

---

## Exercise Sheet 1

complete until Friday, October 21st, 5 pm (Ex. 1) and Monday, October 24th, 9am (Ex. 2 & 3)

**Exercise 1** Get yourself a *large* document collection, preferably HTML, on something that you find personally interesting. Come up with a search question which can be answered from your collection, but not easily so with standard web search (be creative here, but not mean). Post this question on the Wiki, and upload a compressed archive of your collection (follow the instructions which will be given on the Wiki). *The deadline for this part is Friday, October 21st, 5 pm.*

Based on your input we will set up an instance of our autocompletion search engine, and post a link to it on the Wiki. You should then pick three of the questions posted by others and try to answer them with the help of this engine.

**Exercise 2** For your collection, produce a text-file version of an inverted index, as we have seen and discussed it in the first lecture (refer to the Wiki for details). From that text file produce a number of basic statistics: the number of distinct words, the number of documents, the average number of (distinct) words in a document. What is the size of your *compressed* text file relative to the size of the collection? Also, for each word, count in how many documents it occurs and produce a *histogram* of these counts. Come up with at least one more parameter that you think might be interesting.

Make use of the seminar Wiki (which is linked from the seminar homepage) to get help on any problem with these exercises, or to help others if you can. You will most probably run into one practical problem or the other, but where others can easily help you. If you start working on Exercise 2 Sunday evening, it will be too late to get help, so don't do that.

**Exercise 3** Prepare *two slides* (PowerPoint or PDF) with the following contents, and upload them on the Wiki by Monday, October 24th, 9 am (that is, Monday *morning*).

- (a) On the first slide, *introduce yourself*: your name, where you are from, how you came to Saarbrücken ... anything you feel like telling about. In particular, say *which programming language you are familiar with and how well*, and whether you have previously *taken a course on information retrieval or information systems*.
- (b) On the second slide, tell us about the name and contents of your collection, and summarize your statistics from Exercise 1. Briefly report on your experience with the autocompletion engine, and whether you were successful in answering the three questions you picked.