Αλέξανδρος Βεντούρας (ΑΜ: 3160013)
Νίκος Δούσης (ΑΜ: 3160030)
Δικαία Σωτηροπούλου (ΑΜ: 3160172,
email: dikeasot@hotmail.com)

2η Εργασία

Για την 2η εργασία επιλέξαμε να υλοποιήσουμε τον αλγόριθμο αφελή ταξινομητή Bayes.

Δημιουργήσαμε 5 συναρτήσεις με ονόματα *Bayes*, *lexiko*, *IGgetter*, *training* και *targetEmail.*

Αρχικά, οι μεταβλητές *hamPath* και *spamPath* είναι οι μεταβλητές με τα μονοπάτια των φακέλων που περιέχουν τα ham και spam emails αντίστοιχα. Καλούμε τη συνάρτηση ***lexiko*** που δέχεται σαν ορίσματα τα μονοπάτια αυτά και δημιουργεί 2 λεξικά (*spamDirs, hamDirs*) όπου το καθένα περιέχει τις λέξεις που εμφανίζονται στα αντίστοιχα email. Τα *spamDict* και *hamDict* είναι πίνακες με το πλήθος των φορών που εμφανίζεται η κάθε λέξη σε όλα τα email.
Στη συνέχεια καλείται η συνάρτηση ***IGgetter*** με ορίσματα τα *hamDict* και *spamDict*. Στο *mainDict* αποθηκεύονται τα ποσοστά των εμφανίσεων κάθε λέξης. Σε περίπτωση που κάποιες τιμές στα *mainDict, hamDict* και *spamDict* είναι εκτός του διαστήματος [0,1], τις κανονικοποιούμε στο διάστημα αυτό. Τα λεξικά *interLex*, *hamInveInterLex* και *spamIveInterLex* είναι ενδιάμεσα λεξικά που χρησιμοποιούμε για να επεξεργαστούμε τα δεδομένα που υπάρχουν στα λεξικά hamDict και spamDict. Ακόμη, ταξινομούμε το λεξικό *IG* σε φθίνουσα σειρά, ενώ ο πίνακας *mostUsefulWords* που δημιουργήσαμε περιέχει τις 1000 λέξεις που είναι πιο χρήσιμες.
Έπειτα, καλείται η συνάρτηση ***training*** με ορίσματα τα *hamPath* και *spamPath*. Ο πίνακας *data* που δημιουργούμε, έχει τις τιμές των ιδιοτήτων και το C για κάθε email από τα 916 spam και 916 ham που έχουμε λάβει υπόψιν μας στο training. Από το σύνολό αυτών των email (1832), βρίσκουμε τα features που υπάρχουν σε κάθε ένα από αυτά.
Η ***targetEmail*** δέχεται σαν όρισμα το όνομα ενός email και το path στο οποίο βρίσκεται αυτό και ανάλογα με τις λέξεις που περιέχονται στο mail αποφασίζει αν το email ανήκει στα ham ή spam.
Τέλος, καλείται η συνάρτηση ***bayes***, η οποία υλοποιεί τον αφελή αλγόριθμο Bayes. Δέχεται σαν ορίσματα τον πίνακα *data* με τα δεδομένα μας και ένα target email και αποφασίζει σε ποια κατηγορία ανήκει και με ποια πιθανότητα.