Mini Project Report

MuGle: One Step Closer to Google

BY

Miss	Kanrawee	Chiamsakul	6188049
Mr.	Tharit	Chantanalertvilai	6188068
Mr.	Thanyanit	Jongjitragan	6188075

Present

Asst. Prof. Dr. Charnyote Pluempitiwiriyawej

ITCS414 Information Storage and Retrieval

Faculty of Information and Communication Technology

Mahidol University

2020

Question 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:

Time Consumption

	Jaccard	TFIDF
1.	2346	3418
2.	2397	3576
3.	2169	3337
4.	2264	3279
5.	3256	3784
6.	2161	3308
7.	2342	3372
8.	2227	3430
9.	2109	3514
10.	2136	3344
AVG	**2340.7**	**3436.2**

This table is a time consumption (in millisecond) of each algorithm applied with the same *testQueries* on the same documents (LISA), recorded 10 times.

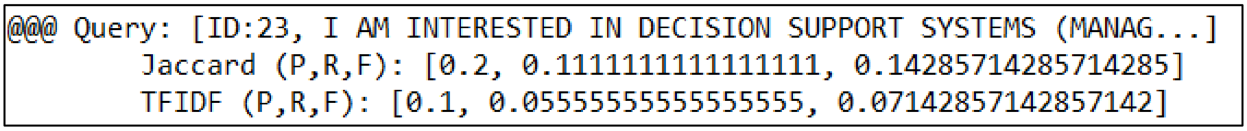
From the table, Jaccard algorithm consumes 31.88% less time than TFIDF algorithm on average.

Thus, Jaccard algorithm clearly better than TF-IDF algorithm in term of time consumption.

Relevance

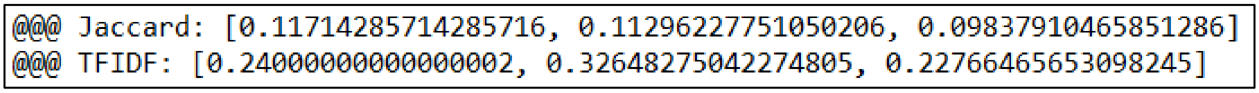
According to precision, recall, and F1 score of each algorithm, based on information from relevance.txt, TFIDF produces a score more than or equal to Jaccard algorithm except query ID 23.

Picture of comparison of both searchers on query ID 23



```
@@@ Query: [ID:23, I AM INTERESTED IN DECISION SUPPORT SYSTEMS (MANAG...]
        Jaccard (P,R,F): [0.2, 0.1111111111111111, 0.14285714285714285]
        TFIDF (P,R,F): [0.1, 0.05555555555555555, 0.07142857142857142]
```

Picture of comparison of both searchers on all queries



```
@@@ Jaccard: [0.11714285714285716, 0.11296227751050206, 0.09837910465851286]
@@@ TFIDF: [0.24000000000000002, 0.32648275042274805, 0.22766465653098245]
```

Thus, in term of relevance of the result, TFIDF algorithm generally produces a better search result.

TFIDF is designed to take into consideration term and document frequency. The process of word value comparison results in a more meaningful result despite longer time consumption.

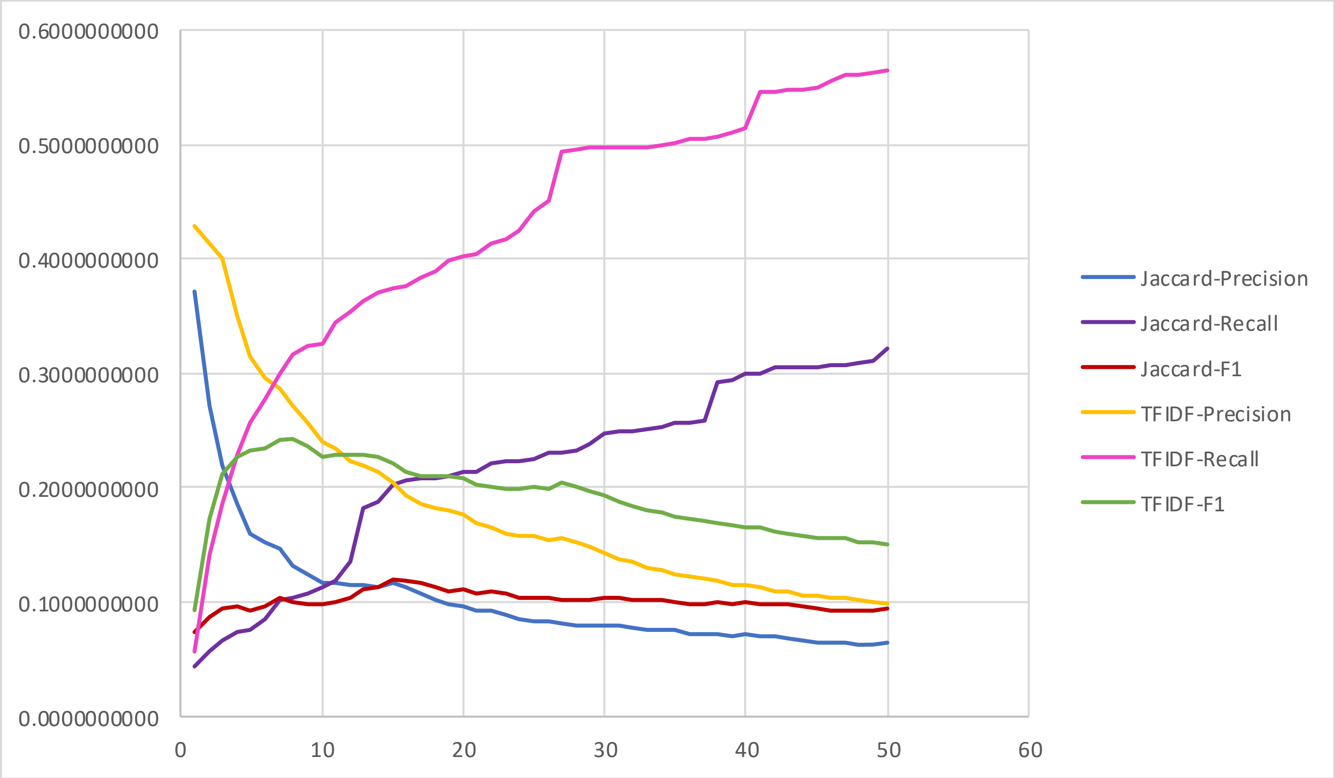
Question 2: Currently, k is fixed at 10. Compute the average precision, recall, F1 for both the search systems for each k (i.e. precision@k, recall@k, and F1@k), where k ranges from 1…50. (You should write a script that automatically does this for you, instead of manually changing k.) Visualize your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

Answer:

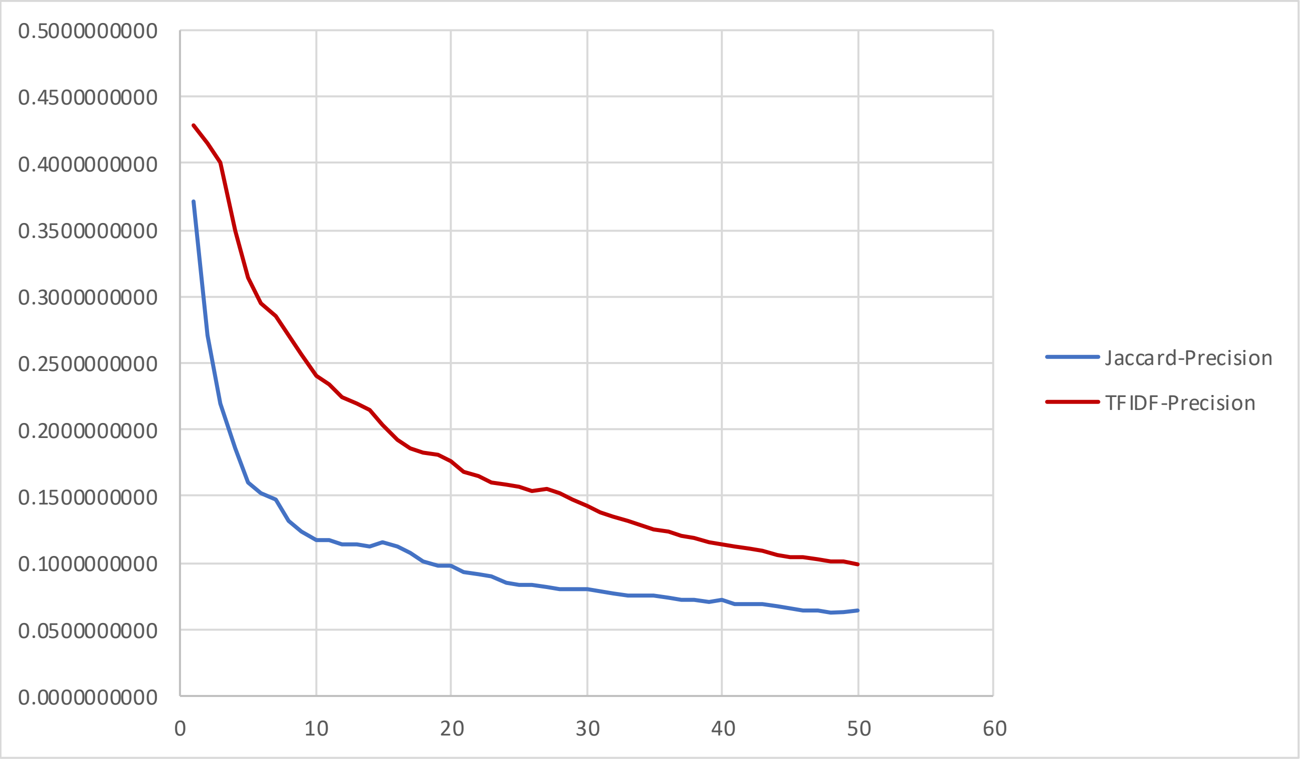
k	Jaccard Search Algorithm			TFIDF Search Algorithm		
	Jaccard-Precision	Jaccard-Recall	Jaccard-F1	TFIDF-Precision	TFIDF-Recall	TFIDF-F1
1	0.3714285714	0.04385718023	0.0738745326	0.42857142860	0.0568854218	0.09299017885
2	0.2714285714	0.0568629896	0.08630778745	0.4142857143	0.1419761442	0.1736862978
3	0.219047619	0.06686298961	0.09361631832	0.4	0.186558948	0.2112999046
4	0.1857142857	0.07454493031	0.09625004347	0.35	0.2294109568	0.2274796971
5	0.16	0.07667131545	0.09265973226	0.3142857143	0.2575010607	0.2316648605
6	0.1523809524	0.08422957032	0.09693476222	0.2952380952	0.2773095912	0.2352536334
7	0.1469387755	0.1026775566	0.1034013271	0.2857142857	0.2995617236	0.2426397485
8	0.1321428571	0.1039197926	0.09935068679	0.2714285714	0.3172089832	0.2426430749
9	0.1238095238	0.1076612892	0.09873887653	0.2571428571	0.3236025297	0.2366496303
10	0.1171428571	0.1129622775	0.0983791047	0.24	0.3264827504	0.2276646565
11	0.1168831169	0.1182234506	0.1004179712	0.2337662338	0.3441236671	0.2296735463
12	0.1142857143	0.1357922172	0.1040043782	0.2238095238	0.3536598452	0.2280182955
13	0.1142857143	0.1830711288	0.1119643959	0.2197802198	0.3631387054	0.2279933185
14	0.112244898	0.1883949975	0.1129628343	0.2142857143	0.3711455455	0.2267542587
15	0.1161904762	0.2024562593	0.1198030552	0.2038095238	0.3740655814	0.2205572409
16	0.1125	0.2070944782	0.1193826492	0.1928571429	0.3753078174	0.2135431971
17	0.1075630252	0.2083367143	0.1165760332	0.1865546218	0.3827856765	0.2110126619
18	0.1015873016	0.2083367143	0.1126178035	0.1825396825	0.3892084463	0.2098428457
19	0.0977443609	0.2095789503	0.1102985066	0.1804511278	0.3990449452	0.2110139631
20	0.09714285714	0.2137461518	0.1108027295	0.1757142857	0.4024135664	0.2078141924
21	0.0925170068	0.2137461518	0.1074669312	0.168707483	0.4047945187	0.2031019755
22	0.09220779221	0.2206344417	0.1085694980	0.1649350649	0.4130887043	0.2014820324
23	0.08944099379	0.2234915845	0.1072446197	0.1602484472	0.4163717567	0.1984273930
24	0.08571428571	0.2234915845	0.1043142077	0.1583333333	0.4253008629	0.1983762115
25	0.0834285714	0.2258725369	0.1030925159	0.1577142857	0.4414410581	0.2009390522
26	0.08351648352	0.2303727985	0.104179385	0.1538461538	0.4512369765	0.1992061825
27	0.08148148148	0.2314716996	0.1026602502	0.1555555556	0.4944190187	0.2046230526
28	0.0795918367	0.2323375005	0.1011120395	0.1520408163	0.4963186465	0.2013527124
29	0.07980295567	0.2377644407	0.1024780341	0.1477832512	0.4968577301	0.1971379911
30	0.08	0.2466257244	0.1042279736	0.1428571429	0.4968577301	0.1924612930
31	0.07926267281	0.2499087767	0.1042234596	0.1382488479	0.4968577301	0.1880099128
32	0.07678571429	0.2499087767	0.1019502456	0.1348214286	0.4973968136	0.1844397528

33	0.07619047619	0.2520351619	0.1015626711	0.1307359307	0.4973968136	0.1803836741
34	0.0756302521	0.2536731465	0.1010642001	0.1277310924	0.499777766	0.1777507873
35	0.07510204082	0.2568096327	0.1009138796	0.1248979592	0.5011383102	0.1750384302
36	0.07301587302	0.2568096327	0.09892209504	0.123015873	0.5044213626	0.1735579541
37	0.07258687259	0.2587092605	0.09863035927	0.1196911197	0.5044213626	0.1700767977
38	0.07218045113	0.2920425938	0.09953376776	0.1180451128	0.5070012625	0.1684637872
39	0.07106227106	0.2931414949	0.09855821834	0.1157509158	0.5095986651	0.1663718945
40	0.07142857143	0.2988104291	0.09986207445	0.1142857143	0.514236884	0.1654190306
41	0.06968641115	0.2988104291	0.09807572774	0.1128919861	0.5451892649	0.1648046007
42	0.0693877551	0.304671235	0.09838499363	0.1102040816	0.5451892649	0.1618136871
43	0.06843853821	0.3055370359	0.09744010224	0.1089700997	0.5471539669	0.1605124232
44	0.06688311688	0.3055370359	0.09579284516	0.1064935065	0.5471539669	0.1577130957
45	0.0653968254	0.3055370359	0.09420199075	0.1047619048	0.5487412685	0.1559193501
46	0.06459627329	0.3064028367	0.09338795074	0.1043478261	0.5546020744	0.1560232037
47	0.06382978723	0.3079901383	0.09277143556	0.103343465	0.5602710086	0.1553846653
48	0.0630952381	0.3092323743	0.0921157743	0.1011904762	0.5602710086	0.152861986
49	0.06297376093	0.3106372588	0.09194768756	0.1002915452	0.5634074948	0.1519344457
50	0.064	0.3216525328	0.0940331997	0.09885714286	0.5646497308	0.1503345016

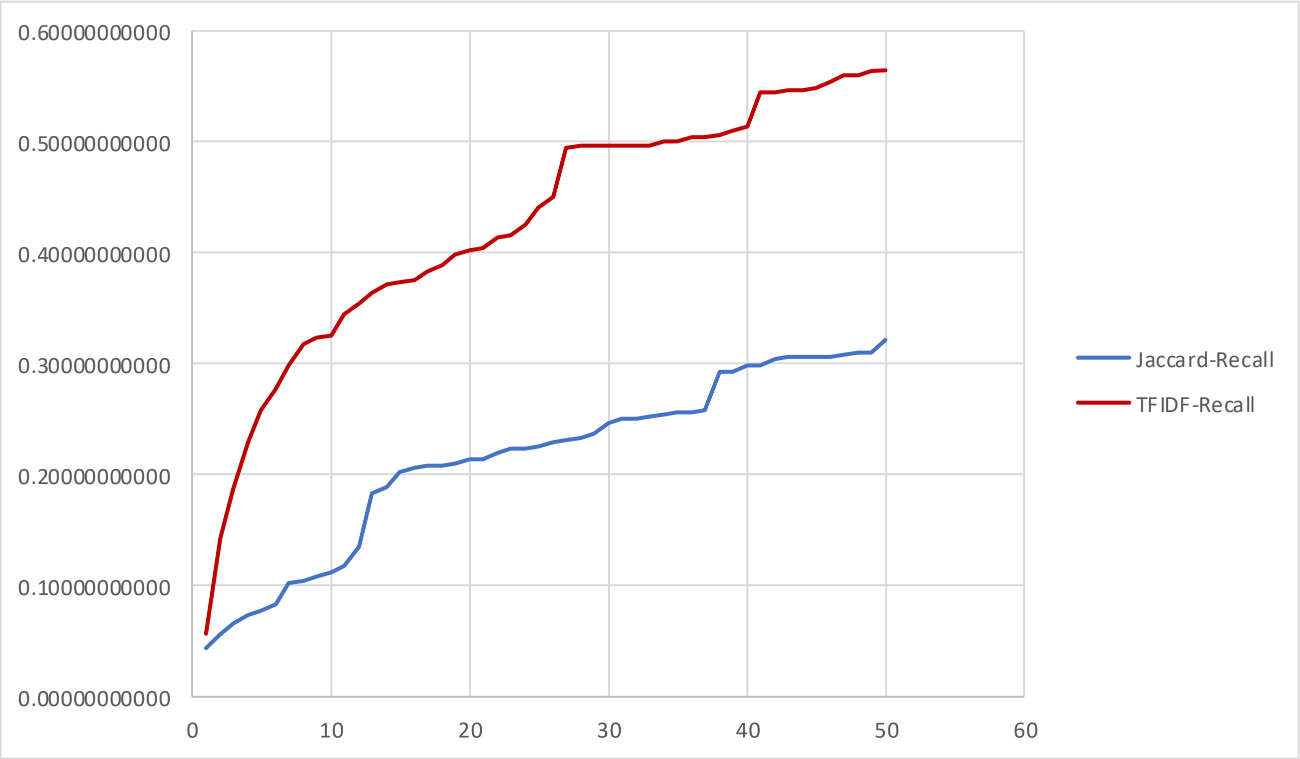
The average precision, recall and F1 of Jaccard and TFIDF



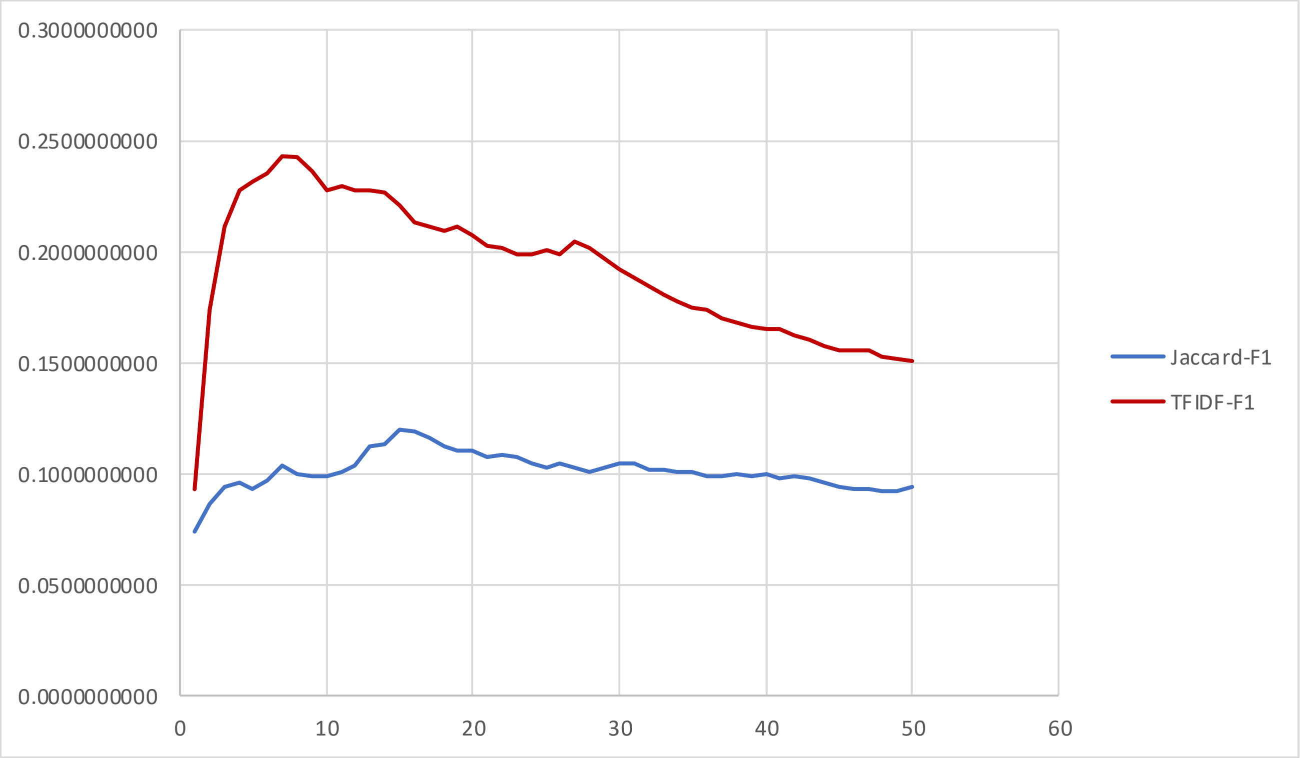
The average precision of Jaccard and TFIDF



The average recall of Jaccard and TFIDF



The average F1 of Jaccard and TFIDF

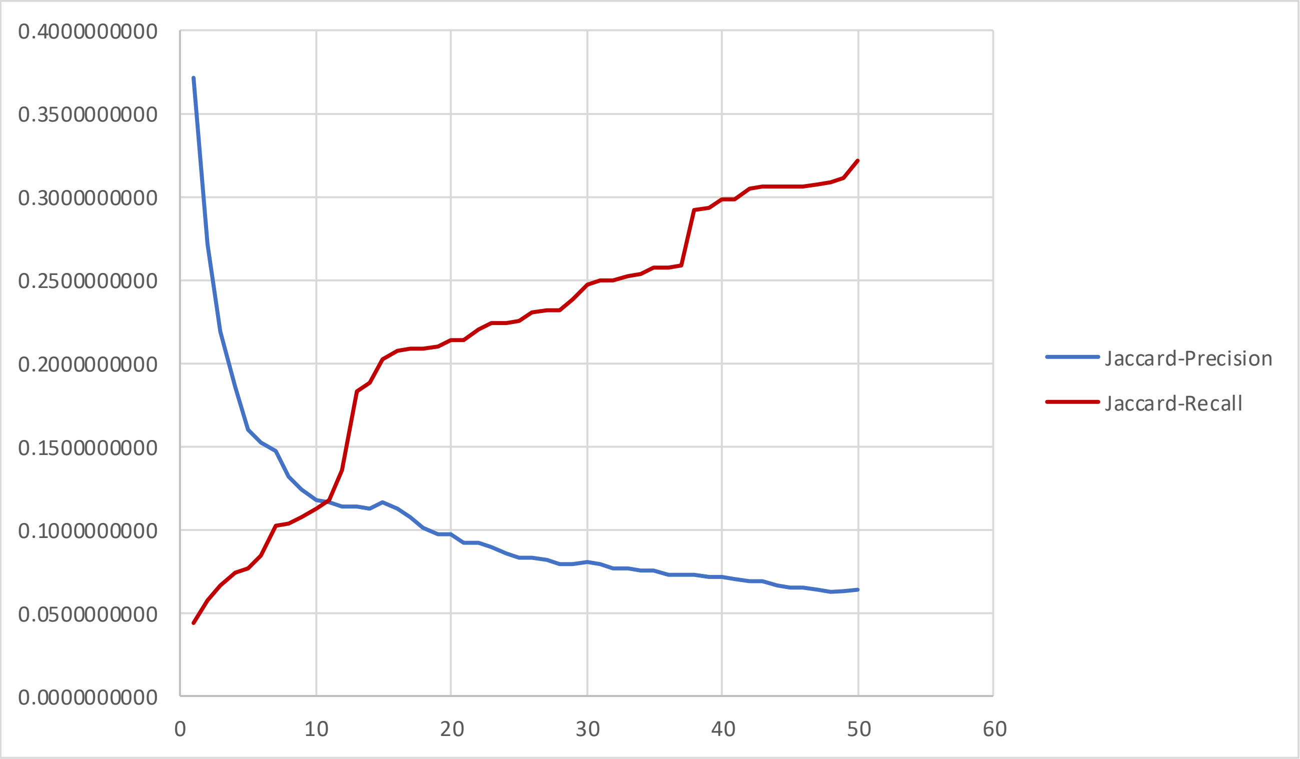


From the above illustrations, TFIDF algorithm has a better score, producing more relevant search results, than Jaccard algorithm.

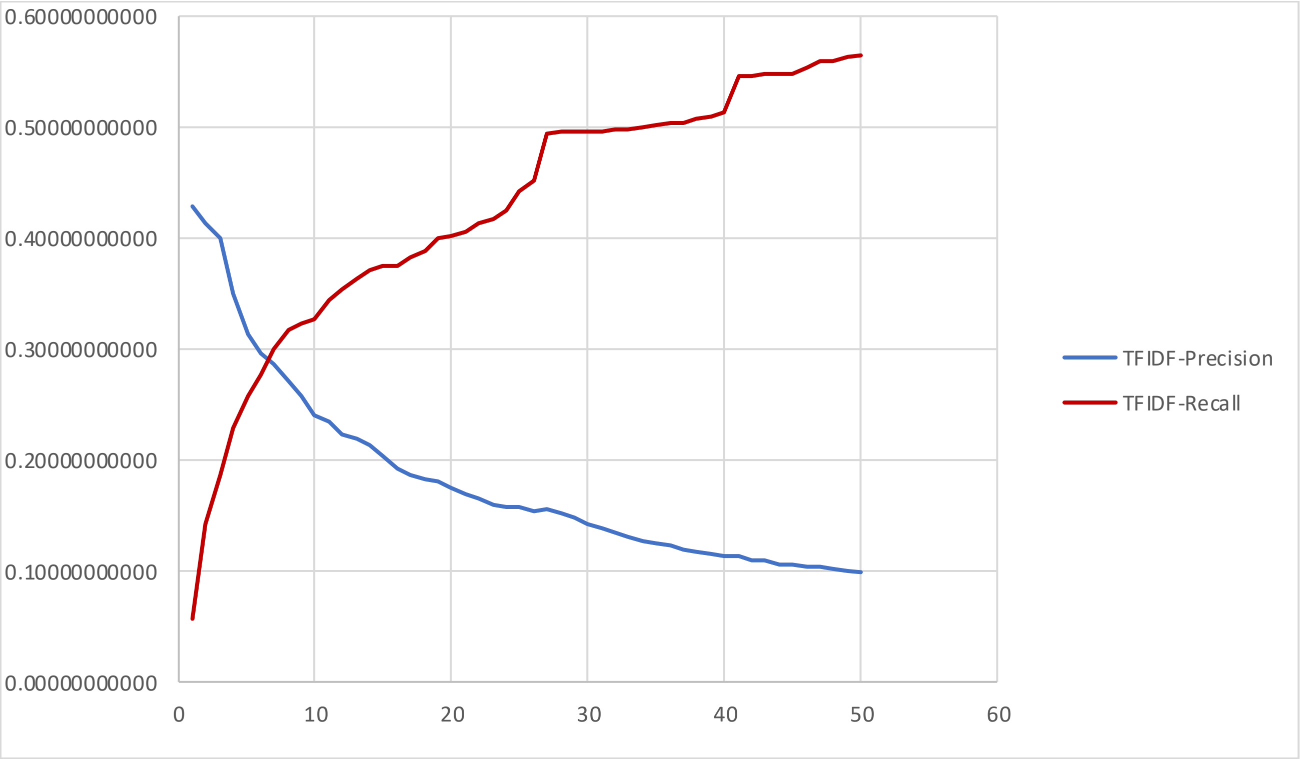
Question 3: From 2.), generate precision vs recall plots for each search system. Explain how you can use these plots to explain the performance of each search algorithm.

Answer:

Jaccard Precision VS Jaccard Recall



TFIDF Precision VS TFIDF Recall



At the first few K, the steeper of the line, the better the result because more steep means more relevant results gathered at the top of the search result list. Ideally, if all relevant search results are gathered before all irrelevant results, the precision line should become perfectly horizontal (steep = 0) before dropping significantly, and the recall line should skyrocket up to one point before become horizontal (steep = 0) and never increase until the end.

Judging from the steep of Jaccard algorithm, the precision line drops significantly at the first few K while the recall line steady going up. This means Jaccard’s top search results may not contain many relevant search results.

For TFIDF algorithm, at the beginning, the precision line drops steadily, and the recall rises up with a decent steep (compared to Jaccard).

From the illustrations, the performance of TFIDF is closer to the ideal than Jaccard.