Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Marjan Celikik

Search Engines WS 09/10

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

UNI FREIBURG

Exercise Sheet 11

complete until Thursday, January 28st

On the Wiki we will provide a text file with a few thousand publication records, each containing the title of the publication (free text), as well as the name of the conference where the publication appeared.

Exercise 1

Write a program that does the following. Take a random 10% of the records as training data (pick every 10th document). From that training data, learn the probabilities of a Naive Bayes classifier for telling which publications appeared at which conference. Use the words of the title as features, just as shown in the lecture.

Exercise 2

Write a program that, using the probabilities learned in Exercise 1, predicts the conference for the remaining 90% of the records. For each conference, identify the most predictive word, that is, for each c the w with the highest $\Pr(W = w|C = c)$.

Exercise 3

Compute the precision and recall figures of your prediction, and display them in a table with one column per conference and three rows: one for the precision, one for the recall, and one for the F-measure.

Exercise 4

Assume you have a random device with k possible outcomes $E_1, \ldots, E_k$. Now you have activated it n times producing a sequence E of outcomes, where outcome E_i occurred n_i times (that is, $n_1 + \cdots + n_k = n$). Assume that the outcomes are independent of each other, and denote the probability of outcome E_i by p_i. Let L be the probability that your sequence E occurred. Prove that L is maximal when $p_i = n_i/n$.

Hint: just proceed as for the heads and tails example in the lecture. Use Lagrangian optimization under side constraints just as shown in the lecture.