ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΜΑΤΙΚΗΣ

Sentiment Analysis

Πτυχιακή εργασία

Δωρίτη Δημήτρη Παναγιώτη

Αθήνα, 2018

ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΜΑΤΙΚΗΣ

Τριμελής Εξεταστική Επιτροπή

Γεώργιος Δημητρακόπουλος
Επίκουρος Καθηγητής, Πληροφορική και Τηλεματική, Ψηφιακής τεχνολογίας

Δημοσθένης Αναγνωστόπουλος
Καθηγητής, Πληροφορική και Τηλεματική, Ψηφιακής τεχνολογίας

Ηρακλής Βαρλάμης
Επίκουρος Καθηγητής, Πληροφορική και Τηλεματική, Ψηφιακής τεχνολογίας

Ο/Η Δωρίτης Δημήτρης Παναγιώτης

δηλώνω υπεύθυνα ότι:

1) Είμαι ο κάτοχος των πνευματικών δικαιωμάτων της πρωτότυπης αυτής εργασίας και από όσο γνωρίζω η εργασία μου δε συκοφαντεί πρόσωπα, ούτε προσβάλει τα πνευματικά δικαιώματα τρίτων.

2) Αποδέχομαι ότι η ΒΚΠ μπορεί, χωρίς να αλλάξει το περιεχόμενο της εργασίας μου, να τη διαθέσει σε ηλεκτρονική μορφή μέσα από τη ψηφιακή Βιβλιοθήκη της, να την αντιγράψει σε οποιοδήποτε μέσο ή/και σε οποιοδήποτε μορφότυπο καθώς και να κρατά περισσότερα από ένα αντίγραφα για λόγους συντήρησης και ασφάλειας.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τον κ. Δημητρακόπουλο για τις συμβουλές, την καθοδήγηση και την υπομονή κατά την εκπόνηση της πτυχιακής μου εργασίας.

Ευχαριστώ πολύ την οικογένειά μου και τους φίλους μου, που πάντα με υποστηρίζουν και με ενθαρρύνουν ό,τι και αν επιχειρώ.

Τέλος, θέλω να ευχαριστήσω τον γάτο μου, που ό,τι και να είναι αυτό που με απασχολεί, αυτός κάθεται και λιάζεται θυμίζοντας μου πως, μάλλον δεν έχει τόση σημασία.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Περίληψη

Η παρούσα πτυχιακή εργασία με τίτλο “Sentiment Analysis” έχει ως βασικό της στόχο τη δημιουργία μιας δικτυακής εφαρμογής, μέσω της χρήσης jsp javascript και jquery, που θα παρέχει στον χρήστη μία εκτίμηση της πολικότητας του κειμένου όσον αφορά τη συναισθηματική του προδιάθεση, καθώς και στατιστικά που θα εξηγούν την απόφαση της. Η Εφαρμογή που εν τέλει δημιουργήθηκε, παρέχει την υπηρεσία αυτή, σε αγγλικά, καθώς και ελληνικά.

Αρχικά, θα αναλυθούν ορισμένες έννοιες, όπως η εξόρυξη δεδομένων, προκειμένου να εξηγήσουμε τους στόχους του γενικού αντικειμένου της συναισθηματικής ανάλυσης, καθώς επίσης και τους τρόπους που αυτή υλοποιείται.

Στην συνέχεια θα, αναφερθούν αναλυτικά, όλα τα εργαλεία και οι τεχνολογίες που χρησιμοποιήθηκαν για την υλοποίηση της εφαρμογής.

Πρώτα θα αναφερθούμε στο WordNet, που αποτέλεσε τη βάση της δημιουργίας του του SentiWordNet 3.0, εργαλείο ειδικά σχεδιασμένο για κατηγοριοποίηση συναισθημάτων και εξόρυξη απόψεων(Baccianella, Esuli, and Sebastiani, n.d.), το οποίο χρησιμοποιήθηκε για την υλοποίηση της εφαρμογής. Σαφώς θα αναλυθεί και η λειτουργία του SentiWordNet.

Θα αναφερθούν και οι γλώσσες προγραμματισμού τις οποίες χρησιμοποίηση η εφαρμογή Sentiment Analysis, δηλαδή η JSP και η Javascript, καθώς και η βιβλιοθήκη της Javascript, Jquery.

Εν συνεχεία, αναλύεται η λειτουργικότητα της δημιουργηθείσας εφαρμογής, η σύνδεσή της με το SentiWordNet καθώς και οι παράγοντες που χρειάστηκε να ληφθούν υπόψιν κατά την υλοποίησή της. Μελετάται ο κώδικας της εφαρμογής και παρουσιάζονται παραδείγματα αποτελεσμάτων ανάλογα με το δοθέν κείμενο.

Για τους σκοπούς της ανάλυσης, έχει πραγματοποιηθεί σύγκριση των αποτελεσμάτων με την ανθρώπινη κρίση η οποία και θα παρουσιαστεί στο τέλος της ανάλυσης μαζί με δυνατές μελλοντικές βελτιώσεις και επεκτάσεις της εφαρμογής σε επόμενες εκδόσεις.

Λέξεις κλειδιά: [Συναισθηματική Ανάλυση, Εξόρυξη Δεδομένων, JSP,Javascript]

Abstract

The present thesis entitled "Sentiment Analysis" aims at creating a web application, through the use of JSP,Javascript and Jquery, which offers to the user an evaluation of the sentimental

polarity of the text they chose to input, as well as charts and statistics that explain the evaluation. The created application, offers the mentioned above feature in both English and Greek.

Firstly, certain terms, such as that of Data Mining, will be analyzed, in order to better explain the targets of Sentiment Analysis as well as the methods of its implementation.

Afterwards, all the tools and technologies that were used in order to create the given application will be thoroughly studied.

First to be analyzed, will be WordNet, a lexical database in english which was the base for creating SentiWordNet 3.0 an enhanced lexical resource for sentiment analysis and opinion mining that is used by the Sentiment Analysis application. The functionality of SentiWordNet 3.0 will also be explained.

The programming languages JSP and Javascript and Javascript's library Jquery, that the “Sentiment Analysis” makes use of will also be analyzed.

Then the function of the created application, and it's connection to SentiWordNet 3.0, as well as the factors that had to be taken into account will be examined. The application's code will be studied, and examples of the results it provides will be shown.

For research purposes and to measure the accuracy of the application, the results of “Sentiment Analysis Application”, are compared to the evaluations of human judgement. The results of the comparison are shown at the end of the thesis as well as recommendation and thoughts on future versions.

Keywords: [Sentiment Analysis, Data Mining, JSP, Javascript]

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ

KDD	Knowledge Discovery from Data

DM	Data Mining
SWN (3.0)	SentiWordNet (3.0)
WN	WordNet
JSP	Java Server Pages

ΕΙΣΑΓΩΓΗ

Τα τελευταία χρόνια, με την έξαρση της κοινωνικής δικτύωσης και την ροπή ενός πολύ μεγάλου ποσοστού του παγκόσμιου πληθυσμού προς την επικοινωνία μέσω του ίντερνετ, η αξία της συναισθηματικής ανάλυσης των μηνυμάτων και πληροφοριών που μεταδίδονται καθημερινώς έχει ανέβει κατακόρυφα.

Άσχετα με την πηγή ή την φύση της εκάστοτε πληροφορίας, μπορούμε να την εντάξουμε σε μία από τις εξής δύο κατηγορίες : Γεγονός ή άποψη.(Bing 2010) Όσον αφορά τα γεγονότα, μπορούμε να υποθέσουμε πως δεν υπάρχει αρνητική ή θετική προδιάθεση καθώς εξ ορισμού πρόκειται για παράθεση συμβάντων. Αντίθετα όμως όταν πρόκειται για απόψεις το περιεχόμενο είναι καλό ή κακό και πολύ σπάνια ουδέτερο(Namrata, Manjunath, and Skiena n.d.), καθώς αποτελούν έκφραση των αντιλήψεων και συναισθημάτων του ατόμου πάνω στο θέμα.

Λόγω αυτού του γεγονότος, οι εφαρμογές της συναισθηματικής ανάλυσης και της εξόρυξης δεδομένων είναι άπειρες. Είτε πρόκειται για την προστασία του ευαίσθητου κοινού από τις πολλές μορφές διαδικτυακού bullying, είτε για την παρακολούθηση συμπεριφορών προκειμένου να εξελίσσονται οι υπάρχουσες εφαρμογές αλλά και οι προσεχείς, η συναισθηματική ανάλυση αποτελεί πολύτιμο εργαλείο για τη βελτίωση της εμπειρίας του ατόμου κατά τη διάρκεια του της πλοήγησης του διαδικτύου. Άλλη μορφή αξιοποίησης των πληροφοριών που παρέχει η συναισθηματική ανάλυση είναι σαφώς η εύρεση ελαττωμάτων και πλεονεκτημάτων σε προϊόντα διαφόρων ειδών. Γνωρίζοντας αν οι αγοραστές/πελάτες έχουν καλή ή κακή άποψη για τις παροχές που προσφέρει, μία επιχείρηση έχει τη δυνατότητα να βελτιώνεται διαρκώς.

Πέραν όμως των οφελών που έγκεινται στην προσωπική εμπειρία ενός χρήστη του διαδικτύου και στο κέρδος επιχειρήσεων, υπάρχουν πολλά οφέλη και στον επιστημονικό τομέα. Αυτό οφείλεται σε μεγάλο βαθμό, στο τεράστιο δείγμα ανθρώπων διαφορετικών ηλικιών, εθνικοτήτων, πολιτισμών και συνθηκών ανατροφής, που συναναστρέφονται καθημερινά μέσω διαδικτύου. Είναι λοιπόν αυτή η ποικιλότητα που καθιστά έρευνες στον τομέα των αντιλήψεων και των συναισθημάτων που διεξάγονται με τη χρήση πληροφοριών από το διαδίκτυο εφικτές και πολλές φορές ακριβείς.

Γλωσσολογική ανάλυση σε υπολογιστικό επίπεδο πραγματοποιείται σε πολλές περιπτώσεις, είτε πρόκειται για απλή αποθήκευση δεδομένων είτε για σύνθετες εφαρμογές τεχνητής νοημοσύνης. Η συναισθηματική ανάλυση, έρχεται και προσθέτει ένα τεράστιο νέο πεδίο δυνατοτήτων στην εξέλιξη όλων αυτών των εφαρμογών.

Μέχρι σήμερα λοιπόν, πολλοί έχουν επιδοθεί στην αναγνώριση των νοημάτων που εμπεριέχονται σε φράσεις και σε κείμενα, έχοντας σαν βασικό οδηγό την συναισθηματική επιρροή των λέξεων σε αναγνώστες.

Πολλές βάσεις δεδομένων έχουν δημιουργηθεί, καταγράφοντας έναν τεράστιο αριθμό λέξεων που έχουν χαρακτηριστεί με κάποιον τρόπο ως θετικές, αρνητικές, ή και ουδέτερες, όλα με σκοπό τη δημιουργία εφαρμογών που εκμεταλλεύονται τις δυνατότητες της συναισθηματικής ανάλυσης.

Έτσι και η παρούσα εφαρμογή Sentiment Analysis(η οποία υλοποιήθηκε κυρίως σε JSP), έχει σαν σκοπό της να παρέχει μία όσο το δυνατόν ακριβέστερη αξιολόγηση των κειμένων που εισάγονται σε αυτήν όσον αφορά πάντα την συναισθηματική τους πόλωση.

Όπως θα δούμε και παρακάτω, η εφαρμογή αξιοποίησε μία πληθώρα υπαρκτών τεχνολογιών και λογισμικών προκειμένου να υλοποιήσει μία ορθή διαδικασία συναισθηματικής ανάλυσης.

Παρά το γεγονός ότι η συναισθηματική ανάλυση, είναι πλέον ένα αντικείμενο το οποίο μελετάται από έναν πολύ μεγάλο αριθμό ατόμων αλλά και οργανισμών, η προοπτική προόδου για τον κλάδο αυτόν είναι τεράστια καθώς ο ανθρώπινος λόγος είναι πολύπλοκος και προκύπτουν διάφορα προβλήματα τα οποία δεν έχουν επιλυθεί ακόμα.

Κάποια από τα εμπόδια στην διαδικασία μίας τέλειας συναισθηματικής ανάλυσης θα φανούν και παρακάτω, καθώς η εφαρμογή Sentiment Analysis κλήθηκε να τα αντιμετωπίσει.

Άσχετα όμως με τα εμπόδια, ο κλάδος της εξόρυξης απόψεων και της συναισθηματικής ανάλυσης, αποτελεί έναν από τους σημαντικότερους κλάδους της πληροφορικής που θα αποτελέσει ένα από τα βασικά θεμέλια για την εξέλιξη όλων των εφαρμογών που για οποιονδήποτε λόγο επεξεργάζονται κείμενο.

ΚΕΦ.1: Βασικοί Όροι Sentiment Analysis

1.1. Data Mining-Opinion Mining

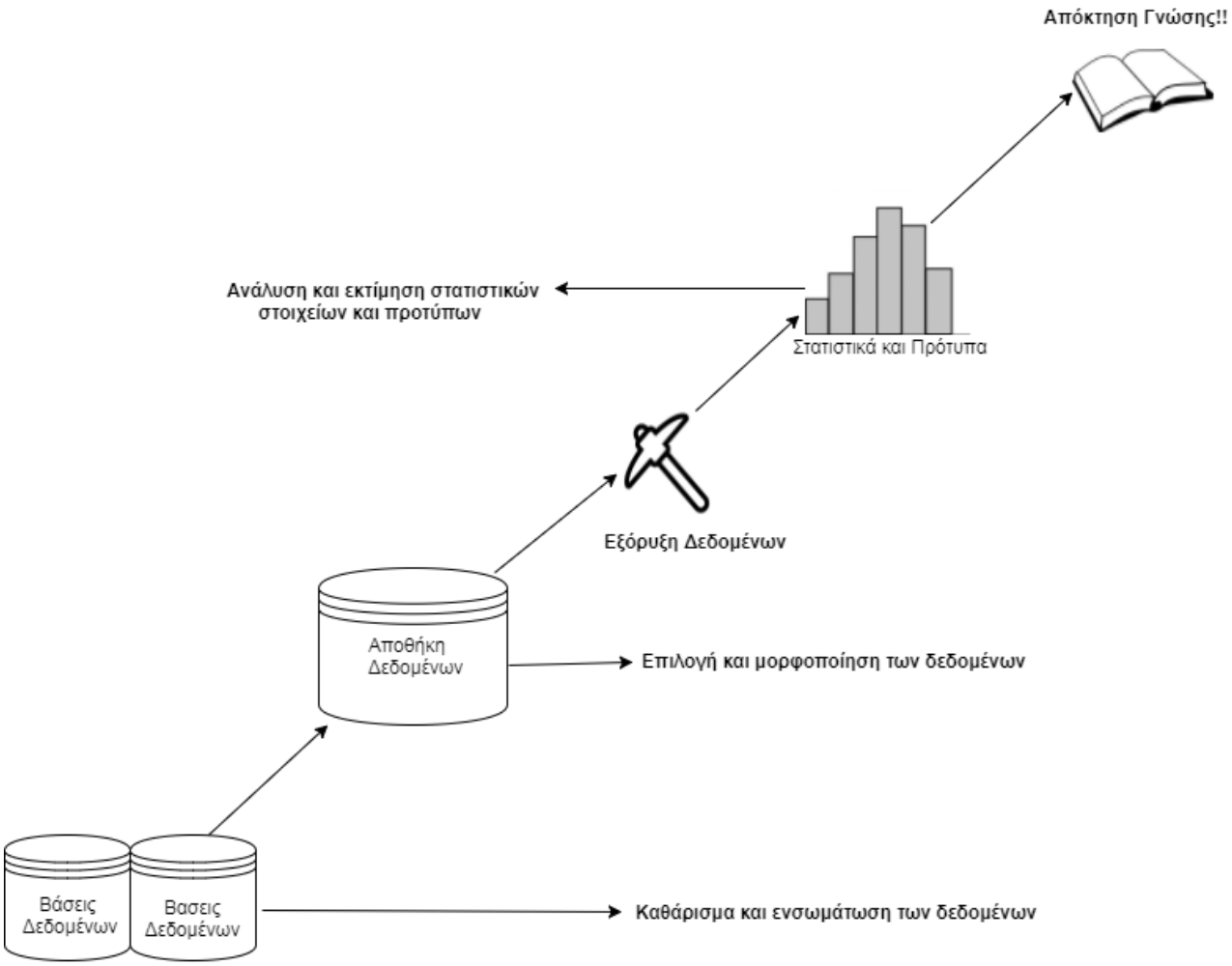
Η διαδικασία εξαγωγής και ανάλυσης συναισθημάτων, που πραγματοποιεί η εφαρμογή, ανήκει σε έναν γενικό κλάδο των computational linguistics, δηλαδή της υπολογιστικής ανάλυσης φυσικής γλώσσας, ο οποίος ονομάζεται Data Mining. Ειδικότερα, ανήκει σε μία υποκατηγορία αυτού του κλάδου, που ονομάζεται Opinion Mining. Παρακάτω, αναλύονται η ερμηνεία και οι στόχοι του κλάδου αυτού της επιστήμης της πληροφορικής.

1.1.1. Data Mining-Ερμηνεία

Ο όρος Data Mining (Εξόρυξη Δεδομένων), καλύπτει σαν έννοια, την διαδικασία εξαγωγής πληροφοριών μέσα από κάποια πηγή κειμένου ή κάποιου άλλου είδους αρχείου. Πρόκειται λοιπόν για μία πολύ γενική και ευρεία έννοια που δεν συλλαμβάνει με ακρίβεια τον στόχο μίας διεργασίας ή ενός project ή τον τρόπο με τον οποίο αυτός ο σκοπός επιτυγχάνεται. Στον φυσικό κόσμο, κατά τη διαδικασία εύρεσης χρυσού ανάμεσα σε πέτρες ή άμμο χρησιμοποιείται η έννοια "Εξόρυξη χρυσού" (ενώ σε άλλες περιπτώσεις μπορεί κανείς να πραγματοποιεί "Εξόρυξη Αλάτων" ή και "Εξόρυξη Πολύτιμων Λίθων")(Jiawei, Jian, and Kamber 2000). Ακόμα όμως και όταν αναφερόμαστε συγκεκριμένα σε εξόρυξη χρυσού, υπάρχουν πολλοί διαφορετικοί τρόποι με τους οποίους αυτή πραγματοποιείται, όπως η μέθοδος Heap Leaching ή το Open Pit Mining. Κατά αντιστοιχία, στον κλάδο του Data Mining σε υπολογιστικό επίπεδο, απαιτείται μία περαιτέρω διευκρίνηση όσον αφορά τα αποτελέσματα που επιθυμεί κανείς να αποφέρει η εξόρυξη δεδομένων καθώς και τον τρόπο με τον οποίο θα την υλοποιήσει.

Μία πιο αναλυτική έκφραση της σημασίας του όρου Data Mining, θα μπορούσε να είναι η "Εξόρυξη Γνώσης από Δεδομένα"(Jiawei, Jian, and Kamber 2000) είναι όμως πολύ μακροσκελής και δύσχρηστη ενώ η έκφραση "Εξόρυξη Γνώσης" παραλείπει εντελώς το πολύ σημαντικό γεγονός πως υλοποιείται μέσω της χρήσης υπολογιστικών δεδομένων(Data).

Πολύς κόσμος, αντιμετωπίζει την εξόρυξη δεδομένων ως ένα συνώνυμο ενός άλλου διαδεδομένου όρου, την "Ανακάλυψη γνώσεων από δεδομένα" (Knowledge Discovery from Data δηλαδή KDD), ενώ άλλοι την θεωρούν ένα κομμάτι της KDD(Jiawei, Jian, and Kamber 2000). Η διαδικασία υλοποίησης της KDD, κομμάτι της οποίας αποτελεί η εξόρυξη δεδομένων, φαίνεται στο παρακάτω σχεδιάγραμμα και αναλύεται λίγο πιο αναλυτικά στη συνέχεια.



Εικόνα 1: Διαδικασία υλοποίησης Knowledge Discovery from Data

Τα βήματα λοιπόν της διαδικασίας υλοποίησης KDD, όπως εμφανίζονται και στην εικόνα, έχουν ως εξής: Πρώτα γίνεται το καθάρισμα των δεδομένων από τις πηγές(στο συγκεκριμένο σχεδιάγραμμα, πρόκειται για βάσεις δεδομένων). Αυτό το βήμα, περιλαμβάνει την αφαίρεση του θορύβου από τα δεδομένα καθώς και την απομάκρυνση "ασυνεπών" δεδομένων(δε μπορεί κανείς εύκολα να εξάγει έγκυρα συμπεράσματα μέσω αυτών). Ύστερα γίνεται η ενοποίηση των δεδομένων, ή αλλιώς η ενσωμάτωση αυτών στην αποθήκη δεδομένων οπού θα πραγματοποιηθούν τα επόμενα βήματα. Όταν λοιπόν συσσωρευτούν τα δεδομένα, η πρώτη διαδικασία που οφείλει να λάβει χώρα, είναι η επιλογή των δεδομένων που περιέχουν, ή μπορεί να περιέχουν, πληροφορίες που αφορούν τον σκοπό του project και στη συνέχεια, η ανάκτησή τους. Έπειτα, τα επιλεγμένα δεδομένα, υφίστανται μορφοποίηση σε μορφές κατάλληλες για την υλοποίηση data mining πάνω σε αυτά. Αφού μορφοποιηθούν,

πραγματοποιείται η εξόρυξη μέσω “έξυπνων μεθόδων” και εξάγονται τα πρότυπα/μοτίβα πληροφοριών. Το τελευταίο κομμάτι της διαδικασίας είναι η ανάλυση και εκτίμηση των πληροφοριών που ανακτήσαμε από τα δεδομένα με βάση κάποια ορισμένα κριτήρια που αφορούν τον αρχικό σκοπό του project. Όταν ολοκληρωθεί και το τελευταίο αυτό βήμα, μπορούμε πλέον να παρουσιάσουμε τη γνώση που εξήχθη από τα δεδομένα, στους χρήστες!

Παρόλο που η διαδικασία που ακολουθείται κατά τη διάρκεια εξόρυξης δεδομένων (με οποιονδήποτε τρόπο και αν υλοποιείται), είναι παρόμοια με αυτή της KDD που περιγράφηκε παραπάνω, υπάρχουν πάντα μικρές διαφορές από εφαρμογή σε εφαρμογή και όπως είδαμε, ο όρος εξόρυξη δεδομένων είναι πολύ γενικός για να μας αποκαλύψει με ακρίβεια την εκάστοτε λογική με την οποία διεξάγεται η εξόρυξη καθώς και το σκοπό μας.

Επειδή λοιπόν είναι απαραίτητες περαιτέρω πληροφορίες για την κατανόηση του ακριβούς σκοπού ενός project data mining, έχουν επινοηθεί πολλοί όροι που ουσιαστικά αποτελούν υποκατηγορίες της έννοιας της εξόρυξης δεδομένων, ένας εκ των οποίων είναι και το Opinion Mining το οποίο και θα αναλυθεί στην συνέχεια.

1.1.2. Data Mining-Επιπλέον Στοιχεία

Ένα βασικό ερώτημα το οποίο οφείλει να απαντηθεί προτού αναλύσουμε περισσότερο τη διαδικασία υλοποίησης είναι σε τι είδους δεδομένα μπορούμε να εφαρμόσουμε Data Mining. Ποια είναι τα δεδομένα δηλαδή από τα οποία μπορούμε να εξάγουμε πληροφορίες χρήσιμες σε οποιοδήποτε επίπεδο. Η απάντηση είναι απλή : **Εξαρτάται**. Στην πράξη οποιοδήποτε είδος ψηφιακών αρχείων και δεδομένων μπορεί να φανεί χρήσιμο ανάλογα με την εφαρμογή(application) . Κατά κύριο λόγο, οι υπαρκτές εφαρμογές χρησιμοποιούν δεδομένα που ανακτώνται από βάσεις δεδομένων συνήθως σε μορφή κειμένου, ήχου ή και εικόνας.

Άλλο ερώτημα το οποίο μπορεί να τεθεί, είναι τι είδους μοτίβα, μπορούν να βρεθούν μέσω της εξόρυξης δεδομένων και κατά πόσο αυτά, είναι ενδιαφέροντα, αναφορικά πάντα με τον στόχο του project. Μ ε τον όρο μοτίβα, αναφερόμαστε σε επαναλαμβανόμενες

συμπεριφορές που μας προσφέρουν κάποιο είδος πληροφορίας. Για να μπορέσουμε να δώσουμε απάντηση στο ερώτημα, πρέπει πρώτα να εξηγήσουμε τον προσανατολισμό διεργασιών data mining και τον όρο της λειτουργικότητας εξόρυξης δεδομένων(Jiawei, Jian, and Kamber 2000). Ο προσανατολισμός των διεργασιών εξόρυξης δεδομένων, αναφέρεται στο αν οι διεργασίες χαρακτηρίζουν τις ιδιότητες των δεδομένων που εισάγουμε ή αν χρησιμοποιούν τα δεδομένα που εισάγουμε με σκοπό να προβλέψουν τις ιδιότητες αυτών που θα εισαχθούν. Ουσιαστικά δηλαδή με την έννοια του προσανατολισμού, ορίζουμε αν ο σκοπός μίας διεργασίας είναι να κατηγοριοποιήσει δεδομένα με βάση τα χαρακτηριστικά τους, ή να προβλέψει τα χαρακτηριστικά καινούριων .Οι λειτουργικότητες, είναι κατά μία έννοια, μέθοδοι με τις οποίες ανακαλύπτουμε και προσδιορίζουμε, πρότυπα/μοτίβα εντός των δεδομένων που έχουμε συλλέξει, τα οποία έχουν κάποια αξία για την έρευνα που διεξάγουμε. Μερικές από αυτές τις λειτουργικότητες είναι οι εξής:

- **Χαρακτηρισμός και διάκρισή.** Κατά την υλοποίηση αυτής της λειτουργικότητας πραγματοποιείται κατηγοριοποίηση εννοιών που μπορεί κανείς να εντοπίσει στα δεδομένα σε έναν αριθμό κλάσεων. Σε κάθε είσοδο δεδομένων, εφαρμόζεται μία περιγραφή που την κατατάσσει σε μία από αυτές τις κλάσεις. Αν για παράδειγμα διεξάγουμε εξόρυξη δεδομένων στο σύνολο των κριτικών που έθεσαν οι πελάτες ενός καταστήματος ηλεκτρικών ειδών, μία κλάση θα μπορούσε να είναι η “Καφετιέρες” ενώ άλλη η “Ψυγεία”. Η κατηγοριοποίηση των αντικειμένων, γίνεται είτε μέσω Χαρακτηρισμού Δεδομένων που υλοποιείται μέσω μελέτης γενικών όρων που αφορούν την εκάστοτε κλάση και στη συνέχεια, συνοψίζοντας τα ευρήματα, είτε μέσω Διάκρισης δεδομένων κατά την οποία εστιάζει κανείς σε βασικές διαφορές ορολογίας μεταξύ των διαφορετικών κλάσεων. Σε ορισμένες περιπτώσεις εφαρμόζονται και οι δύο τεχνικές.

- **Εξόρυξη συχνών μοτίβων σχέσεων και συσχετίσεων.** Όπως συνιστά το όνομα, η λογική αυτής της λειτουργικότητας, βασίζεται στο γεγονός ότι, πολλά τμήματα στα δεδομένα που μελετάμε συχνά επαναλαμβάνονται. Εκτός αυτού κάποιες εκφράσεις συχνά ακολουθούν συγκεκριμένους όρους ή το αντίστροφο. Παραδείγματα όρων που συχνά ακολουθούν ο ένας τον άλλον

μπορεί να είναι η “ποιότητα” και ο “ήχος” σε ηλεκτρονικές συσκευές ή η λέξη “κούπα” και η λέξη “καφές”. Σε άλλες περιπτώσεις επαναλαμβάνονται γενικά μοτίβα. Για παράδειγμα ένας πελάτης σε ένα δικτυακό κατάστημα ηλεκτρονικών ειδών, αφού αγοράσει έναν υπολογιστή, θα ενδιαφερθεί να αγοράσει ένα ποντίκι. Κατά την δημιουργία ενός αλγορίθμου υλοποίησης data mining, λαμβάνει κανείς υπόψιν του αυτές τις συσχετίσεις, φτιάχνοντας, ή ακολουθώντας ήδη υπαρκτές συναρτήσεις που υπολογίζουν την πιθανότητα που υπάρχει, όροι να σχετίζονται μεταξύ τους.

- **Ταξινόμηση και οπισθοδρόμηση για ανάλυση με προβλέψεις.** Όπως συνιστά και το όνομα, η λειτουργικότητα αυτή εφαρμόζεται σε διεργασίες data mining που έχουν ως σκοπό τους να προβλέψουν κάποια αποτελέσματα. Η έννοια της ταξινόμησης βασίζεται στην εύρεση ενός μοντέλου ή κάποιας συνάρτησης, με βάση την οποία ξεχωρίζουμε τα δεδομένα σε κλάσεις και σε έννοιες. Αν τα αντικείμενα των δεδομένων που εξετάζουμε λοιπόν πληρούν κάποιες προϋποθέσεις, κατατάσσονται σε κάποια υπαρκτή κατηγορία δεδομένων. Τα μοντέλα/συναρτήσεις τα οποία χρησιμοποιούνται, περιγράφονται συνήθως με δέντρα αποφάσεων,μαθηματικές φόρμουλες ή και νευρωνικά δίκτυα. Σε αντίθεση με την ταξινόμηση που σαν σκοπό της έχει την πρόβλεψη της κατηγορίας στην οποία ανήκει ένα αντικείμενο και να θέσει κάποια “ταμπέλα” η διαδικασία της οπισθοδρόμησης, αποσκοπεί στην πρόβλεψη συγκεκριμένων αριθμητικών τιμών. Η ανάλυση οπισθοδρόμησης, κάνει χρήση στατιστικών στοιχείων με σκοπό να προβλέψει αριθμούς και να ορίσει την κατανομή των τάσεων με βάση τα υπάρχοντα στοιχεία.

1.2. Opinion Mining-Εξόρυξη Απόψεων

Όπως αναφέρθηκε και προηγουμένως, η έννοια της εξόρυξης απόψεων(Opinion Mining) αποτελεί ένα από τα πολλά παρακλάδια ή αλλιώς μία από τις πολλές υποκατηγορίες της ευρύτερης έννοιας της εξόρυξης δεδομένων. Σε αντίθεση με την γενική έννοια του Data

Mining που μπορεί να αναφέρεται σε πολλών διαφορετικών ειδών διεργασίες, η έννοια του Opinion Mining συνδέεται άρρηκτα με το αντικείμενο του συγκεκριμένου project, το οποίο βασίζεται στην ανάλυση συναισθημάτων. Αν θυμηθούμε την κατηγοριοποίηση του λόγου που πραγματοποιήσαμε προηγουμένως σε φράσεις που παραθέτουν γεγονότα και σε φράσεις που εκφράζουν απόψεις, είναι εύκολο να συμπεράνουμε, πως η διαδικασία του Opinion Mining, μας παρέχει έναν τρόπο να κάνουμε τον διαχωρισμό αυτόν, απομακρύνοντας τα δεδομένα τα οποία δεν είναι κατά οποιονδήποτε τρόπο συναισθηματικά φορτισμένα, οπότε και δε μας ενδιαφέρουν. Η εξόρυξη των απόψεων λοιπόν αποτελεί ένα απαραίτητο βήμα που πρέπει να υλοποιήσουμε προτού μπορέσουμε να προβούμε σε οποιοδήποτε συμπέρασμα σχετικά με την συναισθηματική πολικότητα του κειμένου.

1.3. Sentiment Analysis-Ο Όρος

1.3.1. Sentiment Analysis-Εισαγωγικά Στοιχεία

Η συναισθηματική ανάλυση, αποτελεί ένα παρακλάδι της υπολογιστικής γλωσσικής ανάλυσης κειμένου το οποίο δεν ασχολείται με το περιεχόμενο του κειμένου μα με τις αντιλήψεις και τα συναισθήματα προσπαθώντας να αναγνωρίσει τις υποκειμενικές έννοιες αυτού(Esuli and Sebastiani n.d.). Είναι τομέας της πληροφορικής και της επεξεργασίας φυσικής γλώσσας, που εμπλέκει πολλούς άλλους τομείς όπως η εξόρυξη δεδομένων και έχει μελετηθεί πολύ τα τελευταία χρόνια από πληθώρα ερευνητών. Στο κεφάλαιο αυτό θα εμβαθύνουμε στην έννοια του Sentiment analysis και θα αναλύσουμε πιθανές μεθόδους με τις οποίες μπορεί να πραγματοποιηθεί.

Οι λόγοι για τους οποίους, ο κλάδος της πληροφορικής ασχολήθηκε με την ανάλυση συναισθημάτων ποικίλουν. Ο βασικότερος την συγκεκριμένη στιγμή είναι οι δυναμικές διαφημίσεις. Γνωρίζοντας τις προτιμήσεις ενός χρήστη πάνω σε μία κατηγορία προϊόντων, είναι εύκολο να του παρουσιάσει κανείς άλλα προϊόντα, που βάσει προβλέψεων και εκτιμήσεων έχουν μεγάλες πιθανότητες να του αρέσουν και κατά συνέπεια να οδηγήσουν σε μεγαλύτερα κέρδη.

1.3.2. Sentiment Analysis-Μέθοδοι Υλοποίησης

Από τότε που η συναισθηματική ανάλυση έγινε ένας τόσο σημαντικός τομέας στον κλάδο της υπολογιστικής γλωσσικής ανάλυσης, πολλοί αναλυτές(Bing 2010), έχουν επιδοθεί στην κατηγοριοποίηση και στον ορισμό μεθόδων με τους οποίους μπορούμε να υλοποιήσουμε τη διαδικασία του Sentiment Analysis. Παρακάτω αναφέρονται συντόμως μερικές από αυτές.

- **Sentiment analysis βασισμένη σε χαρακτηριστικά:** Βάσει αυτής της μεθόδου, πρώτα αναγνωρίζονται στο κείμενο οι στόχοι των απόψεων που βρίσκονται σε αυτό και ύστερα, γίνεται αναγνώριση των απόψεων αυτών ως θετικές ή αρνητικές. Παραδείγματος χάρη, στην κριτική ενός προϊόντος, πρώτα θα αναγνωριστεί το χαρακτηριστικό αυτού το οποίο σχολιάζει ο συγγραφέας και μετά θα εκτιμηθεί η πόλωση της γνώμης του. Αυτού του είδους η τεχνική χρησιμεύει σε εφαρμογές που αποσκοπούν στην βελτίωση των προϊόντων μίας επιχείρησης παρέχοντας στοχευμένες πληροφορίες για αυτά όπως για παράδειγμα "Ο αισθητήρας του ποντικιού Χ είναι πολύ κακός".

- **Sentiment analysis με βάση συγκριτικές προτάσεις:** Η τεχνική αυτή βασίζεται στην εξαγωγή πληροφοριών μέσα από την σύγκριση προϊόντων στις φράσεις του κειμένου και όχι στην αξιολόγηση ενός προϊόντος συγκεκριμένα. Ακολουθώντας το προηγούμενο παράδειγμα, μία τέτοια φράση θα μπορούσε να είναι "Ο αισθητήρας του Χ ποντικιού είναι πολύ καλύτερος από τον αισθητήρα του Ψ ποντικιού". Η μέθοδος αυτή παρέχει πληροφορίες όχι για την αξία ή για τα ελαττώματα του προϊόντος βάσει του χρήστη, αλλά για την προτίμησή του ανάμεσα σε δύο προϊόντα, πληροφορία που επίσης είναι πολύ χρήσιμο να ληφθεί υπόψη.

- **Συναισθηματική και υποκειμενική κατηγοριοποίηση:** Η μέθοδος που έχει μελετηθεί περισσότερο και αντιμετωπίζει τη συναισθηματική ανάλυση σαν ένα πρόβλημα απλής κατηγοριοποίησης ενός κειμένου. Στόχος της είναι να εξάγει ένα συμπέρασμα για την θετική ή αρνητική πόλωση του κειμένου σαν σύνολο. Για παράδειγμα εφαρμόζοντας αυτήν την τεχνική σε μία κριτική ενός βιβλίου, οι

πληροφορία που θα πάρουμε είναι αν ο συγγραφέας της κριτικής, έχει θετική ή αρνητική άποψη για το βιβλίο.

- **Αναζήτηση και ανάκτηση απόψεων:** Μία παροχή του κλάδου του Sentiment Analysis, μπορεί να είναι η αναζήτηση μέσω διαδικτύου κάποιου όρου μέσω τις οποίας τα αποτελέσματα που ανακτώνται κατηγοριοποιούνται με βάση τη συναισθηματική πόλωση τους. Θα μπορούσε λοιπόν για παράδειγμα κανείς να αναζητήσει τον όρο "vaccination" και σύμφωνα με τη τεχνική αυτή, πρώτα να ανακτηθούν όλα αποτελέσματα που περιέχουν τον όρο, ύστερα να κατηγοριοποιηθούν και τέλος να εμφανιστούν στον χρήστη ανάλογα με την συναισθηματική τους πόλωση.

1.3.3. Sentiment Analysis-Υποδιαδικασίες

Βάσει των Esuli και Sebastiani, η γενική έννοια της εξόρυξης απόψεων χωρίς να λαμβάνει κανείς υπόψιν τη μέθοδο που χρησιμοποιείται, μπορεί να διαχωριστεί στις εξής τρεις μικρότερες υποδιαδικασίες.

- Καθορισμός υποκειμενικότητας κειμένου. Κατάταξη δηλαδή του περιεχομένου του κειμένου στην κατηγορία των γεγονότων ή σε αυτή των απόψεων. Με αυτό το κομμάτι της συναισθηματικής ανάλυσης ασχολήθηκαν και δημοσίευσαν άρθρο το 2003 οι Βασίλειος Χατζηβασίλογλου και Hong Yu(Hatzivassiloglou and Hong 2003) .

- Καθορισμός προσανατολισμού κειμένου. Δεδομένου πως το περιεχόμενο του κειμένου υπόκειται στην κατηγορία των απόψεων, μπορεί κανείς πλέον να ερευνήσει αν αυτές εκφράζουν αρνητικά ή θετικά συναισθήματα.

- Εκτίμηση έντασης του προσανατολισμού του κειμένου. Αφού αναγνωρίσουμε την θετική ή αρνητική πόλωση των συναισθημάτων του κειμένου μπορούμε να τα κατηγοριοποιήσουμε περαιτέρω σε "πολύ θετικά" ή "λίγο αρνητικά" παραδείγματος χάρη ή να ορίσουμε κάποια μαθηματική κλίμακα κατάταξης.

Άσχετα με το στάδιο της ανάλυσης στο οποίο βρισκόμαστε, βάση της διαδικασίας είναι η εύρεση όρων/λέξεων που ορίζουν τον προσανατολισμό του κειμένου, και μέσω αυτών να πραγματοποιηθεί το machine learning. Λέξεις όπως “καλός”, “κακός”, “ενοχλητικός”, “θαυμάσιος” και ούτω καθεξής, μπορεί κανείς εύκολα να παρατηρήσει πως επηρεάζουν την πόλωση του κειμένου. Από έρευνα σε έρευνα αυτό που αλλάζει, είναι ο τρόπος που επεξεργάζεται ο εκάστοτε ερευνητής τους όρους καθώς και η μέθοδος με την οποία καταλήγει σε μία συλλογή λέξεων αντικειμενικά αρνητικών και αντικειμενικά θετικών.

1.3.4. Sentiment Analysis-Προβλήματα/Προκλήσεις

Είναι πολύ σημαντικό να αναφέρουμε πως όπως και σε οποιαδήποτε άλλη διαδικασία, έτσι και κατά την υλοποίηση της συναισθηματικής ανάλυσης υπάρχουν κάποια εμπόδια τα οποία πρέπει με κάποιον τρόπο να ξεπεραστούν. Όπως και οι πράξεις, έτσι και τα λόγια ή ο τρόπος έκφρασης των ανθρώπων πολύ συχνά διαφέρουν σε πολύ μεγάλο βαθμό παρά το γεγονός ότι αντιμετωπίζουν παρόμοιες αν όχι και ίδιες καταστάσεις. Αυτό από μόνο του αποτελεί ένα πολύ μεγάλο εμπόδιο στην κατανόηση και στην ασφαλή εξαγωγή συμπερασμάτων από κείμενο. Ένα άλλο παράδειγμα ενός τέτοιου προβλήματος, είναι το γεγονός ότι ένας όρος ή φράση που έχει πολύ θετική σημασία σε εντός κάποιου κειμένου, μπορεί να έχει πολύ αρνητική έννοια όταν χρησιμοποιείται σε υπό άλλες συνθήκες. Επίσης, κάτι που καθιστά την εξόρυξη απόψεων και έπειτα την συναισθηματική ανάλυση πολύ δυσκολότερες στην υλοποίηση από άλλες μορφές υπολογιστικής γλωσσικής ανάλυσης κειμένου είναι το γεγονός πως δεν μπορούμε να υποθέσουμε πως αν μπει μία μικρή λέξη ανάμεσα σε δύο άλλες, το περιεχόμενο που εξέφραζαν αυτές οι δύο λέξεις μαζί παραμένει ίδιο. Αν για παράδειγμα κάποιος πει “Τον κύριο Τάδε τον συμπαθώ πολύ” έχει πει κάτι εντελώς διαφορετικό(στην προκειμένη περίπτωση αντίθετο) από κάποιον άλλο που είπε για τον ίδιο κύριο “Τον κύριο Τάδε **δεν** τον συμπαθώ πολύ”. Σε αυτήν την απλή πρόταση είναι εύκολο να θέσουμε κάποιον κανόνα που να λέει πως η λέξη “δεν” αντιστρέφει την συναισθηματική φόρτιση της επόμενης λέξης ή της μεθεπόμενης όπως έγινε στο παράδειγμα, δυστυχώς όμως η επιρροή μίας λέξης στο περιεχόμενο του συνόλου δεν είναι πάντα τόσο προφανής.

Είναι λοιπόν προφανές πως ο κλάδος της ανάλυσης συναισθημάτων είναι πολύ απαιτητικός και σχεδόν πάντα υπάρχει κάτι που δεν μπορεί κανείς να υπολογίσει, ή κάτι το

οποίο μπορεί να βελτιωθεί. Έχουν βρεθεί πολλά τεχνάσματα τα οποία παρακάμπτουν κάποια από τα εμπόδια που αναφέραμε(και πολλά άλλα) και αρκετά από αυτά θα αναλυθούν παρακάτω.

ΚΕΦ.2: Τεχνολογίες υλοποίησης Εφαρμογής Sentiment Analysis

Στο κεφάλαιο αυτό, θα αναλυθούν και θα εξηγηθούν, οι δομές των τεχνολογιών που χρησιμοποιήθηκαν κατά την υλοποίηση της εφαρμογής Sentiment Analysis σε JSP. Συγκεκριμένα θα αναφερθούμε στο WordNet, το οποίο είναι ένα τεράστιο λεξικό της αγγλικής γλώσσας πάνω στο οποίο βασίστηκε η δημιουργία του SentiWordNet 3.0 το οποίο χρησιμοποιεί η εφαρμογή προκειμένου να παρέχει στον χρήστη τις βαθμολογίες των λέξεων. Φυσικά θα εξετάσουμε λεπτομερώς και την δημιουργία του SentiWordNet 3.0 καθώς και την δομή του. Τέλος, θα ασχοληθούμε με τις γλώσσες προγραμματισμού JSP και Javascript καθώς και με μία βιβλιοθήκη της Javascript, την Jquery.

2.1. WordNet-Synsets

Το WordNet, πάνω στο οποίο βασίστηκε το SentiWordNet σε όλες τις εκδόσεις του, είναι μία βάση δεδομένων στην οποία έχει καταγραφεί ένας πολύ μεγάλος αριθμός λέξεων, ταξινομημένος σε ομάδες που ονομάστηκαν Sysnsets. Στην επόμενη παράγραφο αναλύεται η ετυμολογία του όρου Synset. Βασικός σκοπός του WordNet είναι η παροχή ενός εργαλείου χρήσιμο για την γλωσσική ανάλυση κειμένου καθώς και τον κλάδο της τεχνητής νοημοσύνης που ασχολείται με τη φυσική επεξεργασία γλώσσας. Το χαρακτηριστικό του WordNet το οποίο το καθιστά ιδιαίτερο και πολύ σημαντικό για τις προαναφερθείσες διεργασίες είναι το γεγονός πως δεν ασχολείται με την ομοιότητα των λέξεων βάσει των χαρακτήρων που τις αποτελούν, αλλά με την ομοιότητα των εννοιών που εκφράζουν. Χάρη σε αυτό, το WordNet και οι πολλές του εκδοχές χρησιμοποιούνται κατά κόρον σε εφαρμογές που επιχειρούν να αναγνωρίσουν την συναισθηματική πόλωση ενός κειμένου.

Ο όρος Synsets προέρχεται από τη λέξη synonymy, δηλαδή συνωνυμία και τη λέξη sets. Η ονομασία αυτή επιλέχθηκε λόγω του γεγονότος ότι οι λέξεις και φράσεις που εντάχθηκαν

στο WordNet κατατάχθηκαν σε ομάδες συνωνύμων που περιγράφουν μία συγκεκριμένη έννοια. Πέραν των λέξεων, κάθε synset περιέχει και ένα “gloss” μία εξήγηση δηλαδή της έννοιας που περιγράφουν οι λέξεις. Κατά κύριο λόγο ένα synset, αποτελείται από λέξεις που ανήκουν στο ίδιο Pos(Part of speech) είναι δηλαδή όλες ρήματα ή ουσιαστικά ή επιρρήματα ή επίθετα. Για αυτόν τον λόγο, το WordNet είναι χωρισμένο σε τέσσερεις ενότητες καθεμία εκ των οποίων αντιστοιχεί σε ένα μέρος του λόγου.

Εκτός από την συνωνυμία, χρησιμοποιήθηκαν και άλλες σχέσεις για την ένταξη λέξεων σε synsets όπως αυτή της υποκατηγορίας που αντιστοιχεί σε ειδίκευση όπως αυτή που πραγματοποιείται σε αντικειμενοστρεφή προγραμματισμό από τις κλάσεις απογόνους κάποιας άλλης. Ένα παράδειγμα τέτοιας σχέσης είναι η σχέση μεταξύ της λέξης έπιπλο και τραπέζι. Αντίστοιχη αυτής της σχέσης είναι η σχέση σύνολο-μέλος παράδειγμα της οποίας αποτελεί το αυτοκίνητο-μπαταρία.

2.1.1 WordNet-Παραλλαγές

Το WordNet έχει χρησιμοποιηθεί ως βάση προκειμένου να δημιουργηθούν παραλλαγές του, που εκτιμούν ορισμένους παράγοντες όσον αφορά το περιεχόμενο ενός κειμένου. Παρακάτω, αναφέρονται μερικές από αυτές.

- **Wordnet Affect:** Υλοποιήθηκε από τους Carlo Strapparava and Alessandro Valitutti(Strapparava and Valitutti 2004) με σκοπό τη δημιουργία ενός λεξικού αναφοράς όρων που εκφράζουν συναισθηματικά φορτισμένο περιεχόμενο. Η διαδικασία υλοποίησης αποτελείτο από τρία βήματα: Αρχικά συλλέχθηκαν όροι με συναισθηματική πόλωση από εφημερίδες, λεξικά και άρθρα. Εν συνεχεία ορίστηκαν ορίστηκαν μοντέλα έκφρασης περιεχομένου, ένα από αυτά θα μπορούσε να είναι για παράδειγμα το “στοιχεία χαρακτήρα” όπου συμπεριλαμβάνονται όλοι οι συναισθηματικά φορτισμένοι όροι που αφορούν την συμπεριφορά ενός ατόμου. Τέλος εξάγονται συμπεράσματα όσον αφορά την θετική ή αρνητική πόλωση των όρων και κατηγοριοποιούνται βάσει των μοντέλων του δετέρου βήματος. Το WordNet affect, απεδείχθη πολύτιμη προσθήκη στα εργαλεία που έχουν όσοι ασχολούνται με υπολογιστική γλωσσική ανάλυση και έχει ήδη χρησιμοποιηθεί σε πολλές εφαρμογές συναισθηματικής ανάλυσης καθώς και σε

εφαρμογές τεχνητής νοημοσύνης που αφορούν τον "υπολογιστικό χιούμορ" τη δυνατότητα του υπολογιστή δηλαδή να παρέχει προτάσεις που περιέχουν αστείους όρους ή λογοπαίγνια.

- **Wordnet Domains:** Το WordNet domains, δημιουργήθηκε και εκδόθηκε το 2008 και αποτελεί μία ομαδοποίηση των όρων του WordNet σε domains. Οι λέξεις, κατηγοριοποιήθηκαν με βάση τη συσχέτισή τους όσον αφορά τη σημασία τους. Παράγοντες που λήφθηκαν υπόψιν ήταν το πόσο συχνά χρησιμοποιούνται μαζί προκειμένου να εκφράσουν ένα συγκεκριμένο νόημα και κατά πόσο όταν βρίσκονται μόνες τους έχουν παρόμοια σημασία. Η παραλλαγή αυτή του WordNet, χρησιμοποιήθηκε κυρίως σε εφαρμογές γλωσσικής ανάλυσης με σκοπό την κατανόηση του περιεχομένου του κειμένου και λιγότερο σε εφαρμογές συναισθηματικής ανάλυσης.

- **Wordnet Similarity:** Αντίστοιχα με το WordNet Domains, το WordNet similarity ασχολείται με το περιεχόμενο του κειμένου και το αν δύο όροι εκφράζουν παρεμφερές περιεχόμενο, παρέχοντας αριθμητικές εκτιμήσεις που ορίζουν τον βαθμό στον οποίο είναι ίδιοι δύο όροι ή τον βαθμό στον οποίο σχετίζονται(Pedersen, Patwardhan, and Michelizzi 2004).

2.2 SentiWordNet 3.0

Όπως αναφέρθηκε προηγουμένως, το SentiWordNet είναι μία λεκτική πηγή ρητώς σχεδιασμένη για την υποστήριξη εφαρμογών συναισθηματικής κατηγοριοποίησης και εξόρυξης απόψεων(Baccianella, Esuli, and Sebastiani, n.d.). Η πρωταρχική έκδοση του εργαλείου αυτού (SentiWordNet 1.0), δημιουργήθηκε από τους Esuli και Sebastiani το 2006 καθώς η τελευταία και βελτιωμένη έκδοση (SentiWordNet 3.0) από τους Peng και Lee το 2008. Σαν βάση του SentiWordNet, χρησιμοποιήθηκε το WordNet.

Η λογική του SentiWordNet 3.0 βασίζεται στην αντιστοιχία τριών τιμών/βαθμολογιών σε κάθε λέξη ή σύνθετο όρο του WordNet. Οι τρεις αυτές τιμές είναι το Pos(s), το Neg(s) και

τέλος το Obj(s) και ουσιαστικά εκφράζουν τον βαθμό στον οποίο η εκάστοτε λέξη ή όρος είναι θετική, αρνητική ή ουδέτερη αντίστοιχα. Δεδομένου ότι τα ενδεχόμενα να είναι θετικές αρνητικές ή ουδέτερες είναι συμπληρωματικά, οι προαναφερθείσες τιμές έχουν άθροισμα 1 και κάθε μία από αυτές δέχεται τιμές στο κλειστό σύνολο [0,1].

Η ανάθεση των τιμών έγινε μέσω μίας διαδικασίας αποτελούμενης από δύο στάδια. Ένα στάδιο μερικώς εποπτευόμενης εκμάθησης και ένα τυχαίας αναδρομής. Το πρώτο, επιτεύχθηκε με τη χρήση τεσσάρων περαιτέρω διεργασιών, οι οποίες είναι οι εξής:

- Επέκταση σπόρου
- Εκπαίδευση σπόρου
- Κατηγοριοποίηση όρων
- Συνδυασμός ταξινομητών

Κατά τη διαδικασία επέκτασης σπόρου, χρησιμοποιήθηκαν δύο σετ σπόρων, εκ των οποίων ο πρώτος αποτελείτο από όλες τις φράσεις/σύνθετους όρους που περιέχουν 7 παραδειγματικά θετικούς όρους και αντίστοιχα ο δεύτερος από όλες όσες περιέχουν 7 παραδειγματικά αρνητικούς. Οι σπόροι αυτοί, επεκτάθηκαν αυτόματα μέσω αναδρομής σε δυαδικές σχέσεις του WordNet (δικτυακή βάση δεδομένων λέξεων σχεδιασμένη για τη χρήση σε προγράμματα(Miller 1995)) που εκφράζουν το αν οι όροι διατηρούν ή αντιστρέφουν τη συναισθηματική πόλωση του κειμένου.

Στο δεύτερο βήμα οι δύο σύνθετοι όροι που προέκυψαν από το πρώτο, χρησιμοποιήθηκαν σε συνδυασμό με άλλους σύνθετους όρους που θεωρήθηκε πως έχουν ουδέτερη πόλωση, προκειμένου να εκπαιδευτεί ένας τριαδικός ταξινομητής που κατηγοριοποιεί δηλαδή τους όρους ως θετικούς αρνητικούς ή ουδέτερους. Για την εκπαίδευση του ταξινομητή, δε χρησιμοποιήθηκαν οι λέξεις αυτούσιες αλλά τα σχόλια (glosses) των όρων.

Σε αντίθεση με το SentiWordNet 1.0 που χρησιμοποιεί το μοντέλο “bag of words” στο οποίο αγνοείται η γραμματική και η σύνταξη του κειμένου, το SentiWordNet 3.0 κάνει χρήση αποσαφηνισμένων αισθήσεων από το Princeton WordNet Gloss Corpus σύμφωνα με το οποίο ένα σχόλιο είναι μία αλληλουχία σύνθετων όρων από το WordNet. Για αυτό το λόγο, βάσει των

Esuli και Sebastiani η τεχνική δημιουργίας του SentiWordNet 3.0 θα μπορούσε να χαρακτηριστεί “bag of synsets”(Baccianella, Esuli, and Sebastiani, n.d.)

Κατά τη διάρκεια του τρίτου βήματος, ομαδοποιήθηκαν οι λέξεις από το πρώτο βήμα μέσω της χρήσης του ταξινομητή του δευτέρου βήματος.

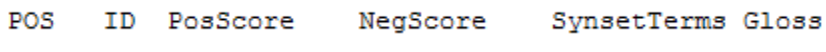
Το τέταρτο βήμα προέκυψε από το γεγονός πως η επέκταση των σπόρων του πρώτου βήματος μπορεί να πραγματοποιηθεί με διαφορετικές τιμές της ακτίνας. Ακτίνα είναι η απόσταση στον γράφο που προκύπτει από τις δυαδικές σχέσεις του WordNet που χρησιμοποιήθηκαν στο πρώτο βήμα, μέσα στην οποία όλες φράσεις/σύνθετοι όροι προστίθενται στους αρχικούς σπόρους λέξεων.

Σύμφωνα με τους Esuli και Sebastiani στο άρθρο τους “Determining Term Subjectivity and Term Orientation for Opinion Mining”(Esuli and Sebastiani n.d.), οι εκτιμήσεις της πόλωσης μίας λέξης είναι πολύ πιο ακριβείς, αν αντί για έναν τριαδικό ταξινομητή χρησιμοποιηθεί μία ομάδα από τριαδικούς ταξινομητές. Για αυτόν τον λόγο, κατά τη δημιουργία του SentiWordNet 3.0 χρησιμοποιήθηκαν 8 διαφορετικοί ταξινομητές, καθένας εκ των οποίων προέκυψε από έναν διαφορετικό συνδυασμό ακτίνας και διαδικασίας machine learning.

Κατά το δεύτερο στάδιο, το WordNet, αντιμετωπίστηκε σαν γράφος και έτρεξαν σε αυτόν μία αναδρομική διεργασία τυχαίου βήματος (random-walk) κατά τη διάρκεια της οποίας οι τιμές που ορίστηκαν στο πρώτο στάδιο για τις πιθανότητες θετικής, αρνητικής, ή ουδέτερης πόλωσης πιθανώς αλλάζουν. Η διαδικασία λαμβάνει τέλος όταν οι τιμές αυτές συγκλίνουν σε μία τιμή.

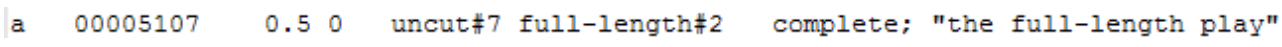
2.2.1 SentiWordNet 3.0-Δομή

Πέραν της διαδικασίας δημιουργίας και των ιστορικών δεδομένων είναι καλό να γνωρίζουμε την τελική μορφή του αρχείου του SentiWordNet 3.0 που διατίθεται ελεύθερα στο http://sentiwordnet.isti.cnr.it/ και αποτελεί και την βάση της Sentiment Analysis σε jsp εφαρμογής.
Πρακτικά, το αρχείο είναι της μορφής txt και περιέχει σε απλό κείμενο όλες τις πληροφορίες που παρέχει. Πρόκειται για ένα μεγάλο λεξικό με πολλά λήμματα, στα οποία έχουν αντιστοιχηθεί κάποιες τιμές που αφορούν την συναισθηματική πολικότητα. Πάνω από τις γραμμές των λημμάτων, υπάρχει αρχικά, η παρακάτω γραμμή με τις ονομασίες των στηλών.



Εικόνα 2: Στήλες SentiWordNet 3.0

Σε αυτό το κεφάλαιο, θα αφοσιωθούμε στο να εξηγήσουμε τι εκφράζει κάθε μία από αυτές τις στήλες και προκειμένου να το πετύχουμε, θα δείξουμε και ένα λήμμα από το SentiWordNet



Εικόνα 3: Λήμμα SentiWordNet 3.0

Το λήμμα λοιπόν αναφέρεται στην λέξη “uncut” και ως συνώνυμο της υπάρχει η λέξη/φράση “full-length” και οι στήλες του SentiWordNet, είναι οι εξής:

- **POS:** Τα αρχικά POS σημαίνουν part of speech δηλαδή μέρος του λόγου, αυτή η στήλη λοιπόν, μας δείχνει αν η λέξη είναι ρήμα, το οποίο συμβολίζεται σε λήμματα ρημάτων με το γράμμα “v”(verb), ή επίθετα το οποίο συμβολίζεται στα λήμματα επιρρημάτων με το γράμμα “a”(adjective), ή ουσιαστικό το οποίο συμβολίζεται στα λήμματα με το γράμμα “n”(noun), ή τέλος τα επιρρήματα τα οποία συμβολίζονται με το γράμμα r. Στο παράδειγμα αυτό σημαίνει πως στη γραμμή θα βρούμε ένα επίθετο(λόγω του γράμματος a με το οποίο ξεκινάει η γραμμή) και πράγματι βρίσκουμε τη λέξη “uncut” ή στα ελληνικά “άκοπος”.

- **ID:** Το id υποδηλώνει το αριθμητικό γνώρισμα του λήμματος δηλαδή την ταυτότητά του. Σε περίπτωση που κάποιος θέλει να ανακτήσει κάποια συγκεκριμένη γραμμή μπορεί να το κάνει του id της. Στο παράδειγμα η ταυτότητα του λήμματος είναι 00005107.

- **PosScore** και **NegScore:** Οι επόμενες δύο στήλες θα εξηγηθούν μαζί κυρίως επειδή έχουν ήδη αναφερθεί. Πρόκειται για την θετική και την αρνητική βαθμολογία του λήμματος αντίστοιχα, που εκφράζουν την πιθανότητα που έχουν **όλες** οι λέξεις και φράσεις που ανήκουν στο ίδιο λήμμα να είναι θετικές ή αρνητικές. Στην προκειμένη περίπτωση, η λέξη άκοπος θεωρείται πως στο 50% των φορών έχει θετική σημασία. Τα νούμερα αυτά δεν έχουν άθροισμα 1 επειδή στο αρχείο δεν καταγράφεται το Obj(s) που είναι ο παράγοντας αντικειμενικότητας και μπορεί κανείς να τον συμπεράνει

αφαιρώντας από το 1 το άθροισμα των δύο υπαρχόντων παραγόντων. Άρα το SentiWordNet 3.0 θεωρεί πως η λέξη “uncut” αλλά και η φράση “full-length” το πενήντα τα εκατό των περιστάσεων που αντιμετωπίζονται σε κείμενο είναι θετικές έννοιες ενώ το άλλο μισό είναι ουδέτερες. Δεν έχουν δηλαδή κάποια επιρροή στην συναισθηματική πολικότητα του κειμένου.

- **SynsetTerms:** Πρόκειται για το σύνολο των όρων που αποτελούν το λήμμα. Τις λέξεις αλλά και τις φράσεις δηλαδή των οποίων η βαθμολογία φαίνεται στην γραμμή που βρισκόμαστε. Όσον αφορά το παράδειγμα, τις έχουμε αναφέρει πολλές φορές ως τώρα και πρόκειται για τη λέξη “uncut” και την φράση “full-length”. Ο αριθμός που μπορεί κανείς να δει δίπλα στην φράση, σχετίζεται με την επόμενη στήλη, αλλά ουσιαστικά είναι μία κατάταξη που εκφράζει πόσο συχνά χρησιμοποιούνται οι λέξεις και φράσεις του λήμματος με τη σημασία που αναφέρεται στην γραμμή.

- **Gloss:** Η στήλη gloss εξηγεί το περιεχόμενο της λέξης που βαθμολογείται. Την ερμηνεία δηλαδή της λέξης που έχει βαθμολογηθεί με αυτά τα σκορ. Γνωρίζουμε πως μία λέξη μπορεί να έχει πολλές σημασίες και για να τις προσδιορίσουμε οφείλουμε να ρίξουμε μία ματιά στα συμφραζόμενα. Στο παράδειγμα η λέξη “uncut” δηλαδή άκοπος χρησιμοποιείται για να υποδηλώσει κάτι που είναι “complete” δηλαδή ολόκληρο. Μία διαφορετική χρήση της λέξης θα μπορούσε να είναι η εξής “Η κορδέλα του δώρου είναι άκοπη(uncut)” στην προκειμένη περίπτωση η λέξη άκοπος δεν εννοεί ακριβώς κάτι ολόκληρο αλλά κάτι το οποίο δεν έχει κοπεί/σχιστεί. Τώρα, μπορούμε να εξηγήσουμε καλύτερα τον αριθμό στο τέλος της λέξης “uncut” από την προηγούμενη στήλη. Αφού είδαμε πως μπορεί να υπάρχουν δύο ή και περισσότερες διαφορετικές ερμηνείες μίας λέξης μπορούμε να φανταστούμε πως μία από αυτές χρησιμοποιείται πιο συχνά από τις υπόλοιπες. Άρα ο αριθμός 7 δίπλα στην λέξη “uncut” του παραδείγματος μας δείχνει πως υπάρχουν τουλάχιστον επτά διαφορετικές ερμηνείες της λέξης “uncut” **μεσα στο SentiWordNet** και η προκειμένη ερμηνεία είναι η έβδομη πιο συχνά χρησιμοποιημένη. Αυτά τα νούμερα δεν έχουν εξαχθεί κατά τη δημιουργία του SentiWordNet 3.0 αλλά υπήρχαν στο WordNet.

2.3 JSP

2.3.1 JSP-Ορισμός

Οι Jsp, ή αλλιώς Σελίδες Σέρβερ σε Java, κυκλοφόρησαν για πρώτη φορά σαν τεχνολογία το 1999 από την εταιρία Sun Microsystems και αποτελούν έναν εναλλακτικό τρόπο παραγωγής δυναμικών ιστοσελίδων σε Java. Ο βασικός λόγος που μία τέτοια τεχνολογία ήταν απαραίτητη και πολύ χρήσιμη, ήταν επειδή κάλυπτε ένα βασικό μειονέκτημα των Java Servlets, που αφορούσε τον χρόνο ανάπτυξης ενός Project. Παρά το γεγονός ότι τα Java Servlets αποτελούν το θεμέλιο του server side προγραμματισμού σε Java, η διαδικασία γραψίματος του κώδικα ενός Java Servlet, η ανάπτυξή(deployment) του καθώς και η εντόπιση σφαλμάτων(debugging) μπορούν να είναι πολύ κουραστικές και χρονοβόρες διαδικασίες. Η JSP(αναφερόμαστε στην τεχνολογία), λοιπόν, υποστηρίζει την τεχνολογία των Servlets, βοηθώντας στην επίλυση αυτού του προβλήματος και απλουστεύοντας την διαδικασία ανάπτυξης ενός Servlet.

Αρχικά, η JSP, δημιουργήθηκε με βάση άλλα ήδη υπάρχοντα μοτίβα τεχνολογιών για sever side προγραμματισμό με σκοπό την παροχή μίας μεθόδου ένθεσης δυναμικού κώδικα με στατικά κομμάτια όπως για παράδειγμα, κομμάτια HTML. Μία γενική εικόνα της λειτουργίας μίας JavaServer σελίδας, είναι η εξής : Όταν μία αίτηση(request) γίνεται από κάποιο περιεχόμενο μίας JavaServer σελίδας, ένας web container ερμηνεύει την JSP, εκτελεί ό,τι ενσωματωμένο κώδικα βρει και στέλνει τα αποτελέσματα σε μία απόκριση(response). Παρακάτω, θα εμβαθύνουμε περισσότερο στην λειτουργικότητα και σε συγκεκριμένα δομικά στοιχεία της τεχνολογίας JSP με βάση το έγγραφο της Oracle για τις λεπτομέρειες της JSP.

2.3.1 JSP-Λειτουργικότητα-Επιμέρους Στοιχεία

Η λειτουργικότητα της JSP, έχει υποστεί πολλές τροποποιήσεις από την πρωταρχική της έκδοση και κάθε φορά προστίθενται λειτουργίες ή τροποποιούνται ήδη υπάρχουσες. Την προκειμένη στιγμή, η τελευταία έκδοση εγγράφου με λεπτομέρειες(specification documentation) που έχει εκδοθεί από την Oracle, είναι η 2.3(JSP 2.3 Specifications) η οποία μπορεί να βρεθεί εδώ : **http://download.oracle.com/otn-pub/jcp/jsp-2_3-mrel2-eval-spec/JSP2.3MR.pdf** .

Για να κατανοήσουμε καλύτερα την λειτουργία της JSP θα εξηγήσουμε τον ορισμό της JSP, καθώς και ορισμένες παροχές της.

1. **JavaServer Pages:** Η βασική σύνταξη και σημασιολογία της JSP, μπορεί να βρεθεί αναλυτικά στο JSP 2.3 Specifications. Ουσιαστικά πρόκειται για μία τεχνολογία που επιτρέπει τον έλεγχο του περιεχομένου ή και της εμφάνισης μίας ιστοσελίδας, μέσω της χρήσης των Java Servlets. Τα Java Servlets, είναι μία κλάση προγραμματισμού σε Java, που χρησιμοποιείται προκειμένου να επεκτείνει τις δυνατότητες ενός Server που φιλοξενεί εφαρμογές που ακολουθούν το μοντέλο αίτησης-απόκρισης. Κατά την διάρκεια της λειτουργίας της, μία σελίδα JSP τεχνολογίας, μεταφράζεται σε σε Java Servlets. Μία βασική σελίδα JSP αποτελείται από απλό κείμενο, στατική σήμανση (για παράδειγμα HTML) και έχει τη δυνατότητα να εκμεταλλευτεί ενσωματωμένα scripts και άλλες λειτουργικότητες για τη δημιουργία δυναμικού περιεχομένου.

2. **Custom Tags:** Η JSP σαν τεχνολογία παρέχει τη δυνατότητα στον χρήστη να δημιουργήσει μικρά δικά του κομμάτια κώδικα σε Java τα οποία συνδέονται με κάποιο στατικό στοιχείο της εκάστοτε ιστοσελίδας και προσφέρουν κάποια λειτουργικότητα. Tag της HTML είναι το <h></h>(που εμφανίζει κάποιο header στην σελίδα), το <body></body> στο οποίο συνήθως περιλαμβάνονται όλα τα στατικά κομμάτια μίας σελίδας και διάφορα άλλα. Η JSP λοιπόν μας δίνει τη δυνατότητα να φτιάξουμε δικά μας tags στατικού περιεχομένου, που προβάλλουν κάτι διαφορετικό από τα ήδη υπάρχοντα. Αυτό θεωρείται ένα από τα ισχυρά σημεία/πλεονεκτήματα της JSP και μπορεί να χρησιμοποιηθεί στη θέση των script σε Java ή και να τα υποστηρίξει.

```
import javax.servlet.jsp.tagext.*;
import javax.servlet.jsp.*;
import java.io.*;

public class HelloTag extends SimpleTagSupport {
   public void doTag() throws JspException, IOException {
      JspWriter out = getJspContext().getOut();
      out.println("Hello Custom Tag!");
   }
}
```

Εικόνα 4: Απλό παράδειγμα Custom Tag (source: www.tutorialspoint.com)

Στο παραπάνω παράδειγμα βλέπουμε τον ορισμό της κλάσης παραγωγής ενός custom tag που τυπώνει στη θέση του στοιχείου που το καλούμε την φράση "Hello Custom Tag!"

3. **Expression Language (Γλώσσα έκφρασης):** Η JSP προσφέρει επίσης την δυνατότητα να προσδιορίσουμε την τιμή κάποιου δυναμικού αντικειμένου και να την αξιοποιήσουμε στο περιεχόμενο ενός στατικού τμήματος ή να την μεταβιβάσουμε μέσω ενός στατικού τμήματος σε μία άλλη λειτουργία JSP(Oracle JSP 2.0 Specification 2013).

Ένα παράδειγμα χρήσης της expression language της JSP είναι το εξής:

${Έκφραση σε expression language}

Πολύς κόσμος, έχει την πεποίθηση πως η σήμανση ${} ουσιαστικά φέρνει, εκεί οπού την χρησιμοποιήσαμε, την τιμή ενός μεμονωμένου στοιχείου, πράγμα το οποίο δεν ισχύει. Εντός του
${}, μπορούμε να εισάγουμε και κάποια λειτουργικότητα, όπως για παράδειγμα την πρόσθεση των τιμών δύο στοιχείων για παράδειγμα ${a+b}, ή γενικά την κλίση μίας μεθόδου όπως αυτής που μετράει το μέγεθος ενός arraylist. Οι τύποι των μεταβλητών που μπορούν να εκφραστούν μέσα σε μία έκφραση expression language, είναι ακέραιοι αριθμοί, double αριθμοί, αλφαριθμητικά, καθώς και οι τελεστές True και False για boolean εκφράσεις.

4.Κύκλος Ζωής JSP: Ένα ακόμη πολύ σημαντικό κομμάτι που πρέπει να εξηγήσουμε, είναι ο κύκλος ζωής μίας σελίδας τεχνολογίας JSP. Τις φάσεις δηλαδή από τις οποίες περνά μία JSP, από την στιγμή που καλείται μέχρι και τον τερματισμό της λειτουργίας της, οι οποίες είναι οι εξής:

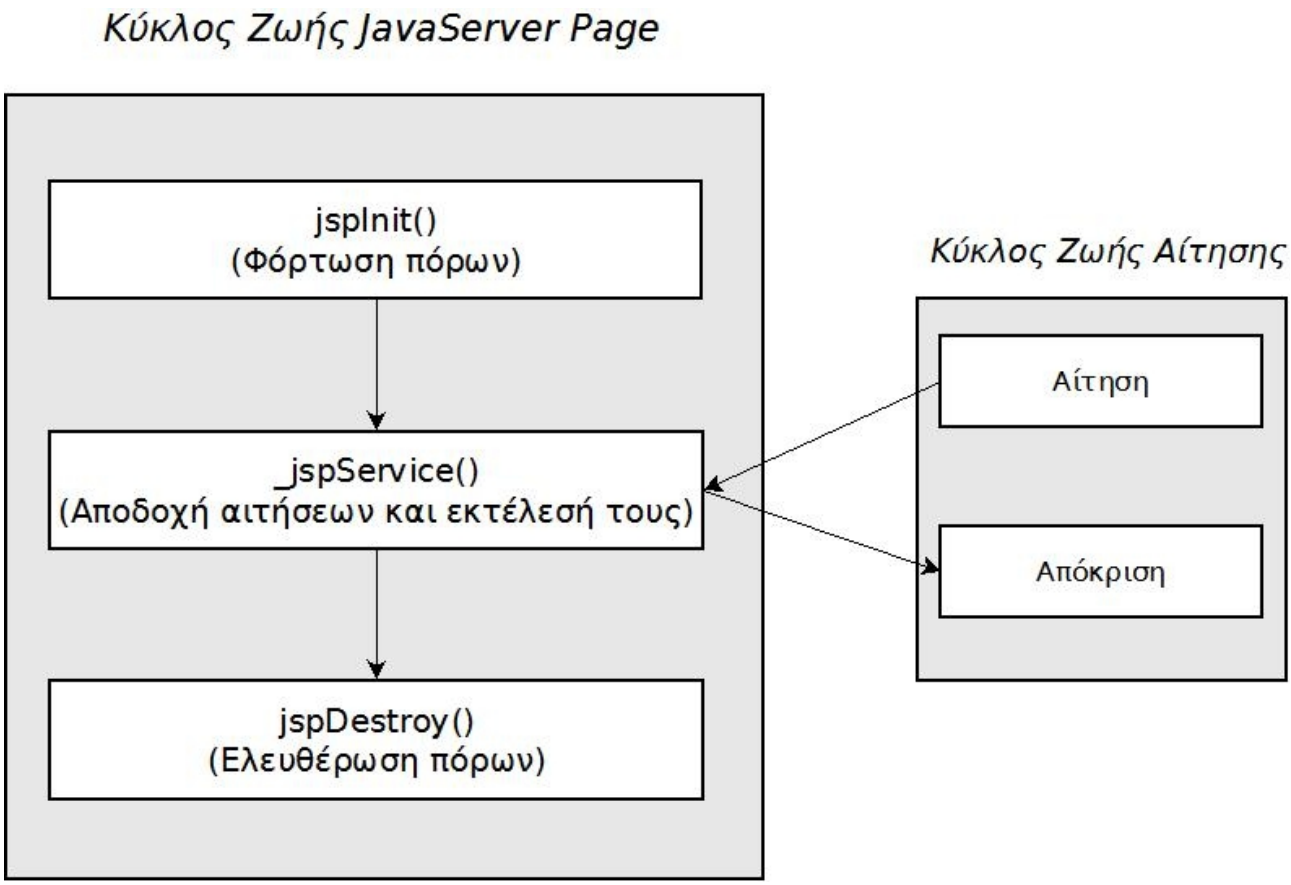
- **Compilation(Μεταγλώττιση):** Όταν το πρόγραμμα περιήγησης του χρήστη, ζητά μία JSP, το Jsp engine, δηλαδή η μηχανή του server στον οποίο είναι ανεβασμένη η JSP, εξετάζει αν η σελίδα έχει ήδη μεταγλωττιστεί και σε περίπτωση που δεν έχει, την μεταγλωττίζει. Η της μεταγλώττισης είναι και η διαδικασία κατά την οποία η JSP μετατρέπεται σε Java Servlet.

- **JSP Initialization(Αρχικοποίηση):** Όταν ο web container φορτώνει μία JSP, επικαλείται την μέθοδο jspInit(), προτού εξυπηρετήσει κάποια αίτηση(Oracle JSP 2.0 Specification 2013). Κατά το κάλεσμα της μεθόδου αυτής, ξεκινά η σύνδεση με πιθανές βάσεις δεδομένων και φορτώνονται όλα τα αρχεία που είναι απαραίτητα για την λειτουργία της σελίδας.

- **Εκτέλεση της JSP:** Αφού βεβαιωθούμε πως η σελίδα έχει μεταγλωττιστεί και αρχικοποιηθεί και έχουν ιδρυθεί όλες οι συνδέσεις με βάσεις και αρχεία, είναι επιτέλους η στιγμή να εξυπηρετήσουμε τις αιτήσεις των χρηστών. Σε αυτήν την φάση λαμβάνει χώρα οποιαδήποτε αλληλεπίδραση με αιτήσεις. Η μέθοδος η οποία επικαλείται και αναλαμβάνει τις αιτήσεις είναι η _jspService(). Είναι σημαντικό να σημειωθεί εδώ πως η JSP τεχνολογία έχει δημιουργηθεί με σκοπό την απλοποίηση της διαδικασίας παραγωγής αντικειμένων HttpServlet. Έτσι λοιπόν, σε αντίθεση με τον προγραμματισμό κατευθείαν σε Java Servlets, δεν υπάρχει διαχωρισμός ανάμεσα σε μία γενική JSP και μία που έχει ως σκοπό να χρησιμοποιηθεί με HTTP. Για αυτόν τον λόγο η μέθοδος _jspService() είναι υπεύθυνη για την δημιουργία και την αποστολή αποκρίσεων και στις επτά μεθόδους HTTP. Παραδείγματα μεθόδων HTTP είναι οι: GET, POST, HEAD, PUT και DELETE.

- **Καταστροφή JSP:** Όταν μία σελίδα αφαιρείται από τον web container, επικαλείται η μέθοδος jspDestroy() προκειμένου να απελευθερωθούν όλοι οι πόροι που καταλάμβανε η JSP.

Παρακάτω απεικονίζεται μία απλή εκδοχή του κύκλου ζωής μίας JavaServer σελίδας



Εικόνα 5: Κύκλος Ζωής JSP

2.4 Javascript

Μία ακόμη γλώσσα προγραμματισμού που χρησιμοποιήθηκε προκειμένου να δημιουργηθεί η εφαρμογή Sentiment Analysis σε JSP είναι η Javascript. Πρόκειται για μία γλώσσα προγραμματισμού που βασίζεται σε scripts(όπως συνιστά και το όνομά της) Εκδόθηκε το 1995 από την netscape(Kurniawan 2002), αλλά άργησε πολύ να εκτιμηθεί σαν εργαλείο δικτυακού προγραμματισμού και για αρκετό καιρό, τα scripts σε Javascript που μπορούσε κανείς να βρει στο διαδίκτυο δεν ήταν συμβατά με πολλά προγράμματα πλοήγησης στο ίντερνετ. Σήμερα όμως χρησιμοποιείται κατά κόρον σε πολλών ειδών δικτυακές εφαρμογές, επηρεάζοντας και άλλες γλώσσες προγραμματισμού, όπως η actionscript(McFarland 2011).

Η Javascript, είναι μία γλώσσα προγραμματισμού που παρέχει τη δυνατότητα στον προγραμματιστή να εμπλουτίσει το περιεχόμενο της ιστοσελίδας του με διαδραστικά πεδία, δυναμικά οπτικά εφφέ και animations(McFarland 2011). Σήμερα, κατά κύριο λόγο, η Javascript,

χρησιμοποιείται για client side προγραμματισμό, αλλά μπορεί να χρησιμοποιηθεί και για server side προγραμματισμό. Αυτό σημαίνει πως οι περισσότερες λειτουργίες που προσθέτει κανείς σε έναν ιστότοπο με Javascript, λαμβάνουν χώρα στον υπολογιστή από τον οποίο προέρχεται η εκάστοτε αίτηση και όχι στον server. Το βασικό χαρακτηριστικό που καθιστά την Javascript πολύ χρήσιμο εργαλείο στον προγραμματισμό δικτυακών εφαρμογών, είναι η αμεσότητα που προσφέρει(McFarland 2011). Είναι πολύ πιο γρήγορο σε έναν ιστότοπο που χρησιμοποιεί javascript, να ελέγξει αν μία φόρμα που ο χρήστης όφειλε να συμπληρώσει είναι ελλιπής και ποιο από τα πεδία της φόρμας μένει να συμπληρώσει και να τον ενημερώσει. Μπορούμε λοιπόν πολύ εύκολα με την Javascript, να ελέγξουμε ή και να διορθώσουμε την πλοήγηση ενός χρήστη στον ιστότοπό μας εμφανίζοντάς του γρήγορα, μηνύματα που αφορούν κινήσεις που πρέπει να κάνει ή σε περιπτώσεις όπου διαχειριζόμαστε ένα ηλεκτρονικό κατάστημα, μηνύματα που αφορούν το κόστος των προϊόντων που υπάρχουν στο καλάθι του. Εκτός από τα μηνύματα λάθους, μία βασική χρήση της Javascript είναι να καθιστά τον ιστότοπό μας πιο ενδιαφέροντα στο χρήστη μέσω της εισαγωγής δυναμικού περιεχομένου. Μπορεί κανείς για παράδειγμα να ρυθμίσει έτσι τον ιστότοπό του με Javascript ώστε ένα μέρος του περιεχόμενου να κρύβεται μέχρι κάποια χρονική στιγμή από την φόρτωση της σελίδας ή μέχρι ο χρήστης να πατήσει κάποιο στοιχείο της σελίδας ή όταν φέρει τον κέρσορα του σε κάποιο συγκεκριμένο σημείο.

Οι δυνατότητες εμπλουτισμού και υποστήριξης ενός ιστοτόπου μέσω της χρήσης της Javascript είναι αναρίθμητες, όμως η χρησιμότητά της δεν περιορίζεται μόνο σε αυτά. Παρά το γεγονός ότι έχουμε αναφερθεί μόνο σε συμπληρωματικές λειτουργίες της Javascript, είναι απολύτως δυνατή η ανάπτυξη μίας ολόκληρης δικτυακής εφαρμογής εξ ολοκλήρου με Javascript.

2.4 JQuery

Παρά το γεγονός ότι η Javascript είναι πλέον ένα πολύ διαδεδομένο εργαλείο για τους σχεδιαστές ιστοτόπων, υπάρχουν δύο μικρά προβλήματα(McFarland 2011). Πρώτον, είναι μία

γλώσσα προγραμματισμού και παρά το γεγονός ότι είναι σχετικά απλή συγκριτικά με άλλες γλώσσες, δεν παύει να έχει κανόνες και σε μερικές περιπτώσεις περίπλοκους κανόνες. Δεύτερον, λόγω της ύπαρξης ενός τεράστιου αριθμού προγραμμάτων πλοήγησης και πολλών διαφορών στην αντιμετώπιση και αναγνώριση της Javascript μεταξύ τους, απαιτείται πολύς χρόνος και κόπος για να δοκιμαστεί μία εφαρμογή που βασίζεται σε μεγάλο μέρος της σε Javascript προκειμένου να λειτουργεί σωστά στην πλειοψηφία των browser.

Τα δύο αυτά προβλήματα έρχεται και αντιμετωπίζει πολύ αποτελεσματικά η Jquery. Η Jquery, είναι μία από τις πιο διαδεδομένες βιβλιοθήκες της Javascript. Στον προγραμματισμό, μία βιβλιοθήκη κάποιας γλώσσας προγραμματισμού, είναι μία συλλογή από μεθόδους, ή από άλλα κομμάτια ήδη γραμμένου κώδικα που σαν σκοπό έχουν την διευκόλυνση της ανάπτυξης μελλοντικών προγραμμάτων. Έτσι και η Jquery, σαν βιβιοθήκη της javascript, περιέχει έναν μεγάλο αριθμό έτοιμων λειτουργιών σε Javascript που απλοποιούν διάφορες διαδικασίες, ή επιλύουν τα προβλήματα συμβατότητας με προγράμματα πλοήγησης. Κάνοντας έτσι το δυναμικό περιεχόμενο που προσθέτουμε σε έναν ιστότοπο σίγουρα προσβάσιμο από τους περισσότερος χρήστες, και από προγραμματιστικής απόψεως συντομότερο κατά δέκα ή πενήντα ή και περισσότερες γραμμές κώδικα ανά λειτουργία!

```
<script>
$(document).ready(function(){

    $("#image").fadeOut(5000);
    $("#image2").delay(5000).fadeIn();

});
```

Εικόνα 6: Απόσπασμα Κώδικα της Εφαρμογής Sentiment Analysis σε JSP που χρησιμοποιεί Jquery

Στην παραπάνω εικόνα, φαίνεται ένα μικρό τμήμα κώδικα που χρησιμοποιεί Jquery και κάνει τα εξής: Η μέθοδος .ready(), δέχεται σαν παράμετρο μία συνάρτηση και την θέτει σε λειτουργία την στιγμή που το Document Object Model(DOM) της σελίδας δημιουργηθεί, ουσιαστικά όταν φορτωθεί η σελίδα. Στην προκειμένη περίπτωση οι ενέργειες που λαμβάνουν χώρα μόλις

φορτωθεί η σελίδα, είναι η **$("#image").fadeOut(5000)** και η **$ ("#image2").delay(5000).fadeIn().** Λειτουργία των .fadeOut() και fadeIn() είναι να εξαφανίζουν και να εμφανίζουν αντίστοιχα ένα στοιχείο της σελίδας που στο κομμάτι αυτό κώδικα είναι δύο εικόνες, η εικόνα "image" και η "image2" . Η πρώτη μεταβλητή που δέχονται, (στο παράδειγμα 5000) εκφράζει το χρονικό περιθώριο μέσα στο οποίο σταδιακά μεταβάλλουν το περιεχόμενο της σελίδας. Άρα με σταθερό ρυθμό η εικόνα "image" εξαφανίζεται από την οθόνη του χρήστη 5000 ms (5 δευτερόλεπτα) από την φόρτωση της σελίδας. Το .delay(5000) εκφράζει πως η "image2" θα αρχίσει να εμφανίζεται στην οθόνη 5 δευτερόλπτα μετά την φόρτωση. Η παροχή αυτών των τριών μεθόδων μόνο , γλιτώνει τον προγραμματιστή αρκετές γραμμές κώδικα.

ΚΕΦ.3: Η Εφαρμογή Sentiment Analysis σε JSP

3.1 Γενική Περιγραφή

Όπως αναφέρθηκε και προηγουμένως, η εφαρμογή Sentiment Analysis βασίστηκε στην αξιοποίηση του SentiWordNet 3.0 με σκοπό την παροχή μίας εύκολα προσβάσιμης δικτυακής πλατφόρμας, που θα παρέχει στον χρήστη έγκυρες αξιολογήσεις όσον αφορά την συναισθηματική πόλωση ενός κειμένου σε αγγλικά ή ελληνικά.

Η εφαρμογή χρησιμοποιεί μία μέθοδο βαθμολόγησης των φράσεων που βασίζεται στα Pos(s) και Neg(s) του SentiWordNet 3.0 που εξηγήθηκαν στο αντίστοιχο κεφάλαιο. Βάσει αυτής της βαθμολόγησης των λέξεων που αναγνωρίστηκαν εκτιμάται η προδιάθεση του κειμένου και παρέχονται στον χρήστη στατιστικά που αφορούν τις λέξεις που αναγνωρίστηκαν. Προκειμένου να αναγνωρίζονται όσο το δυνατόν περισσότερες λέξεις σε αγγλικά και κυρίως ελληνικά, ήταν απαραίτητη η υλοποίηση πολλών γραμματικών κανόνων σε jsp.

Η εφαρμογή Sentiment analysis δίνει την επιλογή γλώσσας στον χρήστη, έπειτα αξιολογεί το δοθέν κείμενο και παρουσιάζει στο χρήστη μία γενική εκτίμηση του κειμένου καθώς και ένα ραβδόγραμμα, ένα διάγραμμα πίτας και μία γραφική παράσταση της πολικότητας του κειμένου σχετικά με την μεμονωμένη εκτίμηση κάθε λέξης ή φράσης που αναγνωρίστηκε.

Τα βήματα που πραγματοποιεί η εφαρμογή λοιπόν είναι τα εξής:

- Εξαγωγή βαθμολογίας των φράσεων του SentiWordNet 3.0.
- Καταχώρηση και διαμόρφωση των φράσεων με βάση τους γραμματικούς κανόνες τις εκάστοτε γλώσσας.
- Αναγνώριση των φράσεων και λέξεων του κειμένου, που αντιστοιχούν σε αυτές που έχουν βαθμολογηθεί βάσει SentiWordNet 3.0.
- Εξαγωγή ενός συμπεράσματος όσον αφορά την συναισθηματική πόλωση του κειμένου, που στηρίζεται σε έναν σταθμισμένο μέσο όρο.
- Παρουσίαση της εκτίμησης στον χρήστη μαζί με ένα ραβδόγραμμα και ένα διάγραμμα πίτας που αναλύουν τις μεμονωμένες εκτιμήσεις των λέξεων.

Κατά την διάρκεια της ανάλυσης της λειτουργίας της εφαρμογής, θα αναφερθούμε σε έναν αριθμό “λεκτικών κατηγοριών” που εκφράζουν μία γενική εκτίμηση της συναισθηματικής πολικότητας του κειμένου, αυτές οι κατηγορίες, εξάγονται από την βαθμολογία δόθηκε στις λέξεις ή σε ολόκληρο το κείμενο από την εφαρμογή.

Οι βαθμολογίες κυμαίνονται από το -1 στο 1 και μία αρνητική βαθμολογία ισοδυναμεί με αρνητική έννοια λέξης/κειμένου ενώ μία θετική με θετική έννοια. Για την διευκόλυνση της κατανόησης των παρακάτω κεφαλαίων, παρατίθεται ο εξής πίνακας:

Λεκτική Κατηγορία	Εύρος Τιμών
Πολύ Θετική	≥0.75, ≤1
Θετική	>0.25, <0.75
Σχετικά Θετική	>0, ≤0.25
Ουδέτερη	0
Σχετικά Αρνητική	≥-0.25, <0
Αρνητική	>-0.75,<0.25
Πολύ Αρνητική	≤-0.75, ≥-1

Πίνακας 1: Λεκτικές Κατηγορίες βάσει βαθμολογίας SWN 3.0

3.2 Έρευνα Παρόμοιων Εφαρμογών

Προτού περάσουμε στην ανάλυση της λειτουργίας της εφαρμογής, οφειλουμε να αναφέρουμε τις εφαρμογές που ήδη υπάρχουν πάνω στο αντικείμενο, και με την ανάπτυξη της τεχνολογίας όλο περισσότεροι ασχολούνται καθημερινά με την ανάλυση συναισθημάτων σε microblogging εφαρμογές και σε άλλων ειδών πλατφόρμες στο ίντερνετ. Βλέποντας την εύκολη πρόσβαση όλων σε πλατφόρμες microblogging και την ελεύθερη δομή μηνυμάτων που αυτές παρέχουν(Pak and Paroubek, n.d.) οι Alexander Pak και Patrick Paroubek ασχολήθηκαν με την ανάλυση συναισθημάτων μέσω του Twitter. Πραγματοποίησαν λοιπόν γλωσσική ανάλυση πάνω σε έναν αριθμό μηνυμάτων μέσω twitter εκτιμώντας στη συνέχεια τα αποτελέσματα. Αντίστοιχη εφαρμογή, υλοποίησαν οι Apoorv Agarwal Boyi Xie Ilia Vovsha Owen Rambow Rebecca Passonneau και εξέδωσαν το άρθρο “Sentiment analysis of twitter data” το 201. Σε αντίθεση με τους Pak και Paroubek και τους Apoorv Agarwal Boyi Xie Ilia Vovsha Owen Rambow Rebecca Passonneau οι Julie Kane Ahkter και Steven Soria, θεώρησαν πως το Twitter δεν είναι ιδανική πηγή εξόρυξης συναισθημάτων και απόψεων, καθώς παρέχει αισθητά μικρότερο πλήθος χαρακτήρων για την έκφρασή τους από το Facebook(Ahkter and Soria, n.d.), κάνοντας δική τους κατηγοριοποίηση εκφράσεων χρησιμοποιώντας ανανεώσεις κατάστασης από αυτό, που παρείχε εως και 420 χαρακτήρες(2010). Κατά τη διάρκεια της διαδικασίας συνέκριναν τα αποτελέσματα τεσσάρων διαφορετικών ταξινομητών, ενός MaxEnt, ενός MaxEnt με labeled

data(LDA), ενός MaxEnt με σήμανση part of speech και τέλος ενός MaxEnt που συνδύαζε τις λειτουργίες των δύο τελευταίων. Όπου MaxEnt αλγόριθμος μέγιστης εντροπίας.

Από τις εφαρμογές που μελετήθηκαν, καμία δεν παρέχει στον χρήστη τη δυνατότητα συναισθηματικής ανάλυσης μέσω διαδικτύου. Όλες έχουν χρησιμοποιηθεί για την διεξαγωγή πειραμάτων και την εξαγωγή μαζικών αποτελεσμάτων.

3.3 Sentiment Analysis-Γενική λειτουργία

Όπως είναι προφανές, η εφαρμογή Sentiment analysis σε jsp είναι στην ουσία ένας ιστότοπος. Ο καλύτερος τρόπος να κατανοήσουμε λοιπόν σε ένα πρώτο επίπεδο την λειτουργία της εφαρμογής είναι να δούμε όλες τις βασικές σελίδες της, την λειτουργικότητα που προσφέρουν και τον τρόπο με τον οποίον αλληλεπιδρούν μεταξύ τους. Η εικόνα παρακάτω, παρουσιάζει ακριβώς αυτό, όλες τις βασικές σελίδες με μικρή εξήγηση του ρόλου τους στο σύνολο και τις αλληλεπιδράσεις τους.

Για να ξεκινήσουμε την ανάλυση του σχήματος ας παρατηρήσουμε πως υπάρχουν τρεις σελίδες προσβάσιμες από τις περισσότερες σελίδες. Οι σελίδες αυτές είναι η σελίδα εισαγωγής αγγλικού κειμένου προς εκτίμηση, η αντίστοιχη αυτής για κείμενο στα ελληνικά και μία ενημερωτική σελίδα με πληροφορίες όσον αφορά τον στόχο της εφαρμογής και τον τρόπο με τον οποίο αυτός επιτυγχάνεται. Αυτό σημαίνει πως παρέχεται στον χρήστη σε όλες τις φάσεις της πλοήγησης στον ιστότοπο η δυνατότητα να εισάγει κείμενο προς εκτίμηση. Από αυτές τις σελίδες η λογική ροή της πλοήγησης, είναι να δώσει ο χρήστης κάποιο κείμενο για να εκτιμηθεί στην γλώσσα που αντιστοιχεί στην σελίδα που βρίσκεται. Στη συνέχεια, το κείμενο στέλνεται στις ανάλογες σελίδες επεξεργασίας του SentiWordNet 3.0 που βάσει αυτού ή του μεταφρασμένου του αντιγράφου, αναγνωρίζουν λέξεις από το κείμενο και αντιστοιχούν σε αυτές μία βαθμολογία. Οι σελίδες αυτές, δεν είναι εμφανείς στον χρήστη και απλώς πραγματοποιούν την λειτουργικότητά τους παραδίδοντας τα αποτελέσματα στις επόμενες σελίδες που παράγουν τα στατιστικά στοιχεία και τις γραφικές παραστάσεις και τις εμφανίζουν στον χρήστη. Αυτή είναι συνοπτικά η λειτουργία των κύριων σελίδων της εφαρμογής. Παρακάτω αναλύεται περαιτέρω ο μηχανισμός λειτουργίας της κάθε σελίδας μεμονωμένα.

3.4 Sentiment Analysis Εφαρμογή-Οι Σελίδες

Αρχική σελίδα εκτίμησης αγγλικού/ελληνικού κειμένου:

Στην προκειμένη παράγραφο, θα εξηγηθεί η λειτουργία και της σελίδας εισαγωγής ελληνικού αλλά και της σελίδας εισαγωγής αγγλικού κειμένου, καθώς διαφέρουν ελάχιστα. Πρόκειται για σελίδες που σαν βασικό σκοπό έχουν το καλωσόρισμα του χρήστη στην εφαρμογή, την παροχή της φόρμας εισαγωγής κειμένου, την υπενθύμιση προς τον χρήστη όσον αφορά την χρήση σωστής ορθογραφίας προκειμένου να μπορέσει στη συνέχεια να γίνει η αναγνώριση των λέξεων και σαφώς τον έλεγχο πως η φόρμα δεν απεστάλη κενή.
Παρακάτω, στην εικόνα, φαίνεται η αρχική σελίδα εισαγωγής αγγλικού κειμένου.

Εικόνα 7: Εμφάνιση σελίδας εισαγωγής αγγλικού κειμένου

Όλες οι εικόνες που χρησιμοποιούνται για την εμφάνιση της εφαρμογής ανήκουν στον Simon Tofield και το https://simonscat.com/

Σελίδα εξαγωγής αποτελεσμάτων αγγλικού κειμένου:

Στην σελίδα αυτή βασίζεται όλη η λειτουργικότητα της εφαρμογής. Σε αυτήν την σελίδα λαμβάνει χώρα η σύνδεση των δεδομένων του SentiWordNet με την εφαρμογή, η ανάλυση των

βαθμολογιών του SentiWordNet με σκοπό την δημιουργία των εκτιμήσεων της εφαρμογής, η αντιστοίχηση των λέξεων του κειμένου που εισήγαγε ο χρήστης με τις λέξεις του SentiWordNet, η εκτίμηση της πολικότητας του συνόλου του όλου κειμένου και φυσικά η μεταβίβαση των αποτελεσμάτων στην σελίδα που δημιουργεί στατιστικά και γραφικές παραστάσεις για να ενημερώσει τον χρήστη. Τα στάδια αυτά μπορούν λοιπόν να εκφραστούν ως εξής:

1. Προσπέλαση του SentiWordNet και Δημιουργία πίνακα αντιστοίχησης λέξεων-βαθμολογιών. Στο βήμα αυτό, η σελίδα διαβάζει τα περιεχόμενα του SentiWordNet και βγάζει μέσους όρους βαθμολογιών για κάθε λέξη που εμφανίζεται περισσότερες από μία φορές. Στη συνέχεια δημιουργεί ένα Hashmap στο οποίο τοποθετεί τα ζεύγη λέξεων και τιμών με κλειδί την εκάστοτε λέξει και τιμή τον δεκαδικό αριθμό που αντιστοιχεί σε μία βαθμολογία

```
HashMap<String, Double> _dict;
```

Εικόνα 8: Το Hashmap λεξικό

2. Ανάγνωση εισαχθέντος κειμένου-Αναγνώριση λέξεων. Σε αυτό το στάδιο η σελίδα προσπελαύνει το κείμενο που ο χρήστης έδωσε προς εκτίμηση και το χωρίζει σε μεμονωμένες λέξεις αλλά και σε φράσεις δύο ή τριών λέξεων προκειμένου να μπορέσουν στο επόμενο στάδιο να αντιστοιχηθούν με φράσεις καταχωρημένες στο SentiWordNet. Οι λέξεις και φράσεις καταχωρούνται στη συνέχεια σε έναν πίνακα.

1. Επόμενο βήμα στην λογική της σελίδας είναι η διεξαγωγή αναζητήσεων στο Hashmap με τις λέξεις του SentiWordNet, με κλειδί της κάθε αναζήτησης μία καταχώρηση του πίνακα με τις λέξεις που συλλέχθηκαν στο βήμα δύο. Έτσι αναγνωρίζονται οι λέξεις και η εφαρμογή είναι πλέον σε θέση να αντιστοιχίσει μία βαθμολογία που εκφράζει την συναισθηματική ή αρνητική πολικότητα της λέξης/φράσης!

2. Σε αυτήν την φάση η σελίδα πραγματοποιεί μία κατανομή των λέξεων σε κατηγορίες που εκφράζουν αριθμητικά και λεκτικά την συναισθηματική φόρτιση

της κάθε λέξης, με σκοπό να μεταβιβάσει αυτά τα στοιχεία στην επόμενη σελίδα που δημιουργεί και εμφανίζει τα στατιστικά στον χρήστη. Για την κατανόηση αυτού, παρατίθεται και εξηγείται το παρακάτω απόσπασμα κώδικα:

```
if ((_dict.get(s + "#n") != null) && x == 0) {

    x++;
    if (_dict.get(s + "#n") >= 0.75) {
        vp = _dict.get(s + "#n") + vp;
        vpi++;
    } else if (_dict.get(s + "#n") < 0.75 && _dict.get(
        p = _dict.get(s + "#n") + p;
        pi++;
    } else if (_dict.get(s + "#n") > 0.0 && _dict.get(s
        sp = _dict.get(s + "#n") + sp;
        spi++;
    } else if (_dict.get(s + "#n") == 0.0) {
        ni++;

    } else if (_dict.get(s + "#n") < 0.0 && _dict.get(s
        sn = _dict.get(s + "#n") + sn;
        sni++;
    } else if (_dict.get(s + "#n") < -0.25 && _dict.get
        neg = _dict.get(s + "#n") + neg;
        negi++;
    } else if (_dict.get(s + "#n") < -0.75) {
        vn = _dict.get(s + "#n") + vn;
        vni++;
    }

    total = _dict.get(s + "#n") + total;
```

Εικόνα 9: Διεξαγωγή ελέγχων για την αναγνώριση πολικότητας

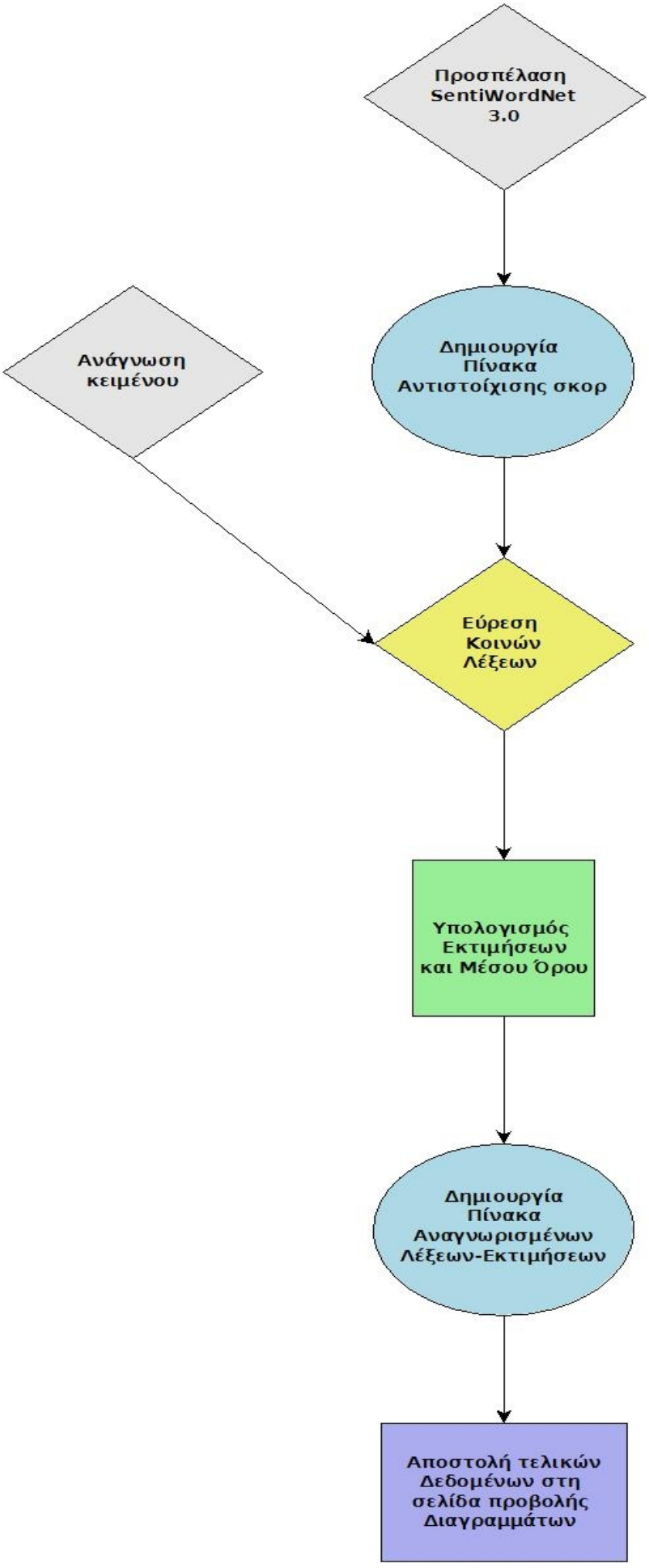
Στο παραπάνω απόσπασμα απλά υλοποιούνται μερικοί έλεγχοι για τον συναισθηματικό προσανατολισμό της λέξης/φράσης. Οι μεταβλητές p, sp, vp, n, neg, sn και vn(οι αντίστοιχες

με i που εκφράζει ακέραιο θα εξηγηθούν στη συνέχεια) εκφράζουν τις περιφραστικές κατηγορίες : Positive, Somehow Positive, Very positive, Neutral, Negative, Somehow Negative και Very Negative αντίστοιχα και αποτελούν **το αθροιστικό σκορ όλων των λέξεων** που ανήκουν σε κάθε μία από αυτές τις κατηγορίες. Οι αντίστοιχες μεταβλητές που έχουν την κατάληξη i εκφράζουν το **πλήθος των λέξεων** που αντιστοιχούν σε κάθε κατηγορία. Όπως εξηγήθηκε στο κεφάλαιο του SentiWordNet 3.0 όσο πιο κοντά βρίσκεται η βαθμολογία μίας λέξης στο 1 τόσο πιο θετική κρίνεται η σημασία της. Αντίστροφα, όσο πιο κοντά είναι η βαθμολογία μίας λέξης στο -1 τόσο πιο αρνητική η έννοιά της. Με αυτήν την λογική, ανάλογα με το σκορ μία λέξη τοποθετείται σε μία από τις παραπάνω κατηγορίες. Αν για παράδειγμα μία λέξη έχει σκορ μεγαλύτερο ή ίσο του 0.75 τότε καταχωρείται στην κατηγορία Very Positive και στην μεταβλητή **vp** προστίθεται το σκορ της συγκεκριμένης λέξης ενώ η μεταβλητή **vpi** αυξάνεται κατά 1. Ανεξάρτητα με την θετική ή αρνητική πολικότητα της λέξης σε μία μεταβλητή total που χρησιμοποιείται για τον υπολογισμό της πολικότητας του συνόλου, δηλαδή την ολική συναισθηματική εκτίμηση, προστίθεται το σκορ της λέξης(Άρα προφανώς αν η λέξη έχει αρνητικό σκορ, η τιμή της μεταβλητής total μειώνεται!).

3. Προκειμένου να μπορέσουμε να δείξουμε στον χρήστη το πως εκτίμησε η εφαρμογή Sentiment analysis την κάθε λέξη του κειμένου(εφόσον αυτή αναγνωρίστηκε) η σελίδα δημιουργεί ένα ακόμα Hashmap αυτή τη φορά με κλειδί τις λέξεις του κειμένου που αναγνωρίστηκαν και τιμή το σκορ που τους ανατέθηκε.

4. Τέλος, μένει να μεταβιβάσουμε στην σελίδα που θα αξιοποιήσει τα αποτελέσματα της διαδικασίας εκτίμησης της πολικότητας. Όλες οι μεταβλητές του βήματος **4** καθώς και το Hashmap που δημιουργήθηκε στο βήμα **5** στέλνονται στην **σελίδα εμφάνισης στατιστικών/αποτελεσμάτων αγγλικού κειμένου**.

Το σύνολο των λειτουργιών της σελίδας αυτής συνοψίζεται στο παρακάτω σχεδιάγραμμα:

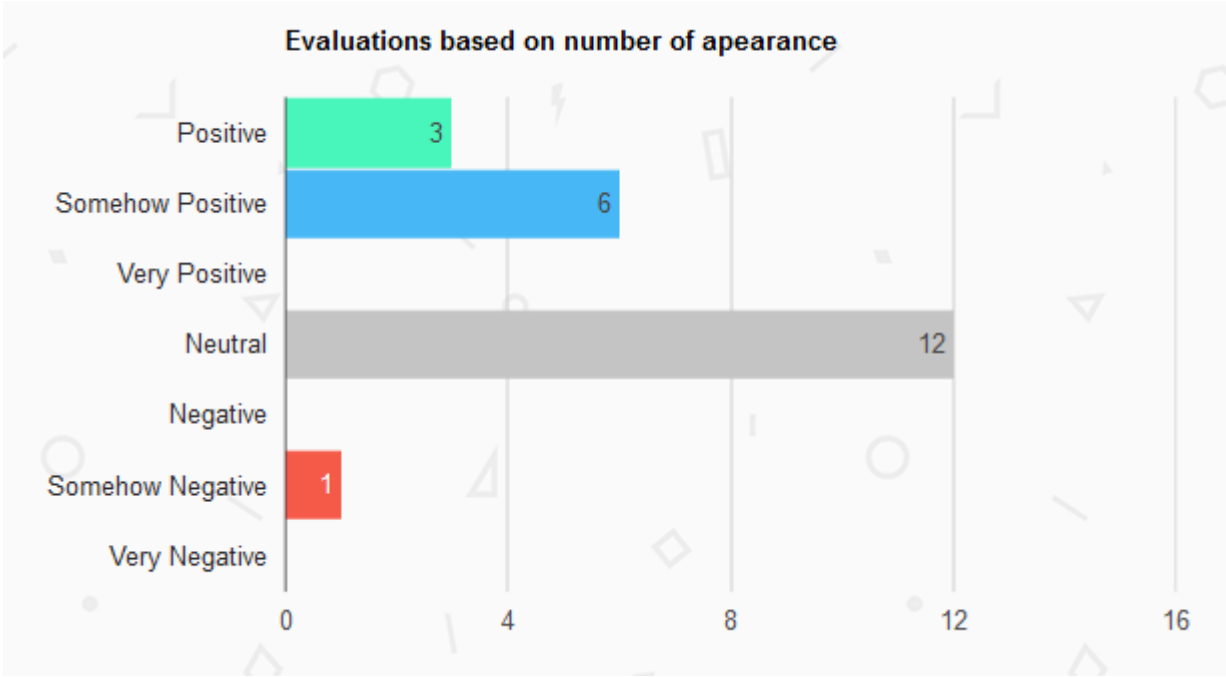


Εικόνα 10: Σύνολο λειτουργιών σελίδας εξαγωγής αποτελεσμάτων αγγλικού κειμένου

Σελίδα εμφάνισης στατιστικών/αποτελεσμάτων αγγλικού κειμένου:

Όπως είναι προφανές, η σελίδα αυτή, είναι υπεύθυνη για την ενημέρωση του χρήστη όσον αφορά τα αποτελέσματα που παρήχθησαν και της γενικής εκτίμησης που δημιουργήθηκε Η σελίδα αυτή αρχικά πληροφορεί τον χρήστη για την λεκτική κατηγορία(όπως Somehow Positive) που ανήκει η το κείμενο που αυτός έδωσε και στη συνέχεια του παραθέτει όλες τις πληροφορίες που αποθηκεύσαμε στις μεταβλητές που μεταδόθηκαν από την προηγούμενη σελίδα. Αυτό επιτυγχάνεται μέσω της δημιουργίας τριών γραφικών παραστάσεων. Οι τρεις γραφικές παραστάσεις είναι οι εξής:

- Ένα **ραβδόγραμμα(bar chart)** το οποίο χρησιμοποιεί τις μεταβλητές πλήθους των κατηγοριών(pi,spi,vpi,ni,negi,sni,vni) με σκοπό να δείξει τον αριθμό των λέξεων που αναγνωρίστηκαν κατανεμημένο στις λεκτικές κατηγορίες, έτσι ώστε να μπορεί ο χρήστης να ξέρει ποια από αυτές τις κατηγορίες πλειοψηφεί στο κείμενο που έδωσε. Αν δηλαδή για παράδειγμα οι περισσότερες λέξεις του κειμένου κρίθηκαν ως πολύ αρνητικές.

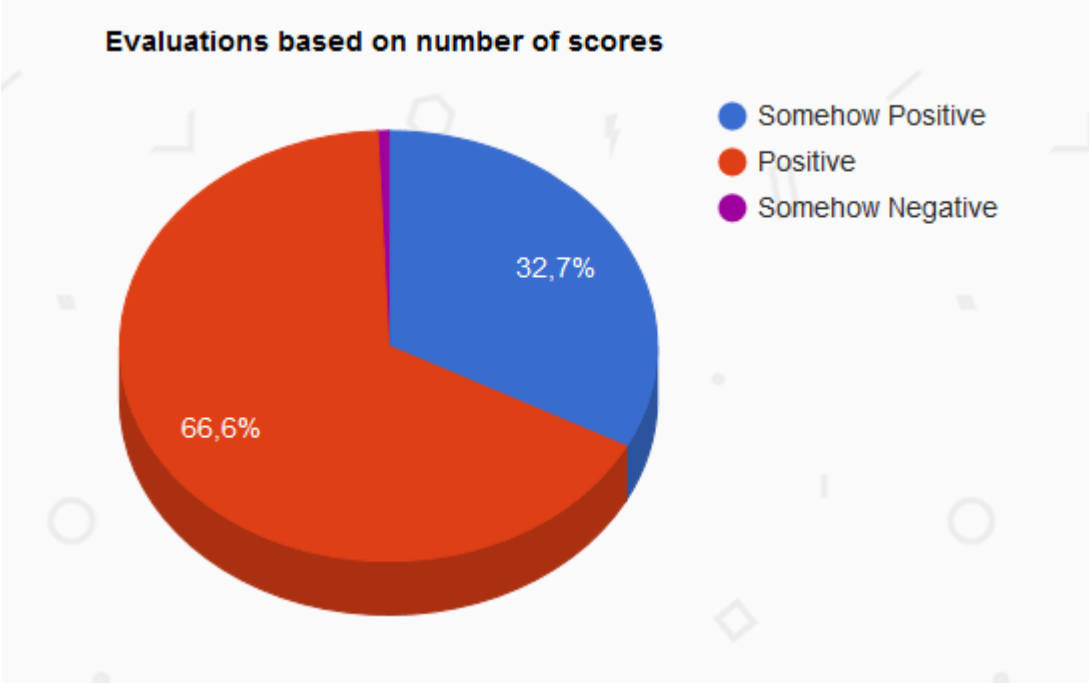


Εικόνα 11: Μορφή ενός ραβδογράμματος

Το παραπάνω ραβδόγραμμα, έχει προκύψει από την εισαγωγή ενός σχολίου από βίντεο στο youtube και οι πληροφορίες που παρέχει είναι πως στις τις επτά κατηγορίες, οι λέξεις του κειμένου ταξινομήθηκαν ως εξής: Στην κατηγορία Positive εισήχθηκαν 3 λέξεις στην κατηγορία

Somehow Positive 6, 12 λέξεις κρίθηκαν ουδέτερες(Neutral) ενώ μία θεωρήθηκε Somehow Negative. Προφανώς δεν βρέθηκαν λέξεις που βάσει της εφαρμογής να ανήκουν στις υπόλοιπες κατηγορίες και το σύνολο των λέξεων που αναγνωρίστηκαν είναι 22.

- **Ένα διάγραμμα πίτας(pie chart)** το οποίο εκμεταλλεύεται τις αθροιστικές μεταβλητές (p,sp,vp,neg,sn,vn) προκειμένου να δείξει στον χρήστη ποια από τις λεκτικές κατηγορίες υπερτερεί των άλλων στο κείμενο που έδωσε. Πόσο δηλαδή επηρέασε η κάθε κατηγορία το τελικό αποτέλεσμα. Μπορεί αριθμητικά μία κατηγορία όπως το “Somehow Negative” να πλειοψηφεί αριθμητικά αλλά παρόλα αυτά λόγω της ύπαρξης μερικών “Very Positive” όρων η εκτίμηση να ήταν θετική και αυτό φαίνεται εδώ.

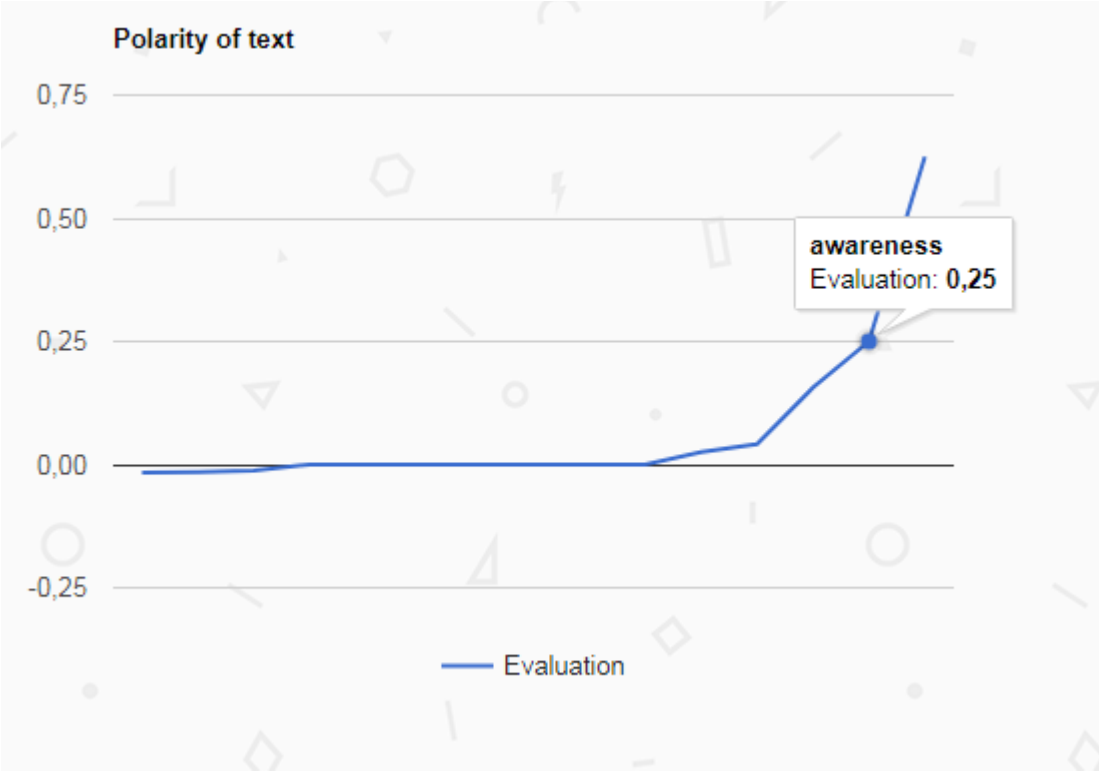


Εικόνα 12: Μορφή ενός διαγράμματος πίτας

Το παραπάνω διάγραμμα πίτας προέκυψε από το ίδιο σχόλιο στο youtube με το ραβδόγραμμα που δείξαμε παραπάνω και μας δείχνει πως η μεταβολή του συνόλου που προκαλείται από τρεις μόνο “θετικές λέξεις” που βρέθηκαν στο κείμενο, είναι κατά πολύ μεγαλύτερη από τη μεταβολή που προξενούν οι έξι “σχετικά θετικές λέξεις” και η μία “σχετικά αρνητική” μαζί.

- Ένα **line chart.** Το τελευταίο διάγραμμα το οποίο παρουσιάζεται στον χρήστη, είναι η **γραφική παράσταση(line chart)** της γενικής εικόνας της συναισθηματικής πόλωσης του κειμένου, η οποία προκύπτει εισάγοντας στον άξονα χχ' τις λέξεις που αναγνωρίστηκαν σε ίδια διαστήματα μεταξύ τους και στον άξονα ψψ' τις τιμές που

τους ανέθεσε η εφαρμογή, περνώντας τον κέρσορά του πάνω στις ακμές του διαγράμματος ο χρήστης μπορεί να δει το ζεύγος τιμών, δηλαδή, την κάθε λέξη σε συνδυασμό με την εκτίμηση πολικότητάς της.



Εικόνα 13: Μορφή μίας γραφικής παράστασης

Και αυτό το παράδειγμα, ανήκει στην εκτίμηση του ίδιου σχολίου με τα προηγούμενα και όπως βλέπουμε ο χρήστης αγγίζοντας την πρώτη ακμή του του διαγράμματος ενημερώνεται πως η λέξη “awareness” που υπήρχε στο κείμενό του βαθμολογήθηκε με 0.25 άρα είναι σχετικά θετική (η λογική βαθμολογίας του SentiWordNet εξηγείται αναλυτικά δίπλα στα διαγράμματα καθώς και στην σελίδα πληροφοριών). Πέραν των αντιστοιχιών, η γραφική αυτή παράσταση παρέχει μία αριθμητική οπτικοποίηση των αποτελεσμάτων.

Εκτός από την προβολή των αποτελεσμάτων μέσω γραφικών παραστάσεων, η σελίδα αναλαμβάνει να ενημερώσει τον χρήστη όπως αναφέρθηκε κι πριν για την λογική με την οποία οι αριθμοί που του εμφανίζονται αντιστοιχούν σε κάποια εκτίμηση της πολικότητας του κειμένου και τέλος του δίνει τη δυνατότητα να γυρίσει πίσω στην σελίδα εισαγωγής κειμένου για να ξαναδοκιμάσει αν αυτός επιθυμεί.

Να σημειώσουμε πως ολόκληρο το σχόλιο στο youtube που χρησιμοποιήθηκε στο παράδειγμα κρίθηκε ως “Πολύ Θετικό” λόγω της ύπαρξης θετικών λέξεων κατά κύριο λόγο.

Σελίδα εξαγωγής αποτελεσμάτων ελληνικού κειμένου

Παρόλο που ο σκελετός της διαδικασίας που πραγματοποιείται σε αυτήν την σελίδα είναι σχεδόν εξ ολοκλήρου ίδιος με αυτόν της αντίστοιχης σελίδας για το αγγλικό κείμενο, υπάρχουν κάποιες πολύ βασικές προσθήκες. Ο λόγος ύπαρξης αυτών, είναι η πολυπλοκότητα των γραμματικών κανόνων της ελληνικής γλώσσας, συγκριτικά με αυτών της αγγλικής. Δεν χρειάζεται να εμβαθύνουμε πολύ σε αυτό, αρκεί να πάρουμε και το πιο απλό ρήμα προκειμένου να το αντιληφθούμε. Το ρήμα “κάνω” για παράδειγμα, στα αγγλικά είναι do και η κλίνοντας το στον ενεστώτα, ή στα αγγλικά present simple, βλέπουμε πως η λέξη δεν αλλάζει καθόλου παρά μόνο στο τρίτο πρόσωπο ενικού. Κάτι τέτοιο δεν ισχύει με το ρήμα κάνω στα ελληνικά. Σε κάθε πρόσωπο το ρήμα έχει διαφορετική κατάληξη. Αυτή είναι μία από τις απλές καταστάσεις, η ελληνική γραμματική έχει τόσους κανόνες και εξαιρέσεις που αυτό και μόνο καθιστά την αναγνώριση των λέξεων στο κείμενο μία ιδιαίτερα δύσκολη διαδικασία.

Προκειμένου λοιπόν να ξεπεραστεί το πρόβλημα αυτό, έγινε χρήση του γεγονότος ότι αυτό που αλλάζει σε ρήματα, ουσιαστικά, επιρρήματα και επίθετα όταν κλείνονται, είναι η **κατάληξη** και σε μερικές περιπτώσεις το **πρόθεμα.** Ο τρόπος με τον οποίο αξιοποιήθηκε η πληροφορία αυτή, είναι ο εξής: Κατά την δημιουργία του Hashmap στο οποίο καταχωρούνται τα ζεύγη λέξης-τιμής που προκύπτουν από την μελέτη του SentiWordNet (όπου κάθε λέξη εισάγεται στον ενικό δηλαδή για παράδειγμα κάθε ρήμα εισάγεται στο πρώτο πρόσωπο ενικού), η κατάληξη της λέξης αφαιρείται από το αλφαριθμητικό στο οποίο είναι αποθηκευμένη, προστίθεται μία καινούρια κατάληξη που αντιστοιχεί στο μέρος του λόγου στο οποίο ανήκει η αρχική και η νέα λέξη που προκύπτει προστίθεται και αυτή στο Hashmap με τιμή ίδια με την αρχική. Η διαδικασία αυτή υλοποιείται για κάθε κατάληξη κάθε χρόνο και κάθε μέρος του λόγου. Για να καταλάβουμε καλύτερα τη διαδικασία θα μελετήσουμε το παρακάτω απόσπασμα κώδικα.

```
if(data[0].equals("v")){
    v.add(index,score);
    String wk = j.substring(0, j.length()-1);
    _temp.put(wk+"εις#"+data[0],v);
    _temp.put(wk+"ει#"+data[0],v);
    _temp.put(wk+"ουμε#"+data[0],v);
    _temp.put(wk+"ειτε#"+data[0],v);
    _temp.put(wk+"ουν#"+data[0],v);
    _temp.put(wk+"ε#"+data[0],v);
    _temp.put(wk+"οντας#"+data[0],v);
    _temp.put(wk+"ανε#"+data[0],v);

    _temp.put(wk+"α#"+data[0],v);
    _temp.put(wk+"ες#"+data[0],v);
    _temp.put(wk+"αμε#"+data[0],v);
    _temp.put(wk+"ατε#"+data[0],v);
    _temp.put(wk+"αν#"+data[0],v);


       _temp.put(wk+"ουν#"+data[0],v);
    _temp.put(wk+"ουνε#"+data[0],v);
    _temp.put(wk+"ε#"+data[0],v);
    _temp.put(wk+"εστε#"+data[0],v);
    _temp.put(wk+"ομε#"+data[0],v);
    _temp.put(wk+"στε#"+data[0],v);
    _temp.put(wk+"οντας#"+data[0],v);
    _temp.put(wk+"ανε#"+data[0],v);
```

Εικόνα 14: Εισαγωγή καταλήξεων και προθεμάτων

Ας ξεκινήσουμε την ανάλυση από τη συνθήκη του if. Όπως αναφέρθηκε και στο κεφάλαιο του SentiWordNet, μία από της πληροφορίες που παρέχει το αρχείο SWN είναι το POS δηλαδή το μέρος του λόγου που είναι η λέξη. Χρησιμοποιώντας λοιπόν αυτήν την παροχή του, βρίσκουμε όλες τις λέξεις που ανήκουν στην κατηγορία "v", δηλαδή verb, άρα ρήμα. Έπειτα δημιουργούμε ένα substring της αρχικής λέξης που αποτελείται από την αρχική λέξη μείον το τελευταίο χαρακτήρα. Η πράξη αυτή γίνεται στην γραμμή "**String wk = j.substring(0, j.length()-1);** ". Τέλος εισάγουμε στο Hashmap το ένα νέο αλφαριθμητικό που αποτελείται από το substring που δημιουργήσαμε στο προηγούμενο βήμα συν την προσθήκη ενός αλφαριθμητικού που περιέχει την κατάληξη. Έτσι, κάνοντας αυτήν την διαδικασία για κάθε χρόνο και κλίση ενός ρήματος, καθιστούμε την εφαρμογή ικανή να κατανοήσει και να εκτιμήσει το περιεχόμενο των περισσότερων μορφών ενός ρήματος.

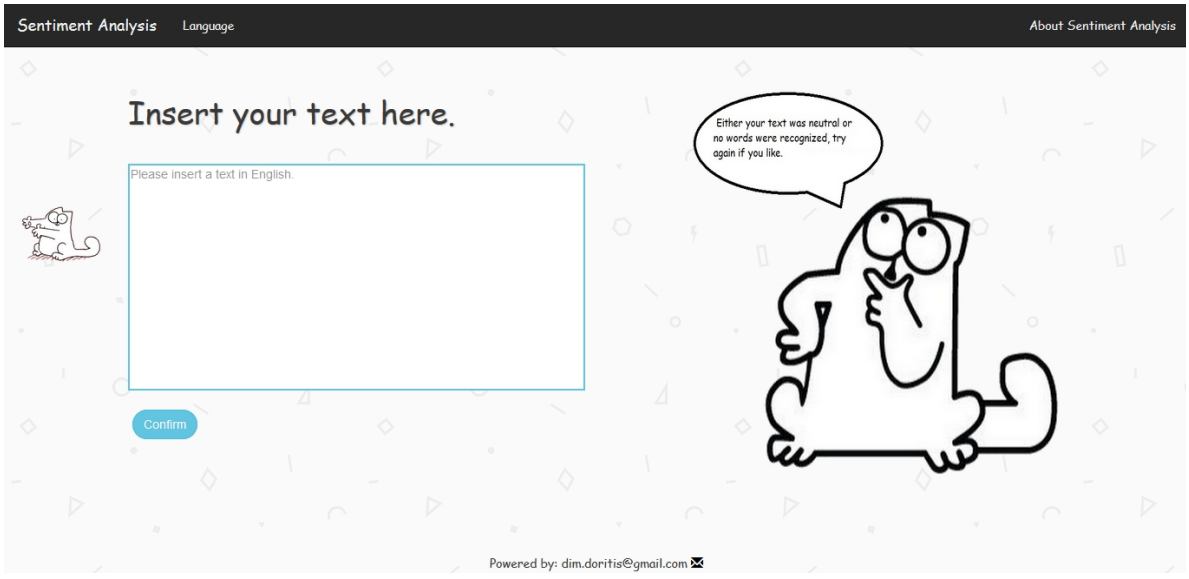
Αντίστοιχες καταλήξεις και προθέματα προστίθενται σε όλα τα μέρη του λόγου που περιλαμβάνει το μεταφρασμένο SentiWordNet σύμφωνα πάντα με τους κανόνες της ελληνικής γραμματικής. Πέραν από τη διαδικασία εκτέλεσης των γραμματικών κανόνων της ελληνικής γραμματικής, η λειτουργία αυτής της σελίδας συμβαδίζει σε λογική με τη σελίδα εξαγωγής αποτελεσμάτων αγγλικού κειμένου.

Σελίδα εμφάνισης στατιστικών/αποτελεσμάτων ελληνικού κειμένου

Η σελίδα αυτή εμφανίζει τα ίδια στατιστικά στοιχεία στον χρήστη με την αντίστοιχη σελίδα για τα αγγλικά και η μόνη διαφορά της είναι πως τα όλα τα αποτελέσματα και οι επιλογές του χρήστη είναι στα ελληνικά.

Σελίδα ενημέρωσης λάθους

Σε περίπτωση που στο κείμενο που έδωσε ο χρήστης, δεν αναγνωρίστηκε καμία λέξη, πράγμα το οποίο μπορεί να είναι αποτέλεσμα κάποιου τυπογραφικού λάθους ή έλλειψης κάποιας λέξης από το SentiWordNet, μόλις το κείμενο περάσει από την διαδικασία εκτίμησης σε κάποια από τις σελίδες εξαγωγής αποτελεσμάτων, ο χρήστης ανακατευθύνεται σε σελίδες που τον ενημερώνουν για την ύπαρξη σφάλματος και τον παραπέμπουν να δοκιμάσει ξανά. Όπως και στις περισσότερες άλλες σελίδες, υπάρχει διαφορετική σελίδα ενημέρωσης λάθους για τις αγγλικές και για τις ελληνικές εκτιμήσεις. Η σελίδα ενημέρωσης λάθους των αγγλικών είναι η παρακάτω.



Εικόνα 15: Εμφάνιση σελίδας ενημέρωσης λάθους

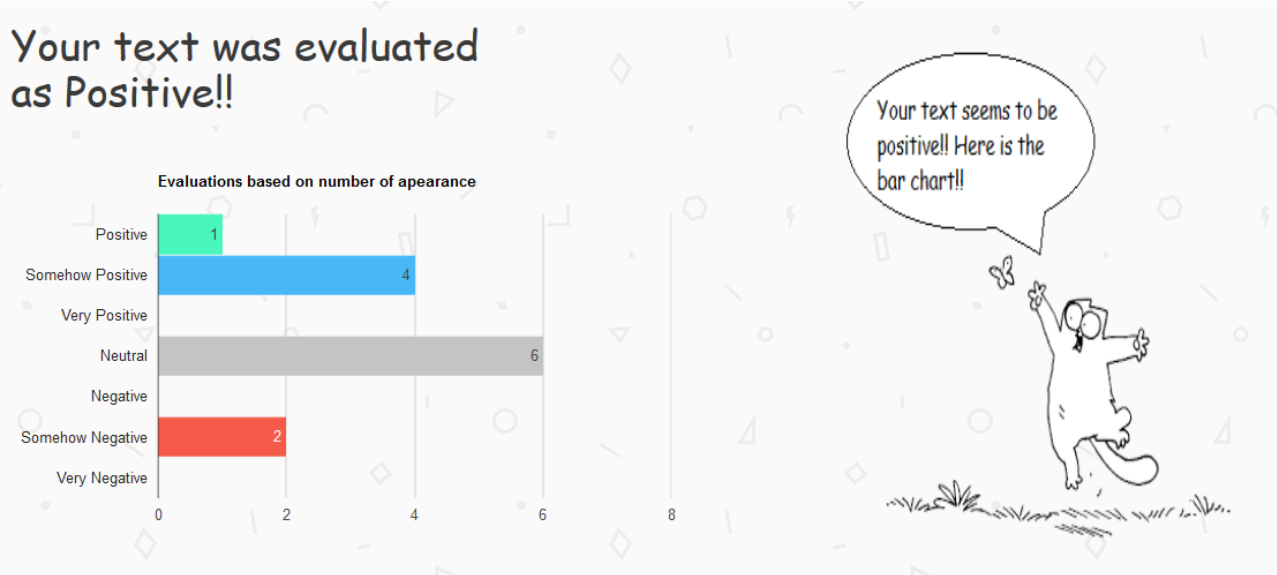
3.5 Sentiment Analysis Εφαρμογή-Ολοκληρωμένα Παραδείγματα

Παράδειγμα 1. Αγγλικό σχόλιο στο Youtube

Στο πρώτο παράδειγμα θα πάρουμε και θα εξετάσουμε ένα σχόλιο από το βίντεο "The Dangers of the Good Child" του καναλιού "The School of Life" (https://www.youtube.com/watch?v=5DTIzzf6ncg). Το σχόλιο είναι το εξής: **"The best kind of people. They just need a little love in adulthood. Without them this world is pure trash."**

Ένας άνθρωπος θα σχημάτιζε εύκολα την εντύπωση πως πρόκειται για ένα θετικό, ίσως και πολύ θετικό σχόλιο. Παρόλα αυτά, υπάρχουν και μερικές λέξεις που έχουν γενικά αρνητικό περιεχόμενο. Ας δούμε λοιπόν την οθόνη με τα αποτελέσματα και τα στατιστικά που εμφανίζει για αυτό το σχόλιο η εφαρμογή Sentiment Analysis.

Το ραβδόγραμμα που εμφανίζει φαίνεται παρακάτω.

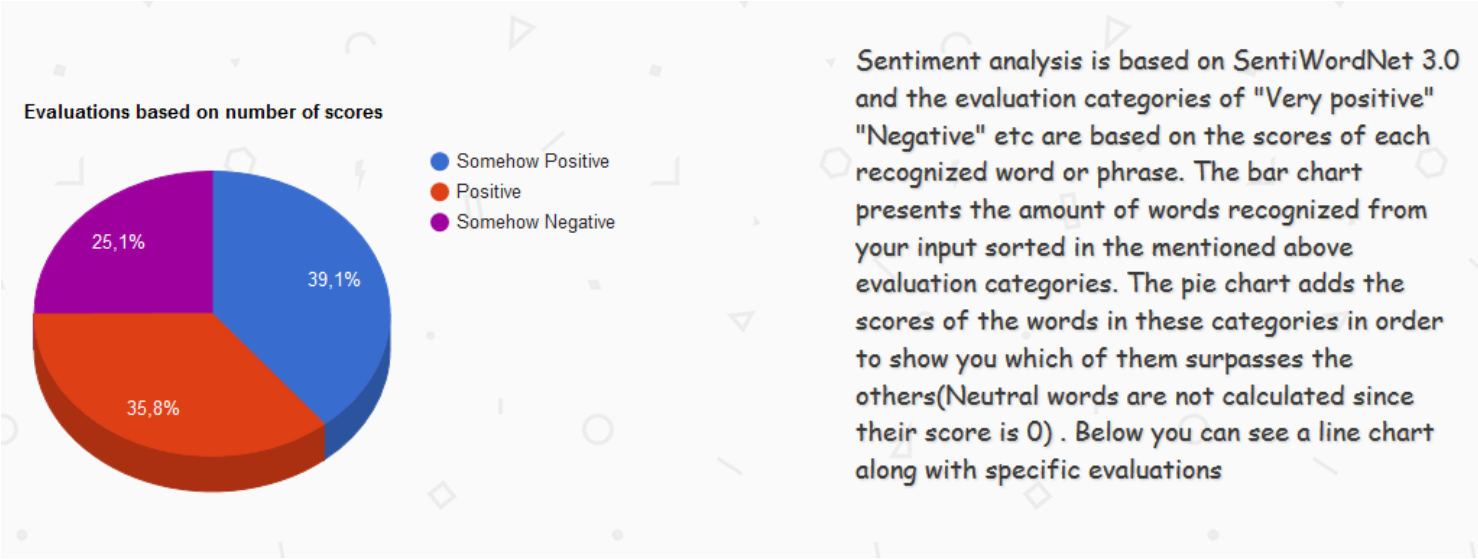


Εικόνα 16: Παράδειγμα 1: Ραβδόγραμμα

Βλέπουμε πως πράγματι, η εφαρμογή εκτίμησε το σύνολο του κειμένου ως θετικό. Συγκεκριμένα αναγνώρισε μία θετική λέξη, τέσσερεις σχετικά θετικές, κι δύο σχετικά αρνητικές. Όλες οι υπόλοιπες λέξεις που αναγνωρίστηκαν κρίθηκαν ουδέτερες επομένως, δεν επηρεάζουν την εκτίμηση της εφαρμογής. Για να μπορέσουμε όμως να αξιολογήσουμε την

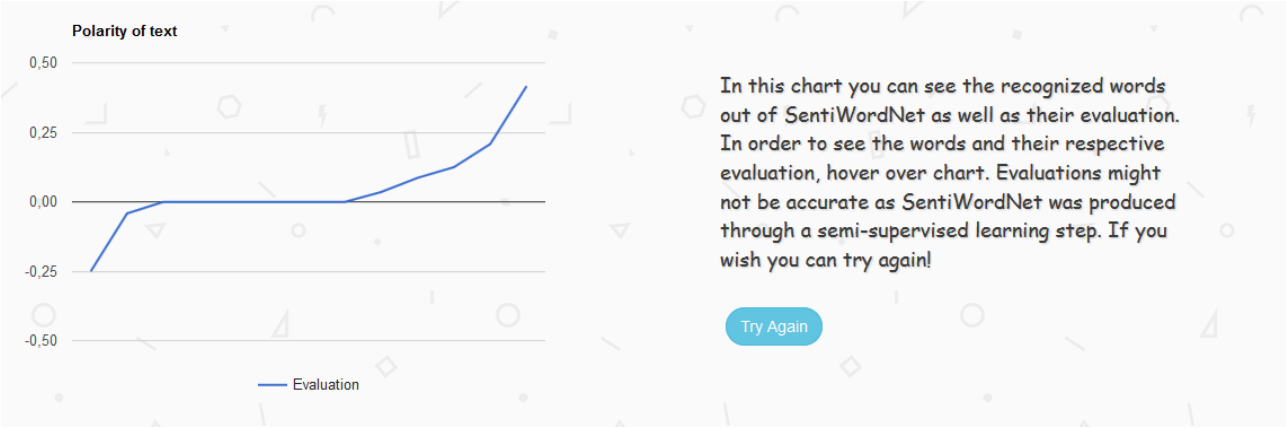
εγκυρότητα του αποτελέσματος που μας έδωσε η εφαρμογή πρέπει να δούμε και τις εκτιμήσεις των λέξεων μεμονωμένα, πράγμα το οποίο θα κάνουμε μόλις δούμε την γραφική παράσταση της πολικότητας του κειμένου.

Τώρα είναι η σειρά του γραφήματος πίτας.



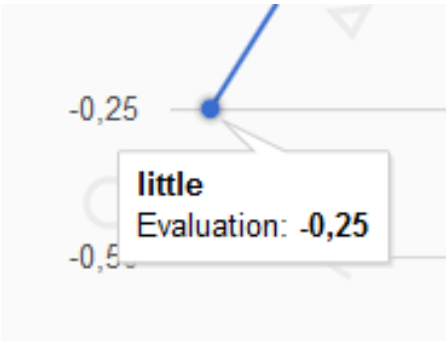
Τώρα που ξέρουμε πως λειτουργεί ο μηχανισμός δημιουργίας του διαγράμματος πίτας, καταλαβαίνουμε πως η αθροιστική επιρροή στην εκτίμηση του συνόλου, των λέξεων που κρίθηκαν ως “Σχετικά Θετικές” που έχουν δηλαδή σκορ μέσα στο σύνολο (0,0.25] είναι μεγαλύτερη από αυτή των λέξεων που ανήκουν στις κατηγορίες “Θετική” και “Σχετικά Αρνητική”. Επίσης παρατηρούμε πως περίπου το 75% της πολικότητας του κειμένου προέρχεται από θετικά βαθμολογημένες λέξεις.

Τελευταίο και πιο σημαντικό στην αξιολόγηση της λειτουργίας είναι το line chart δηλαδή η γραφική παράσταση της πολικότητας του κειμένου.

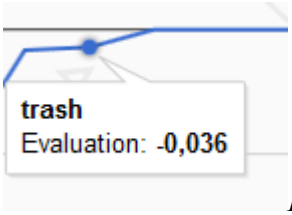


Εικόνα 18: Παράδειγμα 1: Γραφική Παράσταση πολικότητας

Όπως περιμέναμε, η πλειοψηφία της γραφικής παράστασης βρίσκεται πάνω από τον άξονα των χ άρα στα θετικά. Ας εξετάσουμε όμως ποιες είναι οι λέξεις που δημιουργούν τις ακμές κάτω από τον άξονα των χ έχοντας βαθμολογηθεί αρνητικά.



Εικόνα 20: Αξιολόγηση λέξης



Εικόνα 19: Αξιολόγηση λέξης "trash"

Όπως βλέπουμε, οι λέξεις που αξιολογήθηκαν ως αρνητικές στο κείμενο είναι η λέξη “little” με βαθμολογία -0.25 που είναι ακριβώς πάνω στο όριο των “σχετικά αρνητικών” λέξεων και των “αρνητικών” και η λέξη “trash” με σκορ -0.036 που δεδομένου ότι μία λέξη με βαθμολογία 0 είναι ουδέτερη σημαίνει πως είναι ελαφρώς αρνητική. Παρά το γεγονός ότι και οι δύο λέξεις μπορούν γενικά να έχουν ουδέτερο περιεχόμενο, η ανθρώπινη λογική συνιστά πως όταν δεν επηρεάζουν το περιεχόμενο μίας φράσης, συνήθως τον επηρεάζουν αρνητικά. Επομένως μπορούμε να πούμε πως η κρίση των δύο συγκεκριμένων λέξεων ήταν σωστή. Επίσης από το σύνολο των στατιστικών, συμπεραίνουμε πως η ανθρώπινη αντίληψη, συμβαδίζει με αυτήν

της εφαρμογής στο συγκεκριμένο παράδειγμα, οπότε η εκτίμηση του συνόλου μπορεί με ασφάλεια να θεωρηθεί σωστή.

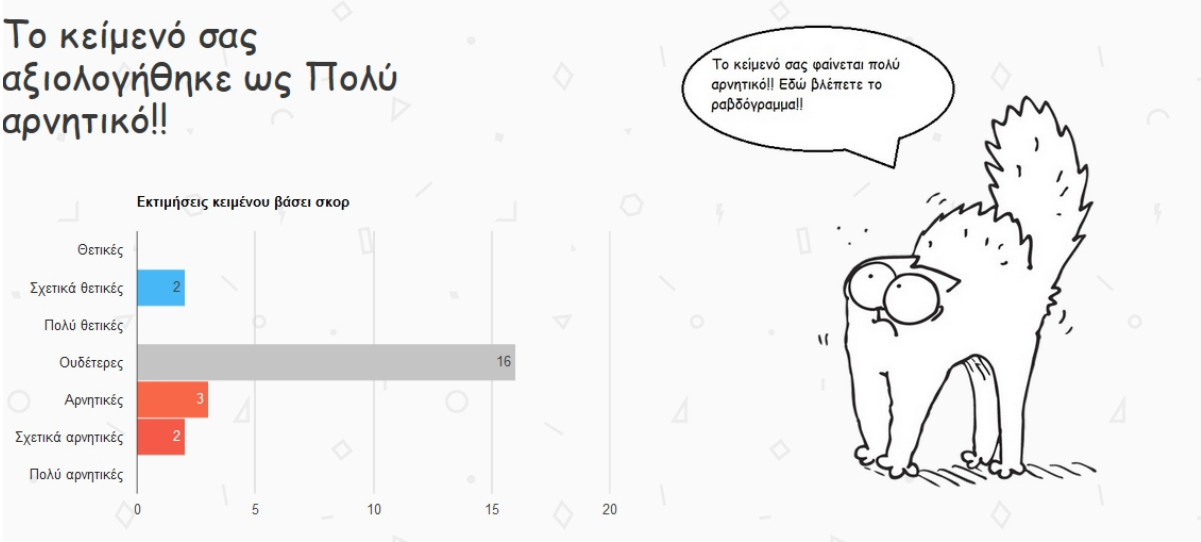
Παράδειγμα 2. Τμήμα ελληνικού πεζού κειμένου

Σε αυτό το παράδειγμα θα δοκιμάσουμε την αντίληψη της εφαρμογής Sentiment Analysis σε jsp με ένα κομμάτι ενός άρθρου/πεζού κειμένου, που αναρτήθηκε στο Healingeffect.gr και αφορά την ασθένεια του καρκίνου (https://www.healingeffect.gr/2016/03/karkinos-3-synesthimatiki-paragontes-pou-odigoun-se-karkino.html). Το τμήμα του κειμένου είναι το εξής:

“Η δυστυχία σχεδόν πάντα δείχνει ότι υπάρχει ένας δρόμος που δεν ακολουθήθηκε. Ένα ταλέντο που δεν καλλιεργήθηκε, ένας εαυτός που δεν αναγνωρίστηκε. Πίσω από τον καρκίνο του τελευταίου σταδίου που εξαπλώνεται ραγδαία υπάρχει μια προσωπικότητα που δεν ζει με τα δικά της κίνητρα, ένας άνθρωπος που αισθάνεται παγιδευμένος μέσα σ’ ένα πρόβλημα χωρίς λύση”

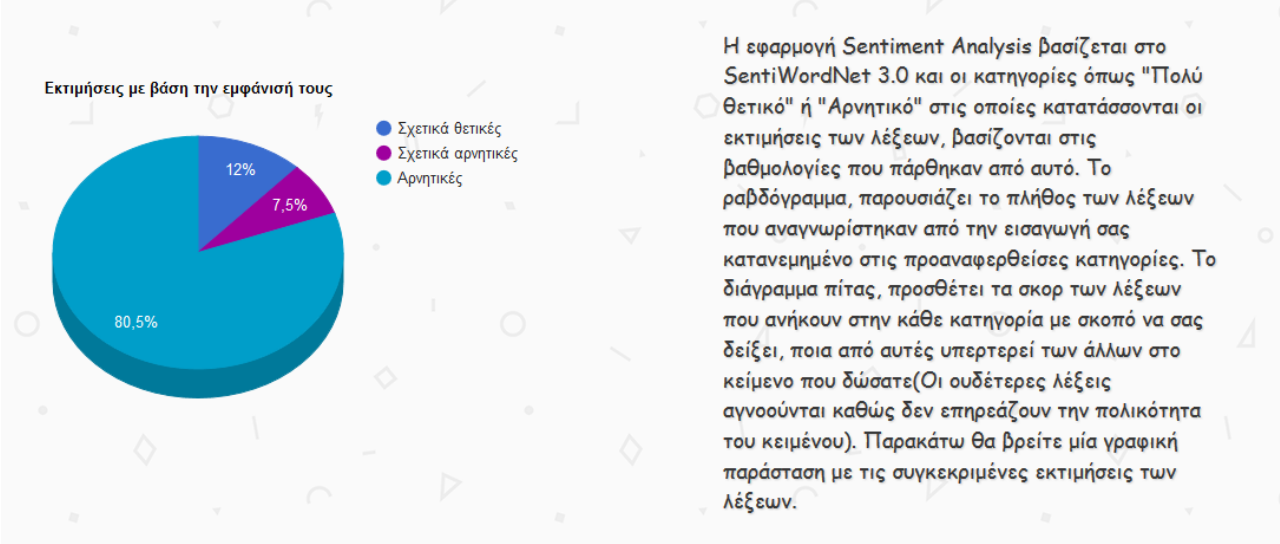
Όπως και με το προηγούμενο παράδειγμα, δεν είναι δύσκολο για έναν άνθρωπο που γνωρίζει τη γλώσσα να αντιληφθεί πως το περιεχόμενο του κειμένου είναι αρκετά λυπηρό, επομένως συναισθηματικά αρνητικό. Ας δούμε λοιπόν τι εμφανίζεται στον χρήστη που εισάγει αυτό το κείμενο στην σελίδα κατάθεσης ελληνικού κειμένου.

Θα ξεκινήσουμε ξανά από το πρώτο διάγραμμα, δηλαδή το ραβδόγραμμα που καταμετρά τις λέξεις ανά κατηγορία. Θα δούμε επίσης και την γενική εκτίμηση του κειμένου που παρέχει η εφαρμογή.



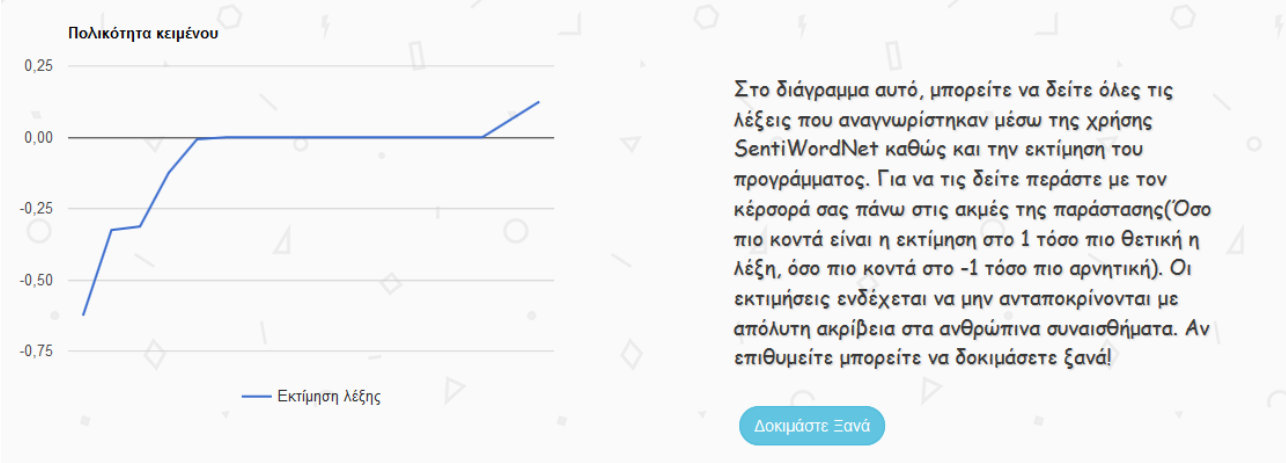
Όπως βλέπουμε, πολλές από τις λέξεις του κειμένου κρίθηκαν ουδέτερες, όμως από τις επτά λέξεις που επηρέασαν την πολικότητα του κειμένου, η εφαρμογή συμπέρανε πως το περιεχόμενο του κειμένου είναι “Πολύ Αρνητικό”. Η κατανομή των λέξεων στις κατηγορίες και συγκεκριμένα η πληθώρα ουδέτερων λέξεων, είναι απολύτως λογική καθώς εισάγαμε ένα αρκετά μεγάλο κείμενο και συνήθως η αρνητική ή θετική συναισθηματική πόλωση ενός κειμένου, προκύπτει από λίγες λέξεις κλειδιά που μας δείχνουν την πρόθεση του συγγραφέα. Παρακάτω θα δούμε ποιες είναι οι λέξεις που κατέβασαν τόσο πολύ τη γενική βαθμολογία του κειμένου, καθώς και ποιες είναι οι δύο “σχετικά θετικές” λέξεις που αναγνωρίστηκαν.

Τώρα θα δούμε ποια από τις κατηγορίες που εντοπίστηκαν καθόρισε σε μεγαλύτερο βαθμό το αποτέλεσμα με τη βοήθεια του γραφήματος πίτας



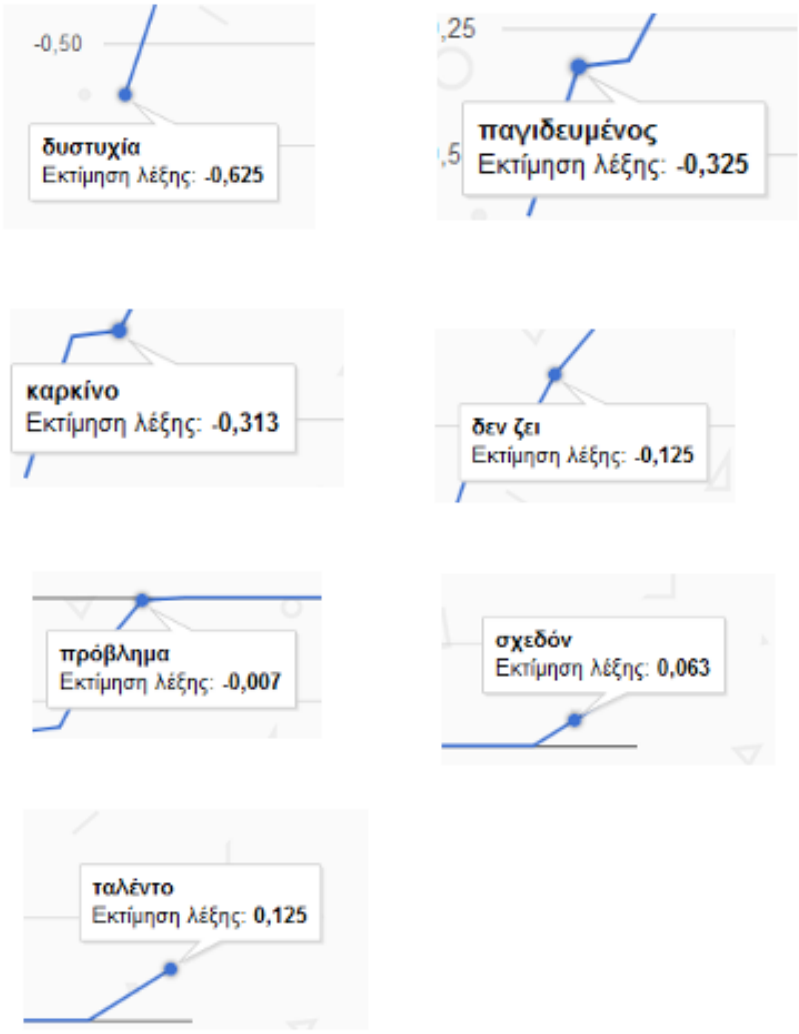
Εικόνα 22: Παράδειγμα 2: Γράφημα Πίτας

Όπως βλέπουμε το 80.5% του γενικού αποτελέσματος δηλαδή της εκτίμησης “Πολύ Αρνητικό” που δόθηκε στο κείμενο, προέκυψε από τις αρνητικές λέξεις. Αυτό σημαίνει ότι κατά απόλυτη τιμή, η βαθμολογία των αρνητικών λέξεων ήταν πολύ μεγαλύτερη και των “σχετικά αρνητικών” αλλά και των “σχετικά θετικών”. Άρα είναι λογικό η συνολική βαθμολογία να έπεσε πιο κοντά στο -1 και για αυτόν τον λόγο το κείμενο να κρίθηκε όπως κρίθηκε.
Μένει τώρα να αναλύσουμε το αποτέλεσμα σε βάθος λέξεων, πράγμα το οποίο θα επιτύχουμε με την γραφική παράσταση της πολικότητας(Line chart).



Εικόνα 23: Παράδειγμα 2: Γραφική Παράσταση

Στην εικόνα βλέπουμε πως υπάρχουν 4 λέξεις που βρίσκονται ξεκάθαρα κάτω από τον άξονα των χ, επομένως έχουν αρνητική βαθμολογία. Υπάρχει και μία λέξη που είναι ακριβώς κάτω από τον άξονα, με αποτέλεσμα να μη φαίνεται καλά σε αυτήν την μακρινή απεικόνιση αλλά θα την δείξουμε παρακάτω μεμονωμένα. Η γραφική παράσταση ταυτίζεται σε μεγάλο τμήμα της με τον άξονα των χ πράγμα το οποίο μας δείχνει αυτό που σχολιάσαμε και παραπάνω, ότι υπάρχουν δηλαδή πολλές ουδέτερες λέξεις στο κείμενο μιας και αυτό είναι σχετικά μεγάλο σε μήκος. Τέλος, βλέπουμε πως σε κάποιο σημείο εμφανίζονται και θετικές λέξεις και για αυτόν το λόγο η συνάρτηση ανεβαίνει πάνω από τον άξονα των χ. Ας δούμε όμως τώρα ποιες είναι οι λέξεις συγκεκριμένα που έδωσαν στο γράφημα αυτήν την μορφή:



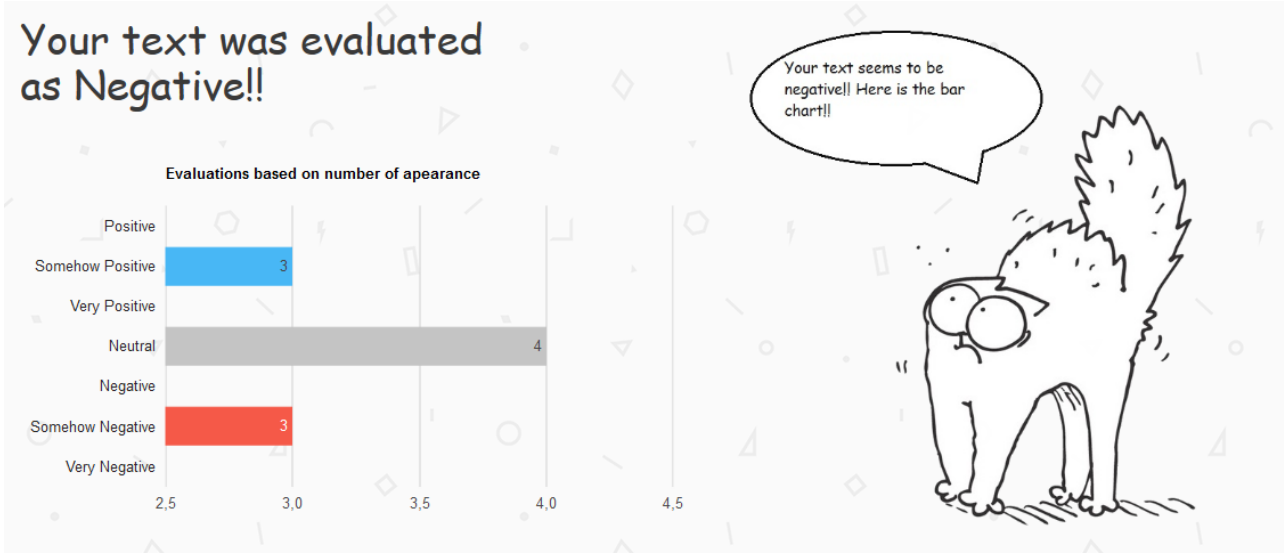
Εικόνα 24: Παράδειγμα 2: Εκτιμήσεις λέξεων

Η εικόνα αυτή, με τις επτά θετικά ή αρνητικά βαθμολογημένες λέξεις μας δίνει τις εξής πληροφορίες. οι αρνητικές λέξεις είναι, η λέξη **“δυστυχία”** με βαθμολογία 0.625 που είναι κοντά στα όρια μίας πολύ αρνητικής λέξης σύμφωνα με την βαθμολογία που εξηγήσαμε παραπάνω, η λέξη **“παγιδευμένος”** με βαθμολογία 0.325 και τέλος, η λέξη **“καρκίνο”.** Οι λέξεις αυτές βάσει και της ανθρώπινης κρίσης έχουν πράγματι αρνητική έννοια συνήθως, οπότε μπορούμε να θεωρήσουμε την εκτίμηση της εφαρμογής σωστή. Οι “σχετικά αρνητικές” λέξεις που εντοπίστηκαν είναι η φράση **“δεν ζει”** και αυτή που βρισκόταν στο γράφημα ακριβώς κάτω από το άξονα τον χ, η λέξη “πρόβλημα”. Για άλλη μία φορά μπορούμε να δούμε πως η κρίση της εφαρμογής ήταν πετυχημένη. Τέλος, οι λέξεις που θεώρησε η εφαρμογή πως έχουν θετική σημασία είναι, η λέξη **“σχεδόν”** και η λέξη **“ταλέντο”.** Όσον αφορά τη λέξη “σχεδόν”, θα μπορούσε κανείς να πει πως δεν επηρεάζει το περιεχόμενο του κειμένου, αλλά

στην πραγματικότητα επηρεάζεται η ίδια από αυτό. Ο άνθρωπος που λέει την φράση “Σχεδόν τελείωσα το πλύσιμο των πιάτων” μάλλον είναι ευτυχισμένος, ενώ ο άνθρωπος που λέει “Χτύπησα τόσο δυνατά το πόδι μου που σχεδόν το έσπασα” πιθανότατα έχει γνωρίσει και καλύτερες μέρες. Βλέπουμε λοιπόν πως πιθανότατα θα ήταν καλύτερα να είχε εκτιμηθεί η λέξη ως ουδέτερη. Θα θεωρήσουμε λοιπόν την εκτίμηση ελαφρώς ατυχή με μία απόκλιση από την θεωρητικά σωστή της τάξεως του 0.063 στα 2(μιας και η μέγιστη απόκλιση μπορεί να είναι από το -1 στο 1). Η λέξη “ταλέντο” όμως έχει πράγματι συνήθως θετική έννοια. Πέραν από την επιτυχία ή αστοχία της εφαρμογής κάτι που μπορούμε να σχολιάσουμε από τα παραπάνω αποτελέσματα σχετικά με την λειτουργία της είναι η φράση “δεν ζει”. Ο λόγος που η εφαρμογή την θεώρησε αρνητική είναι επειδή έχει συμπεριληφθεί ένας κανόνας στην υλοποίησή της που αφορά την ύπαρξη του μορίου “δεν” και την αντιστροφή που επιφέρει στις λέξεις που ακολουθούν.

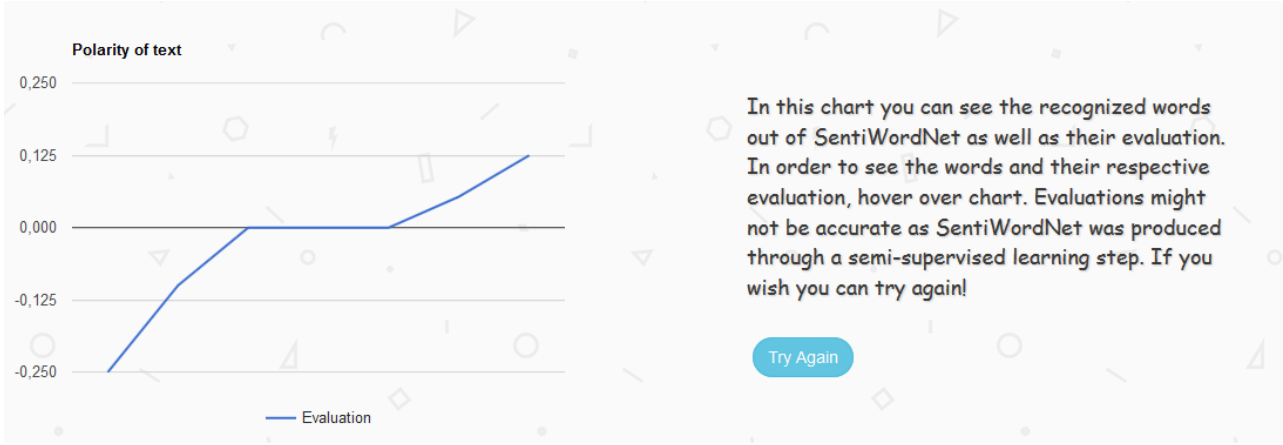
Παράδειγμα 3. Tweet του Barrack Obama 12/08/17

Σε αυτό το παράδειγμα θα χρησιμοποιήσουμε ένα tweet του Barrack Obama που έγραψε στις δώδεκα Αυγούστου του 2017. Το tweet είναι το εξής: ***“People must learn to hate, and if they can learn to hate, they can be taught to love”.*** Το μήνυμα που θέλει να περάσει ο πρώην πρόεδρος των Ηνωμένων πολιτειών της Αμερικής έχει σαφώς θετικό μήνυμα αλλά υπάρχουν λέξεις με αρνητικό περιεχόμενο, ας δούμε τα στατιστικά που θα μας προβάλει η εφαρμογή όταν εισάγουμε αυτό το κείμενο.



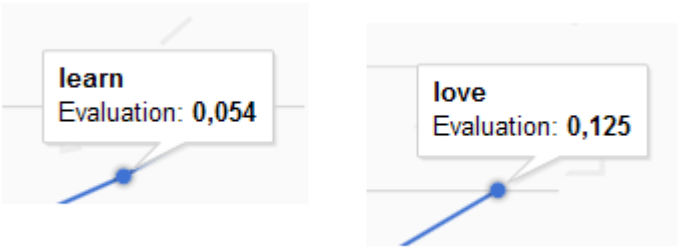
Εικόνα 25: Παράδειγμα 3: Ραβδόγραμμα

Βλέπουμε πως η εφαρμογή έκρινε το κείμενο ως αρνητικό παρόλο που οι λέξεις που αναγνώρισε είναι ισάριθμα χωρισμένες σε κατηγορίες θετικών και αρνητικών λέξεων που σχετίζονται με το ίδιο εύρος απόλυτων τιμών. Αυτό σημαίνει πως οι λέξεις που αναγνωρίστηκαν ως αρνητικές έχουν μεγαλύτερες βαθμολογίες από άποψη απόλυτης τιμής από τις θετικές. Επομένως επηρέασαν περισσότερο την τελική κρίση της εφαρμογής. Αφού εξηγήσαμε αυτό, μπορούμε να παραλείψουμε το γράφημα πίτας και να προχωρήσουμε κατευθείαν στην γραφική παράσταση οπού θα αποκτήσουμε καλύτερη εικόνα των λόγων που οδήγησαν σε αυτό το αποτέλεσμα.



Εικόνα 26: Παράδειγμα 3:Γραφική Παράσταση

Οι "σχετικά θετικές" λέξεις που βρέθηκαν είναι οι εξής:



Εικόνα 27: Εκτιμήσεις λέξεων 'learn,''love"

Ενώ οι "σχετικά αρνητικές" λέξεις είναι οι:

Εικόνα 28: Εκτιμήσεις Λέξεων 'hate,''must"

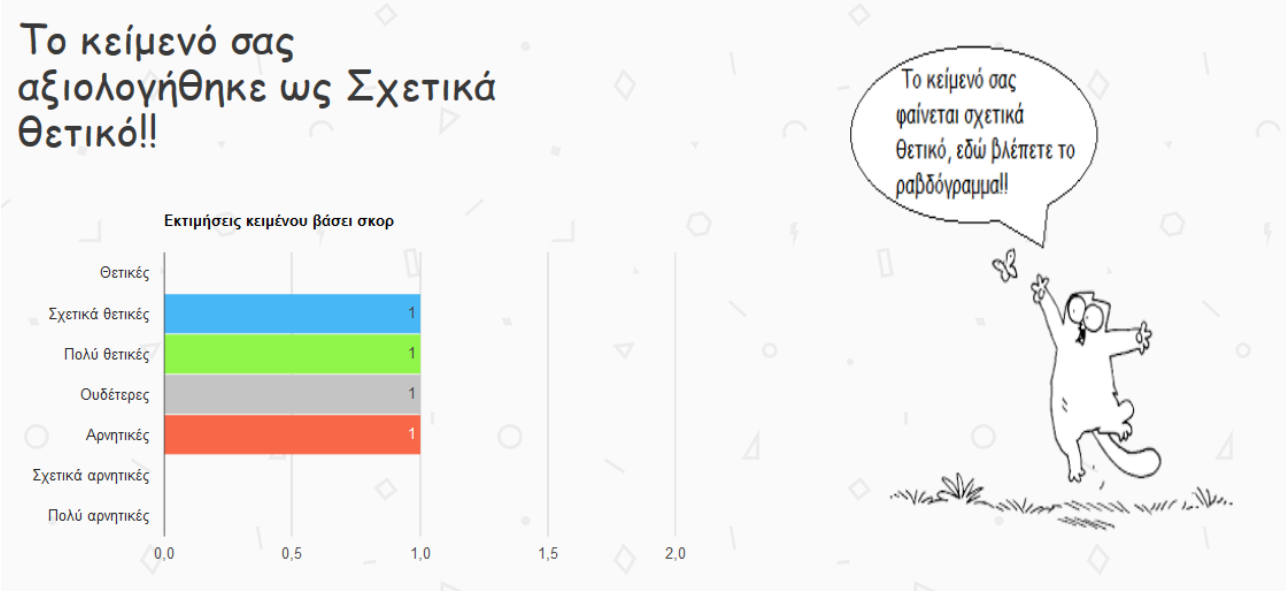
Μία πρώτη παρατήρηση που οφείλουμε να κάνουμε είναι πως ενώ το ραβδόγραμμα μας ενημέρωσε για την ύπαρξη τριών σχετικά αρνητικών λέξεων και τριών σχετικά θετικών λέξεων, στη γραφική παράσταση και οι δύο κατηγορίες φαίνεται να έχουν από δύο μόνο λέξεις. Ο λόγος που συμβαίνει αυτό, είναι επειδή η λέξη learn και η λέξη hate, εμφανίζονται από δύο φορές μέσα στο κείμενο και η θετική ή αρνητική σημασία τους λαμβάνεται υπόψιν δύο φορές. Θα δώσουμε δύο παραδείγματα για να εξηγήσουμε την συμπεριφορά αυτή της εφαρμογής. Ένα πιθανός λόγος που επαναλαμβάνεται στο κείμενο μία λέξη πολλές φορές είναι να την έχει γράψει πολλές φορές συνεχόμενες ο συγγραφέας για λόγους έμφασης. Έχει διαφορά να πει ο μικρός Κωστάκης “Μισώ τις μπάμιες!” από το να πει “Μισώ, μισώ, μισώ, μισώ τις μπάμιες!”. Άλλο ενδεχόμενο και πιθανώς πιο ρεαλιστικό, είναι να αναφέρεται η ίδια λέξη σε διαφορετικά σημεία του κειμένου με διαφορετικά συμφραζόμενα. Αυτό ακριβώς συμβαίνει στο παράδειγμα όπου σε ένα σημείο ο άνθρωπος μαθαίνει να μισεί και σε άλλο μαθαίνει να αγαπά.

Έχοντας εξηγήσει αυτήν την λειτουργία της εφαρμογής ας δούμε τώρα μαθηματικά γιατί το κείμενο κρίθηκε αρνητικό, αυτό συμβαίνει επειδή η λέξη “hate” με αρκετά αρνητική βαθμολογία εμφανίζεται και μετριέται δύο φορές.

Το τελευταίο και πιο σημαντικό που μπορούμε να συμπεράνουμε είναι πως σε αυτό το παράδειγμα η εκτίμηση της εφαρμογής Sentiment Analysis απέχει πάρα πολύ από την πραγματικότητα αυτό όμως δεν οφείλεται σε σφάλμα στην αξιολόγηση των λέξεων μα στο γεγονός πως ο ανθρώπινος λόγος δεν βασίζεται μόνο σε χαρακτήρες και λέξεις για να μεταβιβάσει ένα συναίσθημα αλλά και στις εικόνες και τους συνειρμούς που προκαλεί.

Παράδειγμα 4. Everything Bad is Good for You

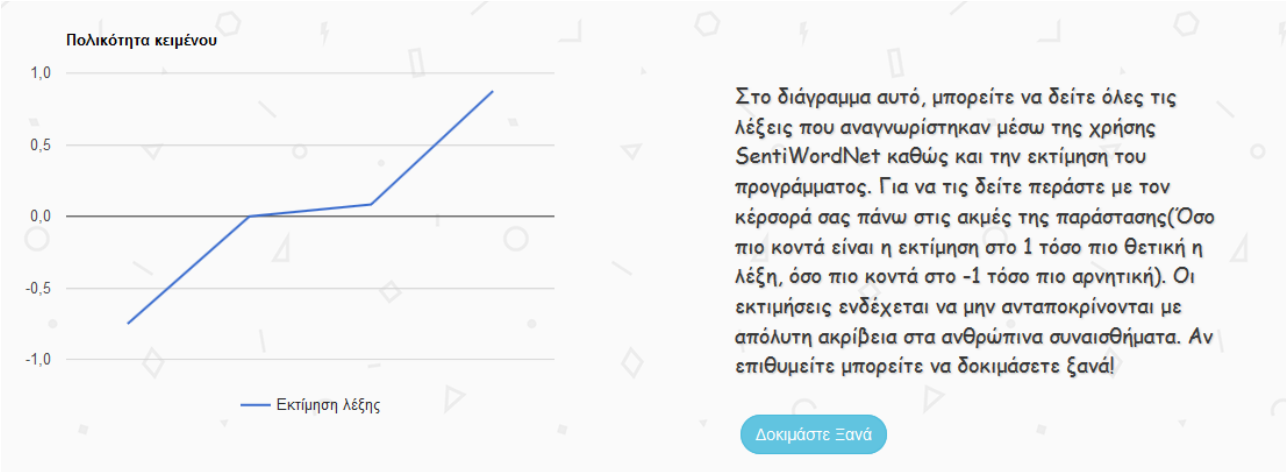
Θα δούμε και ένα τελευταίο παράδειγμα πολύ συνοπτικά στο οποίο ξανά τα νοήματα δεν εκφράζονται μέσα από το πλήθος των λέξεων. Πρόκειται για τον τίτλο ενός βιβλίου “Everything Bad is Good for You” του Steven Johnson(Johnson 2005).
Θα μεταφράσουμε τον τίτλο σε “Οτιδήποτε κακό είναι καλό για σένα” και θα το δώσουμε προς εκτίμηση στην εφαρμογή.



Εικόνα 29: Παράδειγμα 4: Ραβδόγραμμα

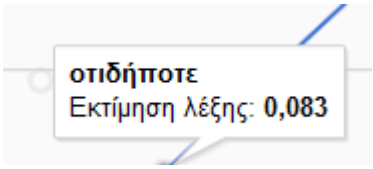
Αυτή τη φορά η εκτίμηση της εφαρμογής Sentiment Analysis ήταν επιτυχημένη επειδή η λέξη "καλός" και "κακός" έχουν παρόμοιες βαθμολογίες από πλευράς απόλυτης τιμής όντας παρόλα αυτά αντίθετες, όμως η λέξη "οτιδήποτε" κρίνεται ως σχετικά θετική βάσει του SentiWordNet, με σκορ 0.083 με αποτέλεσμα η θετική βαθμολογία του κειμένου να υπερτερεί!

Ενδεικτικά, το line chart:



Εικόνα 30: Παράδειγμα 4: Γραφική Παράσταση

Και η βαθμολογία της λέξης "οτιδήποτε"



Εικόνα 31: Εκτίμηση λέξης Ότιδήποτε"

3.6 Εκτίμηση Εγκυρότητας Αποτελεσμάτων

Όπως είδαμε και σε έναν βαθμό στα παραδείγματα, η ακρίβεια των αποτελεσμάτων της εφαρμογής Sentiment Analysis σε jsp, μπορεί να μελετηθεί σε δύο επίπεδα. Το πρώτο είναι το επίπεδο της λέξης, πόσο καλά δηλαδή μπορεί η εφαρμογή να αξιολογήσει την πολικότητα μίας μεμονωμένης λέξης. Το δεύτερο είναι αυτό του συνόλου. Πόσο καλά εκφράζει η εκτίμηση όλου του κειμένου τα συναισθήματα που αυτό περνάει σε κάποιον αναγνώστη. Προκειμένου να μελετηθούν οι δύο αυτοί παράγοντες, ένας μεγάλος αριθμός αριθμός από μικρού μεγέθους κείμενα(συνήθως tweets ή σχόλια σε δικτυακές πλατφόρμες) αλλά και μεγαλύτερου μήκους (συνήθως κρητικές βιβλίων) εισήχθησαν στην εφαρμογή, αφού πρώτα οι όροι του κειμένου αλλά και το ίδιο το κείμενο είχαν χαρακτηριστεί ως θετικά ή αρνητικά βάσει πάντα της ανθρώπινης λογικής.

Συνολικά μελετήθηκαν, πενήντα κείμενα με 1763 **διαφορετικούς όρους** των οποίων το περιεχόμενο κρίθηκε πως έχει κάποια επιρροή στον συναισθηματικό προσανατολισμό του κειμένου. Παρακάτω, φαίνονται τα αποτελέσματα της αξιολόγησης.

Κατηγορία	**Ανθρώπινη Εκτίμηση**	**Εκτίμηση Sentiment Analysis**	**Απόκλιση Συνόλου Κατηγορίας**	**Επιτυχία Κατηγορίας**
Θετικό Κείμενο	32	23	28.5%	60.8%
Αρνητικό Κείμενο	18	27	50%	55.5%
Θετική Λέξη	1148	944	17.7%	65.8%
Αρνητική Λέξη	615	819	33.1%	63.5%

Πίνακας 2: Αξιολόγηση αποτελεσμάτων εφαρμογής

Προτού εξετάσουμε τα στατιστικά που προέκυψαν, είναι σημαντικό να αναφερθεί πως υπήρχαν περιπτώσεις στις οποίες η εφαρμογή έκρινε ως θετικές ή αρνητικές λέξεις που κατά την ανθρώπινη λογική δεν επηρέαζαν άμεσα την πολικότητα του κειμένου. Τα περιστατικά αυτά δεν φαίνονται στον πίνακα παραπάνω γιατί, στην συγκριτική πλειοψηφία αυτών η απόκλιση της βαθμολογίας της εφαρμογής από το 0 απείχε συνήθως λιγότερο από 0,1 στην

κλίμακα [-1,1] και για αυτόν το λόγο η ανάλυση επικεντρώθηκε στο σύνολο των λέξεων που κρίθηκαν σημαντικές από τον άνθρωπο.

Θα περάσουμε τώρα στα στατιστικά. Το ποσοστό Απόκλισης Συνόλου Κατηγορίας, εκφράζει πόσο απείχε ο αριθμός των λέξεων ή κειμένων που έκρινε η εφαρμογή ως αρνητικά ή θετικά φορτισμένα σε σχέση με τον άνθρωπο. Ας πάρουμε για παράδειγμα το σύνολο των θετικών κειμένων. Από τα 50 κείμενα, η ανθρώπινη λογική θεώρησε πως τα 32 είναι θετικά. Η εφαρμογή Sentiment Analysis σε jsp όμως έκρινε μόνο 23 κείμενα ως θετικά. Η διαφορά των δύο αριθμών είναι 9 που αποτελεί το 28.5% του 32.

Η Επιτυχία Κατηγορίας εκφράζει τον αριθμό των λέξεων ή κειμένων που η εφαρμογή έκρινε σωστά. Αυτό σημαίνει πως από τους 1148 όρους που βάσει ανθρώπινων συναισθημάτων είναι θετικοί, η εφαρμογή πέτυχε το 65.8% το οποίο αντιστοιχεί στους 755 από αυτούς. Ενώ από τους 615 αρνητικούς, η εφαρμογή πέτυχε τους 391 δηλαδή το 63.5% περίπου.

Από τα στατιστικά μπορούμε να συμπεράνουμε πως η εφαρμογή, με μέσο όρο επιτυχίας 64.65% στις μεμονωμένες λέξεις, αναγνωρίζει πολύ καλύτερα τον συναισθηματικό προσανατολισμό των λέξεων από ότι των κειμένων, στα οποία σημείωσε ποσοστό επιτυχίας 58.15%. Αυτό οφείλεται κυρίως, σε κάτι που είδαμε και σε μερικά από τα παραδείγματα. Ο ανθρώπινος λόγος πολύ συχνά δεν βασίζεται μόνο στο περιεχόμενο των λέξεων σαν ξεχωριστές οντότητες, αλλά και σε περίπλοκους συνδυασμούς αυτών καθώς και συνειρμούς που προκαλεί στον αποδέκτη.

Το ποσοστό επιτυχίας της εφαρμογής Sentiment Analysis, ανταποκρίνεται στα μέχρι στιγμής υπάρχοντα δεδομένα και οι εκτιμήσεις που παρέχει μπορούν να θεωρηθούν σε μεγάλο βαθμό επαρκής

3.7 Επίλογος-Σκέψεις για μελλοντικές εκδόσεις

Παρά το γεγονός ότι η εφαρμογή Sentiment Analysis, είναι σε θέση να εκτιμά το περιεχόμενο ενός κειμένου με επαρκή ακρίβεια, όπως και οι υπόλοιπες εφαρμογές που προσπάθησαν να υλοποιήσουν κάτι παρόμοιο, υπάρχει προοπτική για εξέλιξη. Στόχος για μελλοντικές εκδόσεις είναι να υλοποιηθούν ακόμα περισσότεροι κανόνες και να ληφθούν υπόψη ακόμα περισσότερα στατιστικά, προκειμένου να βελτιωθούν τα αποτελέσματα που δίνει.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ahkter, Julie Kane, and Steven Soria. n.d. “Sentiment Analysis: Facebook Status Messages.”

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. n.d. “SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.” http://zeynepaltan.info/4-SentiwordNet.pdf.

Bing, Liu. 2010. “Sentiment Analysis and Subjectivity-2010.” University of Illinois at Chicago. http://people.sabanciuniv.edu/berrin/proj102/1-BLiu-Sentiment%20Analysis%20and%20Subjectivity-NLPHandbook-2010.pdf.

Esuli, Andrea, and Fabrizio Sebastiani. n.d. “Determining Term Subjectivity and Term Orientation for Opinion Mining.” Istituto di Scienza e Tecnologie dell’Informazione. Accessed May 17, 2016. http://anthology.aclweb.org/E/E06/E06-1025.pdf.

Hatzivassiloglou, Vasileios, and Yu Hong. 2003. “Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences.”

Johnson, Steven. 2005. *Everything Bad Is Good for You*.

Kurniawan, Budi. 2002. *Java for the Web with Servlets, JSP, and EJB*.

McFarland, David Sawyer. 2011. *JavaScript & JQuery: The Missing Manual*.

Miller, George A. 1995. “WordNet a Lexical Database for English.” Communications of the ACM. http://dl.acm.org/citation.cfm?id=219748.

Oracle JSP 2.0 Specification. 2013. “JSP 2.0 Specification, Final Release.” Oracle. http://download.oracle.com/otn-pub/jcp/jsp-2_3-mrel2-eval-spec/JSP2.3MR.pdf.

Pak, Alexander, and Patrick Paroubek. n.d. “Twitter as a Corpus for Sentiment Analysis and Opinion Mining.” Universit´e de Paris-Sud, Laboratoire LIMSI-CNRS, Bˆatiment 508.

Strapparava, Carlo, and Alessandro Valitutti. 2004. “WordNet-Affect: An Affective Extension of WordNet.” http://s3.amazonaws.com/academia.edu.documents/3436027/wordnet_affect__an_affective_extension_o_73471.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1497452318&Signature=CdfJBoT8A0QS8892428RzW%2FhKmY%3D&response-content-disposition=inline%3B%20filename%3DWordNet-Affect_An_Affective_Extension_of.pdf.Ahkter, Julie Kane, and Steven Soria. n.d. “Sentiment Analysis: Facebook Status Messages.”