DMT - Homework 1

Tran Luong Bang, Juan Mata Naranjo
Master in Data Science

April 7, 2021

	Cranfield Dataset	Time Dataset
Num Indexed Docs	1400	423
Num Queries	225	83
Num Queries in GT	110	80

Table 1: Overview Table

Conf ID	**Text Analyzer**	**Scoring Functions**
1	RegexTokenizer() | LowercaseFilter() | StopFilter(stoplist = STOP_WORDS)	scoring.MultiWeighting(scoring.Frequency(), title=scoring.BM25F(), content=scoring.TF_IDF())
2	RegexTokenizer()| StopFilter(stoplist = STOP_WORDS)| LowercaseFilter() | StemFilter()	scoring.BM25F(K1=1.2, B=.75)
3	StemmingAnalyzer(stoplist=STOP_WORDS)	scoring.TF_IDF()
4	FancyAnalyzer()	scoring.BM25F(K1=2, B=.8)
5	SimpleAnalyzer()	scoring.TF_IDF()
6	StandardAnalyzer()	scoring.Frequency()
7	FancyAnalyzer()	scoring.MultiWeighting(scoring.Frequency(), title=scoring.TF_IDF(), content=scoring.BM25F())
8	StemmingAnalyzer(stoplist=STOP_WORDS)	scoring.BM25F()
9	RegexTokenizer()| StopFilter(stoplist = STOP_WORDS)| LowercaseFilter() | StemFilter()	scoring.MultiWeighting(scoring.Frequency(), title=scoring.TF_IDF(), content=scoring.BM25F(K1=2, B=.8))
10	SimpleAnalyzer()	scoring.BM25F(K1=1.2, B=.75)
11	StemmingAnalyzer(stoplist=STOP_WORDS)	scoring.BM25F(K1=2, B=.8)
12	FancyAnalyzer()	scoring.TF_IDF()

Table 2: Configuration Overview

1 Part 1.1

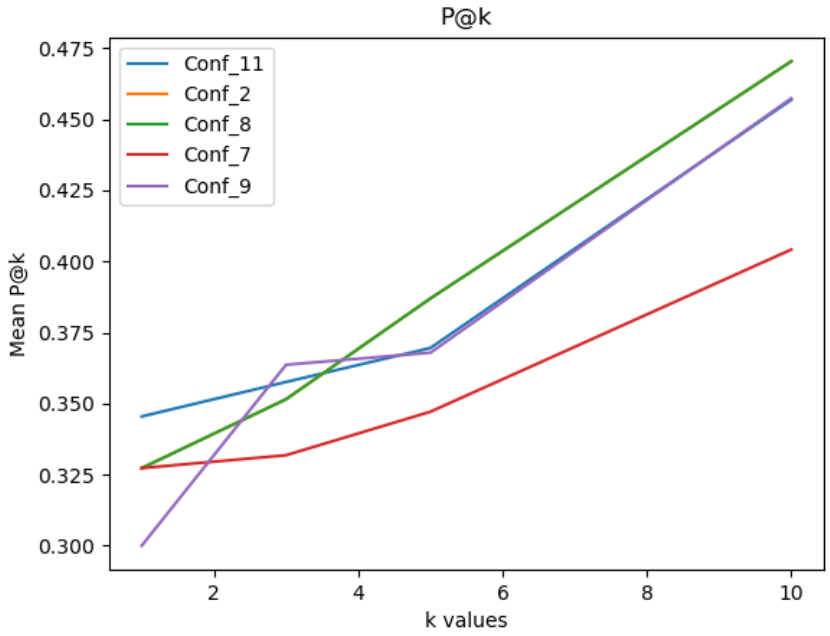
Before we deep dive into each of the individual datasets and their respective results we will start out by first giving a brief overview of the number of documents we have, the number of queries applied over these documents and finally the number of queries for which we have their ground truth at our disposal. We will see that the number of ground truth results we have are much less than the total number of queries. For the evaluation metrics we will assume that the queries for which we don't have ground truth don't exist (we remove them from all evaluation metrics). We will then also outline the different text-analyzer + scoring-functions we have deployed.

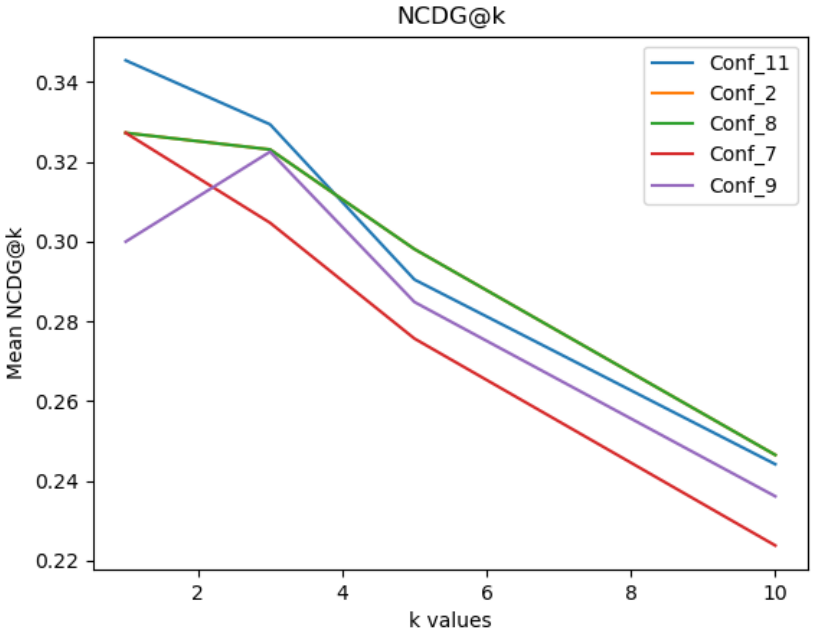
1.1 Cranfield Dataset:

We will now continue by presenting the most relevant results on the Cranfield Dataset. The first result presented is a table containing all the configurations (ordered by the MRR evaluation metric from highest to lowest), and the R-Precision metrics:

Conf ID	MRR	Mean	Min	1st quartile	Median	3rd quartile	Max
11	0.527	0.273	0	0	0.250	0.429	1.000
2	0.522	0.278	0	0	0.250	0.490	1.000
8	0.522	0.278	0	0	0.250	0.490	1.000
7	0.504	0.255	0	0	0.250	0.460	0.667
9	0.499	0.260	0	0	0.250	0.500	1.000
4	0.496	0.264	0	0	0.250	0.448	1.000
10	0.475	0.245	0	0	0.250	0.429	0.667
3	0.416	0.178	0	0	0.143	0.296	1.000
12	0.388	0.177	0	0	0.143	0.286	1.000
1	0.384	0.184	0	0	0.143	0.286	1.000
6	0.303	0.131	0	0	0.000	0.217	1.000
5	0.166	0.074	0	0	0.000	0.103	0.833

Table 3: MRR and R-Precision Overview





1.2 Time Dataset: