Data Mining Technology for Business and Society
Homework 1

Beatrice Nobile (1908315) Katsiaryna Zavadskaya (1847985)

April 2020

1 Search-Engine Evaluation

Each of the tested search engine configurations is made from one of analyzers: SimpleAnalyzer, StandardAnalyzer or LanguageAnalyzer('en'), combined with one of the scoring functions: Frequency, TF_IDF or BM25F. For the Cranfield dataset analysers are used on 'title' and 'content' fields, while for Time dataset only on the 'content' field.

The descriptive information about datasets, queries and Ground-Truth files is presented in the table 1.

Information	Dataset	
	Cranfield	Time
Number of indexed documents	1400	423
Number of queries	222	83
Number of queries in the GT	110	80

Table 1: Dataset information

All the nine search engine configurations were assessed by MRR metric for each of the two datasets, the results are in the next table.

N	Search Engine Configuration	MRR	
		Cranfield	Time
1	Simple Analyzer Frequency	0.0656	0.1612
2	Simple Analyzer TF_IDF	0.1744	0.2449
3	Simple Analyzer BM25F	0.4965	0.6666
4	Standard Analyzer Frequency	0.3180	0.4638
5	Standard Analyzer TF_IDF	0.3895	0.5366
6	Standard Analyzer BM25F	0.5187	0.6797
7	Language Analyzer Frequency	0.3446	0.4624
8	Language Analyzer TF_IDF	0.4114	0.5641
9	Language Analyzer BM25F	0.5095	0.7154

Table 2: MRR metric for all tested search engine configurations

The set of Top-5 search engine configurations according to the MRR table:

Cranfield dataset:

1. Standard Analyzer BM25F
2. Language Analyzer BM25F
3. Simple Analyzer BM25F
4. Language Analyzer TF-IDF
5. Standard Analyzer TF-IDF

Time dataset:

1. Language Analyzer BM25F
2. Standard Analyzer BM25F
3. Simple Analyzer BM25F
4. Language Analyzer TF-IDF
5. Standard Analyzer TF-IDF

For each of the Top-5 configurations for both of the datasets R-Precision distribution table can be found below together with graphs describing average Precision@k and average nDCG@k metrics for k taking $\{1, 3, 5, 10\}$ values.

Dataset	Search Engine Configuration	Mean	Min	1st quartile	Median	3rd quartile	Max
Cranfield	Standard Analyzer BM25F	0.258	0.0	0.0	0.25	0.429	1.0
	Language Analyzer BM25F	0.259	0.0	0.0	0.25	0.46	1.0
	Simple Analyzer BM25F	0.264	0.0	0.014	0.25	0.444	1.0
	Language Analyzer TF_IDF	0.194	0.0	0.0	0.143	0.333	1.0
	Standard Analyzer TF_IDF	0.177	0.0	0.0	0.143	0.286	1.0
Time	Language Analyzer BM25F	0.556	0.0	0.192	0.558	1.0	1.0
	Standard Analyzer BM25F	0.547	0.0	0.321	0.5	0.889	1.0
	Simple Analyzer BM25F	0.54	0.0	0.264	0.5	0.889	1.0
	Language Analyzer TF_IDF	0.309	0.0	0.0	0.333	0.5	1.0
	Standard Analyzer TF_IDF	0.285	0.0	0.0	0.171	0.51	1.0

Table 3: R-Precision distribution table

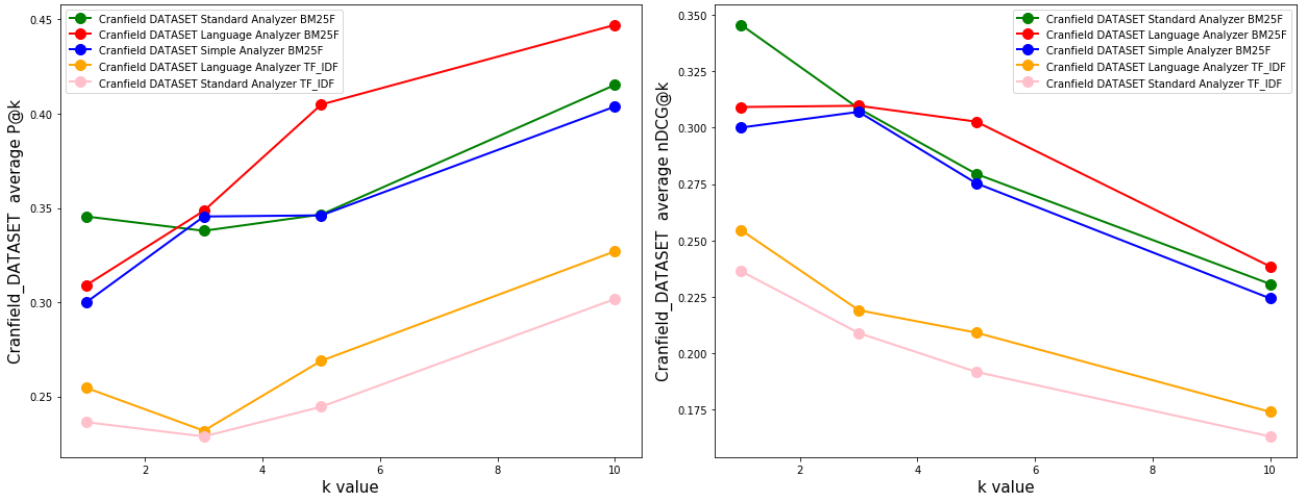

Figure 1: P@k plot and nDCG@k plot for Cranfield dataset

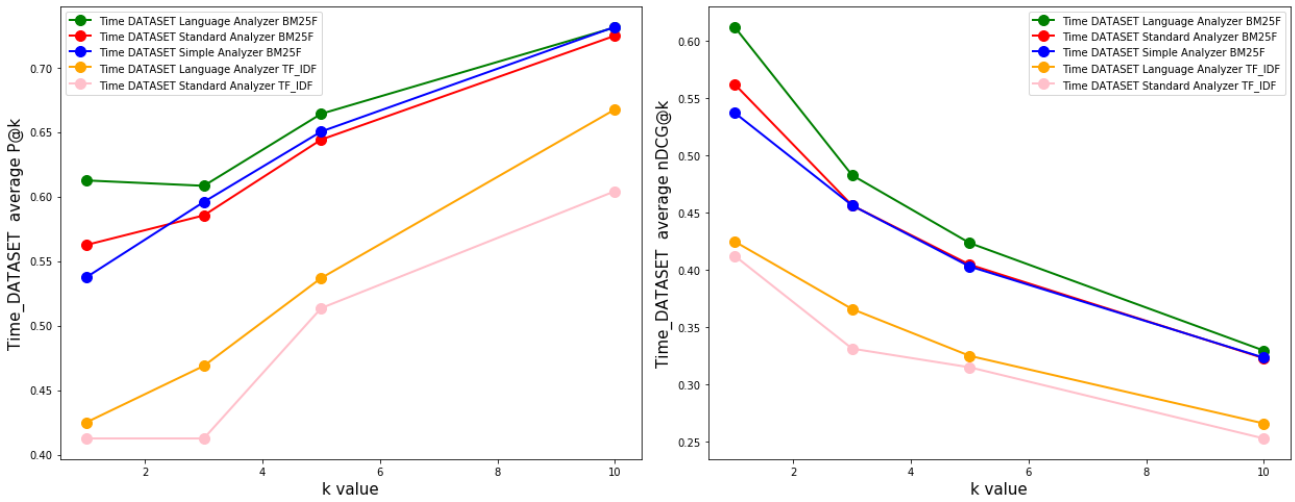

Figure 2: P@k plot and nDCG@k plot for Time dataset

2 Near-Duplicates-Detection

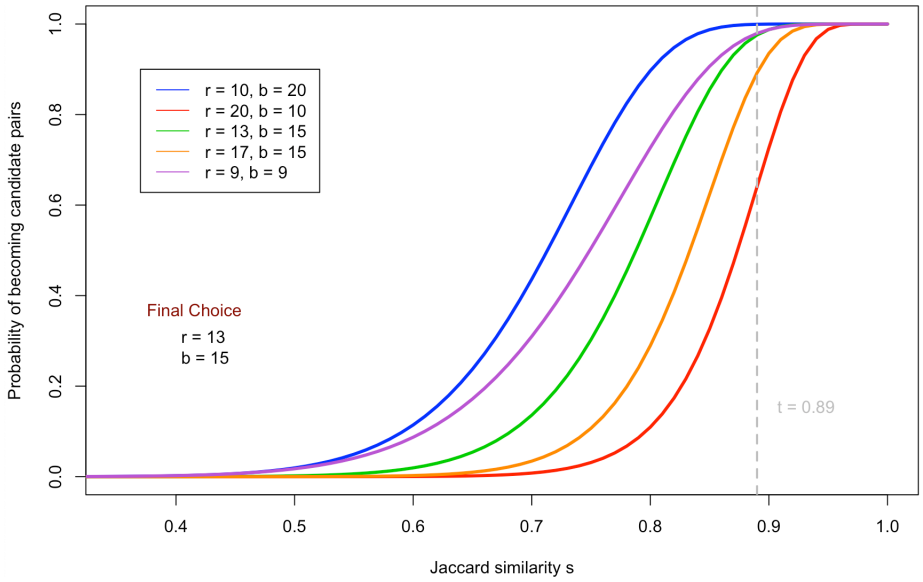

Figure 3: Behaviour of the detector depending on the number of rows and bands

Figure 3 displays different performances of the Near-Duplicates Detector depending on the relationship between the number of rows (r) and the number of bands (b). In particular, we confront the curves on two critical elements: (1) the steepness since we want to achieve a curve that behaves almost like a step function, and (2) how they intersect the threshold line (t = 0.89), since that determines the amount of false positives and false negatives. As we can see, we have the worst performance when the product of r and b is lowest (magenta line), given that we compare documents fewer times. We therefore increase the number of comparisons, and we consider two extreme cases: when r is much smaller than b, and vice versa. In the first case (blue line), we have much thinner chances of missing similar documents, i.e. the probability of false negatives is particularly low. Yet, the probability of false positives is very high, which would make the process particularly inefficient by increasing the number of direct comparisons needed at the end. On the contrary, when r is much greater than b (red line), we make smaller mistakes in terms of false positives, but the false negatives are much higher. It's clear then that there is a trade-off here, and since we can easily recover from a false positive mistake given that we compare directly the candidate pairs, it is much more important to minimize the false negatives.

Finally, by keeping the number of bands b fixed, and increasing the number of rows, we see that the curve shifts to the right, thus closing up the gap with the threshold line t = 0.89. Yet, if the increase in r is too much, we also increase the number of false negatives. We therefore accepted a higher number of false positives in order to reduce the probability of false negatives to an acceptable level (green line). Indeed, the document comparison is done in batches, each batch being a band. That means that if we increase the number of rows per band we make it slightly more difficult for the two documents to end up in the same bucket. This is shown by the orange and green curves, the latter bearing slightly less rows per band, and achieving the required minimization of the false negatives. In conclusion, we chose number of rows to be 13, and the number of bands 15. In the following table (Table 4) we can see more in details the results of this choice in terms of false positives and false negatives rates obtained. Please notice that the false negative rates are extremely small, and that the false positives decrease exponentially fast.

Jaccard Similarity s	False Negatives
0.89	2.415e-02
0.90	1.223e-02
0.95	2.034e-05
1.00	0.000

Jaccard Similarity s	False Positives
0.85	0.855
0.80	0.572
0.75	0.303
0.70	0.136
0.65	0.054
0.60	0.019
0.55	0.006
0.50	0.002

Table 4: Probability of False Positives and False Negatives

The choice of the number of bands and rows, and hence also of the min-hashing sketches, led to an execution time of the Near-Duplicates Detection tool of 2:15 minutes, with a total of 39,009 near-duplicates detected at Jaccard similarity threshold set at 0.89. Below more details on the near-duplicates detection performance.

Jaccard Similarity s	Duplicates with Jaccard Similarity at least s	Duplicates at Jaccard Similarity s
0.89	39009	382
0.90	38627	888
0.91	37739	1015
0.92	36724	744
0.93	35980	759
0.94	35221	817
0.95	34404	648
0.96	33756	685
0.97	33071	617
0.98	32454	612
0.99	31842	316
1.00	31526	31526

Table 5: Results of the Near-Duplicate Detection