Homework 1

Data Mining Technology for Business and Society

Deadline: **10 April 2019 23:59 (Rome Time Zone)**

Exactly Two students for each group

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_2018_2019/DMT4BaS_2018_2019.html

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

Part 1

In this part of the homework, you have to index a collection of documents and improve the search-engine performance by changing its configurations using the provided set of queries and the associated Ground-Truth. For this part of the homework you must use the Whoosh API.

Documents, Queries and Ground-Truth

The documents to index are stored in html files and they are composed by two fields: content and title (please, open them with a text-editor and not with a browser). The title is between the "<title>" tags and the content is between the "<body>" tags. The document-id is the integer number at the end of the html file name. For instance, the file with name "_____42.html" contains the document with ID "42", title "the gyroscopic effect of a rigid rotating propeller on engine… " and content "in many wing vibration analyses it is found necessary…". All documents are stored inside the "DMT4BaS/HW_1/part_1/Cranfield_DATASET/DOCUMENTS" directory.
Queries are stored in the "DMT4BaS/HW_1/part_1/Cranfield_DATASET/cran_Queries.tsv" file and the ground-truth is stored inside the "DMT4BaS/HW_1/part_1/Cranfield_DATASET/cran_Ground_Truth.tsv" file. These two files are linked by the "Query_id" field value.

Constraint on the Performance

The **only** search engine configurations that will be accepted are the ones that satisfy the following constraint: $MRR(Q) \geq 0.32$.
Where Q is the set of the provided queries.

Evaluation Metrics

For each configuration (acceptable and also not acceptable), you must provide the following MRR table:

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

For each acceptable configuration (that are the ones with MRR(Q) ≥ 0.32), you must provide the following information:

.) 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	?.???	?.???	?.???	?.???	?.???	?.???
...	?.???	?.???	?.???	?.???	?.???	?.???

.) The nDCG@k plot:
  .) the x axis represents the considered values for k: you must consider k ϵ [1, 10]
  .) the y axis represents the average nDCG over all provided queries.
  .) Each curve represents a different acceptable search engine configuration.

Information to Provide in the Report

You have to provide in the report the following information:
.) Number of indexed documents and the number of queries.
.) A schematic description of **all** tested search engine configurations.
.) The set of all acceptable search engine configurations: that are the ones with MRR(Q) ≥ 0.32.
.) The MRR table for **all** tested search engine configurations.
.) The R-Precision for each acceptable configuration (that are the ones with MRR(Q) ≥ 0.32).
.) The nDCG@k plot containing all the acceptable configuration (that are the ones with MRR(Q) ≥ 0.32).
You must provide all these information in **at most three pages**.

Part 2

You have to find, in an approximated way, all near-duplicate documents inside the following dataset: /DMT4BaS/HW_1/part_2/dataset/261K_lyrics_from_MetroLyrics.csv .
The dataset contains data on **261K** 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 **≥0.88**.
To complete this part of the homework, you have to use the **Near_Duplicates_Detection_Tool** that is entirely contained inside the directory "DMT4BaS/HW_1/part_2/tools". The file "DMT4BaS/HW_1/part_2/script_for_testing.txt" contains a short description and an example on how to run the **Near_Duplicates_Detection_Tool.**
For creating hash functions, you can use the following software: "DMT4BaS/HW_1/part_2/hash_functions_creator.py".

Details on Shingling

For each lyric of a song, the set of shingles must be a set of natural numbers.
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.

Information to Provide in the Report

You have to provide in the report the following information:
 .) What values for 'r,' 'b' and 'n' did you choose?
 .) The probability to have False-Negatives, in the set of candidate pairs, for the following Jaccard values: 0.88, 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 handle the presence of False-Positives in the set of candidate pairs to be Near-Duplicates?
 .) 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.
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: /DMT4BaS_2019/HW_1/part_1/sw/ (.py files).
2. The software for addressing Part_2: /DMT4BaS_2019/HW_1/part_2/sw/ (.py files).
3. The tsv file containing the Near-Duplicates you found for Part_2: /DMT4BaS_2019/HW_1/part_2/data/ (.tsv files).
4. The final report in **PDF**: /DMT4BaS_2019/HW_1/report.pdf .

The name of the zip file must have this format:
DTMT4BaS_2019__HW_1__STUDENTID_NAME_SURNAME__STUDENTID_NAME_SURNAME.zip
Finally you must send the ".zip" file to fazzone@diag.uniroma1.it with the following email subject:
DMT4BaS_2019__HW_1__StudentID_StudentName_StudentSurname_StudentID_StudentName_StudentSurname.