Homework 1

Data Mining Technology for Business and Society

Deadline: **21 April 2021 23:59 (Rome Time Zone)**

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

The total length of the report cannot exceed 7 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_2020_2021/DMT4BaS_2020_2021.html

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

Part 1.1

You have to index two collections of documents and **improve** the search-engines performance (*the higher the better*) by changing their configurations. You can change a search-engine configuration by using different combinations of Text-Analyzer and Scoring-Function (**for a maximum of 12 combinations in total**). In particular, it is forbidden to use the Frequency scoring function in more than one configuration. The performance must be evaluated 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 **only for SOME** of the queries in the query set: the Ground-Truth.

WARNING: The ground truth does not provide the set of relevant documents for all queries in the query set: YOU MUST TO CONSIDER THIS FACT IN THE COMPUTATION OF THE EVALUATION METRICS!

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/part_1_1/<COLLECTION_NAME>/DOCUMENTS" directories.
Queries are stored in the "DMT/HW_1/part_1/part_1_1/<COLLECTION_NAME>/<COLLECTION_NAME>_Queries.tsv" file and the ground-truth is stored inside the "DMT/HW_1/part_1/part_1_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	?.???
...	?.???

For each configuration, you must provide the following "**R-Precision distribution table**":

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	?.???	?.???	?.???	?.???	?.???	?.???
...	?.???	?.???	?.???	?.???	?.???	?.???

For the **Top-5** search engine configurations, **according to the MRR** evaluation metric, you must provide the following plots:

.) 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 (correctly normalized) 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" for **all** tested search engine configurations.
.) 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".
.) According **only** to the "nDCG@k plot", which is the best search engine configuration? Explain your answer in at **most one half of** a page.

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

Part 1.2

In this part of the homework, you have to solve the “Search Engine Selection Problem” (described in the following section) by performing quantitative analysis to assess the quality of different Search-Engines, using only the ground-truth and their query-results, to select the best one according to the intrinsic characteristics of the problem.

Search Engine Selection Problem

The *“DummyDataScience”* company needs to select the best Search-Engine module, among three modules, for its latest successful app: *“AwesomeSocialApp”*. Data from these three Search-Engine modules are located here: `DMT/HW_1/part_1/part_1_2/dataset` .
For assessing the quality of the Search-Engine modules in a correct way, you have to consider that the app provides in output only four results for each search query. Moreover, these four results are displayed on the smartphone screen in random positions.

By using **at most one** page of the report, you have to show and comment the results of the performed quantitative analysis that justifies your choice of the best Search-Engine module for the app.

Part 2.1

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 **≥ 0.95** .
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/part_2_1/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/part_2_1/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. You **must** use as shingle identifiers natural numbers that span 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 lyrics 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=0.95 **must be > 0.97**.

Details on OUTPUT

Goal: The generation process of near-duplicate pairs you implemented must generate the smallest amount of both False-Negatives and False-Positives.

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 according to the constraints.
 .) How did you reduce the generation of False-Negatives?
 .) How did you reduce the generation of False-Positives?
 .) The Execution-Time of the Near-Duplicates-Detection tool.
 .) The number of Near-Duplicates couples you found.
You must provide all this information on **at most two pages**.

Part 2.2

You have to find 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 TITLE (field "song" in the dataset)** is **1**.

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. You **must** use as shingle identifiers natural numbers that span 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 REPLACE THE CHARACTER "-" WITH A SINGLE SPACE " ", 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 **TITLE (field "song" in the dataset)** of the song.
In the case in which the processed title of the song is shorter than three words, YOU MUST consider the entire title as a single shingle of length less than 3.

Information to Provide in the Report

You have to provide in the report the following information:
.) A **very short** description of the algorithmic approach you applied for addressing the problem.
.) How does your method counteract the generation of False-Negatives?
.) How does your method counteract the generation of False-Positives?
.) The Execution-Time of your method.
.) The number of Near-Duplicates couples you found.
You must provide all this information on **at most ONE page**.

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_1: /DMT_2021/HW_1/part_1/part_1_1/sw/ (.py files).
2. The software for addressing Part_1_2: /DMT_2021/HW_1/part_1/part_1_2/sw/ (.py files).
3. The software for addressing Part_2_1: /DMT_2021/HW_1/part_2/part_2_1/sw/ (.py files).
4. The software for addressing Part_2_2: /DMT_2021/HW_1/part_2/part_2_2/sw/ (.py files).
5. The **COMPRESSED** tsv file you created for addressing Part_2_1 that contains the sets of shingles identifies: /DMT_2021/HW_1/part_2/part_2_1/data/ (compressed .tsv files).
6. The **COMPRESSED** tsv file containing the Near-Duplicates you found for Part_2_1: /DMT_2021/HW_1/part_2/part_2_1/data/ (compressed .tsv files).
7. The **COMPRESSED** tsv file containing the Near-Duplicates you found for Part_2_2: /DMT_2021/HW_1/part_2/part_2_2/data/ (compressed .tsv files).
8. The final report in **PDF**: /DMT_2021/HW_1/report.pdf .

The name of the zip file must have this format:
DMT_2021__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_2021__HW_1__StudentID_StudentName_StudentSurname_StudentID_StudentName_StudentSurname.

p.s.
For any problem, doubt or consideration, please send an email to fazzone@diag.uniroma1.it and siciliano@diag.uniroma1.it.