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 1

complete until Tuesday, October 27th

Exercise 1

Create a user account on the course Wiki (linked from the course web page) and put a short introduction of yourself on the current page together with a link to your favorite cartoon / joke.

Exercise 2

Download (or otherwise get) a collection of text documents of your choice. The collection should have at least 1000 documents, but the more the better.

Exercise 3

Implement a simple parser and an inverted index in the programming language of your choice. Use it to build an inverted index for your collection from Exercise 2.

Exercise 4

Use your inverted index to find two-word queries with one hit, that is, exactly one document that contains both of these words. There are various ways to do this, take the most efficient one you can think of. What is the asymptotic complexity of your algorithm?

Exercise 5

For each word that occurs at least once in the collection, compute its frequency = its number of occurrences. Sort the words by their frequencies and produce a plot showing these frequencies. Estimate the constant factor ε of Zipf's law, as explained in the lecture, for your collection.

Exercise 6

Upload your solutions (including your source code) to the Wiki following the instructions given there.