Information Retrieval

2022/2023

Assignment 3

Submission deadline: **06 January 2023**

For this assignment, you will continue extending your indexing and retrieval system.
Use the datasets from assignment 1 (start with the smaller one).

1. Extend your indexer to store term positions. The index should hold the information shown here:

   term;doc_id:term_weight:pos1,pos2,pos3,…;doc_id:term_weight:pos1,pos2,pos3,…

2. Extend your ranked retrieval method to boost the scores of documents using the minimum window size, that is, the smallest text span in the document that contains all search terms (see note below).
   Use a multiplicative boost factor that has a maximum value *B* when the minimum window size corresponds to the number of distinct terms in the query and should decrease rapidly with the window size (the largest the window, the smallest the boost). For large values of the window size, and when the document does not contain all search terms, the boost factor should be 1.

   Note: Since we are using natural language questions as queries, consider only high IDF terms when finding the *minimum window*. For the question “Which phosphatase is inhibited by LB-100?”, for example, the terms considered when finding the minimum window would be “phosphatase inhibited LB-100” (assuming these terms have high IDF).

3. Evaluate your retrieval engine using the queries and the relevant documents provided (‘questions_with_gs’). Compare the *tf-idf* and BM25 ranking methods, with and without minimum window boost from 2, in terms of the following evaluation and efficiency metrics, considering the top 10, 50 and 100 retrieved documents:

   Precision
   Recall
   F-measure
   Average Precision (AP)
   Query throughput
   Median query latency

   The metrics should be implemented by you and not obtained through existing libraries.
   For precision, recall, f-measure and average precision, report the mean over all queries considering the top 10, 50 and 100 retrieved documents, as illustrated in the table below.
   Note that Average Precision is not the mean of the precision values, but a different metric.

	Top-k		
	10	50	100
Precision			
Recall			
F-measure			
Average Precision			

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