Question-1

Pre_Processing:

1. Loaded dataset
2. Removed Numbers
3. Converted data to lowercase
4. Removed Punctuations,stopwords , 2 length words
5. Lemmatized text

Approach :

Part (a):

1. Written method to find Union and Intersection for document and Query.
2. Written method to compute Jaccard coefficient.

Part (b):

Question-2

Pre-processing

1. Loaded the dataset using Pandas library

2. Retrieved docs with qid:4 using groupby() function

3. Converted the string values to float

Methodology:

1. To find the max_dcg value, we have sorted the relevance_judgement_score in descending order as it will give max_dcg.

2. To find the total number of documents with max_dcg, we are counting the frequency of each relevance_judgement_score and find all the possible permutations.

3. To find the NDCG value, we are using the below formula and calculating DCG and Ideal DCG values. (NDCG = DCG/IDCG)

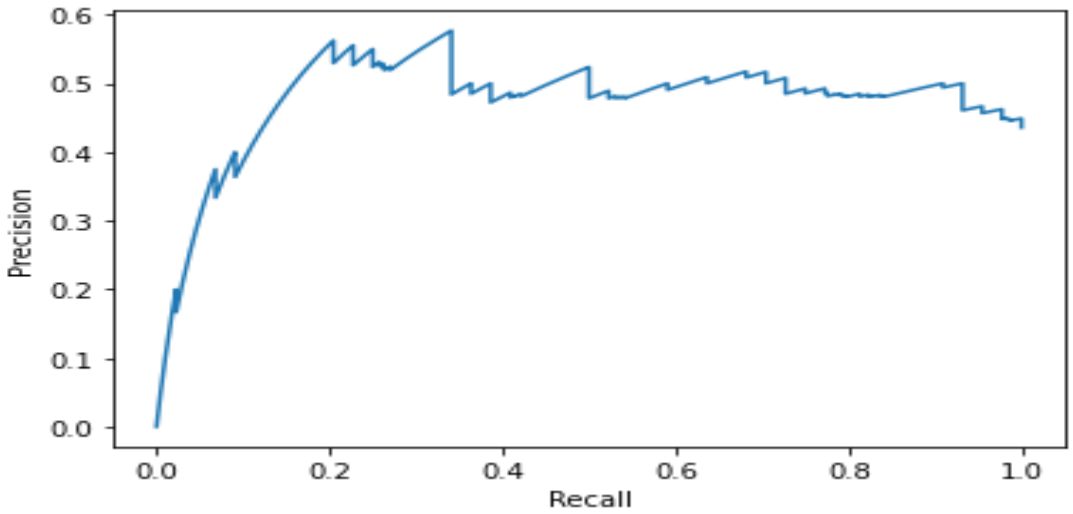
$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log(1+i)}$$

4. To plot the Precision Recall curve, we are finding the precision and recall values for each query using the actual relevance values and Feature_75 values which are normalized between 0-1.

Assumptions:

No assumptions

Precision-Recall Curve



Question-3:

Pre_Processing:

6. Loaded dataset
7. Removed Numbers
8. Converted data to lowercase
9. Removed Punctuations,stopwords , 2 length words
10. Lemmatized text

Approach:

Written a method to perform data splitting
Generated ClassFrequencies and Inverse-Class Frequency

Accuracies : 0.914380714879468
0.823905558288244