Assignment - 3 report - DisTeam

Collection and Query Processing

We processed all the documents using our own html handler which is created using pythons HTMLparser. The parser splits the documents into separate dictionaries. We did not preprocess the text in any way before the indexation.

We extracted the queries using Xml.dom minidom. Our query extraction is detailed in the loadQueries() method in task.py. The said method extracts the metadata id field, and the content of the tag itself. Both of which is stored in a dictionary.

Baseline and improved retrieval systems

The baseline system uses the loadQueries method to make an array of the text and the queries id as identifier. With this we can use the index once for each query. After getting the queries we use a searcher from whoosh to search the index. For the baseline searcher we use TF-IDF weighting and the only difference in the improved one is the useage of the BM25F weighting. We also parse the query content so we search for each word separately. After doing this we append the top 50 results of the search and write them out to a file (improved.out or baseline.out). It was quite a surprise to see how much difference just the ranking of the search made (almost 100 % improvement).

Results

System	P@5	P@10	MRR	MAP	NDCG@10
Baseline	0.384	0.382	0.101	0.498	0.333
Improved	0.684	0.67	0.176	0.847	0.609

Analysis

The biggest difference comes mainly from queries that was either very low or nonexistent in the baseline, but there is also a couple of queries where the baseline actually perform better than the improved version. Top queries are 50, 24 and 15 all with a difference around 0.8. Bottom ones are 39, 18 and 8 with a negative difference around 0.2, 0.15 and 0.1 respectively. See chart below.


Improved
Baseline
50
49
48
47
46
45
44
43
42
41
40
39
38
37
36
35
34
33
32
31
30
29
28
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
0.2
0.4
0.6
0.8
1
1.2