Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Florian Bäurle

Information Retrieval
WS 2012/2013

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

UNI FREIBURG

Exercise Sheet 13

Submit until Wednesday, February 6 at 4:00pm

Complete official Online-Evaluation until Friday, February 8 !

Exercise 1 (20 points)

Complete the official Online-Evaluation for this course. You find a link + instructions on the Wiki. Please take your time, and be honest and concrete. The evaluation is anonymous. Just confirm in your *experiences.txt* that you have completed the evaluation, and we will trust you and give you the points for this exercise. Please complete the evaluation by Friday, February 8 !

The rest of this exercise sheet is about determining whether a simple database query optimization trick (using integer ids instead of string ids) gives a statistically significant performance improvement or not. This is the last exercise sheet for this course, and it's quite useful to really understand the concept of statistical significance.

Exercise 2 (5 points)

Download the latest version of the Freebase TSV data for *acted_in* and *has_won* from the Wiki (no more empty columns or quotes now). Create two variants of the tables, one where person names are replaced by integer ids, and one where they are replaced by hexademical ids. In both cases, the property of the original files should be preserved that each person has a unique id. Consider the advice given towards the end of the lecture on how to do this easily with a combination of *cut*, *sort*, and *join* from the Linux command line.

Exercise 3 (5 points)

Import the modified tables into *Sqlite*. When creating the tables, make sure to use *TEXT* for the hexadecimal ids, and *INTEGER* for the integer ids.

Formulate the straighforward SQL query for the list of actors and all their awards (a simple join of both tables on their first column). Issue this query ten times for both variants of the tables (with hexadecimal ids and with integer ids).

[please turn over]

Exercise 4 (5 points)

Use a Z-test, as explained in the lecture, to determine the statistical significance of the superior performance of integer ids over hexadecimal ids. Do two versions of the test: one where you consider only the first three measurements for each variant, and one where you consider all ten. Determine the corresponding p-values (either via table lookup or using something like Wolfram Alpha) and enter them in the result table on the Wiki.

Exercise 5 (5 points)

Repeat Exercise 4 with the Student's T-test, as explained in the lecture. Also determine the corresponding p-values for this test and enter them in the result table on the Wiki.

As usual, briefly discuss your results in your *experiences.txt* in a new subfolder *exercise-sheet-13*. Along with that, give your usual feedback on the exercise sheet and the lecture. Don't forget to say that you completed the official course evaluation in case you did.