Information Retrieval - CS 7800

Assignment-1

Sai Vijay Kumar Surineela	Manoj Kumar Reddy Avula	Prabhu Charan Murari
U01096171	U01067535	U01099304
surineela.2@wright.edu	avula.25@wright.edu	murari.4@wright.edu

The given dataset contains three files named as- cran.all, query.txt, qrels.txt, readme.txt. The cran file contains document information (marked under. W), document Id (marked under. I) and other informations, for this assignment only document information and Id number are needed. Query file as observed to be in the same format as of cran file but with queries and queries Ids. Qrels has the information of queryIDs and its respective relevant documents. There are a total of 1400 documents and 225 queries in the dataset.

Document and Query processing:

The document matter and ids from cran file are stored into a dictionary, using split () function, and also words in the document are lowered and stopwords are removed with the help of sklearn stopwords, after these words in the doc are tokenised and stored in a list. Similar approach is followed for queries.

There are two approaches that are followed here, binary and vector-space model. In vector-space, Tf-Idf was used to add weights to the tokens. For binary, CounVectoriser() from sklearn is used for vectorising the document and fit_transform() is used on documents and transform() is used for queries.

Cosine Similarity and Euclidean distances are metrics that tries to convey how much a document id related or close to a query, these two metrics are utilized directly by using inbuilt sklearn.metrics.pairwise library. The result is arranged in array of size (225, 1400), with each entry being the score of either cosine or Euclidean of a document wrt a query. np.argsort and slicing are used to sort the top 10 indices that have the highest score, higher the score, higher the relativity for a document wrt to query. This approach is used for Binary, and the Precision, Recall and F1-score are calculated by using method developed to conclude how well the binary search engine developed has performed.

TFIDF had a different approach, weights for each token was added, a new custom class which inherits from CountVectoriser() class is developed as part of this project to full fill the requirement of TF and IDF given in the assignment.

$$TF_{modified(t,d)} = \frac{1 + log(count(t,d))}{1 + log(length(d))}$$

where `count(t,d)` is the raw count of term `t` in document `d`, and `length(d)` is the total number of terms in document `d`.

$$IDF_{modified(t,d)} = \frac{1}{DF(t)}$$

where `DF(t)` is the document frequency, i.e., the number of documents containing the term `t`.

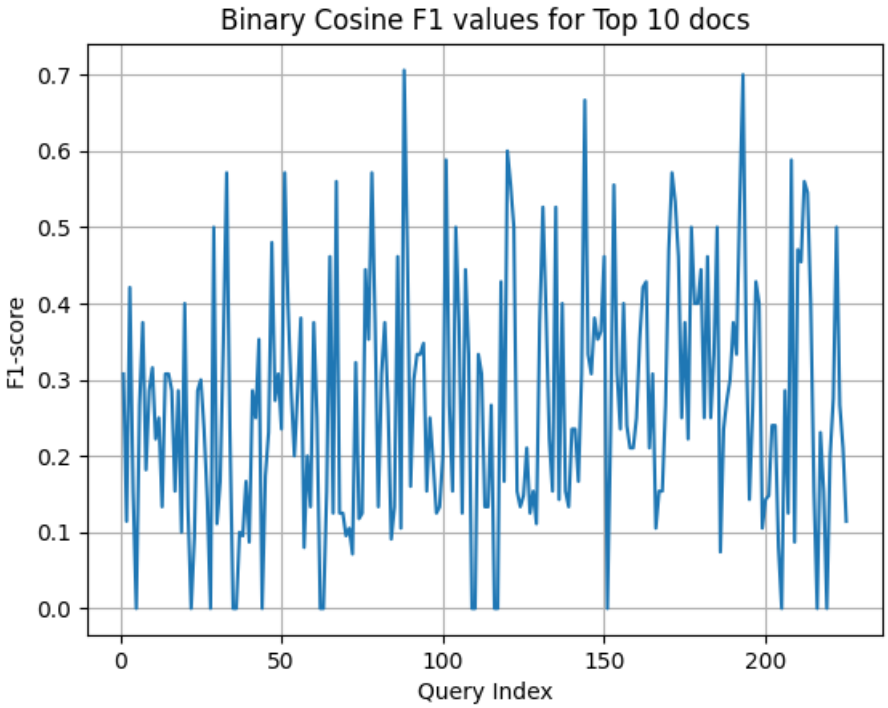
The TFIDF approach is also evaluated on the scores of Precision, Recall, F1-score. A total of 12 are graphs generated wrt to performance metrics and queries for binary and tf-idf for both cosine and Euclidean.

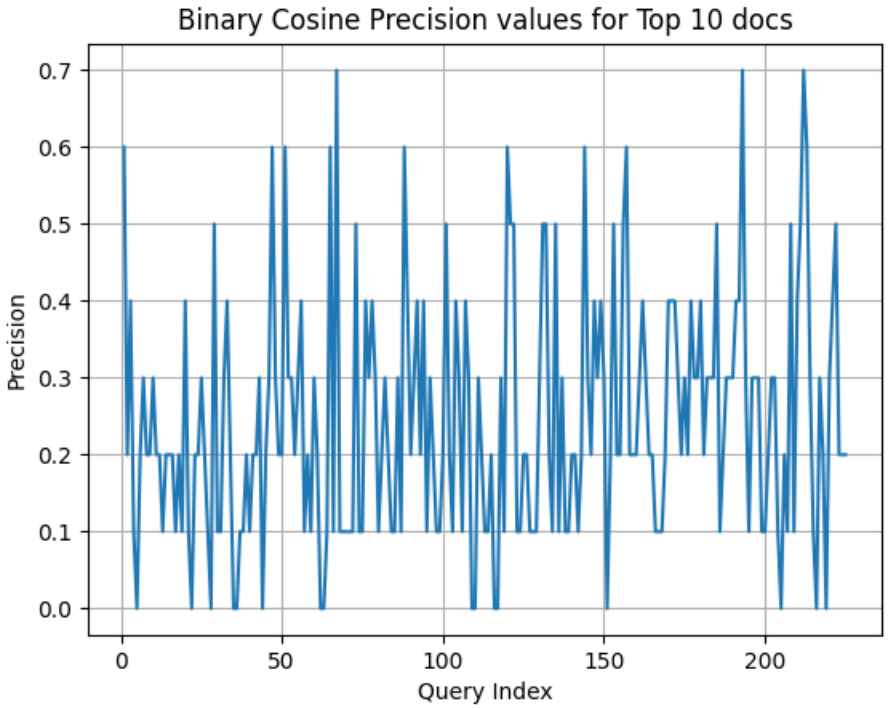
Terminal Results:

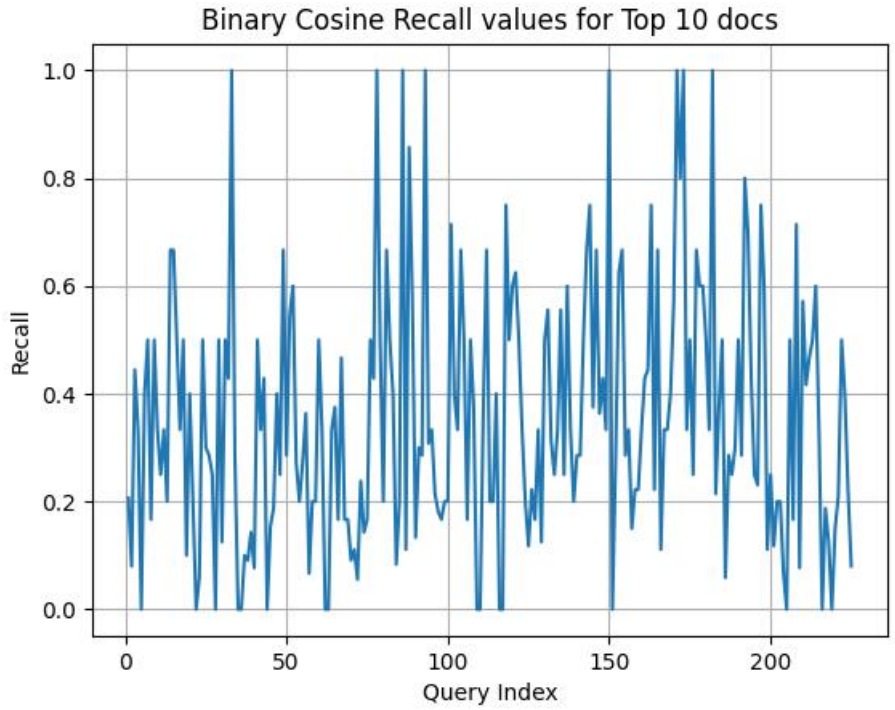
```
{'Binary': {'f': {'cos': (0.26861488575846976, 0.7058823529411764),
                  'euc': (0.0252959745392464, 0.32)},
            'p': {'cos': (0.24400000000000002, 0.7),
                  'euc': (0.023111111111111114, 0.4)},
            'r': {'cos': (0.3534356241268771, 1.0),
                  'euc': (0.03303866759039173, 0.5)}},
 'TFIDF': {'f': {'cos': (0.05108423619647638, 0.5),
                 'euc': (0.009652898556507536, 0.15384615384615383)},
           'p': {'cos': (0.049777777777777775, 0.5),
                 'euc': (0.008444444444444445, 0.1)},
           'r': {'cos': (0.06050421585447386, 0.5),
                 'euc': (0.01325921050626933, 0.3333333333333333)}}}
```

Performance Metrics for Binary Vector and TF-IDF with Cosine Similarity and Euclidean Distance:

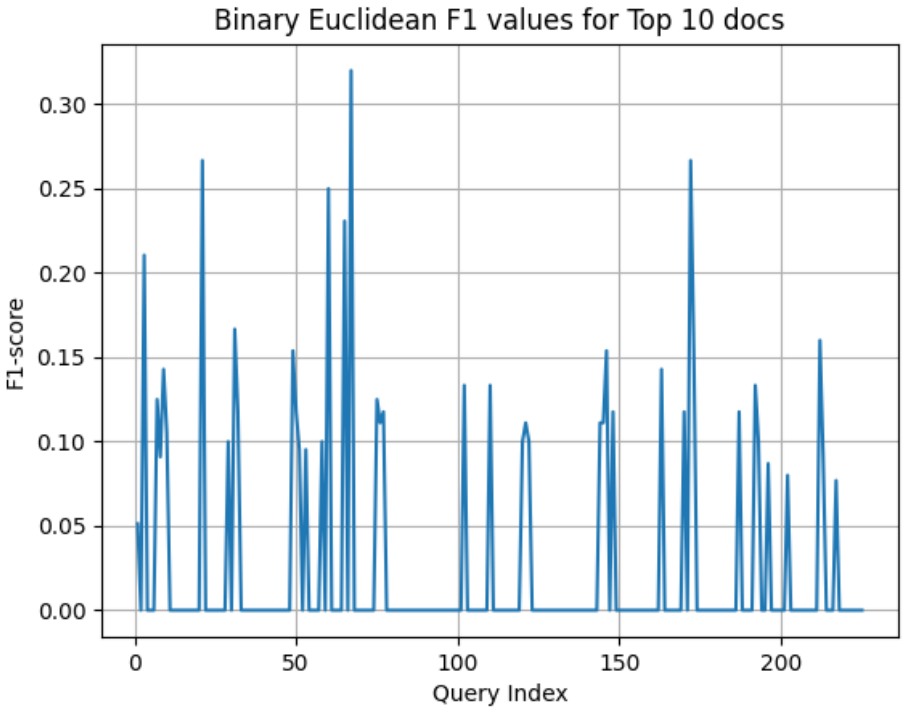
Precision, Recall, and F-Scores for Binary Vector with Cosine Similarity for Top 10 documents:

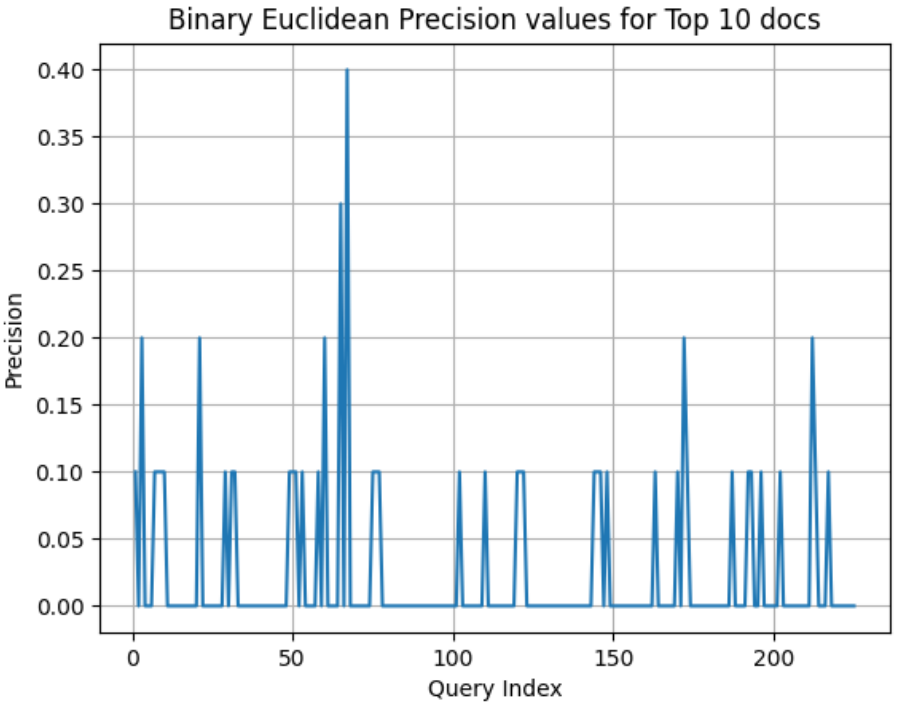


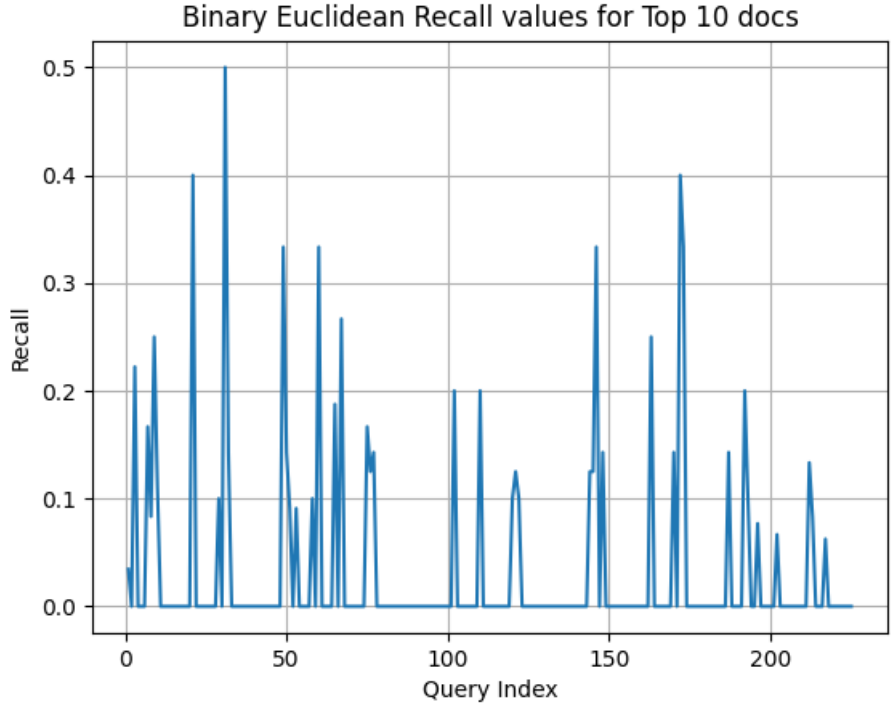




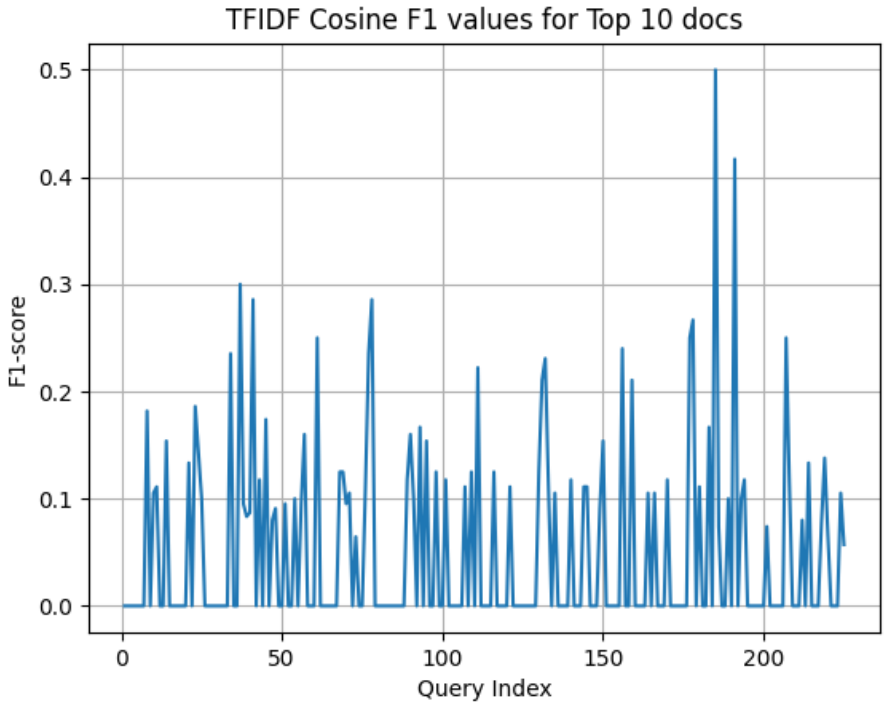
Precision, Recall, and F-Scores for Binary Vector with Euclidean Distance for Top 10 documents:

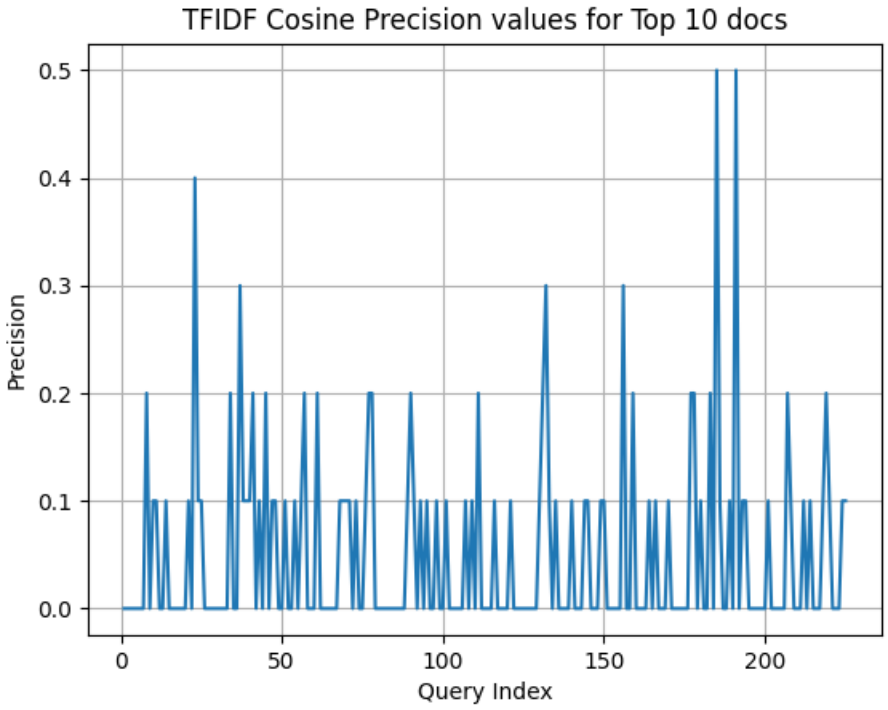


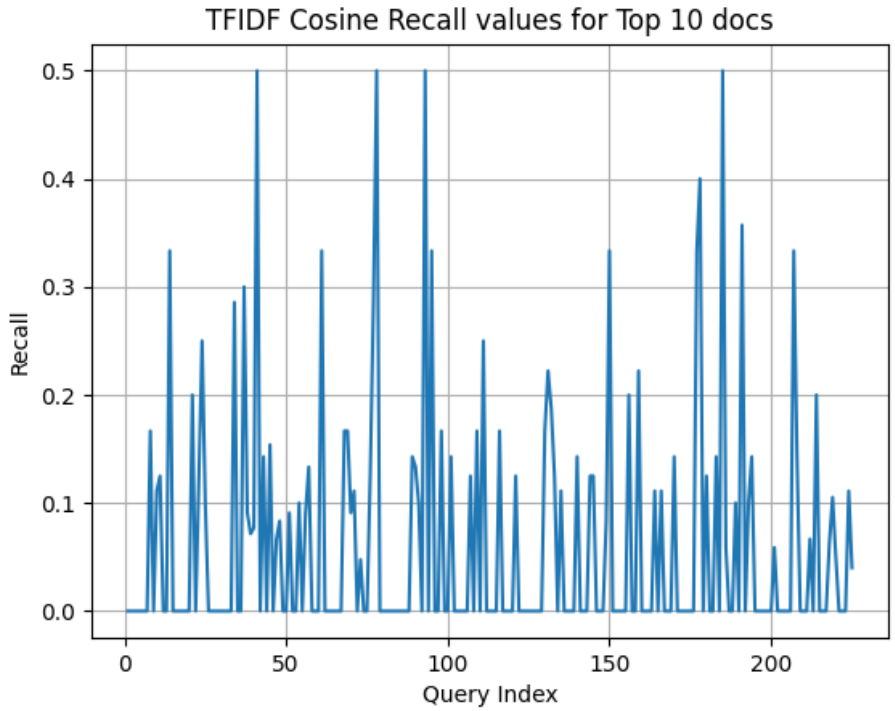

Binary Euclidean Precision values for Top 10 docs
Precision
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
50
100
150
200
Query Index


Binary Euclidean Recall values for Top 10 docs
Recall
0.5
0.4
0.3
0.2
0.1
0.0
0
50
100
150
200
Query Index

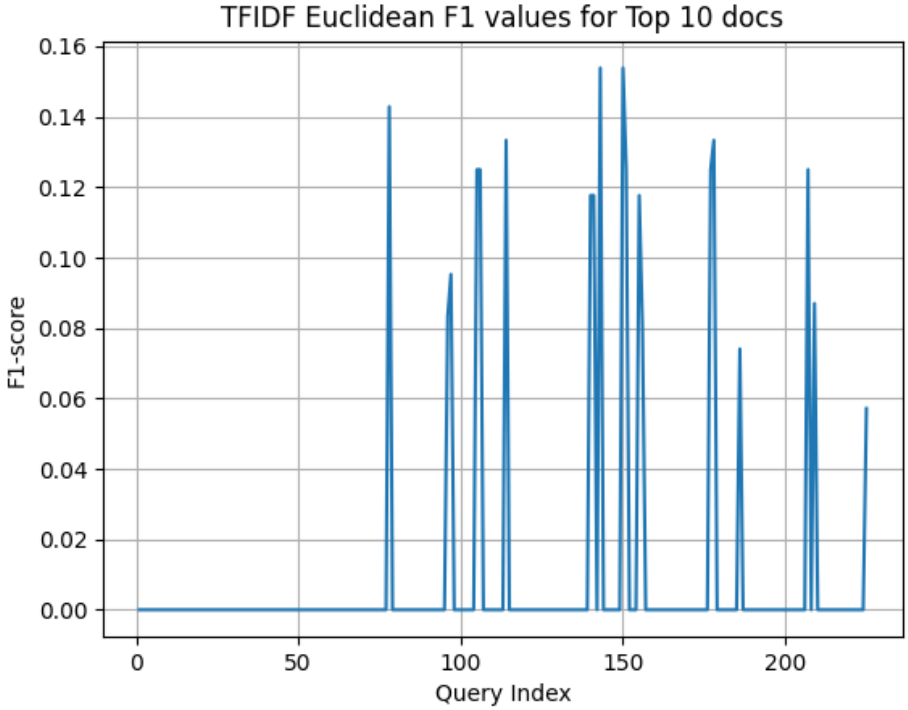
Precision, Recall, and F-Scores for TF-IDF with Cosine Similarity for Top 10 documents:

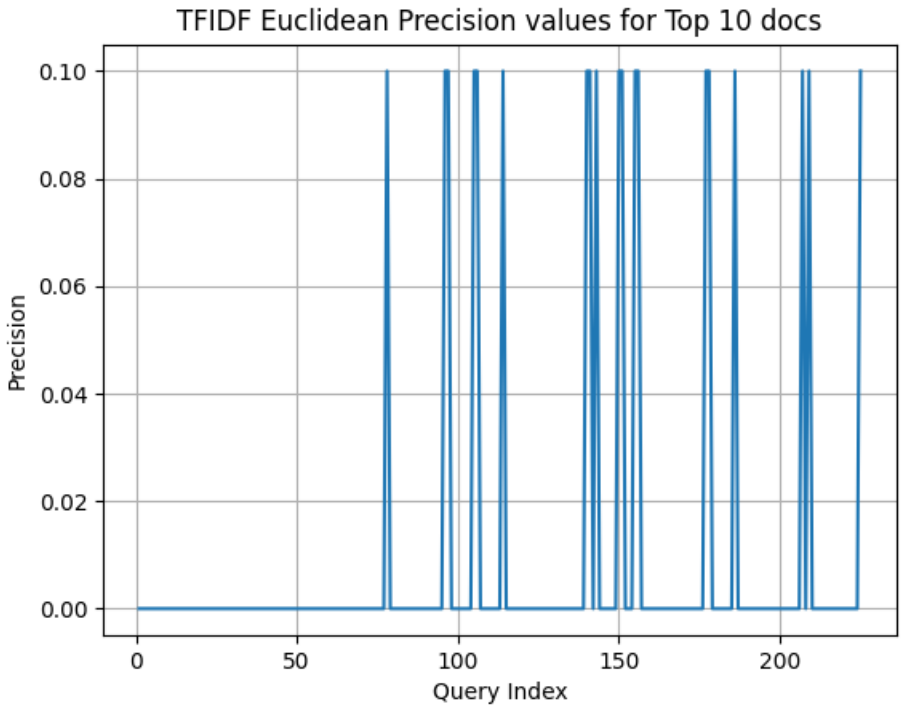


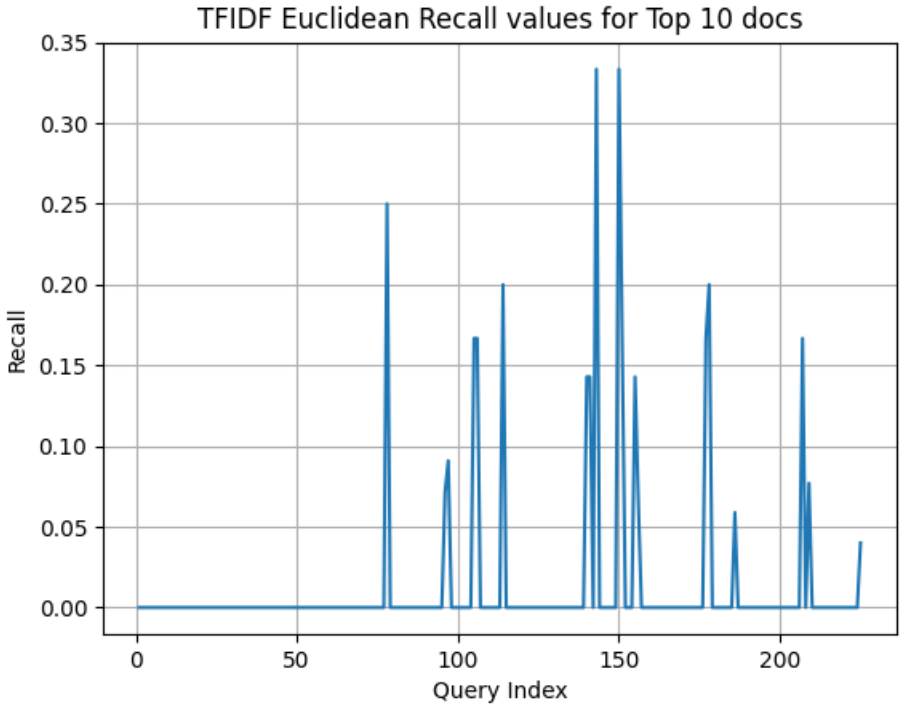




Precision, Recall, and F-Scores for TF-IDF with Euclidean Distance for Top 10 documents:




TFIDF Euclidean Precision values for Top 10 docs
Precision
0.10
0.08
0.06
0.04
0.02
0.00
0
50
100
150
200
Query Index


TFIDF Euclidean Recall values for Top 10 docs
Recall
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
50
100
150
200
Query Index

Thank you.