ΕΡΓΑΣΙΑ 5 – Κατηγοριοποίηση Κειμένου

1. Στον παρακάτω πίνακα περιγράφονται 6 έγγραφα κειμένου. Το σύνολο εκπαίδευσης έχει 5 έγγραφα, και 1 έγγραφο ανήκει στο σύνολο ελέγχου. Για το σύνολο εκπαίδευσης κάθε έγγραφο έχει ετικέτα “ναι” αν ανήκει στην κατηγορία “τηλεόραση”, και “όχι” αν δεν ανήκει. Χρησιμοποιήστε (α) τον κατηγοριοποιητή Naive Bayes (multinomial Naive Bayes) και (β) τον κατηγοριοποιητή Bernoulli Naive Bayes για να προβλέψετε την κατηγορία του 6ου εγγράφου. Γράψτε αναλυτικά τις διαδικασίες υπολογισμού στην κάθε περίπτωση και σχολιάστε τα αποτελέσματα σας.

	docID	Λέξεις του εγγράφου	τηλεόραση
Σύνολο εκπαίδευσης	1	πρόγραμμα πρόγραμμα επεισόδιο σειρά κανάλι	ναι
	2	πρόγραμμα επεισόδιο ταινία επεισόδιο	ναι
	3	επεισόδιο κανάλι επεισόδιο κανάλι ειδήσεις	ναι
	4	επεισόδιο γήπεδο ομάδα	όχι
	5	γήπεδο ομάδα ειδήσεις	όχι
Σύνολο ελέγχου	6	επεισόδιο επεισόδιο γήπεδο γήπεδο γήπεδο ειδήσεις	

2. Θα χρησιμοποιήσετε το RapidMiner για να μελετήσετε ένα πρόβλημα κατηγοριοποίησης. Το σύνολο δεδομένων σας θα αποτελείται από τη συλλογή με Newsgroups άρθρα που μπορείτε να βρείτε στον ιστότοπο http://qwone.com/~jason/20Newsgroups/. Από εκεί θα επιλέξετε να κατεβάσετε την έκδοση: **20news-bydate.tar.gz - 20 Newsgroups sorted by date; duplicates and some headers removed (18846 documents)**
Από τη συλλογή θα επιλέξετε τις εξής κατηγορίες για την εργασία σας:
comp.sys.ibm.pc.hardware
comp.sys.mac.hardware
rec.sport.baseball
rec.sport.hockey
sci.space
misc.forsale
talk.politics.misc
talk.religion.misc
alt.atheism
Τα δεδομένα είναι χωρισμένα σε training (εκπαίδευσης) και test (ελέγχου) για κάθε κατηγορία. Θα χρησιμοποιήσετε τα δεδομένα εκπαίδευσης για την εκπαίδευση του μοντέλου και θα το αποτιμήσετε στα δεδομένα ελέγχου.
Για την κατηγοριοποίηση των εγγράφων θα χρησιμοποιήσετε τον Naive Bayes classifier (με Laplace correlation στο rapidminer για να χειριστούμε μηδενικές πιθανότητες), και ως μέτρο απόδοσης θα μετρήσετε συνολικό accuracy, αλλά θα παρουσιάσετε και τα σχετικά precision και recall ανά κατηγορία.
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.

Για κάθε πρόβλημα θα πρέπει να συγκρίνετε την απόδοση:
α) όταν χρησιμοποιείται διάνυσμα με binary term occurences, term occurences και tf/idf.
β) για διαφορετικά ποσοστά pruning (τουλάχιστον 4 τιμές).
Πιθανόν διαφορετικοί τρόποι προ-επεξεργασίας να δίνουν καλύτερη απόδοση με διαφορετικού είδους διανύσματα. Επιλέξτε μια κατάλληλη προεπεξεργασία για κάθε

διάνυσμα (δεν αποκλείεται να είναι η ίδια--αλλά τεκμηριώστε πώς την επιλέξατε μέσω δοκιμών).
Θα παραδώσετε μια αναφορά, που θα περιγράφει την διαδικασία σας και θα παρουσιάζει τα αποτελέσματα σας. Θα βαθμολογηθείτε και για τον τρόπο που αιτιολογείτε και παρουσιάζετε τις επιλογές σας και τα αποτελέσματά σας.