IT550: Information Retrieval

Assignment 3: Document Retrieval and Evaluation

Instructor: Prasenjit Majumder

Learning Outcome: After this assignment you will learn how to implement a retrieval module and also how to evaluate it

1 Problem description

In document retrieval the documents and queries are converted to vector space and then some matching is done between the documents and queries. Which returns a ranked list of documents. Then evaluation is performed to determine the efficiency the matching function.

2 Implementation

2.1 Dataset

- For this assignment we will use Telegraph news articles, which is in XML format. It contains news on different categories for the year 2004 to 2007. You can download the dataset from this link: Dataset Click
- The Queries are in "en.topics.76-125.2010". The query is of the format shown in Figure 1. Use the sentences enclosed in desc tag for framing your query vector
- The "en.qrels.76-125.2010.txt" contains the documents that are relevant to a query. The format of a qrel is such: Query_No Q0 Document ID Relevance score.
- Relevance score is binary 0 or 1. 1 is for relevant, 0 is for otherwise.
- The documents in the dataset is in the format shown in Figure 2.

```
<top lang='en'>
<num>76</num>
<title>Clashes between the Gurjars and Meenas</title>
<desc>
Reasons behind the protests by Meena leaders against the
inclusion of Gurjars in the Scheduled Tribes.
</desc>
<narr>
The Gurjars are agitating in order to attain the status of a
Scheduled Tribe. Leaders belonging to the Meena sect have
been vigorously opposing this move. What are the main reasons
behind the Meenas' opposition? A relevant document should
mention the root cause(s) behind the conflict between these
two sects.
</narr>
</top>
```

Figure 1: Query Format

```
<DOC>
<DOCNO> </DOCNO>
<TEXT></TEXT>
</DOC>
```

Figure 2: News Format

2.2 Exercise

1. Create the term-document matrix from the text of all documents in the corpus
2. Represent the documents using TF-IDF.
3. Use the text in <title>, <desc> or <narr> in the queries for getting the TF-IDF representation of the queries.
4. The TF-IDF representation should be done using the custom class that you had created in the earlier assignment.
5. Calculate the cosine similarity between the query and document vectors.
6. For each query retrieve the top 10 documents based on cosine similarity.
7. Evaluate this using the Mean Average Precision (MAP)

3 References

- An Introduction to Information Retrieval: Christopher D.Manning ,Prabhakar Raghavan, Hinrich Schütze
- https://towardsdatascience.com/breaking-down-mean-average-precision-map-ae462f623a52

4 Submission

- You have to submit your assignment in Google Classroom with proper comments and explanation of your approach.
- The submission deadline for this assignment in **16th Feb 2021 at 11 PM**