Recuperação de Informação / Information Retrieval

2020/2021 MEI/MIECT, DETI, UA

Assignment 2

Submission deadline: **26 November 2020**

For this assignment, you will create a weighted (tf-idf) indexer and a ranked retrieval method. Use the dataset from assignment 1.

1. Extend your indexer to apply term weighting and implement the following ranking methods.
   1.1. Vector space ranking with tf-idf weights. Use the *lnc.ltc* indexing schema.
   1.2. BM25 ranking. Use k1=1.2 and b=0.75 as default parameters
   1.3. Add a method to write the resulting index to file. Use the following format, or a similar one (one term per line):
   `term:idf;doc_id:term_weight;doc_id:term_weight;…`
2. Evaluate your retrieval engine, comparing both ranking functions.
   2.1. Process the queries (file 'queries.txt') and retrieve the sorted results for each query.
   2.2. Using the relevance scores (file 'queries.relevance.txt') provided, calculate the following evaluation and efficiency metrics, considering the top 10, 20 and 50 retrieved documents:
   a) Mean Precision
   b) Mean Recall
   c) Mean F-measure
   d) Mean Average Precision (MAP)
   e) Mean Normalized Discounted Cumulative Gain (NDCG)
   f) Query throughput
   g) Median query latency

Instructions:

- Use Python or Java (in this case, manage your project with Maven)
- **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 compile and run correctly
- Submit your assignment by the due date using Moodle