Information Retrieval Project - Task 2

Group 2

October 21, 2021

1 Group Details

- Adhikansh Singh - 17CS30002
- Ankit Bagde - 17CS30009
- Kousshik Raj M - 17CS30022
- Shivam Kumar Jha - 17CS30033

2 Approach

The following is a brief description of the approach taken to solve the problems.

- **Task 2A:** We first calculate DF_t for each term t. Next we calculate $TF_{t,d}$ for each term t, document d pair using the inverted index as re processing the documents will be costly. Next we precalculate all query vectors, but we only calculate the document vectors when we need them as preprocessing them all is substantially costly in terms of memory (about 13.5GB). Then for each document, we find the score for each query for all the schemes. Finally, we sort the results and take only the top 50 documents for each query.
- **Task 2B:** We load the gold standard data, and for each query we store all the relevant documents. The rest of the documents are considered irrelevant. Then for each query, we process the results and from the relevancy we calculate the **Precision@K** $\forall 1 \leq K \leq 20$. From this data, we calculate **AP@10** and **AP@20**. Similary, from the relevance scores, we calculate **DCG@K** and **Ideal_DCG@K** $\forall 1 \leq K \leq 20$, which we can use to calculate **NDCG@10** and **NDCG@20**. Finally we calculate the average of all the metrics and store them.

3 Assumptions

- In case of no relevant documents, **Ideal_DCG** scores also become 0, which means **NDCG** at any point is not defined. Hence, we make an assumption that these cases have **NDCG** of 0.
- Since we handle only query ids, we do not know what the actual query is. Hence, in the final results CSV file, the fields are only *Query Id*, *AP@10*, *AP@20*, *NDCG@10*, *NDCG@20.*
- The final row of the CSV file has the average of the metrics, and the *Query Id* field is blank for this row.

3.1 Running Commands

- **Task 2A:-** python PAT2_2_ranker.py *path_to_en_BDNews24_folder path_to_model_queries_2.pth path_to_queries_2.txt*
- **Task 2B:-** No change

4 Software Requirements

- Python version - 3.7
- Python libraries used - *os*, *sys*, *numpy*
- Libraries that need installation - *numpy*

4.1 Running Time

- **Task 2A:-** The ranked retrieval takes approximately **2.25 hours**
- **Task 2B:-** The evaluator runs in approximately **5 seconds**