CONCORDIA UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE AND SOFTWARE ENGINEERING

Information Retrieval and Web Search **COMP 479/6791**

Project 1

Due Date: 18.10.2013

Description:

1. Implement the SPIMI algorithm with disk block merging.
2. Compile an inverted index for Reuters21578 without using any compression techniques (and make memory size a parameter that you can artificially reduce to test your code)
3. Implement the lossy dictionary compression techniques of Table 5.1 in the textbook and compile a similar table for Reuters-21578. Are the changes similar? Discuss your findings. (Note that stemming is not required here, if you run out of time before you get the Porter stemmer to work, that is ok for this assignment, the remaining table is fine.)
4. Implement a simple scheme to retrieve matching documents for a few queries. Techniques from Chapters 1-3 are suitable. Show the queries you used and discuss your findings.

Deliverables:

1. individual project, remember to submit the expectations of originality page
2. well documented code
3. well documented sample runs of the queries posted on 15.10.
4. any additional testing or aborted design ideas that show off particular aspects of your project
5. a project report that summarizes your approach, illustrates your design and discusses what you have learned from the project
6. a four slide overview over your system for the lab demo

Test queries:

1. design three test queries and run them to showcase your code
2. exchange one test query each with three different students in the class, run their queries
3. compare your results during lab time
4. report the experiment in your project report

Submissions: All code has to run on the lab equipment. You have to demo your project during lab time on 21.10-23.10. Submit all your deliverables through the Moodle link no later than 18.10.2013 (Midnight)