Comp 479/6791 # Project 1 **Fall 2016**

Objectives: Implement SPIMI. Implement rudimentary information retrieval using your SPIMI indexer. Test and analyze your system, discuss how your design decisions influence the results.

Due Date: 18.10.2016 noon

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, see below)

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 17.10.

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 three slide overview over your system for the lab demo

Test queries:

1. design three test queries:

   (a) a single keyword query,
   (b) a multiple keywords query returning documents containing all the keywords (AND),
   (c) a multiple keywords query returning documents containing at leat one keywword (OR), where documents are ordered by how many keywords they contain)

   .

2. run your three test queries to showcase your code and comment on the results in your report

3. exchange one test query each with three different students in the class, run their queries

4. compare your results during lab time

5. report the experiment in your project report

Submissions: submit to Moodle. All code has to run on the lab equipment. You have to demo your project during lab time on 18/19/20.10. or as agreed with your lab instructor.

Marks:

Preprocessing (text extraction, tokenization)	1pt	Attr1
Spimi implementation	2pts	Attr1
Inverted index	1pt	Attr1
Memory size limitation	1pt	Attr5
Dictionary compression	2pts	Attr1
single keyword query	1pt	Attr1
multiple keyword AND query	1pt	Attr1
multiple keyword OR query	1pt	Attr1

Memory limitation simulation:

A way to artificially reduce memory size:

Make two parameters to your code: (1) the memory size and (2) the block size

Python: use

... if (sys.getsizeof(blockContent)/1024/1024) $\geq$ blockSizeLimit ...

Java: use

```
...
initialMemory = java.lang.Runtime.getRuntime().getFreeMemory();
...
currentMemory = java.lang.Runtime.getRuntime().getFreeMemory();
usedMemory = initialMemory - currentMemory;
if (usedMemory1024/1024) ≥ blockSizeLimit:
...
```