The Faculty of Information and Communication Technology
Mahidol University

Report
Project 2
MuGle: One Step Closer to Google

Presented by

Wachrayana	Wanprasert	6088082
Chavanont	Sakolpongpairoj	6088157
Panaya	Sirilertworakul	6088164

Section 3

Presented to

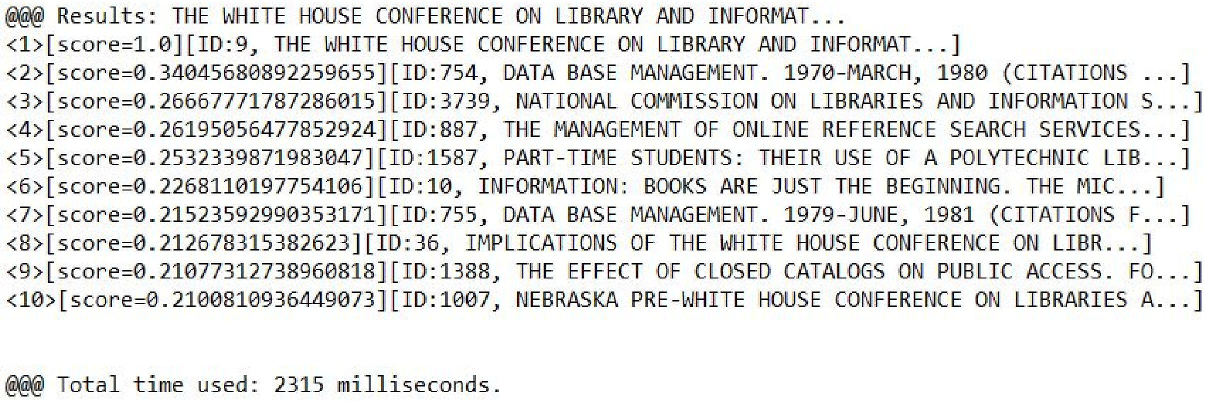
Asst. Prof. Dr. Charnyote Pluempitiwiriyawej

This document is submitted in partial fulfillment of
the requirements of the ITCS414 Information Storage and Retrieval
Semester 1, 2019

1. Which search algorithm (Jaccard vs. TFIDF) is a better search algorithm for the LISA corpus, in terms of relevance and time consumption? Quantitatively justify your reason scientifically and statistically (i.e. avoid using your gut feelings).

Answer:

TFIDF



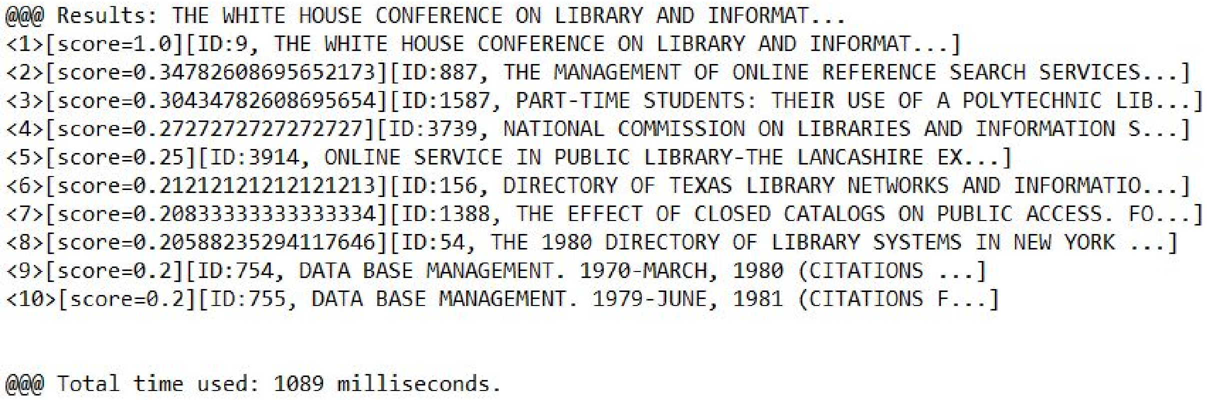
```
@@@ Results: THE WHITE HOUSE CONFERENCE ON LIBRARY AND INFORMAT...
<1>[score=1.0][ID:9, THE WHITE HOUSE CONFERENCE ON LIBRARY AND INFORMAT...]
<2>[score=0.34045680892259655][ID:754, DATA BASE MANAGEMENT. 1970-MARCH, 1980 (CITATIONS ...]
<3>[score=0.26667771787286015][ID:3739, NATIONAL COMMISSION ON LIBRARIES AND INFORMATION S...]
<4>[score=0.26195056477852924][ID:887, THE MANAGEMENT OF ONLINE REFERENCE SEARCH SERVICES...]
<5>[score=0.2532339871983047][ID:1587, PART-TIME STUDENTS: THEIR USE OF A POLYTECHNIC LIB...]
<6>[score=0.2268110197754106][ID:10, INFORMATION: BOOKS ARE JUST THE BEGINNING. THE MIC...]
<7>[score=0.21523592990353171][ID:755, DATA BASE MANAGEMENT. 1979-JUNE, 1981 (CITATIONS F...]
<8>[score=0.212678315382623][ID:36, IMPLICATIONS OF THE WHITE HOUSE CONFERENCE ON LIBR...]
<9>[score=0.21077312738960818][ID:1388, THE EFFECT OF CLOSED CATALOGS ON PUBLIC ACCESS. FO...]
<10>[score=0.2100810936449073][ID:1007, NEBRASKA PRE-WHITE HOUSE CONFERENCE ON LIBRARIES A...]


@@@ Total time used: 2315 milliseconds.
```

Figure1.1 Total time used by TFIDF Searcher

TFIDF is the multiplication between of term frequency and inverse document frequency. For Term Frequency, we use log normalization to calculate term frequency. For Inverse Document Frequency, it is used to determine the priority of each term in the document. In this case, TFIDF has used the total time for 2315 milliseconds.

Jaccard



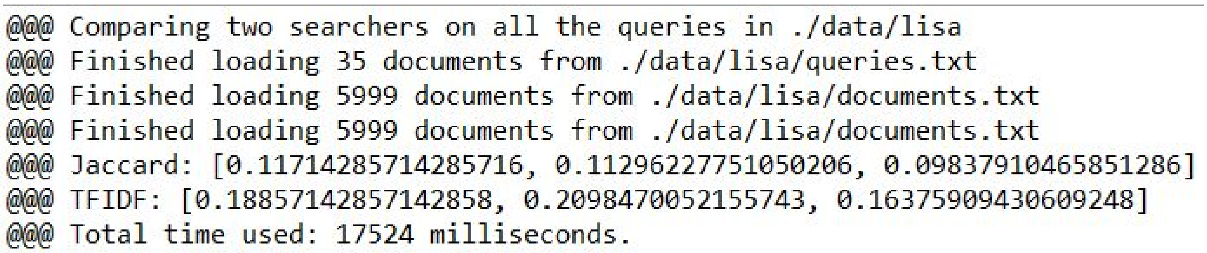
```
@@@ Results: THE WHITE HOUSE CONFERENCE ON LIBRARY AND INFORMAT...
<1>[score=1.0][ID:9, THE WHITE HOUSE CONFERENCE ON LIBRARY AND INFORMAT...]
<2>[score=0.34782608695652173][ID:887, THE MANAGEMENT OF ONLINE REFERENCE SEARCH SERVICES...]
<3>[score=0.30434782608695654][ID:1587, PART-TIME STUDENTS: THEIR USE OF A POLYTECHNIC LIB...]
<4>[score=0.2727272727272727][ID:3739, NATIONAL COMMISSION ON LIBRARIES AND INFORMATION S...]
<5>[score=0.25][ID:3914, ONLINE SERVICE IN PUBLIC LIBRARY-THE LANCASHIRE EX...]
<6>[score=0.21212121212121213][ID:156, DIRECTORY OF TEXAS LIBRARY NETWORKS AND INFORMATIO...]
<7>[score=0.20833333333333334][ID:1388, THE EFFECT OF CLOSED CATALOGS ON PUBLIC ACCESS. FO...]
<8>[score=0.20588235294117646][ID:54, THE 1980 DIRECTORY OF LIBRARY SYSTEMS IN NEW YORK ...]
<9>[score=0.2][ID:754, DATA BASE MANAGEMENT. 1970-MARCH, 1980 (CITATIONS ...]
<10>[score=0.2][ID:755, DATA BASE MANAGEMENT. 1979-JUNE, 1981 (CITATIONS F...]


@@@ Total time used: 1089 milliseconds.
```

Figure1.2 Total time used by Jaccard Searcher

For Jaccard, it is used the formula intersection of each term divided by union of each term to get the similarity score. In this case, Jaccard has used the total time for 1089 milliseconds.

Precision, Recall, and F1



```
@@@ Comparing two searchers on all the queries in ./data/lisa
@@@ Finished loading 35 documents from ./data/lisa/queries.txt
@@@ Finished loading 5999 documents from ./data/lisa/documents.txt
@@@ Finished loading 5999 documents from ./data/lisa/documents.txt
@@@ Jaccard: [0.11714285714285716, 0.11296227751050206, 0.09837910465851286]
@@@ TFIDF: [0.18857142857142858, 0.2098470052155743, 0.16375909430609248]
@@@ Total time used: 17524 milliseconds.
```

Figure1.3 Total time used of Compare Two Searchers On All Queries

In conclusion, it can be described that the TFIDF algorithm has more relevance and quality of searching than the Jaccard algorithms. In contrast, Jaccard algorithm has used less time consuming than the TFIDF because the Jaccard use only intersection and union of each term to find the score of the similarity while TFIDF uses more method such as term frequency, inverse document frequency which they are complex algorithm.

2. Currently, k is fixed at 10. Compute the average precision, recall, F1 for both the search systems for each k (i.e. precision@k, recall@k, and F1@k), where k ranges from 1...50. (You should write a script that automatically does this for you, instead of manually changing k.) Visualize your findings on beautiful and illustrative plots. What conclusions can you make?

Answer:

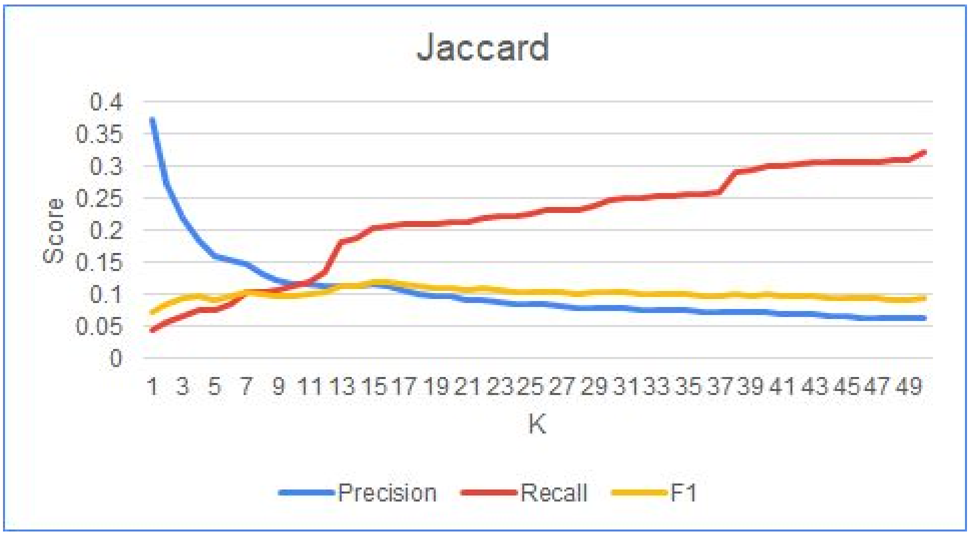


Figure 2.1 Jaccard average of Precision, Recall, F1

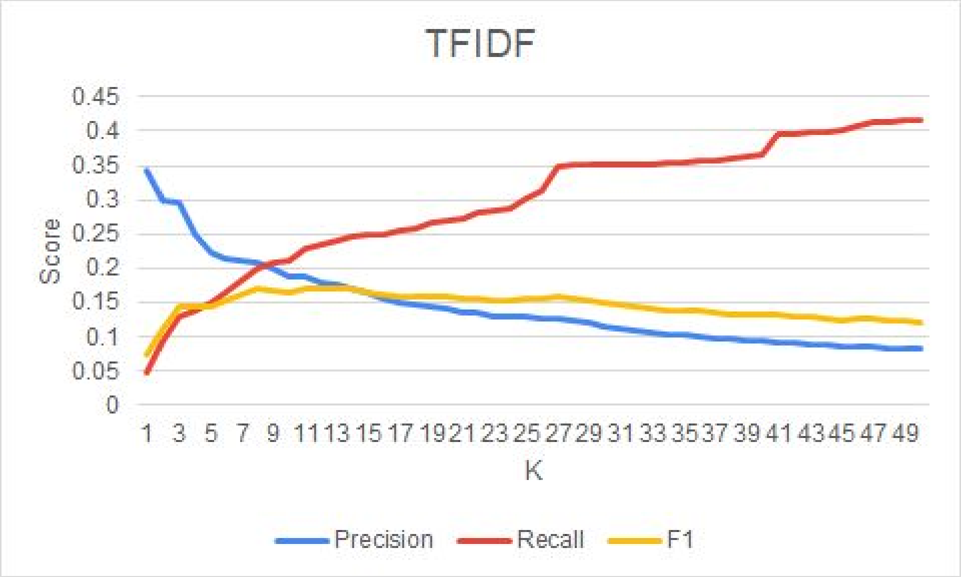


Figure 2.2 TFIDF average of Precision, Recall, F1

We plot the graph that represents about average between Precision, Recall, and F1 of Jaccard and TFIDF in figure 2.1 and figure 2.2. When the value of k is 1 to 50.

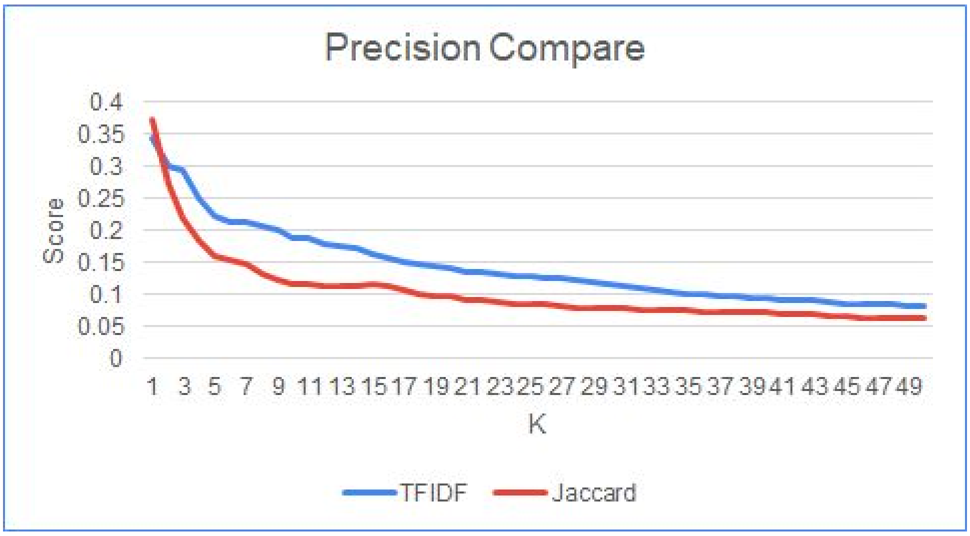


Figure 2.3 Compare between Precision of Jaccard and Precision of TFIDF

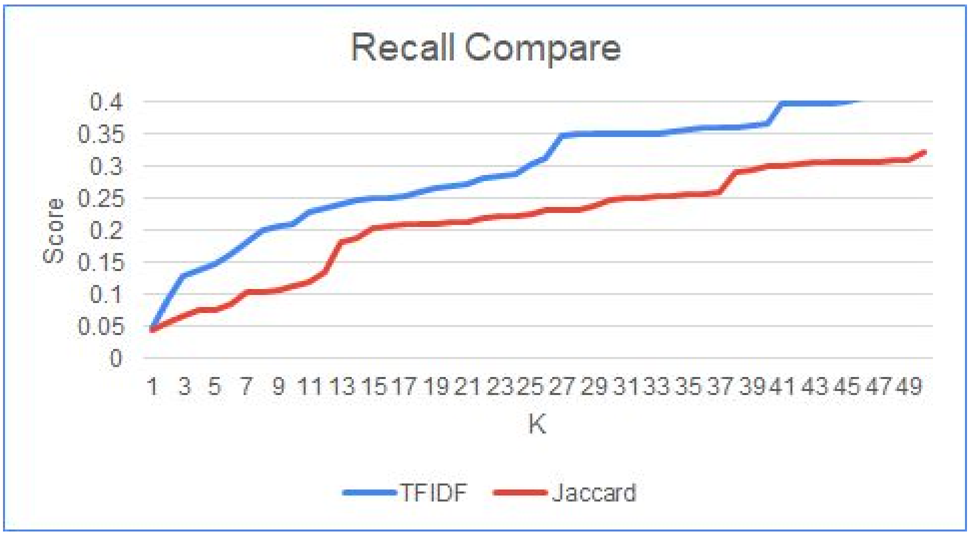
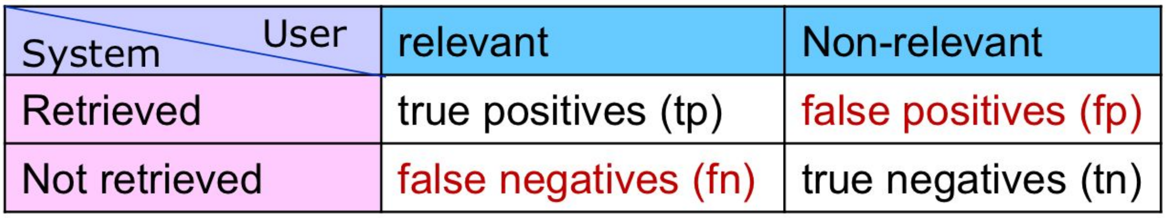


Figure 2.4 Compare between Recall of Jaccard and Recall of TFIDF

For figure 2.3 and figure 2.4 that represent the comparison between Precision and Recall of Jaccard and TFIDF by using formula on the below.



User System	relevant	Non-relevant
Retrieved	true positives (tp)	false positives (fp)
Not retrieved	false negatives (fn)	true negatives (tn)

Precision = #(relevant items retrieved) / #(retrieved items)
= tp / (tp + fp) = Prob(relevant | retrieved)
Recall = #(relevant items retrieved) / #(relevant items)
= tp / (tp + fn) = Prob(retrieved | relevant)

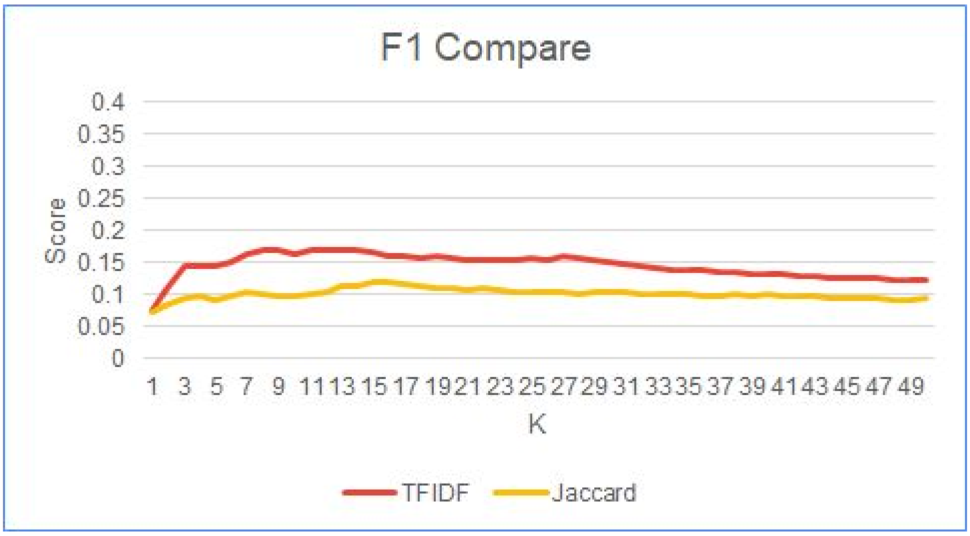


Figure 2.5 Compare between F1 of Jaccard and F1 of TFIDF

The figure 2.5 is showing the comparison between F1 of Jaccard and F1 of TFIDF by using the combination of Precision and Recall with follow by formula:

$$F1 = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall}$$

In conclusion, we take three values which are precision, recall, and F1 of TF-IDF and Jaccard score. It can be seen that for the Jaccard algorithm, the precision of Jaccard will be better than the TF-IDF algorithms if k is range 1-3 but afterward the TF-IDF algorithm is better. Moreover, in recall and F1 score, the TF-IDF has a better score than the Jaccard algorithm in almost every k values.

3. From 2.), generate precision vs recall plots for each search system. Explain how you can use these plots to explain the performance of each search algorithm.

Answer:

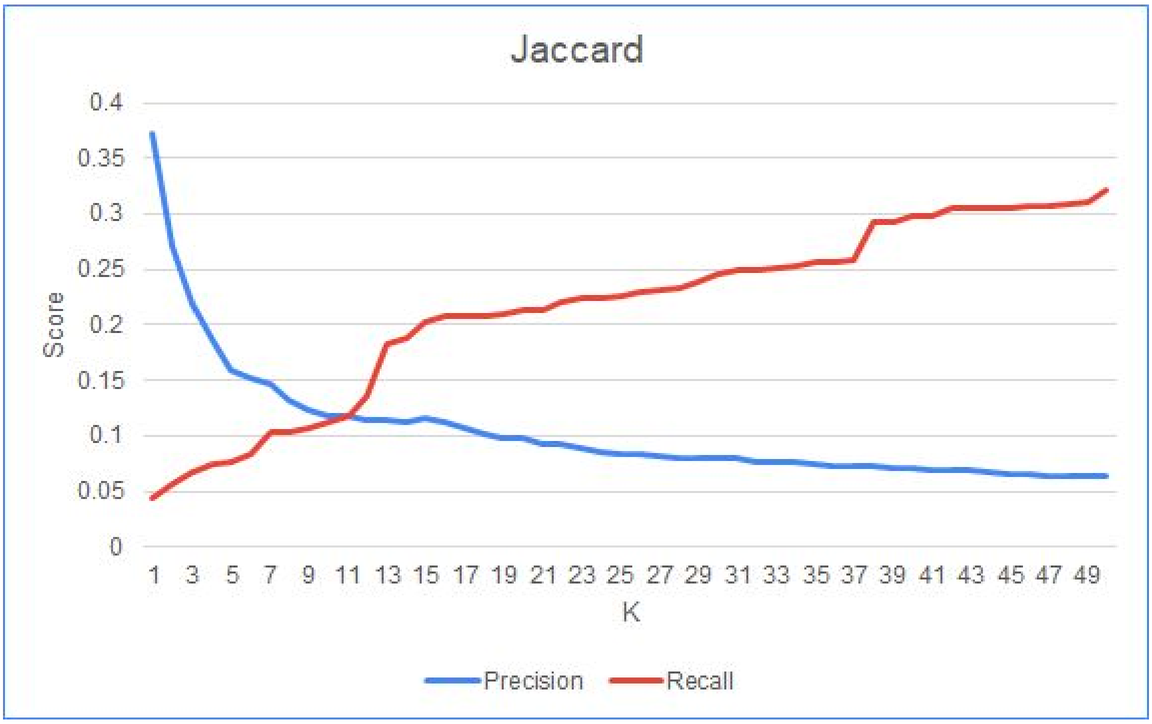


Figure 3.1 Precision and Recall of Jaccard score

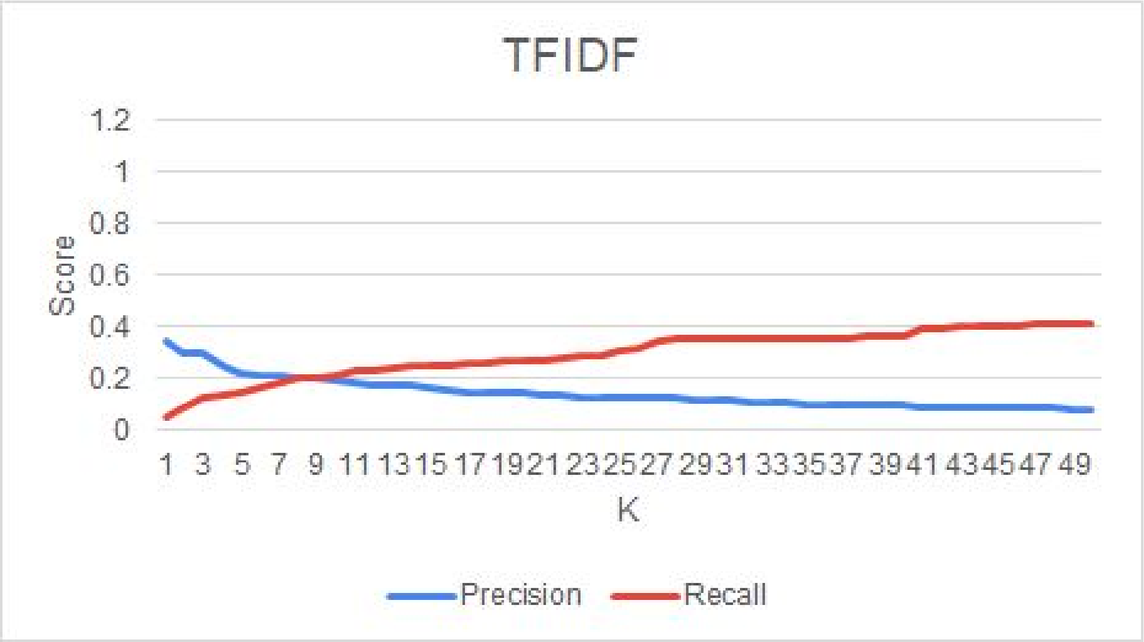


Figure 3.2 Precision and Recall of TFIDF score

Precision is the ratio of correctly predicted positive observations of the total predicted positive observations while Recall is the ratio of correctly predicted positive observations to all observations in the actual class. So, if the precision and recall value is nearing each other, it can be described that the data is quite good. In contrast, if precision and recall value is far away from each other, it can be described that the data is not good. So, in our graph, we can see that there are few that precision and recall value is nearing each other. We can conclude that this data is not quite good.