ΟΝΟΜΑΤΕΠΩΝΥΜΟ: **ΖΕΡΒΑΣ ΜΙΧΑΗΛ**
ΑΡΙΘΜΟΣ ΜΗΤΡΩΟΥ: **ics20015**

# Programming Project 2

Για την υλοποίηση της εργασίας χρησιμοποιήθηκε ολόκληρο το dataset.

**Προεπεξεργασία (train και test set):**
Κατά την προεπεξεργασία αφαιρέθηκαν γραμμές που ξεκινούσαν με συγκεκριμένες λέξεις από την επικεφαλίδα (όπως "From"), αφαιρέθηκαν λέξεις που δεν συνεισφέρουν στο περιεχόμενο (όπως "Subject") και αριθμοί. Για την αφαίρεση των stop words, των σημείων στίξης και για stemming χρησιμοποιήθηκε η βιβλιοθήκη Natural Language Toolkit (NLTK) της Python.
Στη συνέχεια κάθε λέξη κάθε εγγράφου σταθμίστηκε με το tf-idf και βάση αυτού κρατήθηκαν οι 15 πιο σημαντικές λέξεις κάθε εγγράφου.

**Κανόνες συσχέτισης (train set):**
Εδώ με την χρήση της βιβλιοθήκης mlxtend χρησιμοποιήθηκε ο apriori για την εύρεση των συχνών στοιχειοσυνόλων και την εξαγωγή από αυτά, των κανόνων συσχέτισης. Δοκιμάστηκαν διάφορες τιμές support και confidence. Παρατηρήθηκε ότι η μικρότερη τιμή που μπορούσε να πάρει το support χωρίς να γεμίση η ram ήταν 0.003. Η τιμή που τελικά χρησιμοποιήθηκε ήταν support = 0.0031 και παρήχθηκαν 816 συχνά στοιχειοσύνολα. Το confidence δεν επηρέαζε στον ίδιο βαθμό τα αποτελέσματα για μικρές αλλαγές στην τιμή του όπως το support. Για τιμή 0.15 και 0.2 έβγαζε 98 κανόνες συσχέτισης. Τελικά κράτησα την τιμή 0.2.

**Κατηγοριοποίηση κειμένου (train set):**
Επαναυπολογίστηκαν τα συχνά στοιχειοσύνολα και οι κανόνες συσχέτισης, μόνο που στις λέξεις που αντιπροσωπεύουν κάθε έγγραφό προστέθηκε επίσης και η κατηγορία στην οποία ανήκει. Οι κανόνες συσχέτισης φιλτραρίστηκαν έτσι ώστε στο δεξί μέρος του κανόνα να περιέχεται μόνο κάποια από τις κατηγορίες των κειμένων. Έτσι πλέον οι κανόνες έγιναν της μορφής λέξη1, λέξη2 → κατηγορία1. Με τις ίδιες τιμές support και confidence βγήκαν συνολικά 1009 συχνά στοιχειοσύνολα και 168 κανόνες συσχέτισης της μορφής που θέλουμε και οι οποίοι χρησιμοποιήθηκαν και στην αποτίμηση.

**Αποτίμηση:**
Η υλοποίηση της συνάρτησης πρόβλεψης έγινε βασιζόμενη στην κατηγορία που ανήκουν οι περισσότερες (από τις 15) λέξεις κάθε εγράφου. Συγκεκριμένα γίνεται έλεγχος κάθε φορά αν κάθε λέξη του εγγράφου ανήκει σε δεξί μέρος οποιουδήποτε κανόνα συσχέτισης και στο τέλος κρατάμε την κατηγορία που αντιπροσωπεύει τις περισσότερες λέξεις. Επειδή υπήρχαν αρκετές περιπτώσεις που οι λέξεις δεν υπήρχαν στο δεξί μέρος κάποιου κανόνα (στον κώδικα και στα στιγμιότυπα εκτέλεσης αντιπροσωπεύονται από την μεταβλήτή x), τότε το έγγραφο δεν κατηγοριοποιείται. Για την αποτίμηση δημιουργήθηκαν 5 λεξικά: το ένα αποτελείται από το πλήθος των εγγράφων από το test set που ανήκουν σε κάθε κατηγορία, το δευτερο αποτελείται από το πλήθος των προβλέψεων για κάθε κατηγορία και τα άλλα 3 αφορούν τα True Positives, False Positives και False Negatives. Τα αποτελέσματα των μετρικών με macro averaging είναι: Precision = 0.6701 και Recall = 0.3709. Η κατηγορία με το καλύτερο Precision και Recall ήταν η sci.electronics:
Precision of category:  0.9583
Recall of category:  0.05852