Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Patrick Brosi

Information Retrieval
WS 2019/2020

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

UNI FREIBURG

Exercise Sheet 1

Submit until Tuesday, October 29 at **12:00 noon**.

Before you start working on the exercise sheets, please **carefully and completely** read the rules for this course on our Wiki, they are valid for all exercise sheets of this course:

http://ad-wiki.informatik.uni-freiburg.de/teaching/InformationRetrievalWS1920/Rules

Exercise 1 (15 points)

Extend the code from the lecture by the following functionality.

1. Make sure that each inverted list contains a particular record id at most once, even if the respective word occurs multiple time in the same record. Make sure that the whole construction algorithm still runs in time linear in the number of words in the input. (2 points)

2. Write a function *intersect* that computes the intersection of two inverted lists. The function must run in time linear in the total number of elements in the two lists and you must not use a library function. In particular, don't implement the lists as *sets*. (5 points)

3. Write a function *process_query* that, given a keyword query, fetches the inverted lists for each of the keywords and computes their intersection (empty, if there is no inverted list for a keyword), using your *intersect* function. (5 points)

4. Write a *main* function that constructs an inverted index from a given text file (one record per line, file name given as first argument on the command line) and then, in an infinite loop, asks the user for keyword queries and outputs the title and description of up to three matching records. If there are more than three matching records, output those which come first in the given text file. In any case, the order in the output should correspond to the order in the given text file. (3 points)

Optionally (= not mandatory to get full points) highlight the query words in the output, e.g., using ANSI escape codes.

[please turn over with beginner's mind]

Exercise 2 (5 points)

Try your code on the file *movies.txt* provided on the Wiki. Find a query that gives good results (the records shown meet your expectations) and one that does not. Write them in your *experiences.txt* (see below), and very concisely (in one or two sentences) explain why one works and the other doesn't.

Register with our course system Daphne (using your Uni-Account + password for authentication). Make sure that you specify an email adress under which you are reachable during the semester.

Check out a working copy of your folder in the SVN repository of the course, and add your code to a new subdirectory *sheet-01*, and commit it. Make sure that everything runs through without errors on Jenkins. This was briefly demonstrated towards the end of Lecture 1.

Also commit, in that subdirectory, a text file *experiences.txt* where you briefly describe your experiences with the first exercise sheet and the corresponding lecture. As a minimum, say how much time you invested and if you had major problems, and if yes, where.