσης από εταιρείες. Το πρόγραμμα μου θα μπορούσε να βελτιωθεί για αναζήτηση δεδομένων παλαιών ετών, αλλά χρειάζεται επικοινωνία και άδεια από μέλος του Tweeter API.

Βιβλιογραφία

Πρόσβαση την 24η Νοεμβρίου 2016 12:30.

Data mining definition
http://www.investopedia.com/terms/d/datamining.asp
Data Science
http://www.investopedia.com/terms/d/data-science.asp
Big Data
http://www.investopedia.com/terms/b/big-data.asp
Social Data
http://www.investopedia.com/terms/s/social-data.asp
Cloud Computing
http://www.investopedia.com/terms/c/cloud-computing.asp
Cloud Storage
http://www.investopedia.com/terms/c/cloud-storage.asp
Software as a Service-SaaS
http://www.investopedia.com/terms/s/software-as-a-service-saas.asp
Data Breach
http://www.investopedia.com/terms/d/data-breach.asp
Data Loss
http://www.investopedia.com/terms/d/data-loss.asp
Data Anonymization
http://www.investopedia.com/terms/d/data-anonymization.asp
De-Anonumization
http://www.investopedia.com/terms/d/deanonymization.asp
Silk road
http://www.investopedia.com/terms/s/silk-road.asp
Dark Web
http://www.investopedia.com/terms/d/dark-web.asp
Data Analytics
http://www.investopedia.com/terms/d/data-analytics.asp
Descriptive Analytics
http://www.investopedia.com/terms/d/descriptive-analytics.asp
Predictive Analytics
http://www.investopedia.com/terms/p/predictive-analytics.asp
Data mining tutorial
https://www.tutorialspoint.com/data_mining/data_mining_tutorial.pdf
Python 3.4 Docs
https://docs.python.org/release/3.4.0/
MongoDB Docs
https://docs.mongodb.com/
Jupyter notebook docs
http://jupyter.readthedocs.io/en/latest/install.html
Twitter developer docs REST APIs
https://dev.twitter.com/rest/public
Twitter developer docs Streaming APIs
https://dev.twitter.com/streaming/overview
Twitter data mining using python
https://www.youtube.com/user/roshanRush/videos