Recuperação de Informação / Information Retrieval

2021/2022 MEI/MIECT, DETI, UA

Assignment 3

Submission deadline: **01 February 2022**

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

1. Extend your indexer to store term positions.
   Write the index to file using the following format, or a similar one (one term per line):

   ```
   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 that contain two or more of the query terms within a text window. Your boost function should consider the size of the text span and the number of query terms contained within that span.

3. Evaluate your retrieval engine using the queries from Assignment 2 and the relevance scores provided (file 'queries.relevance.txt'). Compare the tf-idf and BM25 ranking functions, with and without the term proximity boost from 2, in terms of the following evaluation and efficiency metrics, considering the top 10, 20 and 50 retrieved documents:

   i. Precision
   ii. Recall
   iii. F-measure
   iv. Average Precision (AP)
   v. Normalized Discounted Cumulative Gain (NDCG)
   vi. Average query throughput
   vii. Median query latency

   Note: Report the mean over all queries, for the top 10, 20 and 50 retrieved documents.

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