Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Claudius Korzen

Information Retrieval
WS 2017/2018

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

UNI FREIBURG

Exercise Sheet 3

Submit until Tuesday, November 14 at **12:00pm**

For this exercise sheet, you can use Java or C++, **but not Python**. Subtle algorithmic improvements and performance tuning make little sense in Python, due to the large overhead of the underlying data structures, in particular of the Python lists/arrays.

Copy the code provided on the Wiki (which is similar to the code of the lecture, without the performance improvements) and proceed from there. It is available in both Java and C++ and contains (1) the method *readFromFile()* for reading posting lists from files and the method *intersectBaseline()* for intersecting posting lists using a baseline algorithm, (2) code for time measurement and (3) tests for the provided methods.

Exercise 1 (20 points)

Imagine you saw a film a long time ago, but you can't really remember its title. The only thing you can remember is that *bowling* and a *rug* played an important role. Based on this information, you want to find out the title of the film (as fast as possible).

1. (10 points) Implement a method *intersect()* that intersects two posting lists by using at least three non-trivial ideas presented in the lecture. The goal is to beat the baseline algorithm for all scenarios. Note that you can also implement several methods and switch between them, e.g. depending on the list sizes (or on any information that you find to be useful). Each implemented method must pass the test case provided for the baseline implementation in the example code!

2. (10 points) Evaluate your algorithm on all three pairwise combinations of the three posting lists *film.txt*, *bowling.txt* and *rug.txt* provided on the Wiki. For each combination, repeat the baseline algorithm and your improved algorithm 5 times, measure the runtimes of both and take the averages. Compute the speedup ratio of your algorithm compared to the baseline algorithm. Report your results in the table on the Wiki, following the rows already there.

3. (Optional) Compute the intersection of *all* three posting lists in order to find out the id of the Wikipedia article that refers to the film you are looking for. Navigate to Wikipedia (*http://en.wikipedia.org/?curid=$id*) and find out the title of the film. Watch the film and count how often the fact that "the rug really tied the room together" is mentioned. Write down the number to your *experiences.txt*.

[the dude wants you to turn over]

As usual, commit your code to a new sub-directory *sheet-03* of your folder in the course SVN. Summarize your experiences and insights in your *experiences.txt*. As a minimum, say how much time you invested and if you had major problems, and if so, where. Please be brief and informative: that makes reading your feedback much more pleasant for us.