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:

- Standard Analyzer BM25F
- Language Analyzer BM25F
- Simple Analyzer BM25F
- Language Analyzer TF-IDF
- Standard Analyzer TF-IDF

Time dataset:

- Language Analyzer BM25F
- Standard Analyzer BM25F
- Simple Analyzer BM25F
- Language Analyzer TF-IDF
- 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	0.75
	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

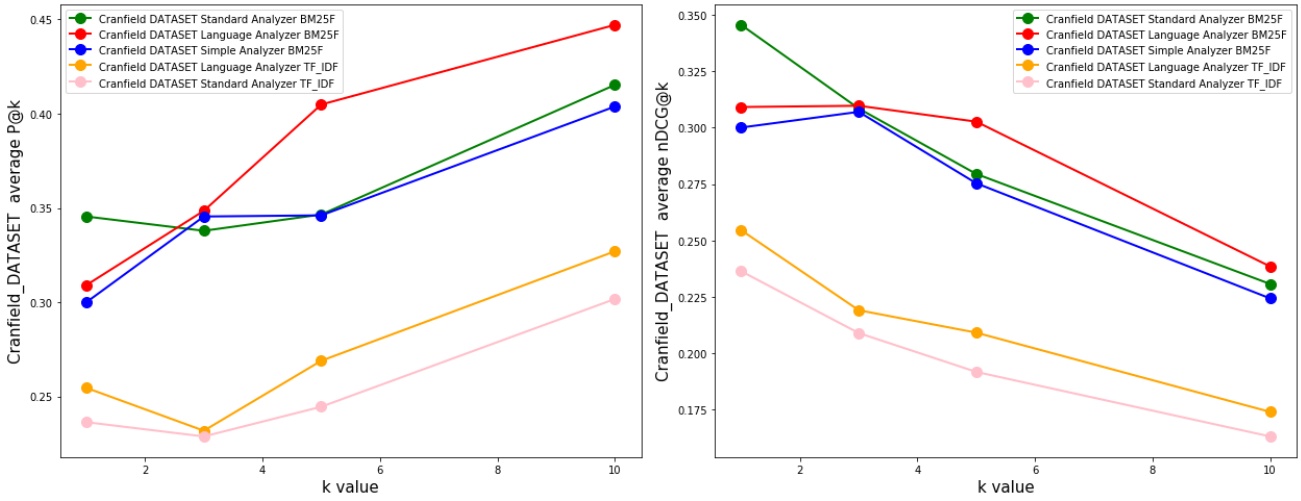

Figure 1: P@k plot and nDCG@k plot for Cranfield dataset

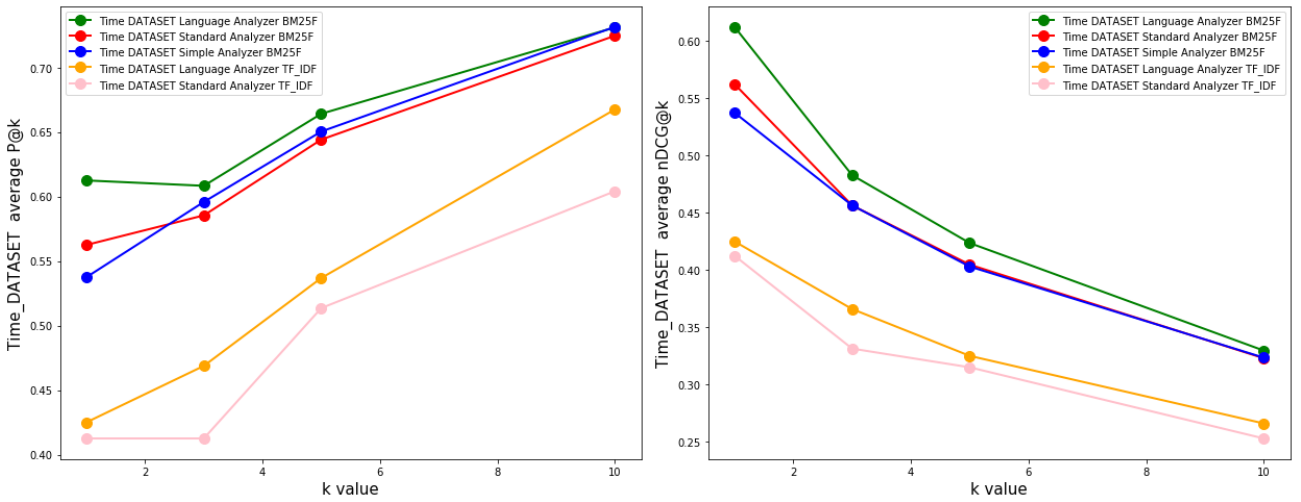

Figure 2: P@k plot and nDCG@k plot for Time dataset