IR-PROJECT

WIKIPEDIA -BASED DOCUMENT RANKING

NAME: K. BALAVARDHAN REDDY

ROLL-NO: S20210010123

EMAIL: balavardhan.k21@iiits.in

Components:

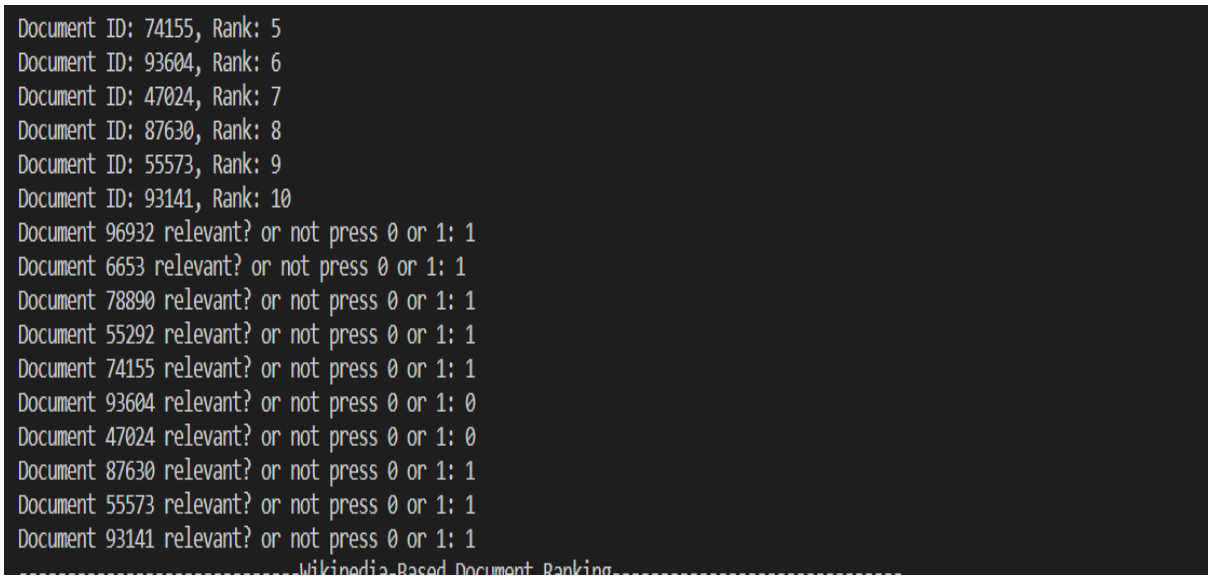
Index (inverted index): In our IR retrieval system we calculate inverted index by building the dictionary for the data set that we are using. To achieve this, we are preprocessing the documents in the dataset by removing all the stop words from the (stopwords.txt) In the inverted index we are storing the word as key, and in which documents the word has been occurred. The document frequency and the word occurred in which documents (document-ID) each document in the dataset has a document-ID.


```
},
"hardwood": {
    "postings": [
        {
            "doc": "5791",
            "freq": "2"
        },
        {
            "doc": "19633",
            "freq": "1"
        },
        {
            "doc": "21930",
            "freq": "1"
        },
        {
            "doc": "25071",
            "freq": "1"
        },
        {
            "doc": "32355",
```

COSINE-SIMILARITY: calculates the similarities between the query and documents by cosine similarity score the document with highest similarity score will retrieve first and so-on… The top 10 documents are retrieved in the decreasing order of cosine similarity score. On entering the query, the documents are retrieved based on the similarity score the score 0 mean document is irrelevant.

Relevance_Feedback:

Collects the feedback for the entered query Each document is relevant to the query or not if document is relevant to the query 1 if it is not relevant to the query 0 it collects the feedback from the user the relevant documents are stored in relevant_docs.json file. Each document score and rank is stored in ranked. Json file.



Precision and Recall Calculation:

Determines the precision and recall metrics for the retrieved documents. Calculates precision and recall values based on the relevance feedback received from the user.

11-Point Interpolated Precision-Recall Curve: Computes and visualizes the 11-point interpolated precision-recall curve using Matplotlib. Displays the precision-recall trade-off for the system's performance evaluation.

```
For Query 1:
Recall at 1 is: 0.125
Recall at 2 is: 0.25
Recall at 3 is: 0.375
Recall at 4 is: 0.5
Recall at 5 is: 0.625
Recall at 6 is: 0.625
Recall at 7 is: 0.625
Recall at 8 is: 0.75
Recall at 9 is: 0.875
Recall at 10 is: 1.0

Precision at 1 is: 1.0
Precision at 2 is: 1.0
Precision at 3 is: 1.0
Precision at 4 is: 1.0
Precision at 5 is: 1.0
Precision at 6 is: 0.8333333333333334
Precision at 7 is: 0.7142857142857143
Precision at 8 is: 0.75
Precision at 9 is: 0.7777777777777778
Precision at 10 is: 0.8
F1 score at 1 is :  0.2222222222222222
F1 score at 2 is :  0.4
F1 score at 3 is :  0.5454545454545454
```

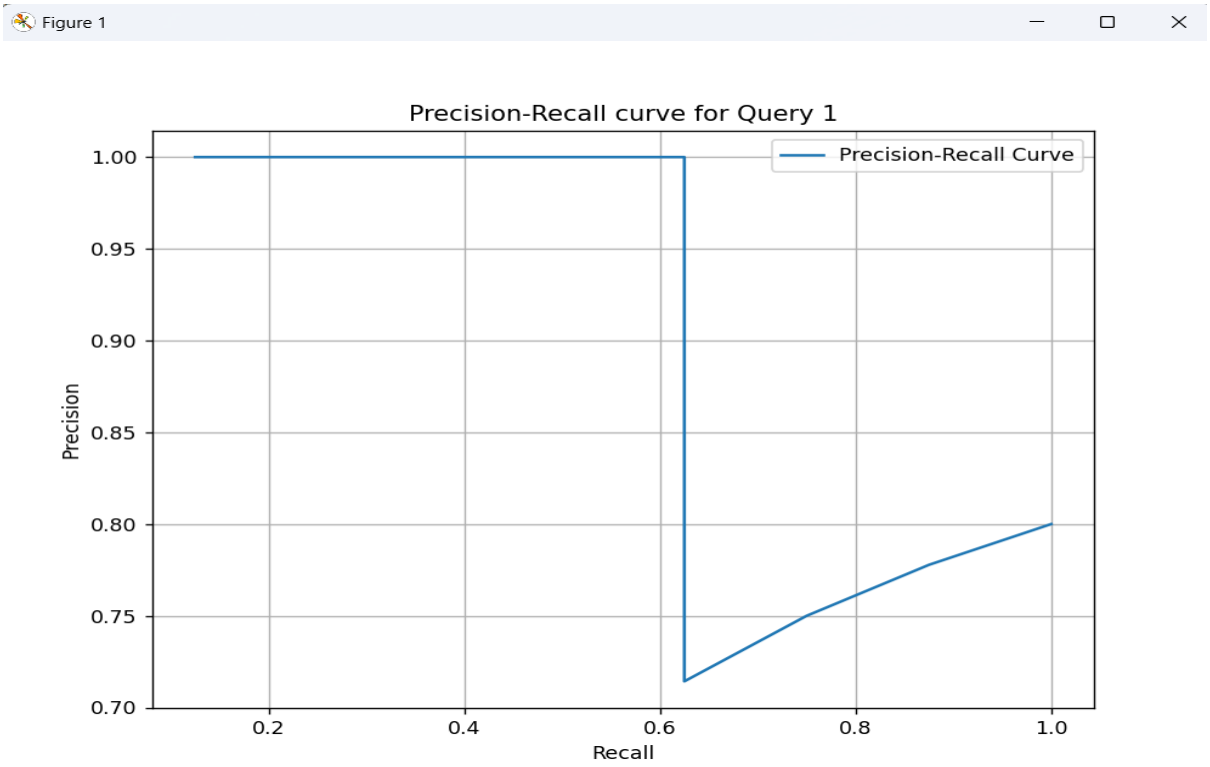
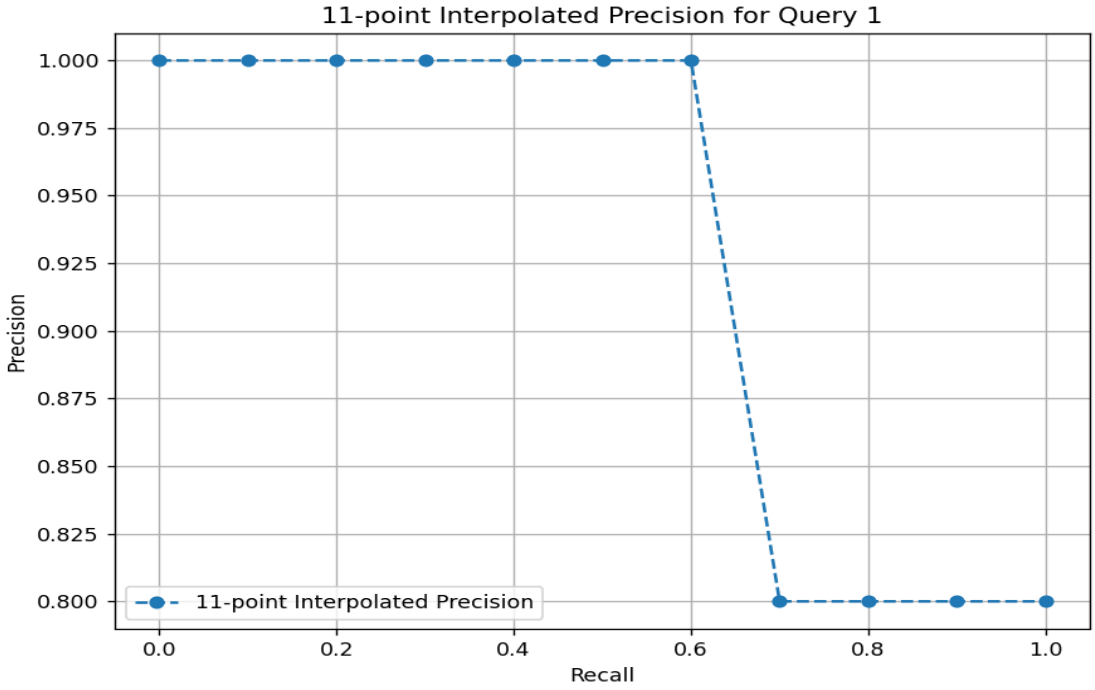


Figure 1
Precision-Recall curve for Query 1
Precision-Recall Curve
Precision
Recall
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.2
0.4
0.6
0.8
1.0


11-point Interpolated Precision for Query 1
11-point Interpolated Precision
Precision
Recall
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0.800
0.0
0.2
0.4
0.6
0.8
1.0