German University in Cairo
Media Engineering and Technology
Dr. Mohamed El Mahdy
Eng. Maged Shalaby

Natural Language Processing and Information Retrieval, Winter 2017
Programming Assignment2 (Individual)

Deadline: 20.10.2017 23:59

Your task in this assignment is to implement and evaluate a ranked search engine as discussed in Lecture 3 and Lab 4 (solutions can be found on the MET website on "Lab 3+Lab 4 (Pt.1) - Sol. Jupytr" , and "Lab 4 (Pt. 2) - Sol. Jupytr".

You will be using the dataset "cranfield.zip" on the MET website for the purpose of this assignment.

Inside, you will find 3 types of files:

a) 1400 text documents (aeronautical engineering abstracts) in the folder "docs", where each file corresponds to a single document

b) 225 queries in the folder "queries", where each file corresponds to a single query

c) Relevance judgments in the file "cranqrel", where each line means that for a certain query, a given document is relevant to it

The goal is to return a ranked list of relevant documents for each query from the text documents. Following that, the Mean Average Precision should be calculated, where the *reference* or *gold standard* is obtained from the relevance judgments.

The relevance judgments file has the following format:

{Query-id} {Doc-id} {Relevance score}

This means that a given document is relevant for the given query. If a query-document pair is not present in the file, then this means that the document is not relevant for the query. You can ignore the relevance score in your calculations.

As an example, taking a subset from the file into Table 1:

Table 1: Subset from cranqrel

1	184	2
1	29	2
1	31	2
2	12	1
2	15	2
2	184	2

This means that for this subset, only documents [184,29,31] are relevant for query 1, and only documents [12,15,184] are relevant for query 2.

Some steps are common between the assignment and what was discussed in the fourth lab. For this assignment the steps should be:

a) Read all document files

b) Transform all documents to TF-IDF

c) Read all queries

d) Transform all queries (using the same vectorizer) using TF-IDF

e) Compute cosine similarity between all pairs of (query, document)

f) Return a ranked list of relevant documents for each query

g) Evaluate the ranked results against the relevance judgments using Mean Average Precision

You are not allowed to use external libraries for the calculation of the Mean Average Precision, however you can check if your calculations are correct using the library `ml_metrics` as shown in Lab 4. You will find an example how to use it in the uploaded "Lab 4 (Pt. 2)" solution. The MAP score for this assignment shouldn't be below 0.1 .

You might find the function `argsort` from the library `numpy` useful in the ranking of results, but note that it returns a sorted 0-based list, while the document ids start from 1.

The original document and query formats were converted for easier processing as the assignment is not about string parsing. If you will read the queries and documents from the original formats, found in folder "original", you will get 2 bonus points.

You should submit your code as a Jupyter notebook file on the assignment on the MET website, by Friday 20.10.2017 at 23:59. Do not submit the dataset; just submit the notebook file.

You should name your file as: Your ID - Your Lab Group - Your Name

Best of luck!