ΕΡΓΑΣΙΑ ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

Μουλαδένιος Κων/νος 2379

Βαρβούτας Κων/νος 2336

Αθανασιάδης Γεώργιος 2331

Το πρόβλημα που μελετάται στην εργασία είναι η ανάλυση κριτικών ταινίας. Έχουμε στη διάθεση μας ένα σύνολο από δεδομένα (train data) τα οποία χαρακτηρίζουν μία πληθώρα κριτικών ως θετικές ή αρνητικές με βάση το περιεχόμενό τους. Υπάρχουν διάφορες τεχνικές που προσεγγίζουν το πρόβλημα οι οποίες χωρίζονται σε supervised (στις οποίες χρησιμοποιούνται τα train data) και unsupervised (στις οποίες ο χαρακτηρισμός των ταινιών πρέπει να γίνει χωρίς τα train data) . Εμείς μελετήσαμε μία supervised και μία unsupervised τεχνική.

SUPERVISED ΤΕΧΝΙΚΗ

Ως supervised τεχνική διαλέξαμε να μελετήσουμε τον **Naive Bayes** κατηγοριοποιητή. Σύμφωνα με αυτόν αναλύουμε τα train data και συγκεντρώνουμε σε ένα hashmap όλες τις λέξεις που εμφανίζονται στα θετικά έγγραφα καθώς επίσης και τη συχνότητα εμφάνισης τους . Το ίδιο κάνουμε και για τα αρνητικά. Έτσι όταν ελέγχουμε τα έγγραφα προς εξέταση υπολογίζουμε για το κάθε έγγραφο την πιθανότητα του να ανήκει στα θετικά και την πιθανότητα του να ανήκει στα αρνητικά και όποια υπερισχύει αντιπροσωπεύει και την κατηγορία στην οποία ανήκει το έγγραφο.

ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ:

Για καλύτερα αποτελέσματα τα έγγραφα δέχονται μία προεπεξεργασία. Καταρχάς δεν υπολογίζουμε τους χαρακτήρες όπως </br> ,) , (, . και μετατρέπουμε όλα τα γράμματα σε πεζά. Στη συνέχεια παίρνουμε τους tokenized όρους από τα έγγραφα και εφαρμόζουμε το stemming ώστε κάθε λέξη να μεταφέρεται στη ρίζα της (χρησιμοποιούμε την κλάση εργαλείο PorterStemmer.java) . Επίσης χρησιμοποιούμε τον Stanford Log-linear Part-Of-Speech Tagger ένα βοηθητικό εργαλείο το οποίο χαρακτηρίζει τις λέξεις ως ρήματα , ουσιαστικά κ.ο.κ . Με αυτό τον τρόπο κρατάμε μόνο τις λέξεις που κρύβουν μέσα τους πληροφορία όπως τα ρήματα , τα επιρρήματα και τα επίθετα (λέξεις τις οποίες βάζουμε στο ανάλογο hashmap).

ΥΠΟΛΟΓΙΣΜΟΣ ΠΙΘΑΝΟΤΗΤΩΝ

Την ίδια προεπεξεργασία δέχονται και τα έγγραφα προς εξέταση. Επομένως παίρνουμε κάθε λέξη που βρίσκουμε στα test data και ελέγχουμε για τη συχνότητα της λέξης στα θετικά έγγραφα και στα αρνητικά έγγραφα.

Η πιθανότητα του εγγράφου να ανήκει στα θετικά έγγραφα είναι ίση με την πιθανότητα η κλάση να είναι positive(δηλαδή το άθροισμα των συχνοτήτων των λέξεων που βρίσκονται στα θετικά έγγραφα (total_pos) προς το άθροισμα των συχνοτήτων όλων των λέξεων και στα θετικά και στα αρνητικά έγγραφα) πολλαπλασιασμένη με τη συχνότητα εμφάνισης τις κάθε λέξης στα θετικά έγγραφα προς την τιμή total_pos .

Με τον ίδιο τρόπο υπολογίζεται και η πιθανότητα το έγγραφο προς εξέταση να ανήκει στα αρνητικά .

UNSUPERVISED ΤΕΧΝΙΚΗ

Ως unsupervised τεχνική επιλέχθηκε προς υλοποίηση συγκεκριμένα μία όπως είναι υποδεδειγμένη στο *paper Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012 (link: https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf,*

βλ. Σελίδα 34, ή εναλλακτικά βλ. το ανηρτημένο στο e-learning του μαθήματος "survey2.pdf"). Η βασική ιδέα της παραπάνω τεχνικής είναι ότι εξάγονται και χρησιμοποιούνται δύο λέξεις εάν τα POS tags τους είναι συμβατά με οποιαδήποτε περίπτωση από αυτές που παρατίθενται στον πίνακα:

	First word	Second word	Third word (not extracted)
1	JJ	NN or NNS	anything
2	RB, RBR, or RBS	JJ	not NN nor NNS
3	JJ	JJ	not NN nor NNS
4	NN or NNS	JJ	not NN nor NNS
5	RB, RBR, or RBS	VB, VBD, VBN, or VBG	anything

Ασφαλώς η παραπάνω συμβατότητα διασταυρώνεται κι εξασφαλίζεται με τη χρήση μιας σειράν εμφωλευμένων εντολών "if". Ο λόγος που η περιπτωσιολογία είναι η συγκεκριμένη, είναι επειδή γίνεται η παραδοχή πως αυτά τα λεκτικά μοτίβα είναι που συνήθως *εκφράζουν απόψεις*.

Για το σύνολο όλων των δοθέντων αρχείων, γίνεται συνοπτικά το εξής:

Κάθε αρχείο διατρέχεται και εξετάζονται ανά τρεις οι λέξεις του. Ελέγχεται εάν οι φράσεις που προκύπτουν από τη διαδοχική αυτή εξέταση πληρούν τις προυποθέσεις του πίνακα, εάν δηλαδή οι φράσεις αυτές

δηλώνουν/εκφράζουν κάποιο opinion σύμφωνα με τις παραδοχές μας. Εάν ναι, κάθε φράση τοποθετείται σε hashmap μαζί με το αντίστοιχο της αρχείο.

Στη συνέχεια με την κατάλληλη επεξεργασία προκύπτουν οι ζητούμενες πιθανότητες.

ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ:

Και πάλι τα έγγραφα δέχονται μία προεπεξεργασία για την εξαγωγή καλών αποτελεσμάτων. Και πάλι δε λαμβάνουμε υπ'όψην τους χαρακτήρες όπως </br> ,) , (, . και μετατρέπουμε όλα τα γράμματα σε πεζά. Χρησιμοποιούμε ασφαλώς τον Stanford Log-linear Part-Of-Speech Tagger προς τον χαρακτηρισμό των λέξεων ως ρήματα , ουσιαστικά, επίθετα κ.ο.κ ούτως ώστε να γίνει η κατάλληλη κατηγοριοποίηση. Έτσι διατηρείται μόνο η χρήσιμη πληροφορία, αυτή δηλαδή που κατά την παραδοχή της υλοποιηθείσας μεθόδου εκφράζει πράγματι opinions. Στη συνέχεια οι λέξεις μπαίνουν στο hashmap/hashset που τους αντιστοιχεί (βλ. pos, neg, index κλπ).

ΥΠΟΛΟΓΙΣΜΟΣ ΠΙΘΑΝΟΤΗΤΩΝ

Η μέθοδος του paper αφορά μηχανές αναζήτησης κι ως εκ τούτου έπρεπε να τροποποιηθεί η μέθοδος υπολογισμού των πιθανοτήτων. Υπολογίζεται ένα γενικό score το οποίο τελικά μας ενδιαφέρει εάν είναι θετικό ή αρνητικό (κι έτσι στο αρχείο predictions.txt τίθεται 1 και 0 αντίστοιχα δίπλα στο αντίστοιχο αρχείο) με γνώμονα την ύπαρξη στο σύνολο των θετικών ή των αρνητικών εγγράφων ή όχι. Υπολογίζεται και πάλι ένα "positive score" κι ένα "negative score" και το τελικό score προκύπτει ως εξής:

X = (pscore * neg.size())/((nscore +0.01) * (pos.size() + 0.01));

Score = (Math.log(X)/Math.log(2));

για κάθε αρχείο.

ΧΡΟΝΟΙ ΕΚΤΕΛΕΣΗΣ

Supervised:

- Train: 160 δευτερόλεπτα
- Test: 137 δευτερόλεπτα

Unsupervised: 363 δευτερόλεπτα