Report Project 2: MuGle (One Step Closer to Google)

Section 1
Dujnapa Tanundet 6088105
Klinton chhun 6088111
Arada Puengmongkolchaikit 6088133

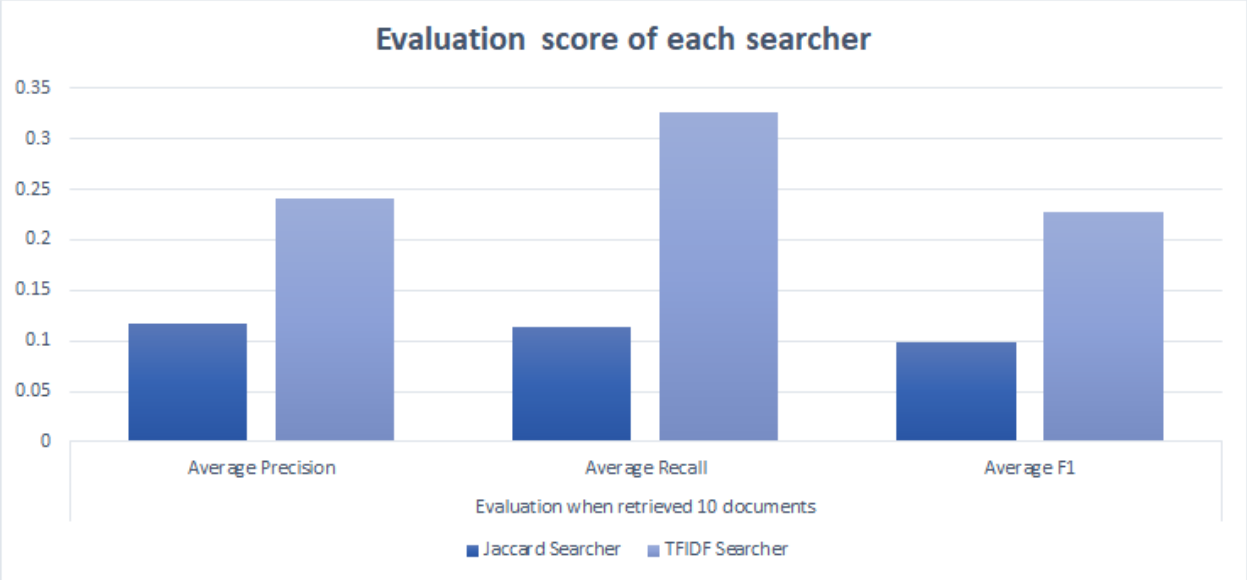
1. **Which search algorithm (Jaccard vs. TFIDF) is a better search algorithm for the LISA corpus, in terms of relevance and time consumption? Quantitatively justify your reason scientifically and statistically (i.e avoid using your gut feelings).**

Tested on Window 10 Pro Processor: Intel(R) Core(TM) i7-8550U CPU @1.80GHz 2.00GHz. RAM 16.0GB. Memory 500GB.	**Searcher**	**Time Used (milliseconds)**
	Jaccard Searcher	1003
	TFIDF Searcher	1082

Table (1) Time consumption of each searcher

Searcher	**Evaluation when retrieved 10 documents**		
	Average Precision	**Average Recall**	**Average F1**
Jaccard Searcher	0.11714285714285716	0.11296227751050206	0.09837910465851286
TFIDF Searcher	0.24000000000000002	0.32648275042274805	0.22766465653098245

Table (2) Evaluation scores of each searcher

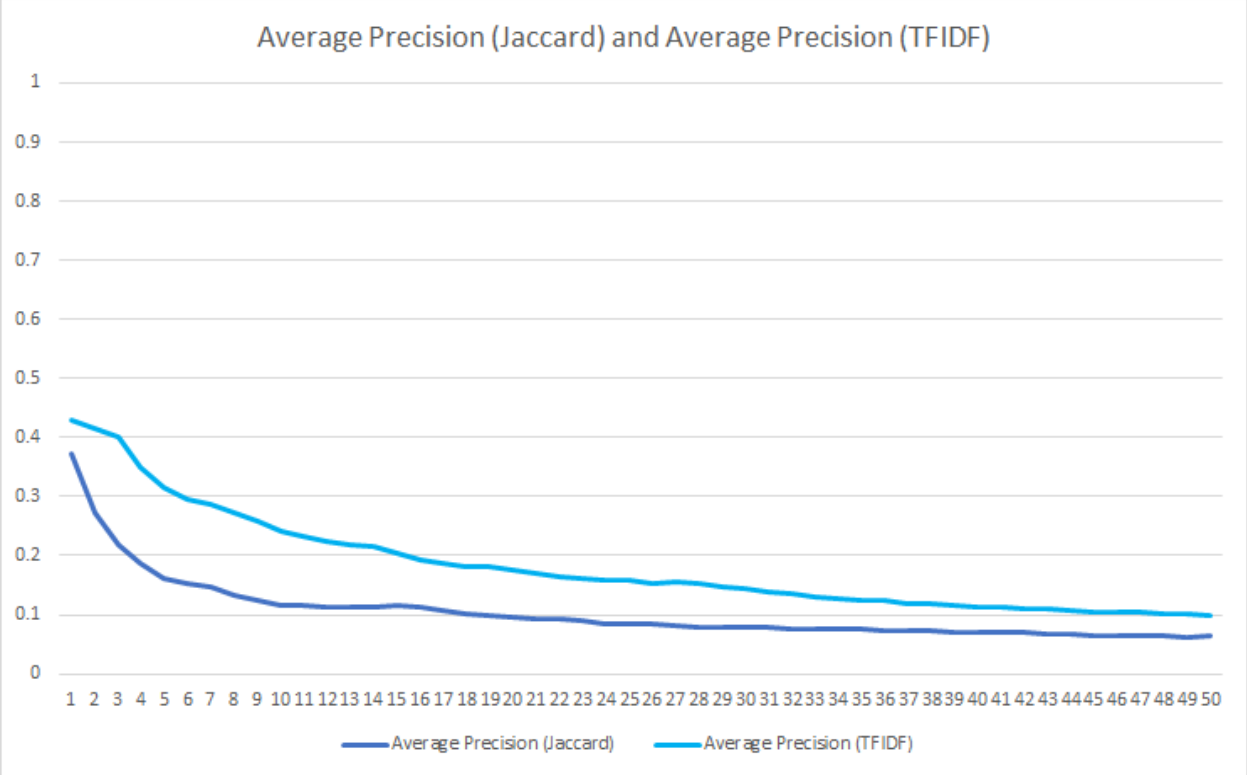


In terms of time consumption, the Jaccard searcher is a better search algorithm than the TFIDF searcher. According to the table above, which represents the amount of time consumption of each algorithm, the TFIDF searcher's time consumption is 79 milliseconds or 7.03% slower than Jaccard searcher's time consumption.

In terms of relevance, the TFIDF searcher is better. The reason is because of the value of Precision, Recall, and F1 is higher than Jaccard which is significantly shown on the table (2) and bar charts.

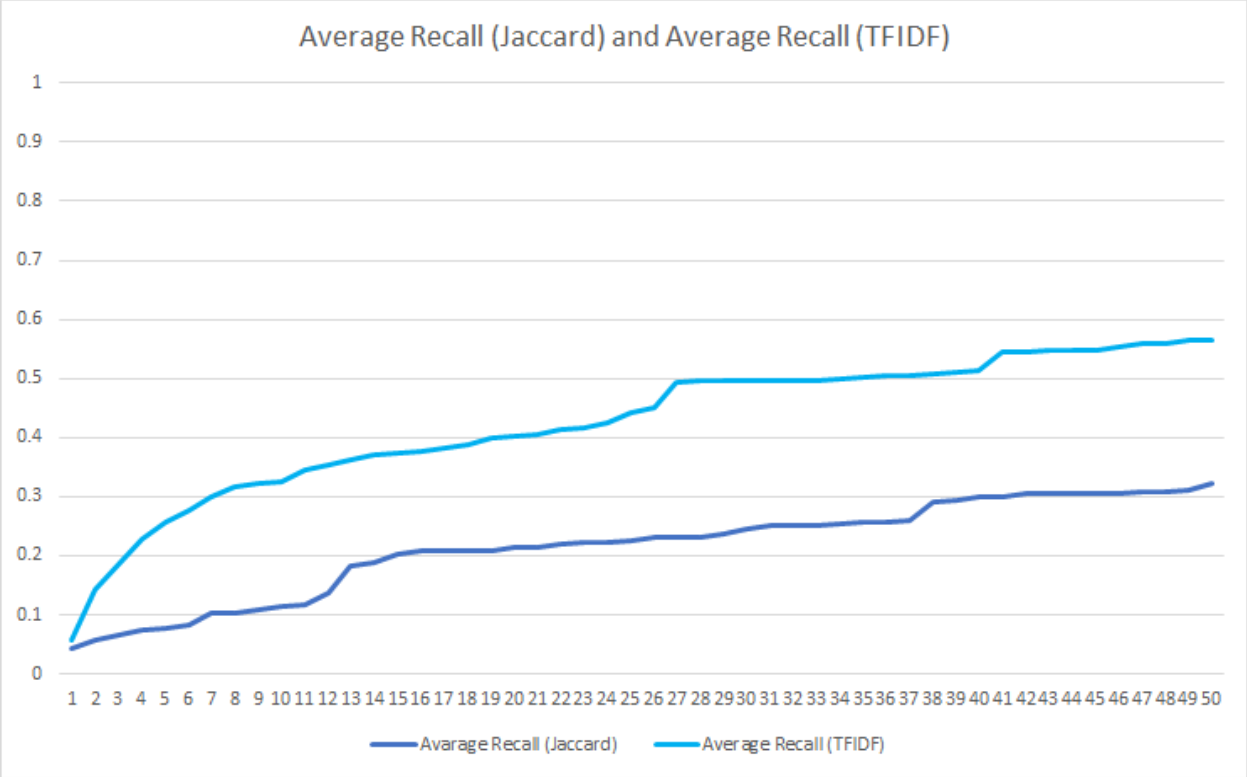
In conclusion, the Jaccard searcher is better in terms of time consumption while the TFIDF searcher is better in terms of relevance.

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 conclusion can you make?**



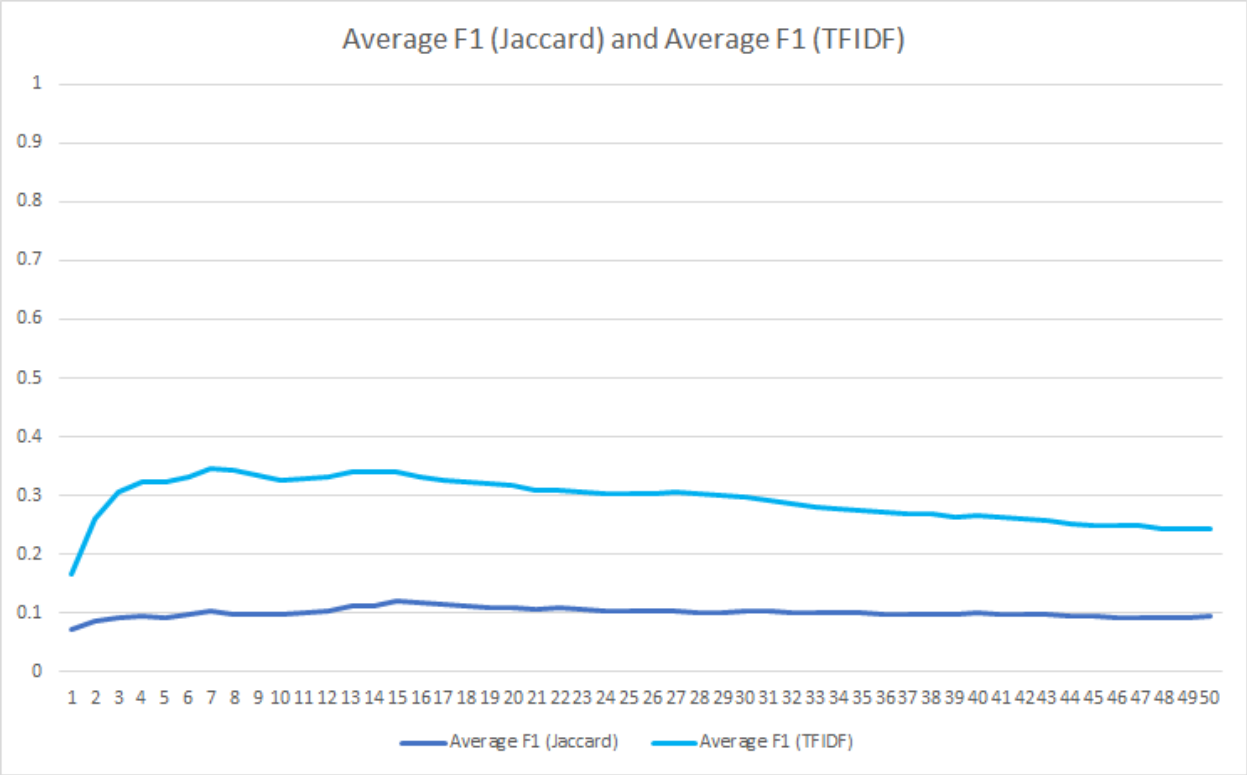
Line Graph (1) Average precision Jaccard and Average Precision TFIDF

According to the line graph (1) which represents the average precision of Jaccard searcher and TFIDF searcher, the TFIDF searcher has an average prediction value higher than the Jaccard searcher. As a result, the TFIDF returned substantially more relevant results among retrieved instances than Jaccard.



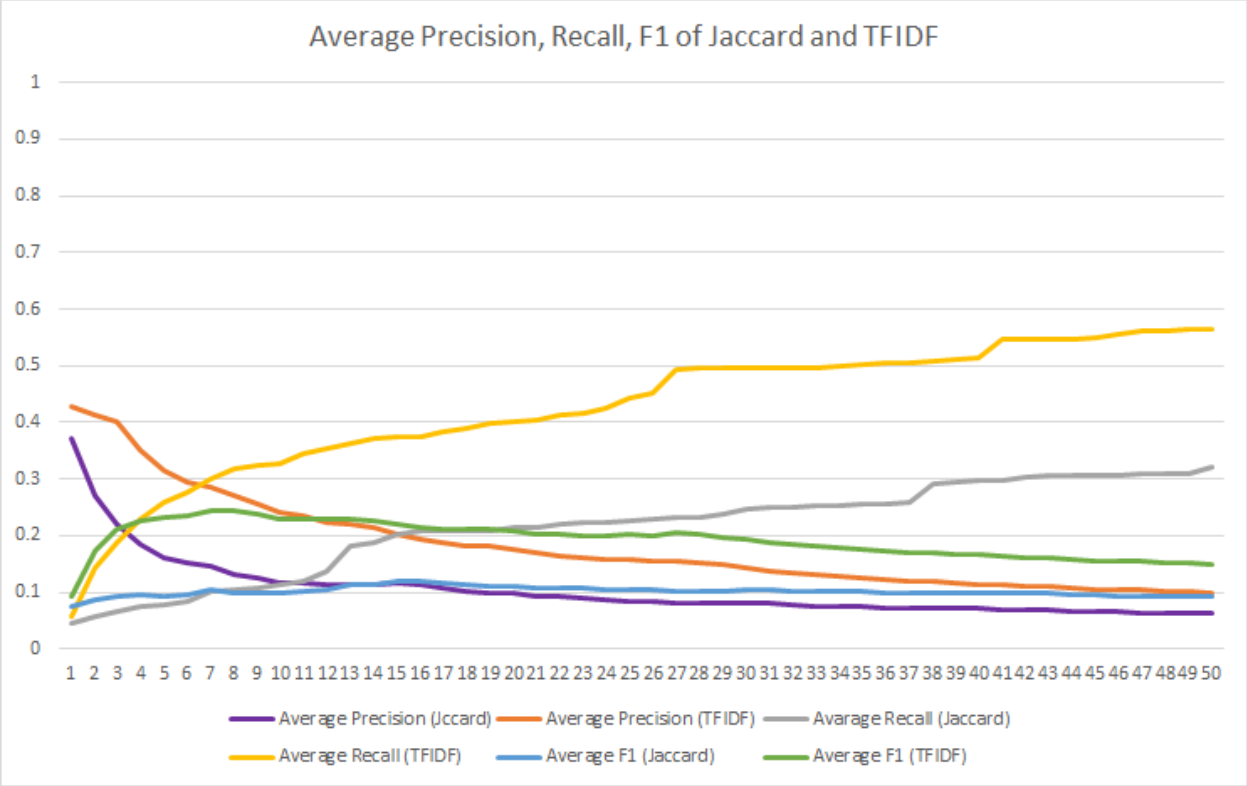
Line graph(2) Average Recall of Jaccard searcher and TFIDF searcher

According to the line graph(2) which shows the average recall of the Jaccard searcher and TFIDF searcher, the value of the average recall of the TFIDF searcher is higher than the Jaccard searcher. Therefore, there are more relevant items that are retrieved by the TFIDF search system than the Jaccard searcher.



Line graph(3) Average F1 of Jaccard searcher and TFIDF searcher

According to the line graph(3) which shows the average F1 of the Jaccard searcher and TFIDF searcher, the weighted average of Precision and Recall of the TFIDF searcher is higher than the Jaccard searcher.

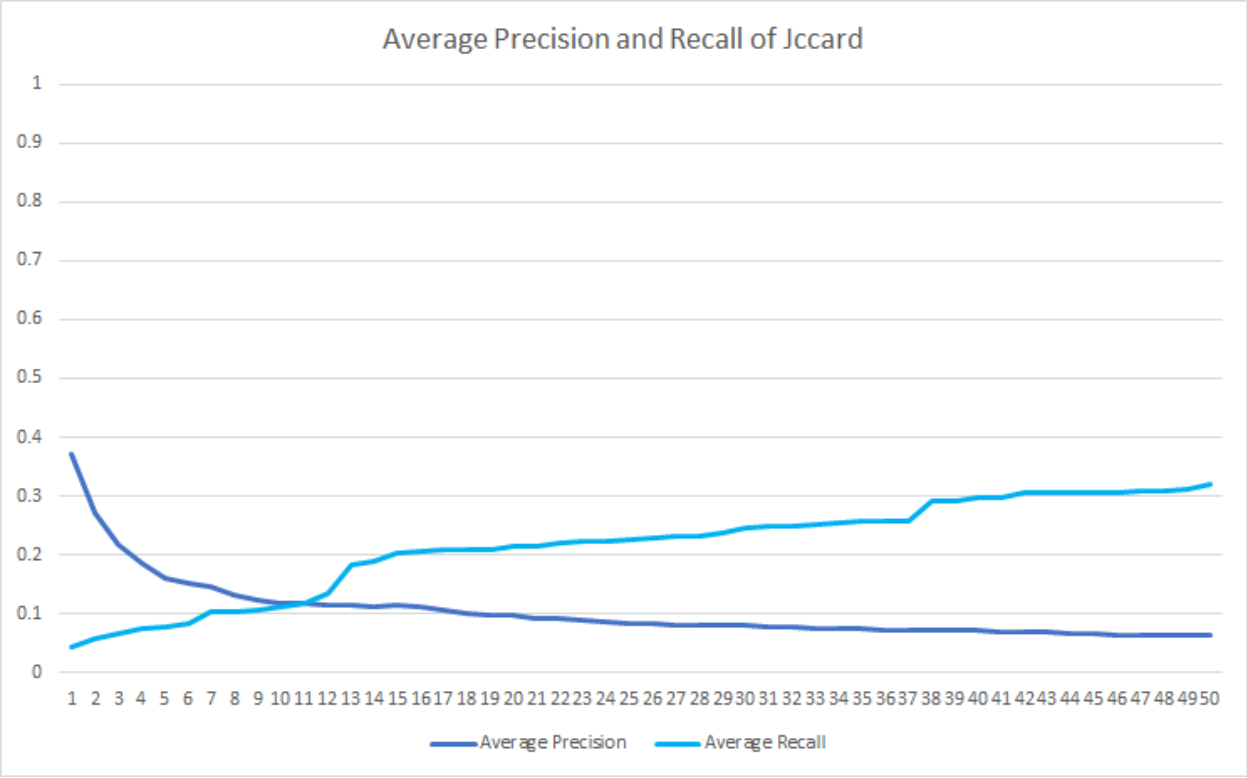


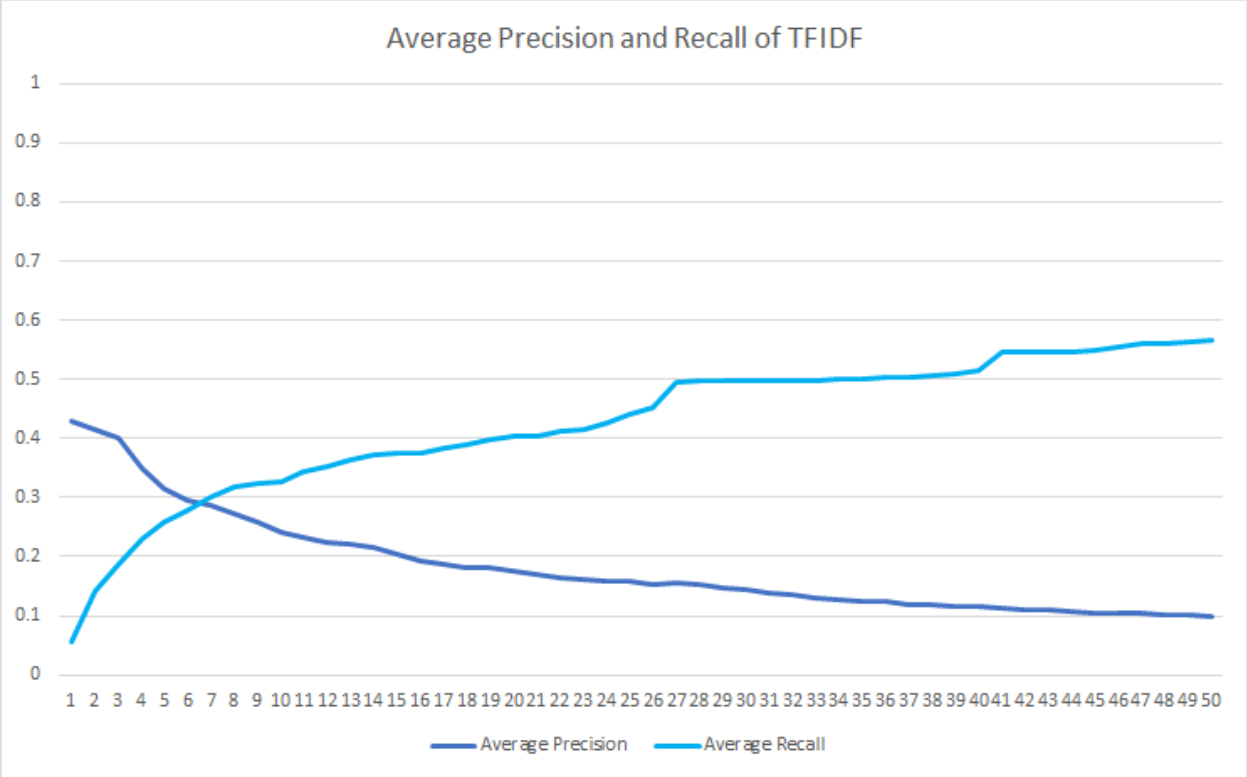
Line graph(4) Average Precision, Recall, and F1 of Jaccard searcher and TFIDF searcher

To sum up, the line graph represents that the overall performance of the TFIDF search system is better than the Jaccard search system, since the average precision, recall, and F1 value of the TFIDF is greater.

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.**

After we generated precision vs recall plots for Jaccard and TFIDF search system, we can illustrate the performance of each search system as follows.





Before talking about the performance, let's talk about the primary purpose of the search system first. The purpose of the search system is to locate the information that the user is seeking. Hence, the system is not concerned with finding a thousand documents that together provide a complete representation of the user's interest. Instead, the task became navigating the user to the one document they looking for.

The performance of the search algorithm can tell by using precision and recall value. The precision value tell how many selected sites are relevant, and the recall value tell how many relevant sites are selected. However, the aforementioned information confirms that in the search engine, Precision is more important than Recall.

According to the graph, the precision value of both the Jaccard and TFIDF search system is much less than the recall value. Moreover, the precision value is decreasing, while the recall value is increasing as the k increase. Therefore, we can conclude that the performance of both of the search algorithm is not really good since the precision value is low and keeps decreasing.

End