Information Retrieval

2022/2023

Assignment 2

Submission deadline: **25 November 2022**

For this assignment, you will add term weighting to your indexing system and implement two ranked retrieval models.
Use the datasets from assignment 1 (start with the smaller one).

1. Extend your indexer to apply term weighting and implement the following document scoring and ranking methods:

   1.1. Vector space ranking with *tf-idf* weights. Implement the *lnc.ltc* indexing schema as default and at least one alternative schema. The indexing schema to use should be passed as a parameter, using the SMART notation.

   1.2. BM25 ranking. Use k1=1.2 and b=0.75 as default values for the parameters and allow specifying other values through command line arguments.

2. Implement the search component of your retrieval system. This *Searcher* should read a previously created index – following one of the models implemented in 1. – and process queries, returning a *paginator* with the results (10 per page).
   Create a test code that starts the search mechanism and continually (in a loop) accepts user queries from the command line and presents the top 10 results, including the document scores.

3. Index, separately, each of the files used in assignment 1 and gather the same statistics:
   a) Total indexing time
   b) Merging time (last SPIMI step)
   c) Number of temporary index segments written to disk (before merging)
   d) Total index size on disk
   e) Vocabulary size (number of terms)

Instructions:

- **Modelling**, code **structure**, **organization** and **readability** will be considered when grading your project
- **Comment** your code; and make sure you include your name and student number
- Write **modular** code
- Favour **efficient** data structures
- Use **parameters**, preferably through the command line
- Make sure all your programs run correctly
- Submit your assignment by the due date using Moodle