Χαροκόπειο Πανεπιστήμιο
Τμήμα Πληροφορικής & Τηλεματικής

Διπλωματική Εργασία

Αυτόματη Εξαγωγή και Διαχείριση Περιεχομένου από Ειδησεογραφικά Site

Χαράτσεβ Φίλιππος

Επιβλέπων

Ηρακλής Βαρλάμης, Λέκτορας

Μέλη Εξεταστικής Επιτροπής

Γεώργιος Δημητρακόπουλος, Λέκτορας

Θωμάς Καμαλάκης, Καθηγητής

Αθήνα
Ιούλιος 2014

Περιεχόμενα

Πίνακας Εικόνων

Σύνοψη

Η πληθώρα ειδησεογραφικών πηγών και η ανάγκη της αγοράς για ανάλυση ειδήσεων σε μεγάλη κλίμακα, έχει καταστήσει αναγκαία την ύπαρξη μηχανισμών που θα συγκεντρώνουν πληροφορία από ειδησεογραφικές πηγές ελαχιστοποιώντας την ανθρώπινη παρέμβαση. Ταυτόχρονα καθιστά πλέον επιτακτική την ολοκλήρωση επιμέρους εργαλείων ανάλυσης και επεξεργασίας κειμένων σε μια ενιαία πλατφόρμα που θα στοχεύει στον τελικό χρήστη αποκρύβοντας τις επιμέρους λεπτομέρειες.

Στόχος είναι να σχεδιάσει και να υλοποιήσει έναν μηχανισμό ο οποίος θα υποστηρίζει αλγορίθμους που θα εντοπίζουν περιεχόμενο ενδιαφέροντος σε ειδησεογραφικά site ,αλλά και γενικότερα site το οποία περιέχουν κείμενο γραμμένο από χρήστες αυτών, και θα δημιουργούν με αυτόματο τρόπο μηχανισμούς εξαγωγής και κατηγοριοποίησης συναισθήματος από κείμενα γραμμένα σε φυσική γλώσσα. Στη συνέχεια θα οργανώνουν και αναλύουν τη συγκεντρωμένη πληροφορία και θα οπτικοποιούν την παραγόμενη γνώση.

Στα πλαίσια της παρούσας εργασίας σχεδιάζεται ένα γενικότερο πλαίσιο διεπαφών με τη χρήση Java Interfaces, το οποίο ορίζει ένα ευέλικτο περιβάλλον εξαγωγής, ανάλυσης και οπτικοποίησης γνώσης. Η αρχιτεκτονική του συστήματος ορίζεται ούτως ώστε να δίνεται η δυνατότητα προσαρμογής του συστήματος σε διαφορετικές ανάγκες μέσω χρήσης διαφορετικών μορφωμάτων κώδικα για κάθε στάδιο της εξαγωγής γνώσης, όπως και διαφορετικούς συνδυασμούς αυτών. Η υλοποίηση των διεπαφών τις οποίες ορίζει το πλαίσιο, από εξωτερικές βιβλιοθήκες οι οποίες έχουν δημιουργηθεί με σκοπό την επίλυση των προβλημάτων που προκύπτουν από το κάθε στάδιο ανάλυσης και εξαγωγής γνώσης, επιτρέπουν ουσιαστικά στη δημιουργία ενός μεγάλου εύρους διαφορετικών προσεγγίσεων ως προς την επίλυση ενός προβλήματος. Δίνεται επίσης η δυνατότητα σύγκρισης των αποτελεσμάτων τα οποία προκύπτουν.

Δημιουργείται μια ενδεικτική υλοποίηση κάθε σταδίου του πλαισίου, ώστε να αποδειχθεί η αποτελεσματικότητα του σχεδιασμού, και μέσω του processing pipeline που ορίζεται, συνδυάζονται και δημιουργούν ένα ολοκληρωμένο σύστημα γνώσης. Για την υλοποίηση κάθε σταδίου αξιοποιούνται βιβλιοθήκες η οποίες παρέχονται δωρεάν στο διαδίκτυο. Μέσω της διαδικασίας δημιουργίας μια ολοκληρωμένης υλοποίησης παράγονται αρχεία με κατηγοριοποιημένη πληροφορία, η οποία μπορεί να αξιοποιηθεί οποιαδήποτε στιγμή ώστε να οπτικοποιηθεί το αποτέλεσμα δίχως την ανάγκη εκ νέου συλλογής και κατηγοριοποίησης των δεδομένων.

1

Εισαγωγή

1.1 Εξόρυξη Γνώμης και Ανάλυση Συναισθήματος

Ο όρος Εξόρυξη Γνώμης Η ανάλυση συναισθήματος ή εξόρυξη γνώμης είναι η μελέτη γνώμης, εκτιμήσεων, συμπεριφορών και συναισθημάτων τα οποία αφορούν προςωπα, γεγονότα και άλλου τύπου οντότητες, με τη χρήση υπολογιστικών μεθόδων. Η μελέτη επικεντρώνεται κυρίως στη ανάλυση γραπτού λόγου, αλλά και σε ανάλυση πολυμέσων.

Η ραγδαία ανάπτυξη των υπηρεσιών κοινωνικής δικτύωσης, αλλά και γενικότερα του διαδικτύου έχει ως αποτέλεσμα την παραγωγή τεράστιου όγκου πληροφορίας, κυρίως με τη μορφή γραπτού κειμένου, η οποία αποτελεί την βάση για τον πειραματισμό και την ανάπτυξη τεχνικών και εργαλείων για την ανάλυση συναισθήματος. Πληροφορία η οποία μπορεί να προέρχεται από σχόλια και κριτικές χρηστών, που τυπώνονται συνεχώς σε blogs, forums και κοινωνικά δίκτυα και περιέχει κομμάτια κειμένου, όπου εκάστοτε εκφράζεται συναίσθημα ή μια γνώμη. Ο εντοπισμός και η εξαγωγή των κειμένων αυτών είναι ένα πολύπλοκο πρόβλημα, λόγο του μεγάλου αριθμού των πηγών αλλά και της ποικιλομορφίας αυτών. Κάθε πηγή τείνει να περιέχει μεγάλο όγκο κειμένου, όπου η εξαγωγή του μέρους όπου πραγματικά υπάρχει κάποιου είδους συναίσθημα είναι δύσκολη πολλές φορές ακόμα και για έναν άνθρωπο. Κυρίως λόγο του ότι ο κάθε άνθρωπος επηρεάζεται άμεσα από παράγοντες οι οποίοι σχετίζονται με προσωπικές απόψεις, προτιμήσεις, εισάγοντας τον παράγοντα της υποκειμενικότητας, ο οποίος είναι σαφώς ανεπιθύμητος. Για το λόγο αυτό υπάρχει η ανάγκη ανάπτυξης αυτόματων συστημάτων τα οποία θα χρησιμοποιούν προκαθορισμένες τεχνικές ώστε να απαλείφονται οι όποιες πιθανές επιρροές, και να υπάρχει η έννοια ενός αντικειμενικού συστήματος εξαγωγής συναισθήματος και γνώμης.

Η χρησιμότητα της ανάλυσης αυτής έγκειται στην επιθυμητή ικανότητα πρόβλεψης ενός μελλοντικού γεγονότος η συμπεριφοράς βάσει συναισθημάτων και γνωμών οι οποίες υπάρχουν ήδη, είτε για ακαδημαϊκούς είτε για εμπορικούς σκοπούς. Όπως για παράδειγμα η αξιολόγηση ενός προϊόντος από την εταιρεία που το παράγει βάσει της γνώμης των καταναλωτών για αυτό και η σύνθεση της κατάλληλης πολιτικής μελλοντικής αξιοποίησης, ή τροποποίηση αυτού ή παρεμφερών προϊόντων.

Υπάρχουν δύο ειδών γνώμες. Οι απλές γνώμες, ή απλώς γνώμες και οι συγκριτικές γνώμες. Μια συγκριτική γνώμη εκφράζει μια σχέση ομοιοτήτων ή διαφορών μεταξύ δύο ή περισσότερων οντοτήτων μ και/ή τη προτίμηση ενός φορέα γνώμης, βασισμένη σε ορισμένα από τα κοινά χαρακτηριστικά των οντοτήτων αυτών[1]. Μια συγκριτική γνώμη εκφράζεται συνήθως με τη χρήση του συγκριτικού ή υπερθετικού βαθμού ενός επιθέτου ή ενός επιρρήματος.

Μια απλή γνώμη, είναι απλώς είναι αρνητικό ή θετικό συναίσθημα ή συμπεριφορά η οποία αφορά μια οντότητα ή ένα από τα χαρακτηριστικά της από έναν φορέα γνώμης. Όσον αφορά τις απλές γνώμες, των οποίων ο εντοπισμός τείνει να είναι σχετικά ευκολότερος, υπάρχει η ανάγκη αποσαφήνισης και κατηγοριοποίησης. Οι βαθμίδες πόλωσης 'Θετικό', 'Ουδέτερο' και 'Αρνητικό' χρησιμοποιούνται συνήθως για τον ορισμό της πόλωσης (polarity) μιας γνώμης. Σε πολλές περιπτώσεις χρησιμοποιούνται επιπλέον βαθμίδες, όπως 'Πολύ Θετικό' και 'Πολύ Αρνητικό' σε περιπτώσεις όπου είναι επιθυμητή η κατηγοριοποίηση απόψεων με μεγαλύτερη λεπτομέρεια. Ο αριθμός των βαθμίδων και η εκάστοτε χρησιμοποιούμενη τεχνική κατηγοριοποίησης επηρεάζουν σε μεγάλο βαθμό την ακρίβεια του τελικού αποτελέσματος. Οι τεχνικές οι οποίες χρησιμοποιούνται είναι συνήθως τεχνικές εξόρυξης γνώσης, δηλαδή αλγόριθμοι όπως ο Naive Bayesian, τα SVM (Support Vector Machines), τα Νευρωνικά δίκτυα κ.α.

Ο εντοπισμός και η κατηγοριοποίηση απόψεων μπορεί να εφαρμοστούν, είτε σε επίπεδο πρότασης είτε σε επίπεδο εγγράφου/κειμένου. Στη πρώτη περίπτωση η αρχική πηγή κείμενο χωρίζεται σε προτάσεις ή φράσεις, στις οποίες εφαρμόζονται οι τεχνικές εντοπισμού ανάλυσης γνώμης, για κάθε πρόταση ξεχωριστά. Το εγχείρημα αυτό απαιτεί σύνθετες τεχνικές εντοπισμού γνώμης, ακολουθώντας συντακτικούς κανόνες για τον εντοπισμό των σχετικών στοιχείων μιας πρότασης όπως το υποκείμενο που εκφέρει την άποψη/συναίσθημα, και το αντικείμενο το οποίο αυτή αφορά. Η κατηγοριοποίηση σε επίπεδο εγγράφου εξάγει μια γενική άποψη με βάσει το σύνολου των προτάσεων του αρχικού κειμένου, δίχως αυτή να αφορά ένα αντικείμενο μιας συγκεκριμένης πρότασης. Το 'αντικείμενο' είναι συνήθως γνωστό όπως για παράδειγμα σε περιπτώσεις απόψεων πολιτών σε ένα forum με θέμα ένα πολιτικό προςωπο. Και στις δύο περιπτώσεις ορίζονται μετρικές οι οποίες αφορούν την συχνότητα εμφάνισης λέξεων και ορισμένες φορές την σχετική τους θέση στο έγγραφο/πρόταση ώστε να καθοριστεί τελικά η πόλωση του συναισθήματος που εντοπίζεται.

Τέλος, όσον αφορά την κατηγοριοποίηση, αυτή γίνεται σε πολλές περιπτώσεις με τη βοήθεια ενός λεξικού του οποίο περιέχει ορισμένες λέξεις ή εκφράσεις χαρακτηρισμού συναισθημάτων για την εκάστοτε γλώσσα. Σε περιπτώσεις όπου είναι επιθυμητή η κατηγοριοποίηση ανεξαρτήτου γλώσσας η κατηγοριοποίηση γίνεται αποκλειστικά με τεχνικές εξόρυξης γνώσης.

1.2 Σκοπός της εργασίας

Η παρούσα εργασία έχει ως σκοπό την δημιουργία ενός πλαισίου χειρισμού κειμένων γραμμένων σε φυσική γλώσσα. Λειτουργία η οποία σαφώς εφαρμόζεται και στην ειδικότερη περίπτωση κειμένων από ειδησεογραφικές πηγές . Το πλαίσιο αυτό θα ορίζει λειτουργίες για την συλλογή κειμένων, την επεξεργασία αυτών και την εξαγωγή συναισθήματος από το περιεχόμενο. Επιπρόσθετα θα περιλαμβάνει και λειτουργία αποτύπωσης του εξαχθέντος συναισθήματος μέσω μεθόδων οπτικοποίησης.

Δίνεται έμφαση στον σχεδιασμό της αρχιτεκτονικής του πλαισίου ώστε να δημιουργηθεί ένα πλήρως παραμετροποιήσιμο σύστημα. Η αρχιτεκτονική του πλαισίου θα ορίζει ανεξάρτητες λειτουργίες με αυστηρά ορισμένες εισόδους και εξόδους. Η συνδυασμός των λειτουργιών σε ένα γενικότερο processing pipeline θα δημιουργεί ένα ολοκληρωμένο σύστημα επεξεργασίας και αποτύπωσης συναισθήματος. Λόγω της ανεξαρτησίας των επιμέρους λειτουργιών, θα επιτρέπεται την αντικατάσταση υλοποιήσεων οποιασδήποτε λειτουργίας με διαφορετικές υλοποιήσεις, οι οποίες θα είναι σχεδιασμένες σύμφωνα με το πρότυπα που ορίζει το πλαίσιο.

Με το τρόπο αυτό δημιουργείται ένα πλήρως modular σύστημα, το οποίο θα μπορεί να προσαρμόζεται στις ανάγκες του εκάστοτε χρήστη, παρέχοντας δυνατότητες σύγκρισης αποτελεσμάτων βάσει διαφορετικών συνδυασμών λειτουργιών, οι οποίες θα υλοποιούνται με τελείως ανεξάρτητα εργαλεία και προσεγγίσεις. Για παράδειγμα φαντάζει πολύ πιθανή η περίπτωση πολλαπλών διαφορετικών υλοποιήσεων της λειτουργίας εξαγωγής συναισθήματος, και η εναλλαγή αυτών στο σύστημα, με σκοπό την σύγκριση αποτελεσμάτων ώστε να επιλεγεί η υλοποίηση η οποία αποδίδει καλύτερα για τις ανάγκες μίας έρευνας.

1.3 Συνεισφορά

Η παρούσα εργασία συστηματοποιεί το processing pipeline από το στάδιο εξαγωγής κειμένων που περιέχουν απόψεις/συναισθήματα χρηστών, έως το στάδιο συγκεντρωτικής παρουσίασης των αποτελεσμάτων, παρέχοντας ένα ολοκληρωμένο πλαίσιο εξαγωγής συναισθήματος από κείμενα γραμμένα σε φυσική γλώσσα.

Λόγω της modular αρχιτεκτονικής που εισάγει το πλαίσιο, επιτρέπει την αξιοποίησης διαφορετικών εργαλείων για την υλοποίηση κάθε σταδίου. Οι πιθανές περιπτώσεις συνδυασμών υλοποιήσεων θα παρέχουν την δυνατότητα αξιοποίησης open source υλοποιήσεων σε συνδυασμό με άλλες open source υλοποιήσεις για μια διαφορετική λειτουργία, αλλά ενδεχομένως και εμπορικών υλοποιήσεων στις οποίες πιθανόν θα έχει πρόσβαση κάθε δυνητικός χρήστης του πλαισίου.Θα παρέχεται ουσιαστικά η δυνατότητα άμεσης και ταχύτατης εξαγωγής συμπερασμάτων αξιοποιώντας ήδη υπάρχουσες υλοποιήσεις λειτουργιών, δίχως να χρειάζεται η δημιουργία ενός εξολοκλήρου νέου συστήματος, η υλοποίηση του οποίου θα απαιτούσε πολύ περισσότερο χρόνο.

Τέλος, εφόσον δημιουργηθούν μια σειρά υλοποιήσεων για κάθε στάδιο και με βάση τα αποτελέσματα των συνδυασμών αυτών, η modular αρχιτεκτονική του πλαισίου θα

παρέχει κίνητρο για τη δημιουργία βελτιώσεων ορισμένων υλοποιήσεων αλλά και τη δημιουργία εναλλακτικών υλοποιήσεων των λειτουργιών. Και αυτό ώστε να συγκρίνονται τα αποτελέσματα και σταδιακά να παράγονται ακριβέστερα συμπεράσματα όπως αυτά ορίζονται από κάθε έρευνα.

2

Υπόβαθρο

2.1 Θεωρητική προσέγγιση

Το πρόβλημα της κατηγοριοποίησης συναισθήματος αντιμετωπίζεται με τεχνικές Supervised Learning αλλά και Unsupervised Learning.

Όσον αφορά το Supervised Learning, το πρόβλημα εκφυλίζεται στην αντιστοίχιση αντικειμένων σε έναν προκαθορισμένο αριθμό κλάσεων , (θετικό, αρνητικό κτλ) χρησιμοποιώντας αρχικά ένα σύνολο δειγμάτων για την εκπαίδευση του εκάστοτε αλγορίθμου, και στη συνέχεια εφαρμόζοντας αυτόν στο σύνολο δειγμάτων του οποίου επιθυμούμε τη κατηγοριοποίηση.

Οι Pang et al. [2], χρησιμοποιούν αλγορίθμους naive Bayesian και SVM ώστε να κατηγοριοποιήσουν σχόλια χρηστών που αφορούν ταινίες, σε δύο κλάσεις, θετικό και αρνητικό. Στη συνέχεια οι Pang, B., L. Lee [3], χρησιμοποιούν περισσότερα μετρικές και τεχνικές με στόχο τη μεγαλύτερη ακρίβεια κατηγοριοποίησης. Η συχνότητα εμφάνισης όρων είναι μια αποδεδειγμένα αποτελεσματική μετρική η οποία χρησιμοποιείται κατά κόρον στη έρευνα, σε συνδυασμό με την εισαγωγή βαρών σε όρους (TF-IDF). Επιπλέον η αξιοποίηση των επιθέτων (adjectives) των προτάσεων αλλά και λέξεων χαρακτηριζόμενων ως φορέων γνώμης/συναισθήματος όπως 'όμορφο', 'καλό' κ.α στις αντίστοιχες γλώσσες, σε συνδυασμό με τις αντίστοιχες εκφράσεις άρνησης, αποτελεί μια ακόμα κύρια παράμετρο κατά την κατηγοριοποίηση. Επιπρόσθετα, ορισμένες έρευνες εισάγουν και τη έννοια της συντακτικής εξάρτησης λέξεων, με τη χρήση δεντρικών δομών εξάρτησης. Οι Dave et al. [4] προτείνουν την εισαγωγή μιας συνάρτησης βαθμολόγησης (score funtcion), βασισμένη στην εμφάνιση ορισμένων λέξεων σε θετικά και αρνητικά σχόλια χρηστών, ενώ οι Tan et al. [5], χρησιμοποιούν όρους ως φορείς συναισθήματος για τον χαρακτηρισμό μέρους των δεδομένων τους, και αξιοποιούν το αποτέλεσμα για την εκπαίδευση ενός νέου αλγορίθμου. Οι Melville et al. [6] προτείνουν την εισαγωγή λεξικολογικής πληροφορίας σε supervised learning συστήματα ώστε να ενισχύσουν την ακρίβεια των αποτελεσμάτων.

Ένα πρόβλημα το οποίο παρατηρείται σε μεγάλο αριθμό ερευνών, και χρίζει ιδιαίτερης προσοχής, είναι το σημασιολογικό πλαίσιο στο οποίο εκπαιδεύεται κάθε αλγόριθμός. Με μεγάλη πλειοψηφία αλγόριθμοι οι οποίο έχουν εκπαιδευτεί και αποδίδουν καλά σε ένα συγκεκριμένο πλαίσιο Α, τείνουν να αποτυγχάνουν πλήρως εάν εφαρμοστούν σε ένα διαφορετικό πλαίσιο Β, ανεξαρτήτου ύπαρξης κοινών λέξεων, παραδείγματος χάριν όπου Α = πολιτική και Β = αθλητικά. Και αυτό γιατί το εκάστοτε πλαίσιο επηρεάζει άμεσα τον τρόπο με τον οποίο εκφέρεται μια άποψη και τον τρόπο με τον οποίο

χρησιμοποιούνται συγκεκριμένες λέξεις για το σκοπό αυτό. Επομένως το συγκεκριμένο πρόβλημα χρήζει έρευνας ώστε να επιτευχθεί σημασιολογική προσαρμογή των αλγορίθμων Pan et al. [7].

Οι μέθοδοι Unsupervised Learning βασίζονται εξολοκλήρου στο γεγονός ότι ορισμένες λέξεις και φράσεις ορίζουν συγκεκριμένα συναισθήματα. Δεν υπάρχει στάδιο εκπαίδευσης αλγορίθμου και στόχος είναι η παραγωγή αυστηρών σχέσεων οι οποίες είναι ικανές να καθορίσουν τα κατηγοριοποίηση των ζητούμενων πηγών. Ενδεικτικά ο Turney, P.[8] προτείνει έναν αλγόριθμο ο οποίος εξάγει επίθετα και επιρρήματα από τις προτάσεις ενός σχολίου χρήστη, καθώς και τις δύο επόμενες ή προηγούμενες λέξεις ώστε να ορίζεται με μεγαλύτερη ακρίβεια η σημασιολογία ενός επιθέτου , όπως για παράδειγμα σε περιπτώσεις άρνησης. Εξάγεται η σχετική θέση των επιθέτων στο κείμενο και τη πρόταση, και ανιχνεύει συνύπαρξη ορισμένων λέξεων. Τελικά παράγεται μια σχέση η οποία είναι εξαρτημένη από τους παραπάνω παράγοντες, η οποία με τη σειρά της παράγει μια αριθμητική τιμή η οποία αντιστοιχεί σε μια κλίμακα κατηγοριοποίησης.

Ένα ακόμα ερευνητικό ζήτημα της ανάλυση συναισθήματος, είναι ο καθορισμός μιας πρότασης ως υποκειμενική ή αντικειμενική. Η διαδικασία αυτή πρέπει να διενεργείται προτού εκτελεστεί η διαδικασία κατηγοριοποίησης συναισθήματος, και έχει νόημα σε περιπτώσεις κατηγοριοποίησης σε επίπεδο πρότασης και όχι εγγράφου λόγω της πιθανής έλλειψης συνέπειας μεταξύ των προτάσεων ενός κειμένου. Όπως επίσης υφίσταται το ζήτημα της ανίχνευσης προτάσεων οι οποίες δεν περιέχουν συναίσθημα, και θα πρέπει να φιλτράρονται προτού διενεργηθούν οι παραπάνω διαδικασίες ώστε να βελτιώνεται η ταχύτητα και η γενικότερη αποδοτικότητα των συστημάτων κατηγοριοποίησης. Ως εξ' ορισμού πρόβλημα κατηγοριοποίησης, η κατηγοριοποίηση υποκειμενικότητας , μπορεί να αντιμετωπιστεί και με τεχνικές Supervised Learning.

Ενδεικτικά οι Wiebe et al[9], χρησιμοποιούν τον αλγόριθμο Naive bayesian για την κατηγοριοποίηση υποκειμενικότητας, εξετάζοντας και άλλους αλγορίθμους σε μεταγενέστερες έρευνες. Οι Yu, H. και V. Hatzivassiloglou[10] χρησιμοποιούν επίσης supervised learning μεθόδους για την κατηγοριοποίηση υποκειμενικότητας. Οι Kim et al[11], προσθέτουν ένα ακόμα επίπεδο και δημιουργούν μοντέλα ώστε να κατηγοριοποιήσουν προτάσεις και εκφράσεις σε είδη συναισθημάτων.

Η κατηγοριοποίηση σε επίπεδο πρότασης ενέχει σημαντικές δυσκολίες. Καταρχάς μεγάλες σε αριθμό λέξεων και πολύπλοκες συντακτικά προτάσεις, είναι πολύ πιθανό να περιέχουν δύο ή περισσότερα διαφορετικά συναισθήματα, τα οποία μπορεί να είναι και υποκειμενικά. Οι Wilson et al[12], προσπαθούν να αντιμετωπίσουν τα προβλήματα αυτά αξιοποιώντας τα συμφραζόμενα των επιθέτων ή επιρρημάτων ανιχνεύοντας άρνηση και αντίθεση στους αντίστοιχους όρους με τη βοήθεια αντίστοιχων λεξικών.

Εκτός των σχολίων χρηστών έρευνες έχουν γίνει και σε κείμενα τα οποία προέρχονται από συζητήσεις σε forums, blogs και κοινωνικά δίκτυα. Στις όποιες περιπτώσεις οι γνώμες των χρηστών συνοδεύονται και επηρεάζονται από την αλληλεπίδραση και το

διάλογο μεταξύ τους, με αποτέλεσμα πολλές προτάσεις να περιέχουν μεγάλο αριθμό λέξεων και δομών λόγου ικανών να προσδιορίσουν απόψεις και συναισθήματα.

Οι Zhai et al[13], παρουσιάζουν μια μέθοδο εντοπισμού προτάσεων αυτού του τύπου από συζητήσεις σε forums, οι οποίες περιέχουν απόψεις χρηστών για συγκεκριμένα αντικείμενα τα οποία αναφέρονται στις συζητήσεις. Οι Hassan et al[14], προτείνουν μια μέθοδο εντοπισμού προτάσεων οι οποίες περιέχουν απόψεις χρηστών που αφορούν άλλους χρήστες της εκάστοτε συζήτησης, προσπαθώντας στη συνέχεια να προβλέψει εάν ορισμένες προτάσεις περιέχουν μια άποψη για ένα αντικείμενο της συζήτησης.

2.2 Στόχος

Η πληθώρα ειδησεογραφικών πηγών στο διαδίκτυο και η ανάγκη της αγοράς για ανάλυση ειδήσεων σε μεγάλη κλίμακα, έχει καταστήσει αναγκαία την ύπαρξη μηχανισμών που θα συγκεντρώνουν πληροφορία από ειδησεογραφικές πηγές ελαχιστοποιώντας την ανθρώπινη παρέμβαση. Ταυτόχρονα καθιστά πλέον επιτακτική την ολοκλήρωση επιμέρους εργαλείων ανάλυσης και επεξεργασίας κειμένων σε μια ενιαία πλατφόρμα που θα στοχεύει στον τελικό χρήστη αποκρύβοντας τις επιμέρους λεπτομέρειες.

Οι ενδιαφέρουσες ιδιαιτερότητες της περίπτωσης των [[ειδησεογραφικών άρθρων]] είναι η έλλειψη της έννοια του opinion spam. Ένα άρθρο περιέχει ρητά μια είδηση και την άποψη του αρθρογράφου, δίχως να υπάρχει νόημα περαιτέρω αναπαραγωγής των περιεχομένων του άρθρου και των απόψεων που περιέχει, αποκλείοντας περιπτώσεις αυτοματοποιημένου spam όπως αυτές που συναντώνται σε forums και blogs forums. Μια ακόμα ιδιαιτερότητα είναι η μορφή του λόγου στα ειδησεογραφικά άρθρα, η οποία τείνει να μην περιέχει σε μεγάλο βαθμό λέξεις που χαρακτηρίζονται ως φορείς συναισθήματος, γεγονός που δεν ισχύει σε σχόλια και κριτικές χρηστών, όπου η συχνότητα εμφάνισης λέξεων φορέων είναι σημαντικά μεγαλύτερη. Επομένως παρουσιάζει τελικά ενδιαφέρον η σύγκριση των αποτελεσμάτων τα οποία θα προκύψουν, σε σχέση με πιθανά αποτελέσματα ανάλυσης απόψεων από πηγές όπως forums και κοινωνικά δίκτυα, τα οποία θα αφορούν κοινά αντικείμενα, και ενδεχομένως σε κοινό χρονικό ορίζοντα.

Η υπάρχουσα βιβλιογραφία δεν περιέχει αναφορές και υλοποιήσεις οι οποίες αφορούν την εξαγωγή και ανάλυση συναισθήματος από [[ειδησεογραφικές πηγές.]] Σε αυτά τα πλαίσια, προτείνεται μια έρευνα με στόχο να σχεδιαστούν και να υλοποιηθούν τεχνικές που θα εντοπίζουν περιεχόμενο ενδιαφέροντος σε ειδησεογραφικά site και θα δημιουργούν με αυτόματο τρόπο μηχανισμούς εξαγωγής περιεχομένου. Στη συνέχεια θα οργανώνουν και αναλύουν τη συγκεντρωμένη πληροφορία και θα οπτικοποιούν την παραγόμενη γνώση.

2.3 Σχετικές πλατφόρμες

Το LingPipe είναι μια πλατφόρμα η οποία παρέχει τεχνικές κατηγοριοποίησης κειμένων, αρκετές γλώσσες οι οποίες υλοποιούν ξεχωριστά μοντέλα. Επίσης παρέχει λειτουργίες εντοπισμού λεξικολογικών εννοιών όπως και τεχνικές ανάλυσης συναισθήματος από κείμενα σε φυσική γλώσσα. Παρέχει ένα μεγάλο υποσύνολο των λειτουργιών του δωρεάν, ενώ Επιπρόσθετες λειτουργίες παρέχονται με εμπορικούς σκοπούς. Το LingPipe χρησιμοποιείται στα πλαίσια στης παρούσας εργασίας, για την υλοποίηση ενός demo με βάση τις λειτουργίες κατηγοριοποίησης και ανάλυσης συναισθήματος που ορίζονται από το γενικότερο πλαίσιο.

Το OpenNLP είναι μια βιβλιοθήκη Java που βασίζεται σε τεχνικές machine learning για την ανάλυση κείμενο σε φυσική γλώσσα. Παρέχει επίσης λειτουργίες όπως εντοπισμός προτάσεων, tokenization, εντοπισμό ονομάτων όπως και εντοπισμός συντακτικών εννοιών σε προτάσεις(υποκείμενο,ρήμα, αντικείμενο).

Το R Project αποτελεί μια open source πλατφόρμα στατιστικής ανάλυσης και παραγωγής γραφημάτων. Η επέκταση του είναι το TM package (test mining), η οποία προσφέρει ένα framework λειτουργιών text mining με βάση το R. Προσφέρει επεκτάσιμες λειτουργίες για τη διαχείριση εγγράφων σε μια πλειάδα formats, όπως και για την ενσωμάτωση ανεξάρτητων υλοποιήσεων αλγορίθμων. Προσεγγίζει τη λογική σχεδιασμού του framework που παρουσιάζεται στη παρούσα εργασία, ωστόσο περιορίζεται σε επίπεδο text mining.

Το NLTK 3.0 αποτελεί μια open source πλατφόρμα για την ανάπτυξη εφαρμογών σε Python, με σκοπό την ανάλυση κειμένων γραμμένων σε φυσική γλώσσα, παρέχοντας διεπαφές για τη χρήση πολλών λεξικολογικών πηγών όπως το WordNet, όπως και ένα σύνολο βιβλιοθηκών οι οποίες υποστηρίζουν τεχνικές ανάλυσης κειμένου, όπως classification, tagging, tokenization κ.α.

Το Gate είναι ένα project το οποίο περιλαμβάνει ένα περιβάλλον ανάπτυξης εφαρμογών. Παρέχει επίσης ένα web application, και υποστηρίζει ένα γενικότερο framework ώστε ο χρήστης να έχει πρόσβαση στα παραπάνω με σκοπό την αξιοποίηση των λειτουργιών ανάλυσης φυσικής γλώσσας της οποίες παρέχει. Το project αναπτύσσεται εδώ και 15 χρόνια και παρέχει μια πλειάδα components λειτουργιών ανάλυσης φυσικής γλώσσας, διαχείρισης εγγράφων κ.α. Η πλειονότητα των λειτουργιών που παρέχονται είναι δωρεάν, με ορισμένες να χρησιμοποιούνται και μέσω εμπορικών συμφωνιών.

Το Stanford NLP Group παρέχει open source components για την επίλυση κοινών προβλημάτων ανάλυσης φυσικής γλώσσας. Components η πλειοψηφία των οποίων μπορεί να αξιοποιηθούν από εφαρμογές γραμμένες σε γλώσσες όπως οι Python, Ruby, Perl, Javascript κ.α.

2.4 Προτεινόμενη προσέγγιση

2.4.1 Συλλογή

Αρχικά προτείνεται η δημιουργία ενός μηχανισμού αυτόματης συλλογής περιεχομένου (Web Crawler). Η πλειονότητα των διαδικτυακών ειδησεογραφικών πηγών υπακούει σε έναν συγκεκριμένο τύπο δομής. Η δομή αυτή αποτελείτε κατά κύριο λόγο από σελίδες κατηγοριών και σελίδες άρθρων. Οι σελίδες κατηγοριών περιέχουν υπερσυνδέσμους (hyperlinks) προς σελίδες άρθρων, οι οποίοι υπερσύνδεσμοι περιέχουν συνήθως τον τίτλο για το άρθρο στο οποίο οδηγούν. Οι σελίδες άρθρων είναι αυτές που περιέχουν την επιθυμητή πληροφορία, δηλαδή τον τίτλο μιας είδησης, το κυρίως κείμενο ή σώμα της είδησης, και πιθανών μια σχετική εικόνα αλλά και σχόλια χρηστών που αφορούν το περιεχόμενο της είδησης. Ο Web Crawler θα είναι ικανός να επισκέπτεται ειδησεογραφικά sites ανά τακτά χρονικά διαστήματα να εντοπίζει τις σελίδες κατηγοριών συλλέγοντας τα links προς σελίδες άρθρων. Στη συνέχεια να επισκέπτεται μεμονωμένα κάθε σελίδα άρθρου, και να συλλέγει το σχετικό περιεχόμενο. Τα ειδησεογραφικά sites προς επίσκεψη, θα δίνονται ως είσοδος στον Crawler με τη μορφή λίστας υπερσυνδέσμων προς τις κεντρικές σελίδες αυτών. Ο μηχανισμός θα πρέπει να είναι ικανός να εντοπίζει αυτόματα τους υπεσυνδέσμους προς σελίδες κατηγοριών οι οποίοι εμφανίζονται στην κεντρική σελίδα. Ο εντοπισμός αυτών από ένα κοινό χρήστη είναι εύκολος λόγο του ότι οι σύνδεσμοι αυτοί φέρουν χαρακτηριστικούς τίτλους κατηγοριών με νοηματικό περιεχόμενο, όπως για παράδειγμα 'Πολιτική', και 'Αθλητικά'. Ο Web Crawler θα αξιοποιήσει το γεγονός ότι οι σύνδεσμοι αυτοί βρίσκονται σχεδόν σε όλες τις περιπτώσεις σε μορφές μενού (HTML element), ουσιαστικά έχοντας έναν κοινό πρόγονο σε μία δεντρική αναπαράσταση των στοιχείων της κεντρικής σελίδας. Άρα ο μηχανισμός θα είναι ανεξάρτητος από νοηματικούς παράγοντες και επομένων ανθρώπινης παρέμβασης.

Ο σχεδιασμός του Web Crawler ώστε να συλλέγει αποτελεσματικά περιεχόμενο αποτελεί δύσκολο εγχείρημα για δύο κυρίως λόγους: 1) Η μορφή των ειδησεογραφικών sites αλλάζει συνεχώς (δομή HTML), κάτι το οποίο σημαίνει ότι ο Web Crawler πρέπει να παραμετροποιηθεί κατάλληλα ώστε να λαμβάνει υπόψιν τυχόν μεταβολές. 2) Ορισμένα sites έχουν αρκετά πολύπλοκη δομή, η οποία ενδεχομένως να οδηγήσει τον Web Crawler είτε στο να συλλέξει επιπλέον αχρείαστη πληροφορία μη σχετική με ειδήσεις, είτε να συλλέξει μέρος της είδησης. Άρα πρέπει να ληφθεί επιπλέον μέριμνα για τέτοιου είδους περιπτώσεις, εισάγοντας μηχανισμούς ικανούς να συσχετίζουν κομμάτια πληροφορίας σχετικά με τον τίτλο της εκάστοτε είδησης.

2.4.2 Ανάλυση

Εφόσον πραγματοποιηθεί η συλλογή της πληροφορίας, είναι απαραίτητη η δημιουργία αλγορίθμων με σκοπό τον εντοπισμό συναισθήματος/απόψεων μέσα στο σώμα του κάθε ειδησεογραφικού άρθρου. Για τον σκοπό αυτό προτείνεται αρχικά μια μέθοδος υπολογισμού συχνότητας εμφάνισης όρων (TF-IDF), και με τη βοήθεια ενός λεξικού αντιπροσωπευτικών λέξεων φορέων συναισθήματος, να υπολογίζεται η πόλωση συναισθήματος (Θετικό, Αρνητικό, Ουδέτερο), σε επίπεδο εγγράφου αλλά και σε επίπεδο πρότασης. Σημειώνεται ότι το 'Ουδέτερο' αντιστοιχεί στην έλλειψη συναισθήματος, και όχι στην ενδεχόμενη ουδετερότητα η οποία θα μπορούσε να προκύψει από μια πιθανή συνύπαρξη θετικών και αρνητικών απόψεων/συναισθημάτων. Η χρήση αλγορίθμων κατηγοριοποίησης είναι μια επιπλέον επιλογή. Οι μέθοδοι unsupervise learning ενδείκνυνται σε αυτή τη περίπτωση λόγο της δεδομένης έλλειψης αρχικού συνόλου δεδομένων εκπαίδευσης. Γεγονός το οποίο συνηγορεί και στον αρχικό στόχο ενός πλήρους αυτοματοποιημένου συστήματος, καθώς η δημιουργία ενός συνόλου εκπαίδευσης θα απαιτούσε ανθρώπινη παρέμβαση ώστε να εγγυηθεί η αξιοπιστία του από πλευράς ορθών συναισθημάτων. Σε κάθε περίπτωση θα πρέπει να ληφθεί μέριμνα για τον εντοπισμό αρνήσεων, με κατάλληλες τεχνικές όπως αυτές περιγράφονται στη βιβλιογραφία καθώς έχουν εμφανώς ικανοποιητικά αποτελέσματα σε αυτόν το τομέα.

Η ανάλυση συναισθήματος σε επίπεδο πρότασης και σε επίπεδο εγγράφου παρουσιάζουν εξίσου ενδιαφέρον. Κάθε άρθρο αφορά ένα γενικό αντικείμενο το οποίο αφορούν οι περιεχόμενες απόψεις/συναισθήματα, τα οποία εξετάζονται σε επίπεδο εγγράφου και βάσει όλων των προτάσεων που περιέχει το άρθρο. Σε επίπεδο πρότασης, είναι δυνατός ο εντοπισμός πόλωσης σε πολύ συγκεκριμένες οντότητες/αντικείμενα άποψης, οι οποίες δεν ταυτίζονται απαραίτητα με τη κύρια οντότητα, είναι όμως πολύ πιθανά άμεσα συσχετιζόμενες με αυτή.

Κάθε προσέγγιση θα παράγει διαφορετικά αποτελέσματα των οποίων η μετέπειτα σύγκριση παρουσιάζει ιδιαίτερο ενδιαφέρον, καταρχάς ως προς τη συνέπεια του συναισθήματος το οποίο θα ανιχνεύεται και τους πιθανούς λόγους για τους οποίους οι έξοδοι των δύο μεθόδων ταυτίζονται ή όχι.

2.4.3 Οπτικοποίηση

Τέλος, θα υλοποιηθεί ένας μηχανισμός παρουσίασης των αποτελεσμάτων. Οι αριθμοί, τα μεγέθη τα οποία θα προκύπτουν και η πληροφορία η οποία αυτά συνεπάγονται θα έχει μεγαλύτερο νόημα να παρουσιαστεί με τη μορφή γραφημάτων και διαγραμμάτων, ώστε να είναι ευκολότερη η σύγκριση μεταξύ διαφόρων αποτελεσμάτων. Όπως επίσης θα δίνεται η δυνατότητα να προβληθούν και να εντοπιστούν συμπεράσματα τα οποία θα ήταν πιθανών λιγότερο εμφανή με τη μορφή αριθμητικών τιμών, γεγονός το οποίο είναι ο εξ ορισμού στόχος της οπτικοποίησης. Τα αποτελέσματα θα μπορούν να παρουσιάζονται συναρτήσει διαφόρων παραγόντων όπως της πόλωσης των απόψεων οι οποίες αφορούν ένα αντικείμενο συναρτήσει των διαφόρων πηγών άρθρων στα

οποία αναφέρεται το αντικείμενο, ή χρονική εξέλιξη των απόψεων για το αντικείμενο, λαμβάνοντας υπόψιν τις ημερομηνίες ανάρτησης των σχετικών άρθρων. Μια ακόμα επιλογή θα αποτελεί η συσταδοποίηση (clustering) των άρθρων ανά κατηγορία και η συγκεντρωτική παρουσίαση της πόλωσης των απόψεων που αφορούν ένα ή περισσότερα άρθρα, είτε ανά πηγή είτε γενικότερα στο σύνολο των ειδήσεων.

3

Σχεδίαση

3.1 Ορισμός Πλαισίου

Προτείνεται ο σχεδιασμός ενός πλαισίου το οποίο θα παρέχει στον εκάστοτε χρήστη την δυνατότητα της εξαγωγής δεδομένων από πηγές στον διαδίκτυο και της αποθήκευσης και διαχείρισης αυτών. Θα παρέχεται η δυνατότητα εφαρμογής αλγορίθμων εξόρυξης γνώσης πάνω στα δεδομένα αυτά με σκοπό την εξαγωγή συναισθήματος και τελικώς η οπτικοποίηση των όποιων αποτελεσμάτων τα οποία θα οδηγούν στην πιθανή εξαγωγή συμπερασμάτων.

Προτείνεται αρχιτεκτονική η οποία θα επιτρέπει την προσάρτηση ανεξάρτητων μορφωμάτων κώδικα (modules), ανάλογα με της ανάγκες και προτιμήσεις του χρήστη. Κάθε module θα είναι υπεύθυνο για μία μεμονωμένη λειτουργία του συστήματος, αντλώντας από αυτό, και αντίστοιχα παρέχοντας σε αυτό κατάλληλα επεξεργασμένες πληροφορίες. Οι πληροφορίες αυτές θα προσαρμόζονται στη γενικότερη ροή δεδομένων του συστήματος ώστε να επεξεργαστούν από τα υπόλοιπα modules και να παραχθεί το τελικό αποτέλεσμα. Ο αυστηρός ορισμός της μορφής των δεδομένων εισόδου και εξόδου του κάθε module από το σύστημα, θα καθιστά το σύστημα ως σύνολο πλήρως τροποποιήσιμο και ευέλικτο.

Κάθε δυνητικός χρήστης θα έχει τη δυνατότητα υλοποίησης νέων modules, τα οποία ενδεχομένως να καλύπτουν καλύτερα τις ανάγκες του, και να τα προσαρτήσει στο σύστημα. Η πιθανή αντικατάσταση ενός ή περισσότερων modules θα είναι πλήρως θεμιτή, δίχως να διαταράσσεται η ροή δεδομένων και η ακεραιότητα του τελικού αποτελέσματος.

Μια τέτοιου είδους αρχιτεκτονική θα επιτρέπει την χρήση ποικίλων μεθόδων, τεχνικών, ενδεχομένως και τεχνολογιών αλλά και συνδυασμών αυτών, οι οποίες εφόσον προσαρμοστούν κατάλληλα στο προτεινόμενο πλαίσιο, θα επιτρέπουν την εξαγωγή ποικίλων συμπερασμάτων και συγκριτικής μελέτης αυτών. Παρέχοντας με τον τρόπο αυτό ένα πλήρως προσαρμόσιμο περιβάλλον εξαγωγής, διαχείρισης, επεξεργασίας και παρουσίασης αποτελεσμάτων που αφορούν την γενικότερη διαδικασία εξόρυξης συναισθήματος.

3.2 Υλοποίηση Πλαισίου

Για τις ανάγκες υλοποίησης του εν λόγω πλαισίου θα χρησιμοποιηθούν Java Interfaces. Τα Interfaces θα ορίζουν έναν τρόπο ενσωμάτωσης των modules στο σύστημα. Κάθε Interface ορίζει μεθόδους οι οποίες αντιστοιχούν στις βασικές λειτουργίες τις οποίες θα πρέπει να περιέχει το εκάστοτε module ώστε να μπορεί να προσαρμοστεί στο σύστημα.

Ορίζονται Java Classes (αντικείμενα) τα οποία αντιστοιχούν στην αναπαράσταση των δεδομένων στο σύστημα. Τα εν λόγω αντικείμενα αντιπροσωπεύουν έναν κοινό τρόπο αναφοράς στα δεδομένα για όλα τα modules. Οι λειτουργίες κάθε module θα αναφέρονται σε, είτε θα δημιουργούν τα αντίστοιχα αντικείμενα ώστε να επιτελέσουν την γενικότερη λειτουργία του module και να τροφοδοτήσουν το σύστημα με την κατάλληλη πληροφορία σε μια αναγνωρίσιμη από αυτό μορφή.

Η δημιουργία ενός module αντιστοιχεί στην υλοποίηση των λειτουργιών του αντίστοιχου Interface σε μία κλάση Java η οποία και θα αποτελεί το τελικό module. Κάθε υλοποίηση θα πρέπει να μεριμνά για τον μετασχηματισμό των δεδομένων σε αντικείμενα τα οποία αναγνωρίζει το σύστημα, όπως και τα υπόλοιπα modules τα οποία υλοποιούνται, χωρίς να υπάρχουν αλλαγές ή απώλεια πληροφορίας.

3.3 Ορισμός Δομικών Modules

3.3.1 Web Crawler

Το module το οποίο θα είναι υπεύθυνο για την συλλογή των πληροφοριών προς επεξεργασία θα αποτελείται από ένα Web Crawler. Ο Web Crawler θα συλλέγει τις πληροφορίες από ένα σύνολο προκαθορισμένων πηγών στο διαδίκτυο, και θα τις αποθηκεύει σε αρχεία με συγκεκριμένη δομή ανεξάρτητα από το περιεχόμενο σε κάθε περίπτωση. Ουσιαστικά θα δημιουργούνται δομημένα έγγραφα τα οποία τηρούν ορισμένους μορφολογικούς κανόνες.

Κάθε πηγή στο διαδίκτυο, η οποία αφορά αποτύπωση γνώμης με μορφή γραπτού κειμένου, φέρει κοινά χαρακτηριστικά στην παρουσίαση του κειμένου. Σε κάθε περίπτωση το κύριο σώμα κειμένου το οποίο περιέχει την γνώμη ενός χρήστη, συνοδεύεται από ένα τίτλο οποίος αφορά το περιεχόμενο του και μία ημερομηνία καταχώρησης του κειμένου στη διαδικτυακή πηγή από το χρήστη. Σε ορισμένες περιπτώσεις το κείμενο συνοδεύεται και από πολυμέσα όπως εικόνες και βίντεο, τα οποία αφορούν το περιεχόμενο και πιθανόν να εκφράζουν και την γνώμη που περιέχεται στο κείμενο. Ο Web Crawler πρέπει να είναι σε θέση να συλλέγει και τα αντίστοιχα πολυμέσα, εάν το απαιτεί η εκάστοτε υλοποίηση. Η συλλογή αυτών θα αντιστοιχεί στη συλλογή των υπερσυνδέσμων που τα αφορούν.

Σε κάθε περίπτωση η είσοδος του εν λόγο module πρέπει να αποτελείται από ένα σύνολο οδηγιών προς τη συλλογή των επιθυμητών πληροφοριών από ένα σύνολο

διαδικτυακών πηγών, και έξοδος ένα σύνολο αρχείων τα οποία θα περιέχουν την πληροφορία σε μια εύκολη προς επεξεργασία από το σύστημα μορφή.

3.3.2 Files

Στάδιο στο οποίο περιέχονται τα κείμενα με τη μορφή αρχείων, όπως αυτά προκύπτουν από την έξοδο του web cralwer, είτε ως ένα επεξεργασμένο dataset.

3.3.3 Repository

Module το οποίο είναι υπεύθυνο για την επεξεργασία του συνόλου των αρχείων τα οποία περιέχουν τα κείμενα με τις απόψεις χρηστών, την αποθήκευσή τους σε ένα repository, και την κατά απαίτηση ανάκτηση του συνόλου, ή υποσυνόλου αυτών, προς περαιτέρω επεξεργασία.

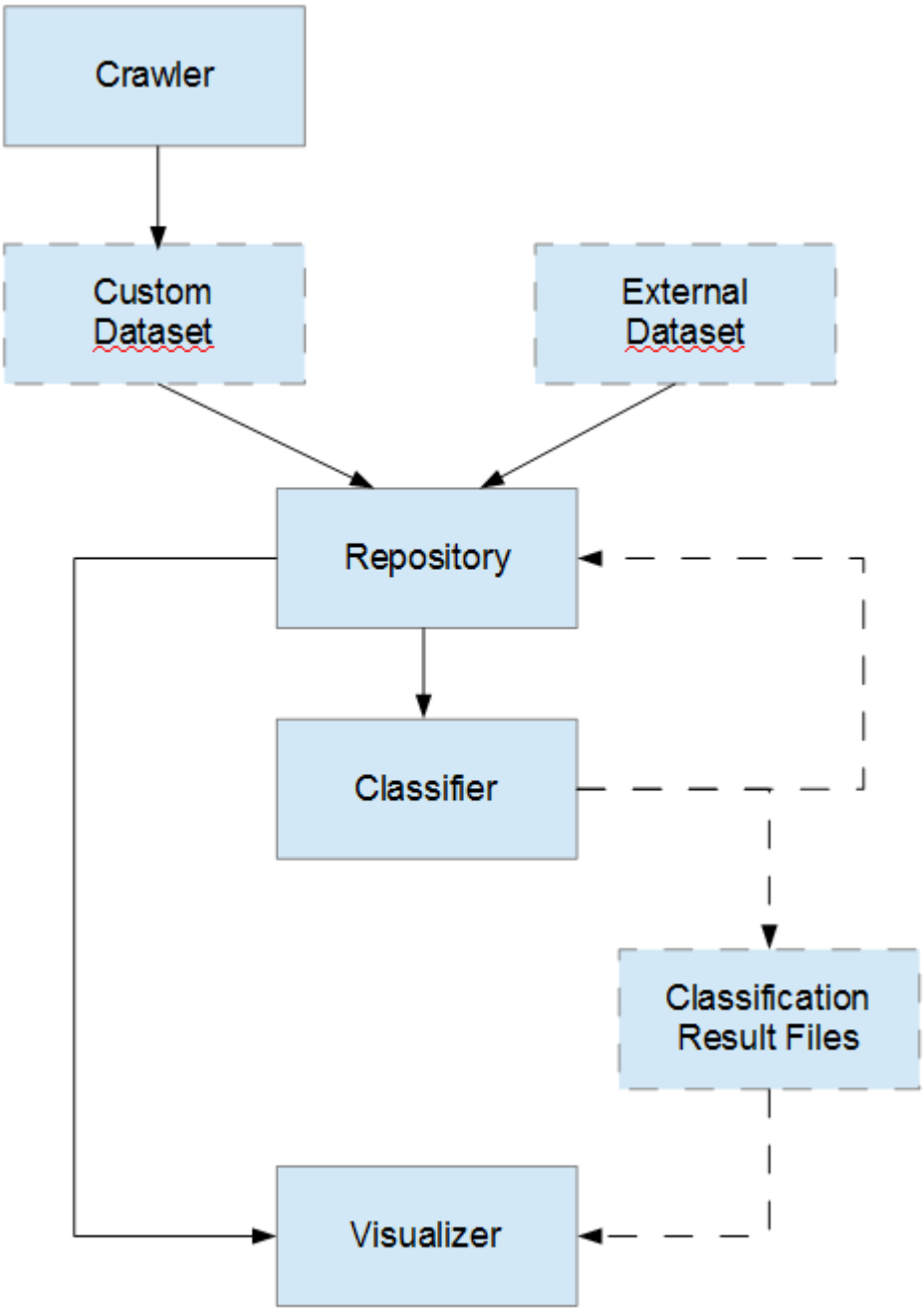
Η αποθήκευση του περιεχομένου των αρχείων σε repository απαιτεί τον μετασχηματισμό του περιεχομένου σε μορφή αναγνωρίσιμη από την αρχιτεκτονική του συστήματος. Στο σημείο αυτό απαιτείται ορισμός της δομής αυτής, η οποία πρέπει να είναι αρκετα γενικευμένη ώστε να καλύπτει όλες τις πιθανές περιπτώσεις συνδυασμών εισερχόμενης πληροφορίας, όπως για παράδειγμα την ύπαρξη ή μη πληροφορίας που αφορά πολυμέσα ή ημερομηνία καταχώρησης. Εφόσον ολοκληρωθεί ο μετασχηματισμός, η πληροφορία μπορεί πλέον να αποθηκευτεί στο επιλεγμένο repository, ώστε να γίνει διαθέσιμη προς επεξεργασία από τα αντίστοιχα modules του συστήματος. Η ανάκτηση πληροφορίας από το repository πρέπει να Υλοποιείται ως μια παραμετροποιημένη διαδικασία ώστε να επιτρέπει την ενδεχόμενη επιλογή υποσυνόλων αυτών βάσει κριτηρίων όπως για παράδειγμα την επιλογή κειμένων που περιέχουν συγκεκριμένες λέξεις ή εκφράσεις ή έχουν κατοχυρωθεί σε ένα συγκεκριμένο χρονικό φάσμα.

3.3.4 Classifier

Η εκτίμηση συναισθήματος από τα κείμενα τα οποία περιέχονται στο repository θα γίνεται από ένα module κατηγοριοποίησης. Το module αυτό θα αξιοποιεί την εκάστοτε υλοποίηση ενός ή περισσότερων τεχνικών κατηγοριοποίησης κειμένων σε φυσική γλώσσα. Η είσοδός του module θα είναι ένα σύνολο κειμένων, με την κοινά αναγνωρίσιμη προς το σύστημα μορφή η οποία έχει οριστεί, το οποίο έχει προκύψει από το γενικότερο σύνολο των κειμένων τα οποία έχουν καταχωρηθεί στο repository. Η έξοδος θα αποτελεί ένα αρχείο με προκαθορισμένη μορφή το οποίο θα φέρει την πληροφορία που προκύπτει από την κατηγοριοποίηση και αφορά την εκτίμηση συναισθήματος, για το περιεχόμενο των κειμένων.

3.3.5 Visualizer

Το module οπτικοποίησης, θα αναπαριστά τα αποτελέσματα που έχουν προκύψει από την έξοδο του module κατηγοριοποίησης, σε ορισμένες γραφικές αναπαραστάσεις. Θα επεξεργάζεται την πληροφορία των αρχείων που περιέχουν την εκτίμηση συναισθήματος για το εκάστοτε σύνολο κειμένων που αφορά γνώμες χρηστών, και θα την αποτυπώνει σε γραφήματα, όπως για παράδειγμα χρονικό ορίζοντα εάν παρέχεται ημερομηνία καταχώρησης σχολίων, είτε σε αναπαραστάσεις συχνότητας εμφάνισης όρων. Με το τρόπο αυτό θα επιτελείτε και ο βασικός στόχος της έννοια της οπτικοποίησης, δηλαδή η εξαγωγή συμπερασμάτων τα οποία δεν μπορούν να εξαχθούν άμεσα μέσω υπολογιστικής ανάλυσης, κυρίως σε περιπτώσεις μεγάλου όγκου πληροφορίας.



Εικόνα 1. Αλληλεπίδραση Λειτουργιών

4

Υλοποίηση

4.1 Δομικές Κλάσεις

4.1.1 SentimentFramework.Basics

Ορίζονται κλάσεις οι οποίες αντιπροσωπεύουν περιορισμούς ως προς την ανάκτηση εγγράφων από το Repository.

- Filter

Κλάση η οποία αντιπροσωπεύει έναν περιορισμό και αποτελεί την βάση για τον ορισμό σύνθετων περιορισμών

- DateFilter

Κλάση η οποία αποτελεί επέκταση της Filter και αντιπροσωπεύει έναν περιορισμό που αντιστοιχεί σε ένα χρονικό εύρος. Ορίζεται από δύο ημερομηνίες οι οποίες αντιπροσωπεύουν την αρχή και το τέλος ενός του χρονικού διαστήματος στο οποίο αντιστοιχεί ένα σύνολο εγγράφων.

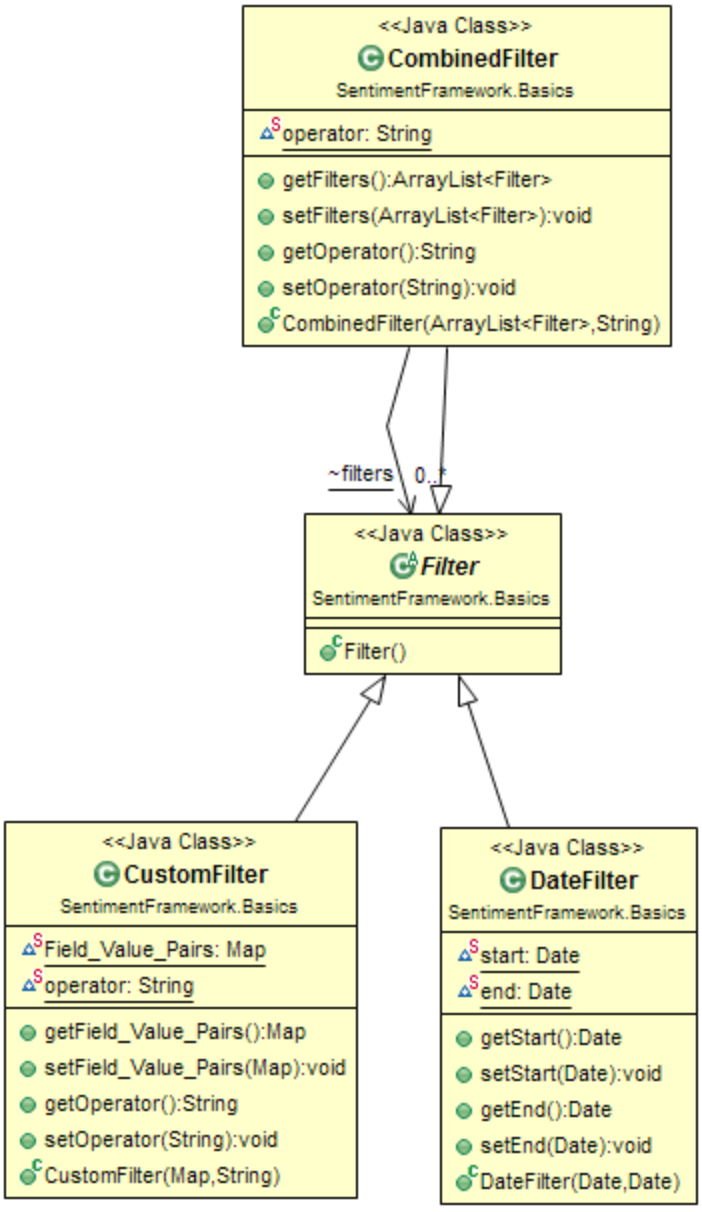
- CustomFilter

  Κλάση η οποία αποτελεί επέκταση της Filter και αντιπροσωπεύει ένα σύνολο λεκτικών περιορισμών. Αποτελείται από μια δομή τύπου key-value pair, όπου ως κλειδί ορίζεται το πεδίο αναζήτησης και ως τιμή ορίζεται ο λεκτικός περιορισμός. Η δομή αυτή συνοδεύεται από μια παράμετρο τελεστή ο οποίος αντιστοιχεί σε εισόδους τύπου AND/OR, ώστε να δίνεται η δυνατότητα δημιουργίας περισσότερο σύνθετων συνθηκών αναζήτησης στο Repository.

- CombinedFilter

  Κλάση η οποία αποτελεί επέκταση της Filter και αντιπροσωπεύει ένα συνδυασμό περιορισμών τύπου Filter ώστε να επιτρέπεται η δημιουργία συνθηκών που συνδυάζουν λεκτικούς περιορισμούς με περιορισμούς χρονικού

εύρους. Ορίζεται ως ένα σύνολο αντικειμένων τύπου Filter, και μια παράμετρος τελεστή όπως αυτή περιγράφεται στη περίπτωση της κλάσης CustomFilter.



Εικόνα 2. Διάγραμμα Κλάσεων- SentimentFramework.Basics

4.1.2 SentimentFramework.Basics.text

Ορίζονται κλάσεις οι οποίες αντιπροσωπεύουν λεκτικές δομές τις οποίες αναγνωρίζει και χειρίζεται το σύστημα. Ουσιαστικά αποτελούν την αναπαράσταση των κειμένων του dataset και υποδομών αυτών στη μνήμη.

- TextualElement

  Διεπαφή η οποία αντιπροσωπεύει ένα λεκτικό αντικείμενο και αποτελεί τη βάση για τον ορισμό των υπολοίπων δομών αναπαράσταση κειμένου.

- Sentence

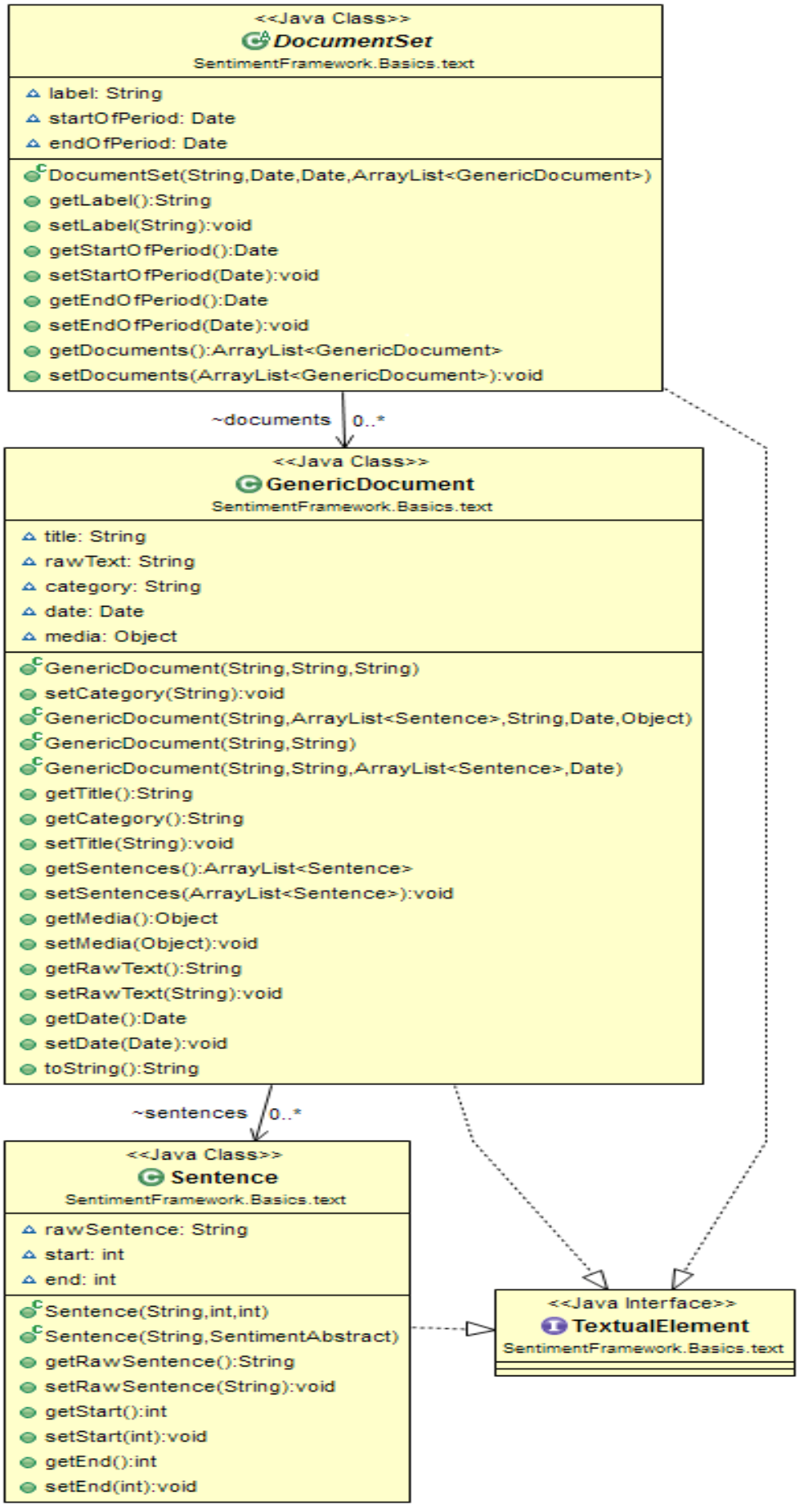
  Κλάση η οποία αντιπροσωπεύει μια δομή πρότασης. Αποτελείται από μια μεταβλητή τύπου String η οποία αντιπροσωπεύει το περιεχόμενο μιας πρότασης, και από δυο ακέραιες μεταβλητές οι οποίες αντιπροσωπεύουν τη θέση της πρότασης μέσα στην ευρύτερη δομή κειμένου στην οποία περιέχεται.

- GenericDocument

  Κλάση η οποία αντιπροσωπεύει μια δομή άρθρου κειμένου ή γενικότερα ενός εγγράφου. Αποτελείται από μεταβλητές τύπου String οι οποίες αντιστοιχούν στον τίτλο ενός εγγράφου, στο σώμα κειμένου του εγγράφου, στην ημερομηνία υποβολής του εγγράφου σε έναν ιστότοπο ή στη ημερομηνία συλλογής αυτού από τον Crawler και από ένα γενικότερο αντικείμενο το οποίο αντιπροσωπεύει πληροφορία η οποία αφορά πολυμέσα που ενδεχομένως να συνοδεύουν το έγγραφο. Περιέχονται επίσης επιπλέον μεταβλητές οι οποίες αντιστοιχούν στην αναπαράσταση του κύριος κειμένου με μορφή ενός συνόλου αντικειμένων τύπου Sentence, και στο όνομα κατηγορίας στην οποία αντιστοιχεί ένα έγγραφο. Κατηγορίες οι οποίες είτε αφορούν την άποψη ενός χρήστη (θετικό, αρνητικό) είτε για παράδειγμα τις κατηγορίες όπως αυτές παρουσιάζονται σε ειδησεογραφικούς ιστοτόπους (πολιτική, πολιτισμός κ.α).

- DocumentSet

Κλάση η οποία αντιπροσωπεύει ένα σύνολο αντικειμένων τύπου GenericDocument. Αποτελείται από μια δομή λίστας τύπου GenericDocument, από μια ετικέτα η οποία χαρακτηρίζει το σύνολο των αντικειμένων στη λίστα, όπως επίσης από δυο αντικείμενα τύπου ημερομηνίας τα οποία αντιπροσωπεύουν την αρχή και το τέλος μιας περιόδου στην οποία αντιστοιχούν χρονικά τα έγγραφα που περιέχονται στη δομή.


<<Java Class>>
DocumentSet
SentimentFramework.Basics.text
label: String
startOfPeriod: Date
endOfPeriod: Date
DocumentSet(String,Date,Date,ArrayList<GenericDocument>)
getLabel():String
setLabel(String):void
getStartOfPeriod():Date
setStartOfPeriod(Date):void
getEndOfPeriod():Date
setEndOfPeriod(Date):void
getDocuments():ArrayList<GenericDocument>
setDocuments(ArrayList<GenericDocument>):void
~documents
0..*
<<Java Class>>
GenericDocument
SentimentFramework.Basics.text
title: String
rawText: String
category: String
date: Date
media: Object
GenericDocument(String,String,String)
setCategory(String):void
GenericDocument(String,ArrayList<Sentence>,String,Date,Object)
GenericDocument(String,String)
GenericDocument(String,String,ArrayList<Sentence>,Date)
getTitle():String
getCategory():String
setTitle(String):void
getSentences():ArrayList<Sentence>
setSentences(ArrayList<Sentence>):void
getMedia():Object
setMedia(Object):void
getRawText():String
setRawText(String):void
getDate():Date
setDate(Date):void
toString():String
~sentences
0..*
<<Java Class>>
Sentence
SentimentFramework.Basics.text
rawSentence: String
start: int
end: int
Sentence(String,int,int)
Sentence(String,SentimentAbstract)
getRawSentence():String
setRawSentence(String):void
getStart():int
setStart(int):void
getEnd():int
setEnd(int):void
<<Java Interface>>
TextualElement
SentimentFramework.Basics.text

Εικόνα 3. Διάγραμμα Κλάσεων- SentimentFramework.Basics.text

4.1.3 SentimentFramework.Basics.sentiment

Ορίζονται κλάσεις οι οποίες αντιπροσωπεύουν τις δομές οι οποίες αντιστοιχούν σε διαφορετικούς τρόπους αναπαράστασης του συναισθήματος το οποίο εξάγεται από τα κείμενα των χρηστών.

- SentimentAbstract
  Διεπαφή η οποία αποτελεί τη βάση ορισμού περισσότερο σύνθετων δομών αναπαράστασης συναισθήματος.

- PolaritySentiment
  Κλάση η οποία αντιπροσωπεύει δομή συναισθήματος η οποία αντιστοιχεί σε θετικές και αρνητικές απόψεις χρηστών.

Ορίζονται οι λειτουργίες:

1. PolaritySentimentDir

Δέχεται ως είσοδο ένα αντικείμενο τύπου DocumentSet, και δημιουργεί φακέλους οι οποίοι περιέχουν κείμενα χρηστών κατηγοριοποιημένα βάσει περιεχομένου σε θετικά και αρνητικά.

2. PolaritySentimentMap
   Δέχεται ως είσοδο ένα αντικείμενο τύπου DocumentSet, και δημιουργεί ένα αντικείμενο τύπου HashMap, το οποίο περιέχει κείμενα χρηστών κατηγοριοποιημένα βάσει περιεχομένου σε θετικά και αρνητικά. Ως keys ορίζονται οι κατηγορίες αρνητικό,θετικό και ως values σύνολα εγγράφων τα οποία αντιστοιχούν στις κατηγορίες αυτές.

- MultiClassSentiment
  Κλάση η οποία αντιπροσωπεύει δομή συναισθήματος η οποία αντιστοιχεί σε ένα δυναμικό σύνολο κατηγοριών κατά τις οποίες πρέπει να κατηγοριοποιηθούν τα κείμενα χρηστών.

Ορίζονται οι λειτουργίες:

1. MultiClassSentimentDir
   Δέχεται ως είσοδο ένα αντικείμενο τύπου DocumentSet, και δημιουργεί φακέλους οι οποίοι περιέχουν κείμενα χρηστών κατηγοριοποιημένα βάσει περιεχομένου στο σύνολο κατηγοριών κατά τις οποίες πρέπει να κατηγοριοποιηθούν τα κείμενα χρηστών.

2. MultiClassSentimentMap

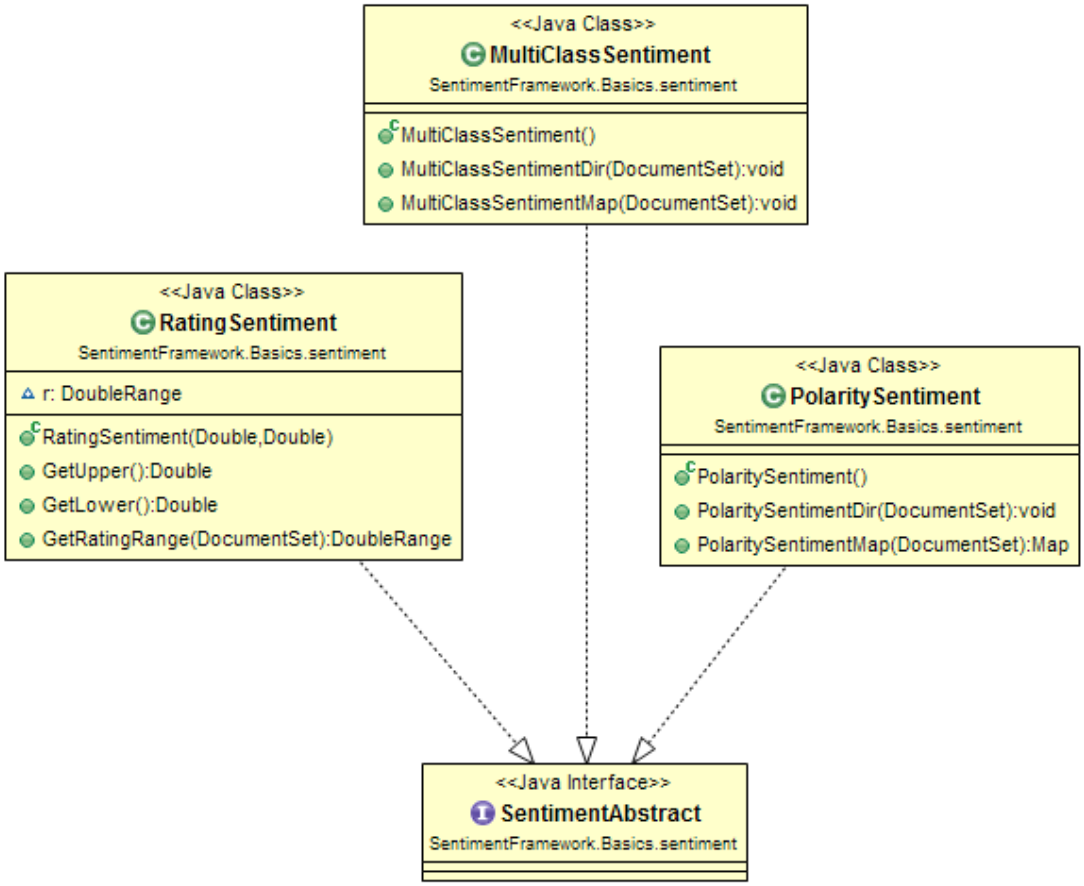
Δέχεται ως είσοδο ένα αντικείμενο τύπου DocumentSet, και δημιουργεί ένα αντικείμενο τύπου HashMap, το οποίο περιέχει κείμενα χρηστών κατηγοριοποιημένα βάσει περιεχομένου στο σύνολο κατηγοριών κατά τις οποίες πρέπει να κατηγοριοποιηθούν τα κείμενα χρηστών. Ως keys ορίζονται οι κατηγορίες και ως values σύνολα εγγράφων τα οποία αντιστοιχούν στις κατηγορίες αυτές.

- RatingSentiment
  Κλάση η οποία αντιπροσωπεύει δομή συναισθήματος η οποία αντιστοιχεί στο εύρος του συνόλου των βαθμολογιών χρηστών, που αφορούν ένα εννοιολογικό αντικείμενο το οποίο σχολιάζεται από τους χρήστες στα κείμενα τους. Οι βαθμολογίες ορίζονται ως αριθμοί κινητής υποδιαστολής (Double).
  Ορίζονται οι λειτουργίες:

  1. GetRatingRange
     Δέχεται ως είσοδο ένα αντικείμενο τύπου DocumentSet, και επιστρέφει ένα αντικείμενο τύπου org.apache.commons.lang.math.DoubleRange, το οποίο αντιστοιχεί στο εύρος στο οποίο κυμάνθηκαν οι βαθμολογίες των χρηστών

  2. GetUpper
     Επιστρέφει την μεγαλύτερη βαθμολογία που δόθηκε από ένα χρήστη.

  3. GetLower
     Επιστρέφει την μικρότερη βαθμολογία που δόθηκε από ένα χρήστη.



Εικόνα 4. Διάγραμμα Κλάσεων- SentimentFramework.Basics.sentiment

4.1.4 SentimentFramework.Crawling

Interface: Crawling.java

Interface το οποίο αντιστοιχεί στη διαδικασία συλλογής δεδομένων για την δημιουργία ενός dataset, το οποίο θα αναλυθεί από το σύστημα.

Ορίζονται οι λειτουργίες:

- fetchData
  Λειτουργία η οποία επιστέφει ένα σύνολο εγγράφων τύπου GenericDocument, το οποίο αντιστοιχεί στη μετασχηματισμένη για της ανάγκες του συστήματος πληροφορία, η οποία αφορά απόψεις χρηστών. Η έλλειψη παραμέτρων εισόδου ουσιαστικά δεν ορίζει περιορισμούς ως προς το σύνολο των εγγράφων τα οποία θα επιστραφούν. Σύνολο το οποίο θα εξαρτάται και θα διαμορφώνεται από την εκάστοτε υλοποίηση της λειτουργίας.

- FetchDataParam
  Λειτουργία η οποία επιστέφει ένα σύνολο εγγράφωντύπου GenericDocument, το οποίο αντιστοιχεί στη μετασχηματισμένη για της ανάγκες του συστήματος πληροφορία, η οποία αφορά απόψεις χρηστών. Ως εισόδους ορίζεται ένα σύνολο παραμέτρων σε μορφή String, οι οποίες αντιστοιχούν σε κριτήρια και περιορισμούς που πρέπει να ληφθούν υπόψιν από τον Crawler κατά τη συλλογή των απόψεων χρηστών. Κριτήρια όπως για παράδειγμα η συλλογή μόνο εκείνων των εγγράφων τα οποία περιέχουν ένα συγκεκριμένο λεκτικό αντικείμενο, ή έχουν καταχωρηθεί σε συγκεκριμένο ιστότοπο κ.α

4.1.5 SentimentFramework.Repository

Interface: RepositoryManager

Interface το οποίο αντιστοιχεί στη λειτουργία διαχείρισης του Data repository, όπου θα αποθηκεύονται το περιεχόμενο και πληροφορίες για τα κείμενα τα οποία περιέχουν απόψεις χρηστών

Ορίζονται οι λειτουργίες:

- getData
  Λειτουργία η οποία επιστρέφει ένα σύνολο εγγράφων από το repository, βάσει των περιορισμών τύπου Filter που δέχεται ως είσοδο. Η έξοδος είναι τύπου δομής Hashmap, επομένως ζευγάρια key-value, όπου key ορίζεται η κατηγορία στην οποία αντιστοιχεί μια ομάδα εγγράφων, και αντίστοιχα ως value ορίζεται μια ομάδα εγγράφων σε μορφή λίστας.

- storeData
  Λειτουργία η οποία δέχεται ως είσοδο ένα σύνολο εγγράφων τύπου GenericDocument, και εισάγει τη πληροφορία που περιέχουν τα έγγραφα στο repository.

- loadData
  Λειτουργία η οποία δέχεται ως είσοδο ένα φάκελο ο οποίος περιέχει αρχεία τα οποία αντιστοιχούν το καθένα σε μια άποψη χρήστη. Επεξεργάζεται την πληροφορία, μετατρέποντας ουσιαστικά κάθε αρχείο σε ένα αντικείμενο τύπου GenericDocument το οποίο περιέχει την αντίστοιχη πληροφορία. Επιστρέφει ένα σύνολο αντικειμένων GenericDocument, το οποίο σύνολο αντιστοιχεί στη συνολική πληροφορία που περιέχει ο φάκελος.

4.1.6 SentimentFramework.Clustering

Interface: Clusterer

Interface το οποίο αντιστοιχεί στη λειτουργία συσταδοποίησης των κείμενων τα οποία περιέχουν γνώμες χρηστών.

Ορίζονται οι λειτουργίες:

- getClusters
  Λειτουργία η οποία δέχεται ως είσοδο ένα αντικείμενο τύπου DocumentSet, και επιστρέφει ένα σύνολο αντικειμένων DocumentSet τα οποία αντιστοιχούν στις συστάδες τις οποίες δημιουργεί η λειτουργία από τα έγγραφα που περιέχονται στο DocumentSet εισόδου.

4.1.7 SentimentFramework.Sentiment

Interface: SentimentAnalyst

Interface το οποίο αντιστοιχεί στη λειτουργία εξαγωγής άποψης/συναισθήματος από τα κείμενα τα οποία περιέχουν γνώμες χρηστών.

Ορίζονται οι λειτουργίες:

- getSentimentFromSet

  Λειτουργία η οποία δέχεται ως είσοδο ένα αντικείμενο τύπου DocumentSet, και επιστρέφει ένα αντικείμενο τύπου SentimentAbstract το οποίο αντιστοιχεί στην άποψη των χρηστών η οποία αφορά τα έγγραφα που περιέχονται στο αντικείμενο DocumentSet.

- getSentimentFromDoc

  Λειτουργία η οποία δέχεται ως είσοδο ένα αντικείμενο τύπου GenericDocument, και επιστρέφει ένα αντικείμενο τύπου SentimentAbstract το οποίο αντιστοιχεί στην άποψη ενός χρήστη η οποία αφορά το περιεχόμενο του εγγράφου που αντιστοιχεί στο αντικείμενο GenericDocument.

- getSentimentFromSentence

  Λειτουργία η οποία δέχεται ως είσοδο ένα αντικείμενο τύπου Sentence, και επιστρέφει ένα αντικείμενο τύπου SentimentAbstract το οποίο αντιστοιχεί στην άποψη ενός χρήστη η οποία αφορά το περιεχόμενου μίας πρότασης ή γενικότερα ενός μικρού σε έκταση κειμένου το οποίο αντιστοιχεί στο αντικείμενο Sentence.

4.1.8 SentimentFramework.Visual

Interface: Visual

Interface το οποίο αντιστοιχεί στη λειτουργία οπτικοποίησης των κατηγοριοποιημένων δεδομένων, με τη χρήση ορισμένων τεχνικών οπτικοποίησης κειμένων και γενικότερων στατιστικών γραφημάτων

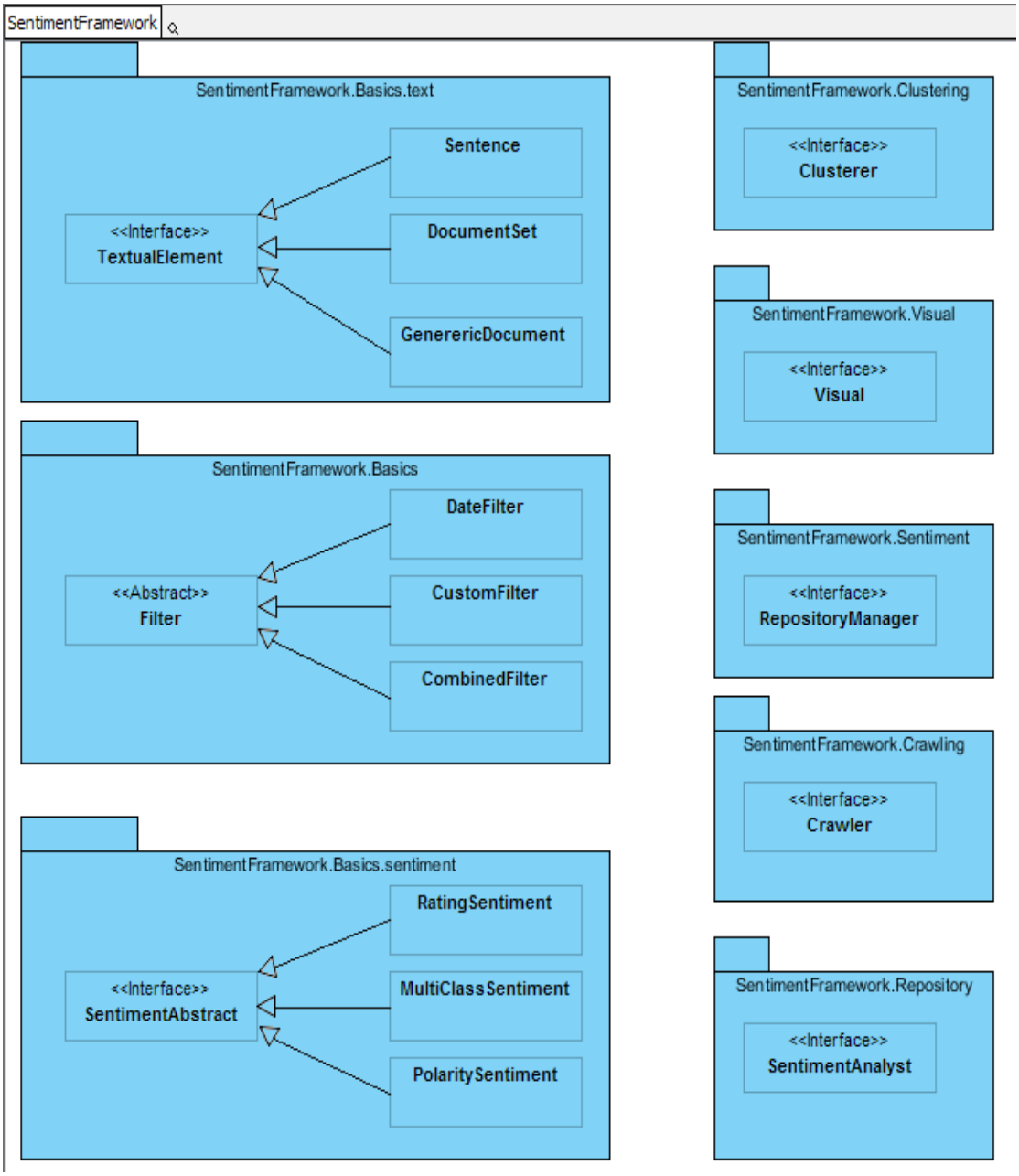
Ορίζονται οι λειτουργίες:

- getTagcloud
  Λειτουργία η οποία δέχεται ως είσοδο ένα αντικείμενο τύπου DocumentSet, και παράγει μια οπτική αναπαράσταση τύπου Tag Cloud η οποία αντιστοιχεί στην αποτύπωση των όρων με τη μεγαλύτερη συχνότητα εμφάνισης στα κείμενα που περιέχονται στο DocumentSet.

- GetWordFrequency

Λειτουργία η οποία δέχεται ως είσοδο ένα αντικείμενο τύπου DocumentSet, και παράγει δυο Tag Clouds τα οποία περιέχουν τους όρους με τη μεγαλύτερη συχνότητα εμφάνισης στα κείμενα που περιέχονται στο DocumentSet, όπου το κάθε Tag Cloud θα αντιστοιχεί σε μία κατηγορία, ή μια ομάδα εφάμιλλων νοηματικά κατηγοριών.

- getBarChart
  Λειτουργία η οποία δέχεται ως είσοδο ένα αντικείμενο τύπου DocumentSet, και παράγει μια οπτική αναπαράσταση τύπου Bar Chart, όπου κάθε μπάρα αντιπροσωπεύει τις ανα κατηγορία γνώμες χρηστών για μια συγκεκριμένη χρονική περίοδο.



Εικόνα 5. Διάγραμμα Πακέτων- SentimentFramework

5

Ενδεικτική Υλοποίηση

Δημιουργείται μια ενδεικτική υλοποίηση αξιοποιώντας ορισμένα από τα components του framework. Για την υλοποίηση κάθε component, χρησιμοποιούνται βιβλιοθήκες Java οι οποίες διατίθενται δωρεάν στο διαδίκτυο. Όσον αφορά τα δεδομένα τα οποία περιέχουν απόψεις χρηστών σε φυσική γλώσσα, χρησιμοποιείται ένα dataset το οποίο περιέχει κριτικές χρηστών που αφορούν ταινίες. Ως εκ τούτου δεν δημιουργείται component το οποίο αντιστοιχεί σε web crawler.

5.1 Data

5.1.1 Dataset

Αξιοποιείται το Large Movie Review Dataset (http://ai.stanford.edu/~amaas/data/sentiment/) , το οποίο περιέχει κριτικές χρηστών που αφορούν ταινίες. Το παρόν Dataset περιέχει 50000 κριτικές σε μορφή κειμένου, το 50% των οποίων ενδείκνυται για την εκπαίδευση του εκάστοτε αλγορίθμου και το υπόλοιπο για σκοπούς testing/evaluation.

5.1.2 Data parsing

Δημιουργούνται βοηθητικές κλάσεις οι οποίες είναι υπεύθυνες για την προσπέλαση των αρχείων του dataset και την μετατροπή των δεδομένων σε κατάλληλη μορφή ώστε να μπορούν να καταχωρηθούν στο repository.
Ορίζονται οι κλάσεις:

- SimpleParser
  Κλάση η οποία περιέχει μεθόδους για την εξάγει των δεδομένων από αρχεία μορφής .txt, τα οποία περιέχουν κείμενο σε φυσική γλώσσα. Αξιοποιείται το όνομα αρχείου και το περιεχόμενο.

- XMLParser
  Κλάση η οποία περιέχει μεθόδους για την εξάγει των δεδομένων από αρχεία μορφής .xml, τα οποία ενδεχομένως περιέχουν περισσότερες πληροφορίες όπως για παράδειγμα μια ημερομηνία για κάθε σχόλιο χρήστη. Αξιοποιείται το σύνολο του περιεχομένου ενός αρχείου το οποίο μπορεί να περιέχει περισσότερες από μια κριτικές, και αναλύονται τα σχετικά elements.

5.2 Apache Lucene

Το Apache Lucene είναι μια βιβλιοθήκη η οποία παρέχει λειτουργίες indexing εγγράφων, όπως και λειτουργίες αναζήτησης στα καταχωρημένα έγγραφα. χρησιμοποιείται ως repository των κειμένων τα οποία περιέχουν τις κριτικές των χρηστών.

5.2.1 Λειτουργίες Διεπαφής

Υλοποιείται η διεπαφή RepositoryManager.java του πακέτου SentimentFramework.Repository όπως και οι αντίστοιχες λειτουργίες:

- getData
- storeData
- loadData

Σε κάθε στάδιο οι πληροφορίες που περιέχονται στα αρχεία μετατρέπονται σε τύπο GenericDocument, όπως αυτό περιγράφεται στο πακέτο SentimentFramework.basics.text. Αντίστοιχα δημιουργούνται έγγραφα τύπου lucene document στα οποία καταχωρούνται ο πληροφορίες με τη μορφή Textfields για κάθε σχετικό πεδίο, όπως το όνομα αρχείου και το περιεχόμενο αυτού. Τελικά τα lucene documents καταχωρούνται στο repository.

5.2.2 Επιπρόσθετες Λειτουργίες

Ορίζονται οι λειτουργίες:

- updateDocument
  Λειτουργία η οποία ενημερώνει ένα έγγραφο το οποίο είναι καταχωρημένο στο repository ,με νέες τιμές στα πεδία αυτού. Δέχεται ένα αντικείμενο τύπου GenericDocument το οποίο φέρει την νέα πληροφορία.

- GetFilter
  Λειτουργία η οποία δέχεται ως είσοδο ένα αντικείμενο τύπου Filter, όπως αυτό ορίζεται στο πακέτο SentimentFramework.Basics, το αναλύει και παράγει στη έξοδο ένα String το οποίο αντιστοιχεί σε ένα Lucene Query, ώστε να γίνεται η κατάλληλη αναζήτηση στο repository.

5.3 LingPipe

Το LingPipe είναι μια πλατφόρμα με λειτουργίες όπως αναζήτηση ονομάτων σε κείμενα, κατηγοριοποίηση αποτελεσμάτων αναζήτησης σε κοινωνικά δίκτυα κ.α. Παρέχεται δωρεάν ένα υποσύνολο των λειτουργιών, για ανάλυση και κατηγοριοποίηση κειμένων γραμμένων σε φυσική γλώσσα, με τη μορφή βιβλιοθήκης Java, η οποία και αξιοποιείται για την ολοκλήρωση του component κατηγοριοποίσης της παρούσας υλοποίησης.

5.3.1 Λειτουργίες Διεπαφής

Υλοποιείται η διεπαφή PolaritySentiment.java του πακέτου SentimentFramework.Basics.sentiment όπως και οι αντίστοιχες λειτουργίες:

- PolaritySentimentDir
- PolaritySentimentMap

Στο στάδιο της κατηγοριοποίησης η διαδικασία δέχεται ως είσοδο ένα αντικείμενο τύπου DocumentSet, το οποίο περιέχει τα κείμενα προς κατηγοριοποίηση. Αναλόγως τη χρήση μίας εκ των δύο παραπάνω μεθόδων, είτε παράγεται ένας φακέλους με αρχεία με κατηγοριοποιημένη πλέον την πληροφορία, είτε ενημερώνεται το repository με τις ορθές κατά την κατηγοριοποίηση, κατηγορίες στις οποίες ανήκει το κάθε κείμενο. Η διεπαφή ορίζει κατηγοριοποίηση σε δυο κατηγορίες, με στόχο τη χρήση σε περιπτώσεις νοηματικών κατηγοριών όπως 'θετικό' και 'αρνητικό', οι οποίες αφορούν απόψεις/γνώμες για μια νοηματική οντότητα την οποία σχολιάζουν οι χρήστες στα κείμενα τους.

5.3.2 Επιπρόσθετες Λειτουργίες

Ορίζονται οι λειτουργίες:

- train
  Λειτουργία η οποία είναι υπεύθυνη για τη διαδικασία εκπαίδευσης του αλγορίθμου κατηγοριοποίησης κειμένων που ορίζει η δομή com.aliasi.classify.DynamicLMClassifier του lingpipe.
- Evaluate

Λειτουργία η οποία είναι υπεύθυνη για τη διαδικασία επαλήθευσης των δεδομένων σύμφωνα με τον classifier ο οποίος έχει εκπαιδευτεί με τη λειτουργία train.

5.4 JfreeCharts/OpenCloud

Το JFreeChart είναι μια δωρεάν βιβλιοθήκη Java η οποία επιτρέπει τη δημιουργία ενός μεγάλου εύρους γραφημάτων. Το Opencloud είναι μια δωρεάν βιβλιοθήκη Java η οποία επιτρέπει τη δημιουργία αναπαραστάσεων νεφών λέξεων(Tag Cloud), σε μορφή HTML. Οι δυο αυτές βιβλιοθήκες αξιοποιούνται για την υλοποίηση του component οπτικοποίησης της παρούσας υλοποίησης.

5.4.1 Λειτουργίες Διεπαφής

Υλοποιείται η διεπαφή Visual.java του πακέτου SentimentFramework.Visual όπως και οι αντίστοιχες λειτουργίες:

- getTagCloud
  Λειτουργία η οποία δέχεται ως είσοδο ένα αντικείμενο τύπου DocumentSet, το οποίο περιέχει κατηγοριοποιημένα δεδομένα, μετασχηματίζει τα δεδομένα και μέσω της Κλάσης TagCloud, και επιτρέπει την αποτύπωση σε αντικείμενο Java Swing Jpanel, ενός Tag Cloud με τους όρους με τη μεγαλύτερη συχνότητα εμφάνισης στα κείμενα χρηστών.

- getWordFrequency
  Λειτουργία η οποία δέχεται ως είσοδο ένα αντικείμενο τύπου DocumentSet, το οποίο περιέχει κατηγοριοποιημένα δεδομένα, μετασχηματίζει τα δεδομένα και μέσω της Κλάσης FrequentWords, και επιτρέπει την αποτύπωση σε αντικείμενο Java Swing Jpanel, δύο Tag Cloud με τους όρους με τη μεγαλύτερη συχνότητα εμφάνισης στα κείμενα χρηστών. Κάθε Tag Cloud περιέχει μόνο όρους από κείμενα βάσει της κατηγορίας στην οποία ανήκουν.

- getBarChart
  Λειτουργία η οποία δέχεται ως είσοδο ένα αντικείμενο τύπου DocumentSet, το οποίο περιέχει κατηγοριοποιημένα δεδομένα, μετασχηματίζει τα δεδομένα και μέσω της Κλάσης StackedBarChart, και επιτρέπει την αποτύπωση σε αντικείμενο Java Swing Jpanel, ενός γραφήματος με μπάρες οι οποίες αντιπροσωπεύουν τις γνώμες των χρηστών. Η λειτουργία χρησιμοποιείται μόνο σε περιπτώσεις όπου τα δεδομένα περιέχουν ημερομηνίες, ώστε να παράγονται μπάρες οι οποίες να αφορούν κείμενα χρηστών ανά συγκεκριμένες χρονικές περιόδους.

5.4.2 Επιπρόσθετες Λειτουργίες

Ορίζονται οι λειτουργίες:

- getSetFromFiles
- getSetFromRepository

Λειτουργίες οι οποίες επιστρέφουν στην έξοδο ένα αντικείμενο τύπου DocumentSet, δεδομένου ενός φακέλου ο οποίος περιέχει κατηγοριοποιημένα δεδομένα, είτε μέσα από το repository.

5.4.3 Κλάση StackedBarChart

5.4.3.1 Λειτουργία createDataset

Η μέθοδος δέχεται ως παράμετρο εισόδου έναν δισδιάστατο πίνακα πραγματικών αριθμών. Ανατίθενται οι κατάλληλες τιμές σε ορισμένα περιγραφικά στοιχεία της οπτικοποίησης και επιστρέφεται αντικείμενο τύπου org.jfree.data.general.DatasetUtilities, το οποίο δημιουργείται συναρτήσει των στοιχείων αυτών και του δισδιάστατου πίνακα.

5.4.3.2 Λειτουργία createChart

Ανατίθενται οι κατάλληλες τιμές σε ορισμένα περιγραφικά στοιχεία της οπτικοποίησης, και τροποποιούνται γραφικά στοιχεία.

Επιστρέφεται αντικείμενο τύπου org.jfree.chart.JFreeChart, στο οποίο έχουν προσαρμοστεί τα δεδομένα οπτικοποίησης τα οποία η μέθοδος λαμβάνει ως παράμετρο εισόδου.

5.4.4 Κλάση Tag Cloud

5.4.4.1 Λειτουργία createDataset

Υπεύθυνη για τη προσαρμογή των επεξεργασμένων δεδομένων κριτικών των χρηστών στην τελική οπτικοποίηση. Δέχεται ως παράμετρο εισόδου μεταβλητή τύπου γραμματοσειράς, η οποία περιέχει όλους τους όρους οι οποίοι εμφανίζονται στις κριτικές χρηστών. Δημιουργείται αντικείμενο τύπου org.mcavallo.opencloud.Cloud, στο οποίο προσαρμόζονται οι όροι.

Επαναληπτικά, για κάθε όρο που περιέχεται στο αντικείμενο τύπου Cloud, παράγεται κατάλληλα κώδικας HTML, ώστε κάθε όρος να εμφανίζεται με μέγεθος αναπαράστασης ανάλογο με τη συχνότητα εμφάνισής του.

Ο κώδικας HTML προσαρμόζεται σε αντικείμενο τύπου γραμματοσειράς, και επιστρέφεται από τη μέθοδο.

5.4.5 Κλάση FrequentWords

5.4.5.1 Λειτουργία createDataset

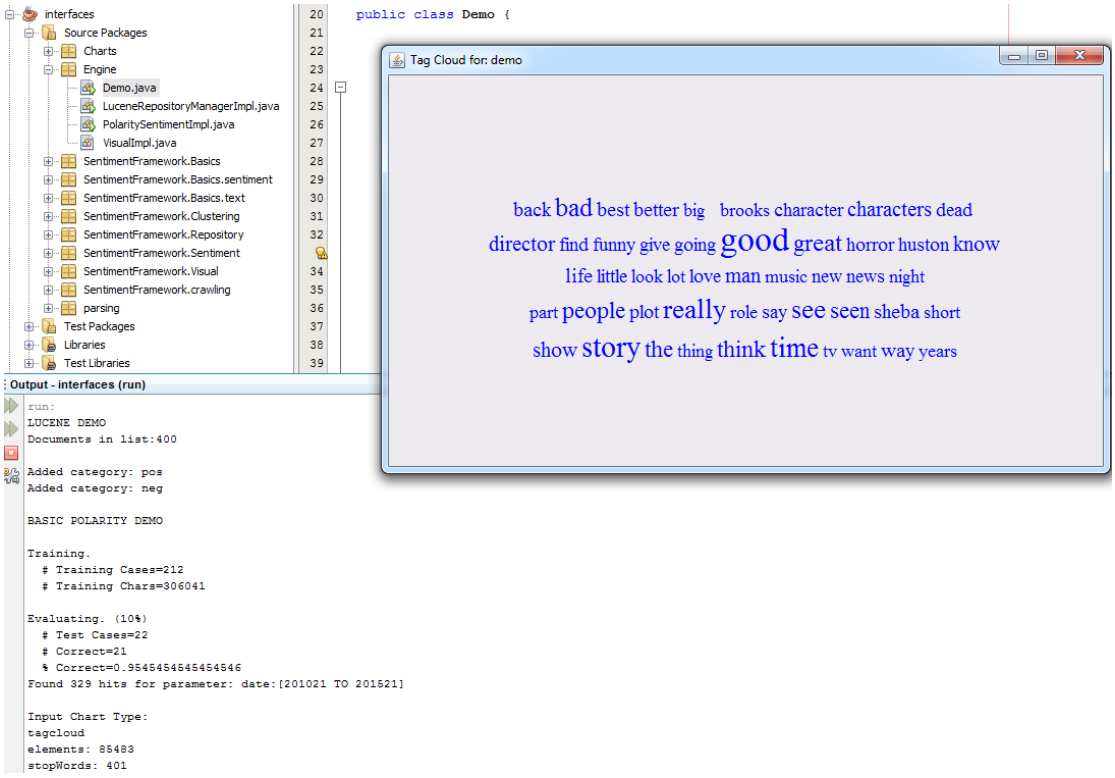
Υπεύθυνη για τη προσαρμογή των επεξεργασμένων δεδομένων κριτικών των χρηστών στην τελική οπτικοποίηση. Δέχεται ως παράμετρο εισόδου αντικείμενα τύπου org.mcavallo.opencloud.Cloud, τα οποία περιέχουν όλους τους όρους οι οποίοι εμφανίζονται στις κριτικές χρηστών. Το ένα περιέχει τους όρους από τις Θετικές κριτικές, ενώ το άλλο τους όρους που βρίσκονται στο σύνολο των Αρνητικών κριτικών.

Επαναληπτικά, για κάθε όρο που περιέχεται στα αντικείμενα τύπου Cloud, παράγεται κατάλληλα κώδικας HTML, ώστε κάθε όρος να εμφανίζεται με μέγεθος αναπαράστασης ανάλογο με τη συχνότητα εμφάνισής του. Επίσης, το σύνολο των «θετικών» όρων παίρνει διαφορετικό χρώμα, από τους «αρνητικούς» όρους.

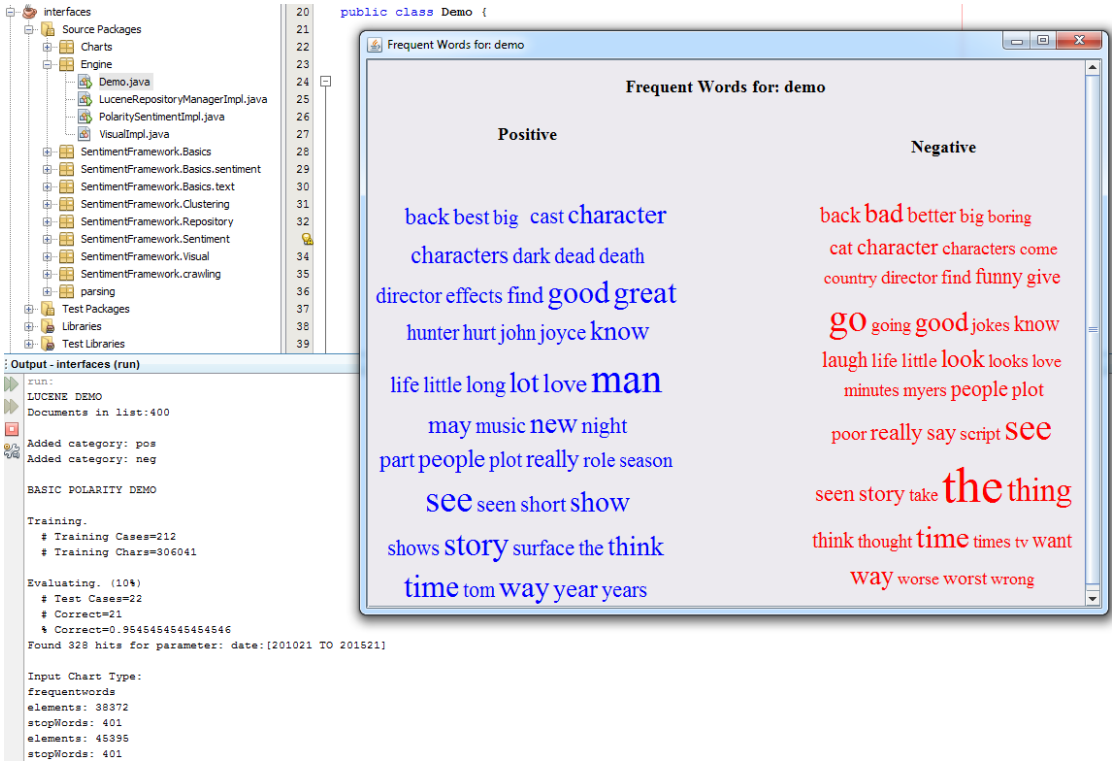
Ο κώδικας HTML προσαρμόζεται σε αντικείμενο τύπου γραμματοσειράς, και επιστρέφεται από τη μέθοδο.

5.4.6 Παραδείγματα Εκτέλεσης

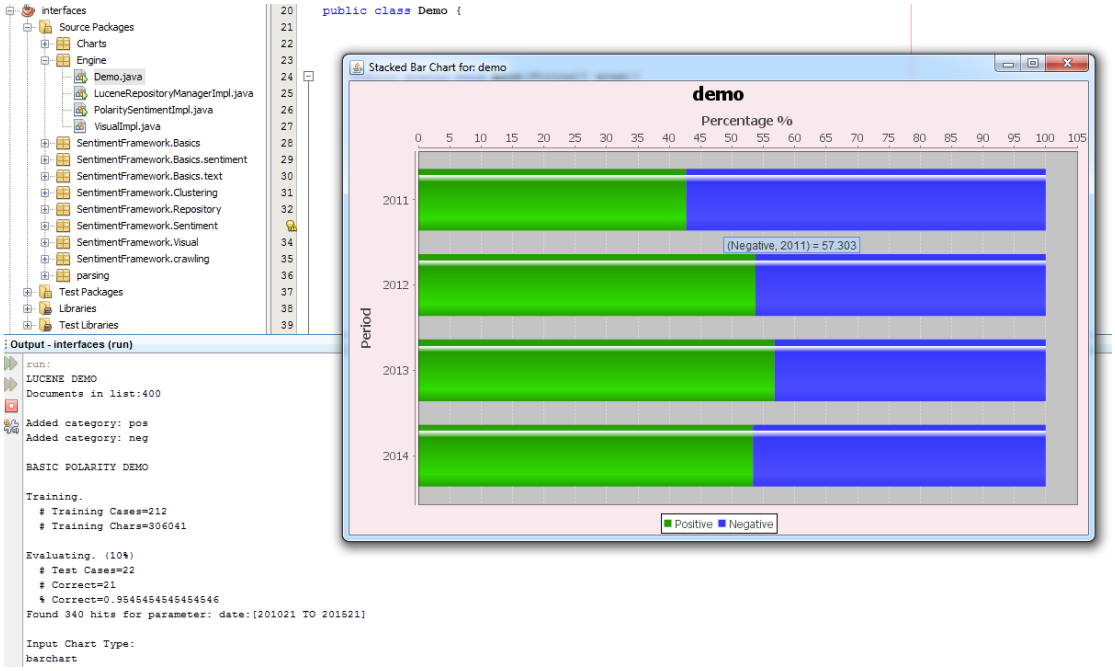
Παρατίθενται εικόνες εκτέλεσης της υλοποίησης, με είσοδο τετρακόσια (400) έγγραφα, τα οποία αντιπροσωπεύουν απόψεις χρηστών.



Εικόνα 6. Εκτέλεση - TagCloud



Εικόνα 7. Εκτέλεση – FrequentWords



Εικόνα 8. Εκτέλεση – StackedBarChart

6

Συμπεράσματα

Μέσω της διαδικασίας υλοποίησης του Framework το οποίο σχεδιάστηκε στα πλαίσια της παρούσας εργασίας έγινε φανερή η ουσιαστική ανάγκη κατασκευής μιας αυτού του είδους πλατφόρμας. Μια πλατφόρμας η οποία θα μπορεί μέσω του σχεδιασμού της, να αξιοποιεί οποιαδήποτε ερευνητική προσπάθεια η οποία αποτυπώνεται στη δημιουργία αξιοποιήσιμου open source κώδικα και βιβλιοθηκών, στους τομείς της εξόρυξης γνώσης και ανάλυσης συναισθήματος.

Το παρόν Framework παρέχει τη δυνατότητα πρακτικής αξιοποίησης γενικότερων μορφωμάτων κώδικα και τον συνδυασμό αυτών, σε ένα συγκεκριμένο processing pipeline, ώστε να επιτρέπεται συγκριτική μελέτη της απόδοσης των συνδυασμών των εκάστοτε μορφωμάτων είτε σε σχέση με άλλους συνδυασμούς είτε σε σχέση με άλλες ολοκληρωμένες πλατφόρμες εξόρυξης και ανάλυσης συναισθήματος.

Επιπλέον, μέσω της σύγκρισης των αποτελεσμάτων τα οποία παρουσιάζονται στο στάδιο της οπτικοποίησης, θα δίνεται η δυνατότητα περαιτέρω πειραματισμού ως προς την ανεύρεση των ιδανικών τεχνικών ανάλυσης συναισθήματος. Ανάλογα με την μελέτη περίπτωσης κάθε έρευνας, η εναλλαγή των components του συστήματος με εφάμιλλα αυτών components, θα οδηγεί ιδανικά σε μεγαλύτερη ακρίβεια αποτελεσμάτων.

Τέλος, το γεγονός της ύπαρξης ενός ευέλικτου πλαισίου το οποίο δίνει τη δυνατότητα σχεδόν άμεσης εφαρμογής εξωτερικών components και την ανάλυση αποτελεσμάτων, χωρίς να χρειάζεται κάθε φορά ένας εκ του μηδενός σχεδιασμός και υλοποίηση ενός συστήματος μόνον για το σκοπό της εκάστοτε έρευνας, αποτελεί σημαντικό βοήθημα και σίγουρα θα συμβάλει στην εξοικονόμηση χρόνου ως προς την τελική εξαγωγή συμπερασμάτων.

Βιβλιογραφικές Αναφορές

[1] Jindal, N., B. Liu. Identifying comparative sentences in text documents, 2006.

[2] Pang, B., L. Lee, S. Vaithyanathan. Thumbs up?: sentiment classication using machine learning techniques, 2002.

[3] Pang, B., L. Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2008.

[4] Dave, K., S. Lawrence, D. Pennock. Mining the peanut gallery: Opinion extraction and semantic classication of product reviews, 2003.

[5] Tan, S., Y. Wang., X. Cheng. Combining learn-based and lexicon-based techniques for sentiment detection without using labeled examples, 2008.

[6] Melville, P., W. Gryc., R. D. Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classication, 2009.

[7] Pan, S., X. Ni, J. Sun, Q. Yang, Z. Chen. Cross-domain sentiment classication via spectral feature alignment, 2010.

[8] Turney, P. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classication of reviews, 2002.

[9] Wiebe, J., R. Bruce, T. O'Hara. Development and use of a gold-standard data set for subjectivity classications, 1999

[10] Yu, H. and V. Hatzivassiloglou. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences, 2003.

[11] Kim, S., P. Pantel, T. Chklovski, M. Pennacchiotti. Automatically assessing review helpfulness. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2006.

[12] Wilson, T., J. Wiebe, P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis, 2005.

[13] Zhai, Z., B. Liu., L. Zhang., H. Xu., P. Jia. Identifying evaluative sentences in online discussions, 2011.

[14] Hassan, A., V. Qazvinian., D. Radev. What's with the attitude? Identifying sentences with attitude in online discussion, 2010.

Σχετικά Εργαλεία

Apache Lucene (https://lucene.apache.org/core/)

LingPipe (http://alias-i.com/lingpipe/index.html)

JFreeChart (http://www.jfree.org/jfreechart/)

OpenCloud (http://opencloud.mcavallo.org/)