Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Björn Buchhold

Information Retrieval
WS 2013/2014

http://ad-wiki.informatik.uni-freiburg.de/teaching

UNI FREIBURG

Exercise Sheet 10

Submit until Tuesday, January 21 at 4:00pm

This exercise sheet is about classification using Naive Bayes. See the Wiki for a code skeleton of the class *NaiveBayes* for further specifications and implementation advice. You must write unit tests for your code, otherwise it is very likely that your results are bogus. You can use the simple 6-document example from the lecture for your unit tests.

Exercise 1 (5 points)

Write a method *readDocumentsAndLabel* that reads the documents and labels from a file and builds the document vectors. Note that for Naive Bayes, you just need the sequence of word ids for each document, and no BM25 scores or anything fancy like that. You might want to reuse you code from the first exercise sheet for reading the lines and splitting them into words.

Exercise 2 (5 points)

Write a method *train* that computes the counts n_c and n_{wc} for a given (training) set of documents.

Exercise 3 (5 points)

Write a method *predict* that computes the most likely class for each document from a given (testing) set. Use the n_c and n_{wc}, learned from the training set, as explained in the lecture.

Exercise 4 (5 points)

Write a program *NaiveBayesMain* that takes a file name as single argument. The program should consider every tenth document (starting with the first) as training documents, and all others as testing documents. Compute the precision of the predictions for the testing documents. Run your program on the data set linked on the Wiki. Report the training time, prediction time, and precision in the table linked on the Wiki.

Add your code to a new sub-directory *exercise-sheet-10* of your folder in the course SVN, and commit it. Make sure that *compile*, *test*, and *checkstyle* run through without errors on Jenkins. As usual, also commit a text file *experiences.txt* where you briefly describe your experiences with this exercise sheet and the corresponding lecture. As a minimum, say how much time you invested and if you had major problems, and if yes, where.