Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Florian Bäurle

Information Retrieval
WS 2012/2013

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

UNI FREIBURG

Exercise Sheet 10

Submit until Wednesday, January 23 at 4:00pm

This exercise sheet is about implementing Naive Bayes classification for text documents and applying it to our example collection. Consider the explanations in the lecture and the code design suggestions on the Wiki.

Exercise 1 (5 points)

Write a class *NaiveBayes* with a method *readFromCsvFile* that constructs for each document a set the ids of the words it contains. Note: for the parsing, you can re-use much of your readFromCsvFile method from the previous exercise sheet. However, there is now one line per document (not per sentence), and there is an additional column for the class label.

Exercise 2 (5 points)

Add a method *train* that learns the prior probabilities $\Pr(C = C)$ and $\Pr(W = w|C = c)$ from a given subset of the documents.

Exercise 3 (5 points)

Add a method *predict* that computes the most likely class label for each from a given subset of the documents.

Exercise 4 (5 points)

Write a program *NaiveBayesMain* that reads the CSV file linked on the Wiki (with one line per document, and three columns: URL, class label, contents), takes every tenth of those documents with only one class label as the training set, and predicts the labels for the remaining documents. Output the percentage of correct predictions (a prediction for a document with several class labels is correct if the prediction is one of those labels), and briefly discuss the results in your *experiences.txt*. Add your prediction percentage to the result table linked on the Wiki, along with the time needed for reading, training, and prediction.

Commit your code to our SVN, in a new sub-directory *exercise-sheet-10*, and as usually make sure that Checkstyle and Tests run through without errors on Jenkins. Also commit a text file *experiences.txt* with your feedback. As a minimum, say how much time you invested and what are the symptoms of the current influenza virus.