IR Challenge

Homework Submission Guidelines

1. **Due date: 20/01/18**
2. The challenge can be done in groups of **up to** 2 students
3. **Submission instructions are on the last page**
4. Challenge submission should be done via moodle in the corresponding area (by **only** one of the students)
5. Late submission will not be considered
6. Questions / clarifications and more in the dedicated discussion sub-forum.

The challenge

1. For this challenge we will use the Azure environment.
2. ROBUST index is located in /data/IRCompetition/ROBUSTindex
3. **Note: Do not** copy the index to your home directories!

4. The files are located in /data/IRCompetition/
5. Inside the folder you will find the following files:
   a) "queriesROBUST.xml" – 249 queries.
   b) "qrels_50_Queries" – the ROBUST relevance judgments for the first 50 queries.

Goal:
Achieving the highest retrieval effectiveness as measured using **MAP**.

Task:
There are 249 queries, but we provide the relevance judgments for the first 50 queries only (50 queries - train, 199 queries – test). You are required to return a ranked list of documents from the collection ordered in the decreasing probability of relevance.
For each test query (of the 199 queries), you submit a ranking of the top **1,000 documents**. Your retrieval will be evaluated based on the 199 test queries.

The relevance judgments of the first 50 queries can be used to train your retrieval method.

You may submit up to **3** different result lists (runs).
Each file name is of the form: run_i.res (i is set to value in {1,2,3}).
The format of each file is the standard 6 columns TREC format:

```
630 Q0 ZF08-175-870  1 0.7 run1
630 Q0 ZF08-306-044  2 0.5 run1
630 Q0 ZF09-477-757  3 0.3 run1
630 Q0 ZF08-312-422  4 0.1 run1
630 Q0 ZF08-013-262  5 -0.3 run1
 etc.
```

where:

- the first column is the topic number.
- the second column is currently unused and should always be "Q0".
- the third column is the official document identifier.
- the fourth column is the rank the document is retrieved
- the fifth column shows the score (integer or floating point) that generated the ranking. This score MUST be in descending (non-increasing).
- the sixth column is called the "run tag" (can be set to your run_i)

All runs must be **compressed (zip)**

Tools:

1. You can use Indri toolkit / Lemur toolkit.
2. You can write your own ranking algorithm using any programing language that you feel most comfortable or to use the those learned in class.
3. **Important:** your algorithm must be reproducible.

Good Luck