Information Retrievel Term Project - Individual Report

Aditya Basu, 19IE10002

Project Code: DEFAULT-IR-TP, Group 9

Assignment 3

Parts completed

- Part 3A of DEFAULT-IR-TP was to measure relevance feedback. Altaf Ahmad and I worked on this part. After Altaf completed the initial portion of modifying the queries using Rocchio's algorithm. After we obtained the modified queries, the cosine similarities between documents were calculated. Once having gotten the scores, ranking of relevant documents were done. I worked on the evaluation part using mean Average Precision (MAP@20) and Normalized Discounted Cumulative Gain (NDCG) by writing the "calcNDCG20_map20" function.

$$IDCG_p = \sum_{i=1}^{|REL_p|} \frac{2^{rel_i} - 1}{log_2(i+1)} \qquad nDCG_p = \frac{DCG_p}{IDCG_p}$$

- The documentation of the project was written and maintained by me. This includes the README file containing the running commands, files generated, running time, python version for 2A, 2B, 3A, 3B and how these parts relate to one another. More importantly, the the problems we faced in each part has also been briefly mentioned in the README.

Challenges faced

For a certain number of queries, while calculating the scores of NDCG@20 - there were 0 documents which matched the gold standard documents. This meant there would be a problem in calculating the Discounted Cumulative Gain (DCG) for these values. the value I was getting was "np.NaN" (which is a null value). This led to a zero division error and handled was initially handled by a try except block in Python.

Design methodology

Initially, nested try-except blocks were used to handle all possible errors but later some were removed after careful scrutiny.
Initially explicitly written for-loops was replaced to use the power of vectorization as much as possible - this helped in reducing the overall running time of the code.