Group 13
PART-2

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 2A (TF-IDF Vectorisation)

Created a function to preprocess sentence which removes stop words, punctuation marks and performs lemmatization to generate tokens from the corpus.For the preprocessing we used libraries like regex for punctuation removal and nltk library for stop words removal and lemmatization.

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.

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. For creating the term frequency, we have divided files into chunks of 5000 files and we iterate through each chunk, processing each file in that chunk, saving it to disk (we have created a directory "./temp_tf") and freeing up memory before moving to the next chunk. After processing all the chunks, we then merged the chunks using the code to generate a single term frequency. This optimization helped us in processing all the files with limited computational facility available with us. Now, for the three different schemes calculated the the TF-IDF vector from the tf and df for each (query, doc) pair and then calculated the score by using the cosine similarity metric and ranked the documents according to the score for all the three schemes respectively and score the top50 documents ranked according to the scores into the respective csv files.

Task 2B (Evaluation)

In the part firstly, we read the gold standard values from the qrels.csv provided in the dataset for the top 20 data values of the documents from the rank list obtained in the part-2A for query and checking whether the document is relevant or non-relevant based on the judgement obtained from the qrels.csv file.

Now,for all the three csv obtained from the part 2A in which the TF-IDF were calculated based on different schemes namely lnc.itc, Lnc.Lpc, anc.apc and based on the relevance for the gold standard, calculated the average precision and normalised discounted cumulative gain for the 10th and 20th position.Then, taking the average of the mAP@10, mAP@20 , NDCG@10and NDCG@20 across all the queries and for all the three cases and then written the results into a text file as mentioned.