Προγραμματιστική Εργασία 2 - Κατηγοριοποίηση Κειμένου

Καλείστε να λύσετε ένα πρόβλημα κατηγοριοποίησης κειμένου, και συγκεκριμένα ένα πρόβλημα ανάλυσης συναισθήματος όπου κάθε κείμενο μπορεί να ανήκει σε μία εκ των δύο κατηγοριών “θετικό” (positive) ή “αρνητικό” (negative). Θα χρησιμοποιήσετε δεδομένα κειμένου που αποτελούν κριτικές ταινιών και μπορείτε να βρείτε στον ιστότοπο http://www.cs.cornell.edu/people/pabo/movie-review-data/. Από εκεί θα επιλέξετε να κατεβάσετε την έκδοση: **polarity_dataset_v2.0 - (includes README v2.0): 1000 positive and 1000 negative processed reviews. Introduced in Pang/Lee ACL 2004. Released June 2004.**

Για την κατηγοριοποίηση των εγγράφων θα χρησιμοποιήσετε παραλλαγές του Naive Bayes Classifier και με δύο τρόπους: α) με χρήση κατάλληλου λογισμικού (RapidMiner) και β) με ανάπτυξη δικού σας κώδικα.

Για το α), χρησιμοποιείστε στο RapidMiner τον κατηγοριοποιητή Naive Bayes Classifier με Laplace Correlation (για να χειριστούμε μηδενικές πιθανότητες).
Χρησιμοποιείστε 5-fold cross validation για την εκπαίδευση και αποτίμηση του μοντέλου σας και μετρείστε: accuracy, αλλά και weighted mean recall και weighted mean precision (ή απλά recall και precision) καθώς και το αντίστοιχο confusion matrix.
Συγκρίνετε την απόδοση όταν χρησιμοποιείται ως μοντέλο εγγράφου διάνυσμα με:

1. binary term occurrences (δυαδικές εμφανίσεις όρων)
2. term occurrences (πλήθος εμφανίσεων όρων)

Για το β) αναπτύξτε κώδικα, σε όποια γλώσσα προγραμματισμού προτιμάται (συνιστάται η χρήση python) ο οποίος θα υλοποιεί:

1. Multinomial Naive Bayes Classifier με μοντέλο εγγράφων διάνυσμα με term occurrences
2. Bernoulli Naive Bayes (με μοντέλο εγγράφων με δυαδικές εμφανίσεις όρων)

Χρησιμοποιείστε και σε αυτήν την περίπτωση των ίδιο τρόπο εκπαίδευσης και αποτίμησης με το α.

Για την προεπεξεργασία των δεδομένων μπορείτε να συνδυάσετε όσο βήματα επεξεργασίας θέλετε με στόχο να πετύχετε την καλύτερη απόδοση, δλδ. tokenize, stemming, filtering, stop word removal, κτλ και με διαφορετική παραμετροποίηση για το καθένα όπως και διαφορετικά επίπεδα pruning.
Μπορείτε να χρησιμοποιήσετε διαφορετική προεπεξεργασία για τα 1 και 2, αλλά κατά το δυνατό τα α.1, β.1 και α.2, β.2 θα πρέπει να ακολουθούν την ίδια προεπεξεργασία.

Θα παραδώσετε ένα zip αρχείο με:

- Την διαδικασία στο rapidminer (extract .rpm file)
- Τον κώδικα σας και readme αρχείο με οδηγίες για την εκτέλεση του κώδικα
- Μια αναφορά (pdf) στην οποία θα περιγράφονται τεκμηριωμένα οι επιλογές προεπεξεργασίας που έχετε κάνει και θα παρουσιάζονται και θα συγκρίνονται τα αποτελέσματα σας, τόσο μεταξύ των δύο μοντέλων (binary και term occurrences) όσο και των δύο τρόπων υλοποίησης με κώδικα και λογισμικό. Θα αξιολογηθεί τόσο η τεκμηρίωση των επιλογών σας όσο και η παρουσίαση των αποτελεσμάτων σας.