Programming Project 2

Στόχος της εργασίας είναι η εξόρυξη γνώσης από συλλογές κειμένων με τη χρήση κανόνων συσχέτισης.

Τα δεδομένα που θα χρησιμοποιήσετε αντιστοιχούν σε δεδομένα από 20 Newsgroups (http://qwone.com/~jason/20Newsgroups/) τα οποία αντιστοιχούν σε 20 κατηγορίες (κάποιες με συνάφεια και κάποιες μη συναφείς μεταξύ τους), έτσι ώστε να σχηματίζονται και 6 πιο γενικές κατηγορίες (για λεπτομέρειες στον ιστότοπο).
Συγκεκριμένα, θα χρησιμοποιήσετε το http://qwone.com/~jason/20Newsgroups/20news-bydate.tar.gz αρχείο για τα δεδομένα σας.

Θα αναπτύξετε πρόγραμμα σε όποια γλώσσα επιθυμείτε (αλλά συνίσταται έντονα η χρήση python) που θα υλοποιεί τις παρακάτω φάσεις:

- **Προεπεξεργασία:** σε αυτή τη φάση θα πρέπει να επεξεργαστείτε τα κείμενα σας που ανήκουν στο σύνολο εκπαίδευσης, για να αντιστοιχίσετε σε κάθε άρθρο ένα σύνολο από σημαντικές λέξεις.
  - Αφαιρέστε επικεφαλίδες που δεν συνεισφέρουν στο περιεχόμενο, αφαιρέστε σημεία στίξης, συχνές λέξεις χωρίς σημασιολογία, και κάντε stemming αν το θεωρείτε απαραίτητο ή και όποιο άλλο βήμα θέλετε.
  - Σταθμίστε κάθε λέξη με κάποιο βάρος της επιλογής σας (συχνότητα εμφάνισης, tf-idf, αριθμός εμφανίσεων, κτλ)
  - Κρατήστε ως έγγραφο το top-k των σημαντικών του λέξεων σύμφωνα με τη στάθμιση που κάνατε
- **Κανόνες συσχέτισης:** Ανιχνεύστε κανόνες συσχέτισης στα έγγραφα (σύνολα λέξεων) μετά την προεπεξεργασία δεδομένου support και confidence threshold. Εφαρμόστε a priori θεωρώντας κάθε έγγραφο ως μια συναλλαγή και κάθε λέξη ως ένα item. Δηλαδή θα παράγετε κανόνες τύπου, π.χ.: λέξη1, λέξη2 → λέξη3.
- **Κατηγοριοποίηση κειμένου:** Επεκτείνετε το dataset σας χρησιμοποιώντας και την κατηγορία των εγγράφων, έτσι ώστε να ανήκει και η κατηγορία στις λέξεις του εγγράφου και βρείτε κανόνες που συσχετίζουν την εμφάνιση λέξεων με την κατηγορία στην οποία ανήκουν. Στο δεξί μέρος των κανόνων αυτών μας ενδιαφέρει πάντα η κατηγορία του εγγράφου.
- **Αποτίμηση:** Χρησιμοποιήστε τους κανόνες της προηγούμενης φάσης στο σύνολο ελέγχου και μετρήστε την απόδοση του συστήματος. Μετρήστε precision/recall ανά κατηγορία και macro-average.

Στις δύο τελευταίες φάσεις μπορείτε να δουλέψετε με τις 6 γενικές κατηγορίες αν θέλετε. Θα πρέπει να πειραματιστείτε με διαφορετικές τιμές support και confidence για να πάρετε κάποιο αποτέλεσμα που σας ικανοποιεί. Επίσης είναι δική σας επιλογή πόσες λέξεις θα κρατήσετε ανά έγγραφο. Αν εξάγετε πολλούς κανόνες κατηγοριοποίησης μπορείτε να επιλέξετε τους καλύτερους για χρήση στην αποτίμηση. Αποτελεί δική σας επιλογή πώς θα τους συνδυάσετε αν θέλετε. Αν έχετε θέμα με τον όγκο των δεδομένων χρησιμοποιήστε δειγματοληψία για να μειώσετε.

Τι θα παραδώσετε:

- τον κώδικα σας
- μία σύντομη αναφορά (pdf) στην οποία θα περιγράφετε τις παραμέτρους και τις επιλογές που κάνατε για την υλοποίηση και θα παρουσιάζετε τα αποτελέσματα των τριών τελευταίων φάσεων με όποιο τρόπο θεωρείτε καλύτερο (και jupyter notebooks γίνονται δεκτά).