Μυρτώ Χριστίνα Ελευθέρου – 3170046

Χρυσόστομος Ισκάκης – 3170052

Χρυσούλα Οικονόμου – 3170127

2η Εργασία

Φόρτωμα αρχείων train

Χρησιμοποιούμε το αρχείο imdb.vocab, στο οποίο όλες οι λέξεις είναι ταξινομημένες κατά φθίνουσα σειρά, σύμφωνα με την συχνότητα που εμφανίζονται σε όλα τα αρχεία και αποθηκεύουμε στο HashMap vocabulary τις m-n συχνότερες λέξεις.

Το αρχείο labeledBow.feat περιέχει για όλα τα αρχεία train, τον συνολικό αριθμό που εμφανίζεται κάθε λέξη σε κάθε αρχείο. Χρησιμοποιώντας το αρχείο αυτό, δημιουργούμε το HashMap frequency, στο οποίο αποθηκεύουμε το id του review και μία λίστα με τις λέξεις που περιλαμβάνει το συγκεκριμένο review.

Οι posVectors και negVectors, περιλαμβάνουν τα διανύσματα μορφής 0 ή 1, για τις θετικές και τις αρνητικές κριτικές αντίστοιχα, ενώ ο vectors περιλαμβάνει όλες τις κριτικές. Ο classes, δείχνει για κάθε αρχείο, σε ποια κατηγορία ανήκει(θετικές ή αρνητικές).

Τέλος, δημιουργούμε δύο HashMap negWords και posWords, τα οποία περιλαμβάνουν τον συνολικό αριθμό που εμφανίζεται κάθε λέξη στα θετικά και στα αρνητικά reviews.

Αλγόριθμος ID3

- Για τον αλγόριθμο ID3, δώσαμε ως υπερπαραμέτρους n=100 και m=1000. Στην κλάση ID3, η μέθοδος entropies(), υπολογίζει για κάθε λέξη την εντροπία της έτσι ώστε να μπορέσουμε να υπολογίσουμε στην συνέχεια το information gain κάθε λέξης.

- Στην μέθοδο informationGain(), αφού υπολογίσουμε πρώτα την συνολική εντροπία, υπολογίζουμε στην συνέχεια το IG.
- Στην μέθοδο maxIG, βρίσκουμε το maximum IG, έτσι ώστε να βρούμε ποια είναι η λέξη που έχει το μεγαλύτερο information gain, και με βάση αυτή να δημιουργήσουμε τα καινούργια παιδιά του δέντρου. Για κάθε κατάσταση, το ig υπολογίζεται εκ νέου.
- Η pureClass ελέγχει αν όλες οι λέξεις μίας κατάστασης, ανήκουν σε μία κατηγορία.

Η κλάση TreeNode περιλαμβάνει τους κόμβους για την δημιουργία του δέντρου.

Στην κλάση Recursive, έχουμε την αναδρομική συνάρτηση για την δημιουργία του δέντρου.

Στην κλάση fixLL σχηματίζονται οι νέοι πίνακες (δυσδιάστατος reviewsXfeatures και classes) του κάθε παιδιού.

Αλγόριθμος Naive Bayes

Για τον αλγόριθμο Bayes, δώσαμε ως υπερπαραμέτρους n=80 και m=600. Στην κλάση NaiveBayesClassifier, έχουμε τις μεθόδους Train() και Evaluate(). Αρχικά, ξεκινάμε με την κλάση Main, η οποία δημιουργεί ένα αντικείμενο b της κλάσης NaiveBayesClassifier. Το αντικείμενο αυτό καλεί αρχικά την μέθοδο Train() στην οποία γίνεται το φόρτωμα των αρχείων όπως αναφέρθηκε παραπάνω, και στη συνέχεια δημιουργούνται δύο πίνακας με θετικές και αρνητικές κριτικές αντίστοιχα, σε μορφή διανυσμάτων (<0,1,0,1,1,..>). Αμέσως μετά καλείται η μέθοδος Evaluate() στην οποία γίνεται η διαδικασία του Testing. Πιο συγκεκριμένα, φορτώνεται το αρχείο labeledBow.feat που βρίσκεται στον φάκελο test και υπολογίζεται η πρόβλεψη του αλγόριθμου Bayes για την κατηγορία της κάθε κριτικής (θετική ή αρνητική). Εφόσον υπολογιστούν όλες οι προβλέψεις για όλες τις κριτικές, στο τέλος υπολογίζεται η ακρίβεια του αλγορίθμου(accuracy = 52%) , η οποία ουσιαστικά είναι η διαίρεση των κριτικών για τις οποίες έγινε σωστή πρόβλεψη με το πλήθος όλων των κριτικών.

Random Forest :

Έχει ακριβώς τις ίδιες βάσεις με τον ID3 και επιπλέον δημιουργούνται πολλά δέντρα και βάσει πλειοψηφίας αποτελεσμάτων επιλέγεται η κατηγορία της κριτικής.