Group 13
PART-3

We have discussed the project among ourselves and we have developed the workflow for the same but for developing the codes, we have divided the work among the group as clearly mentioned in the individual report.

Short description of approach:

Task 3A (Relevance Feedback)
Firstly, we imported all the libraries which were required for implementing the following task.
Then we used csv.dictreader function of python to store the csv mapping of cord_id and paper_id to a python dictionary and read all the documents using glob library.

Now, we created a dictionary “top20_retrieved” which takes the results from the part 2A of the assignment and contains the top 20 ranked documents corresponding to the given query. Now, we have also created a dictionary named goldStandard_rankedList which reads the information from the gold standard file (qrels.csv) and for each query we removed the duplicate entries of same doc id, taking only the one with highest iteration value.Thus, after processing, we get a goldStandard_rankedList which gives the gold standard relevance for a given query and for a given document. Now, we have defined the rochhio function which takes the value of alpha,beta,gamma, query, and dictionary of relevant and non-relevant and gives the modified query as output.

Then, used pickle read the inverted index from the **model_queries_13_.bin** built in the Assignment 1. Now, for each key in the inverted index,calculated the document frequency of the each tokens by calculating the length of its inverted index.
Now, created a dictionary tf and calculated the term frequency of each tokens by iterating through all the documents in the corpus.
Now, we created a ranking function named rankfunc which created the td-idf of the modified query and read the tf-idf of the documents from the file td-idf bin file generated in the part-2 of the assignment and created the rank list based on the scores obtained by the modified query on the documents in the corpus.

Now, according to the Relevance and Psuedo-relevance feedback as instructed in the problem statement we created a dictionary for relevant document and irrelevant documents and using the rochhio function as defined earlier we we modified the query based on both the schemes of relevance feedback and for different values of alpha,beta and gamma and then using the ranking function (rankfunc) defined earlier we ranked the documents according to the modified queries for both of the feedback schemes and based on the new ranklist of the documents for each we calculated the average MAP@20 and average NDCG20 for all the queries.

Task 3B (Identifying words from pseudo relevant documents that are closer to the query)

Now, In this part we considered the tf-idf for the top 10 ranked document for a given query and for each word in the query we calculated the average values of the tf-idf across the top10 document and then each query ranked the top 5 word having the highest values of average tf-idf across the top 10 documents and then stored the results in the csv file in given format.