Homework 1

Data Mining Technology for Business and Society

Deadline: **28 April 2020 23:59 (Rome Time Zone)**

Having **TWO** students per group is **RECOMMENDED**.

Having one student per group is allowed only if coordination is not possible.

The total length of the report cannot exceed 5 pages.

It is forbidden to print or store this document, you can only read this document online.

It is forbidden to submit software written with Python-Notebook.

Only ".py" software is considered as a valid solution.

The software must be commented.

Data and software are available at:

http://www.diag.uniroma1.it/~fazzone/Teaching/Data_Mining_Technology_for_Business_and_Society_2019_2020/DMT4BaS_2019_2020.html

The homework is composed of two parts: Search-Engine Evaluation and Near-Duplicates-Detection.

Part 1

You have to index two collections of documents and improve the search-engines performance by changing their configurations. Using the provided sets of queries, and the associated Ground-Truths. For this part of the homework you must use the Whoosh API.

The Two Collections of Documents

The two different collections of documents are: `Cranfield_DATASET` and `Time_DATASET`. They consist of:

..) a set of html documents.

..) a set of queries.

..) a set of relevant documents identifiers for each of a subset of the query set - the Ground-Truth.

The ground truth does not provide the set of relevant documents for all queries in the query set.

Documents, Queries and Ground-Truth

The documents to index are stored in html files and they are composed of two fields: **content** and **title** (please, open them with a text-editor and not with a browser). The content of the "title" field is located between the "`<title>`" tags and the content of the "content" field is located between the "`<body>`" tags. The **document-id** is the integer number at the end of the html file name. For instance, for the `Cranfield_DATASET`, the file with name "`_____42.html`" contains the document with ID "`42`", title "`the gyroscopic effect of a rigid rotating propeller… `" and content "`in many wing vibration analyses it is found necessary…`". All documents are stored inside the "`DMT/HW_1/part_1/<COLLECTION_NAME>/DOCUMENTS`" directories.

Queries are stored in the "DMT/HW_1/part_1/<COLLECTION_NAME>/<COLLECTION_NAME>_Queries.tsv" file and the ground-truth is stored inside the "DMT/HW_1/part_1/<COLLECTION_NAME>/<COLLECTION_NAME>_Ground_Truth.tsv" file. These two files are linked by the "Query_id" field value.

An Important consideration for Time_DATASET.

The content of the field "title" is not informative. The content of this field **must not** be taken into consideration.

Evaluation Metrics

For each configuration, you must provide the following **"MRR table"**:

Search Engine Configuration	**MRR**
conf_x	?.???
conf_y	?.???
conf_z	?.???
...	?.???

Only for the Top-5 configurations in the "MRR table" (the ones with the best five MRR values), you must provide the following information:

.) **"R-Precision distribution table"**, with the following format:

Search Engine Configuration	Mean (**R-Precision**_Distrbution)	min(**R-Precision**_Distrbution)	1° quartile (**R-Precision**_Distrbution)	MEDIAN(**R-Precision**_Distrbution)	3° quartile (**R-Precision**_Distrbution)	MAX(**R-Precision**_Distrbution)
conf_w	?.???	?.???	?.???	?.???	?.???	?.???
conf_t	?.???	?.???	?.???	?.???	?.???	?.???
conf_z	?.???	?.???	?.???	?.???	?.???	?.???
...	?.???	?.???	?.???	?.???	?.???	?.???

.) The **"P@k plot"**, where:
  .) the x axis represents the considered values for k: you must consider k ϵ {1, 3, 5, 10}
  .) the y axis represents the average (correctly normalized) P@k over all provided queries.
  .) Each curve represents one of the Top-5 search engine configurations (according to the "MRR table").

.) The **"nDCG@k plot"**, where:
  .) the x axis represents the considered values for k: you must consider k ϵ {1, 3, 5, 10}
  .) the y axis represents the average nDCG over all provided queries.
  .) Each curve represents one of the Top-5 search engine configurations (according to the "MRR table").

Information to Provide in the Report

For both `Cranfield_DATASET` and `Time_DATASET`, you have to provide in the report the following information:

.) Number of indexed documents and the number of queries.
.) Number of queries in the Ground-Truth.
.) A schematic description of **all** tested search engine configurations.
.) The "MRR table" for **all** tested search engine configurations.
.) The set of all Top-5 search engine configurations according to the "MRR table".
.) The "R-Precision distribution table" with data from the Top-5 search engine configurations according to the "MRR table".
.) The "P@k plot" with data from the Top-5 search engine configurations according to the "MRR table".
.) The "nDCG@k plot" with data from the Top-5 search engine configurations according to the "MRR table".

You must provide all this information in **at most three pages**.

Part 2

You have to find, in an approximated way, all near-duplicate documents inside the following dataset: `/DMT/HW_1/part_2/dataset/250K_lyrics_from_MetroLyrics.csv` .
The dataset contains data on **250K** songs.
Two songs are considered near-duplicates if, and only if, the Jaccard similarity between their associated sets of shingles computed **only on their lyrics** is **≥0.89**.
To complete this part of the homework, you have to use the **Near_Duplicates_Detection_Tool** that is entirely contained inside the directory "`DMT/HW_1/part_2/tools`". The file "`DMT/HW_1/part_2/script_for_testing.txt`" contains a short description and an example on how to run the **Near_Duplicates_Detection_Tool**. Moreover, the file "`DMT/HW_1/part_2/dataset/1K__test_sets_for_LSH.tsv`" contains a representation of 1000 documents as sets of `shingle_IDs` and can be used **only** for testing the **Near_Duplicates_Detection_Tool**.
For creating hash functions you can use the following software: "`DMT/HW_1/part_2/hash_functions_creator.py`".

Details on Shingling

For representing a song as a set of shingles identifiers in a correct way, you have to assign a natural number IDENTIFIER to each distinct shingle you generated by processing all 250K documents. I suggest you use as shingle identifier a natural number that spans from 0 to the number of distinct shingles you generated minus one: `0, 1, 2, 3, ... , number_of_all_observed_distinct_shingles-1`.
Before shingling a document, it is required to remove punctuations and convert all words in lower-case, moreover, stopword removal, stemming and lemmatization are forbidden. The length of each shingle **must be 3**.
You have to shingle **only** the lyric of the song.

Details on Sketching

Constraint 1: Each set of shingles, that represents an original document, must be sketched in a Min-Hashing sketch with a length of at most **300**.

Details on LSH

Constraint 2: The probability to have as a near-duplicate candidate a pair of documents with Jaccard=0.89 **must be > 0.97**.

Information to Provide in the Report

You have to provide in the report the following information:
 .) The number of rows and the number of bands that you chose.
 .) The probability to have False-Negatives, in the set of candidate pairs, for the following Jaccard values: 0.89, 0.9, 0.95 and 1.
 .) The probability to have False-Positives, in the set of candidate pairs,
    for the following Jaccard values: 0.85, 0.8, 0.75, 0.7, 0.65, 0.6, 0.55 and 0.5.
 .) How did you reduce the probability to have False-Negatives?
 .) The Execution-Time of the Near-Duplicates-Detection tool.
 .) The number of Near-Duplicates couples you found.
 .) The number of Near-Duplicates couples you found with an approximated Jaccard similarity value **of at least** 0.89, 0.90, 0.91, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98, 0.99, 1.
You must provide all these information in **at most two pages**.

Where/What To Send

At the end of the process, you have to create a **zip** file with **ONLY** the following data:
1. The software for addressing Part_1: /DMT_2020/HW_1/part_1/sw/ (.py files).
2. The software for addressing Part_2: /DMT_2020/HW_1/part_2/sw/ (.py files).
3. The **COMPRESSED** tsv file you created for addressing Part_2 that contains the sets of shingles identifies: /DMT_2020/HW_1/part_2/data/ (compressed .tsv files).
4. The **COMPRESSED** tsv file containing the Near-Duplicates you found for Part_2: /DMT_2020/HW_1/part_2/data/ (compressed .tsv files).
5. The final report in **PDF**: /DMT_2020/HW_1/report.pdf .

The name of the zip file must have this format:
DMT_2020__HW_1__StudentID_StudentName_StudentSurname_StudentID_StudentName_StudentSurname.zip
Finally you must send the ".zip" file to fazzone@diag.uniroma1.it with the following email subject:
DMT_2020__HW_1__StudentID_StudentName_StudentSurname_StudentID_StudentName_StudentSurname.