CE205 Databases and Information Retrieval

Assignment 2

Evaluation of Information Retrieval System

TASKS

(a) Dataset [20%]

Identify a publicly available dataset containing a set of documents. Provide a detailed description of this dataset. Some publicly available datasets are given below (you can describe one of them or feel free to find a dataset of your choice).

- The Signal Media One-Million News Articles Dataset
  https://research.signal-ai.com/newsir16/signal-dataset.html
- 20 Newsgroups Dataset
  http://qwone.com/~jason/20Newsgroups/
- Reuters-21578 Dataset
  https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection

(b) Construct a Bag-of-Words Model for Shakespeare's Sonnets [30%]

Shakespeare's sonnets can be found here

http://www.shakespeares-sonnets.com/Archive/allsonn.htm

It can also be found in MATLAB

https://ch.mathworks.com/help/textanalytics/ug/data-sets-for-text-analytics.html

https://ch.mathworks.com/help/textanalytics/ug/extract-text-data-from-files.html

Consider the first ten sonnets and create a bag-of-words model using MATLAB (*do not use built-in functions; tokenizedDocument and bagOfWords*).

- ❖ What is the vocabulary size (number of unique tokens)?
- ❖ Show the document vectors for each sonnet (*take screenshots and paste them into the report*).

Briefly write the process inside the report that you follow to construct the bag-of-words, and append the MATLAB code (.m file) at the end of the report.

(c) Evaluation of an IR System [30%]

Assume that an information retrieval (IR) system has returned *D* number of documents (replace *D* with a number of your choice) for a given query and *R* number of documents (replace *R* with a number that makes sense) are relevant among them. Determine the number

of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Then, evaluate the effectiveness of the IR system by computing the following measures.

- Precision
- Recall
- F1-score

❖ Briefly explain (also provide formula) that how these measures will help you to evaluate the IR system.

❖ Precision and Recall tend to be in an inverse relationship; as Precision goes up, Recall comes down, and vice versa. Why is this? *(use values of TP, FP, TN, and FN to justify your answer)*

❖ How does the F1-score attempt to overcome the limitations of precision and recall?

❖ What is Average Precision (AP) and mean AP? What are the purposes to compute them?

(d) Precision-Recall Curve [10%]

To obtain a robust estimation of an IR system's performance, Precision-Recall Curve is a commonly used approach. Briefly discuss it and implement using MATLAB.

(e) Structure of the Report [10%]

❖ The report should cover all important aspects of the given tasks.

❖ The answers to the tasks must be brief (*preferably, use bullet points where possible*).

❖ The task/subtask must be mentioned before providing the answer.

SUBMISSION

- Submit one pdf file to FASER called: **id_ce205_assignment2.pdf**
- The submission deadline is **Monday, 24 January 2022, 11:59 (mid-day).** Always, keep an eye on FASER for the submission deadline.

PLAGIARISM

You should work individually on this assignment. Anything you submit is assumed to be entirely your work. The usual Essex policy on plagiarism applies:

http://www1.essex.ac.uk/plagiarism/

Best of Luck!