ifis
Institut für Informationssysteme
Technische Universität Braunschweig

Information Retrieval and Web Search Engines
Summer Semester 2010

Prof. Dr. Wolf-Tilo Balke and Joachim Selke

Homework Assignment 5

Due to July 8, 2010
(36 points in total)

Note: We still highly recommend to use MATLAB for working on this assignment. However, you are free to use any tool and programming language you want to (as long as you explain what you did and hand in any relevant code). Of course, you are also free to use any MATLAB libraries you like.

Note 2: If you encounter any significant performance problems, let us know! We can give you hints on how to write more efficient code or give you pointers to the right tools.

Exercise 5.1 (Document Classification)

The documents of the Reuters-21578 collection have been classified into 120 categories/topics. The MATLAB data file *reuters-21578-stemmed-with-topics.mat*, which is available for download at `http://www.ifis.cs.tu-bs.de/webfm_send/478`, contains the whole (stemmed) Reuters data set as well as a list of all topic titles (cell array *topics*) and an assignment of topics to documents (sparse binary matrix *ToD*). Moreover, the data file contains two reweighted versions of the term–document matrix *TDStemmed*: TF–IDF (sparse matrix *TDStemmed_TFIDF*) and log entropy (sparse matrix *TDStemmed_LE*).

This exercise is about evaluating different binary classification algorithms on the Reuters collection.

a) Randomly split the document collection into a training and a test set. The test set should contain about 30% of the whole collection.

   Hint: The MATLAB function *randperm* might be useful. (2 points)

b) Choose three topics that you think are well-suited for evaluating classifiers. Please explain your reasoning.

   Hint: For example, the topics *bfr*, *castorseed*, and *citruspulp* would be a poor choice. (3 points)

c) Evaluate the classification performance of the Rocchio classifier on the Reuters data set (use cosine similarity to measure distances). Use the training/test data just created as well as your selection of topics. Try the following three document representations: raw term frequencies (matrix *TD*), TF–IDF (matrix *TD_TFIDF*), and log entropy (matrix *TD_LE*). For each topic and each kind of document representations, measure the classifier's performance on the test set using the balanced F-measure (α = 0.5). (5 points)

d) Repeat exercise (c) for naive Bayes classification. (5 points)

e) Repeat exercise (c) for kNN classification. Use at least three different values for k. (8 points)

f) Repeat exercise (c) for (soft-margin) support vector machines. Use two different kernels: a

linear kernel and a Gaussian radial basis function kernel with parameter $\gamma = 1$ (sometimes this parameter is denoted by σ).

Hint: Depending on your tools, training SVM classifiers could take a long time. If you encounter those problems, you may reduce the size of your training set appropriately. (10 points)

g) Summarize your findings of the previous exercises. Which classifiers work best? Which document representation seems to be most useful for purposes of classification? (3 points)