Information Retrieval Evaluation in the Field of History of Europe

CE205 Assignment 2 2019-20

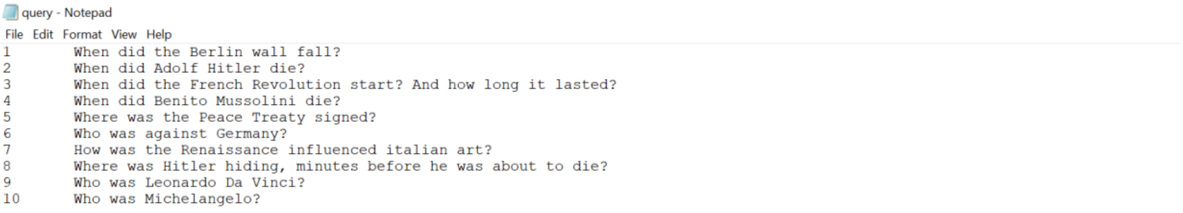
(<1703690>)

1. Topic and Queries

The first topics that I decided to query through the Whoosh! Information Retrieval system, with the field of History of Europe were: Europe post WWII, Europe before the French Revolution and Italy in the Renaissance. The reason why I decided to choose these topics, was because I had previously studied them, so it gave me an idea of what kind questions to generate and, later how to develop the rightest queries. This also helped me to clarify some of the doubt I had, in terms what to choose, since Europe is a continent and to decide what to research, would have been cumbersome. Another factor that was essential in the final decision for my topics was the fact that the selected topics are well-known by everyone, which makes it easier to name. To make my topics more relevant, I have used several sources to broaden my understanding of the selected topics and I also used the Whoosh 2.7.4 documentation to have a better understanding of the query structure.[i]

The preliminary questions that I devised are:

1. "When did the Berlin wall fall?"
2. "When did Adolf Hitler die?"
3. "When did the French Revolution start? And how long it lasted?"
4. "When did Mussolini die?"
5. "Where was the Peace Treaty signed?"
6. "Who was against Germany?"
7. "How the Renaissance was influenced Italian art?"
8. "Where was Hitler hiding, minutes before he was about to die?"
9. "Who was Leonardo Da Vinci?"
10. "Whom was Michelangelo the closest to?"



```
1       When did the Berlin wall fall?
2       When did Adolf Hitler die?
3       When did the French Revolution start? And how long it lasted?
4       When did Benito Mussolini die?
5       Where was the Peace Treaty signed?
6       Who was against Germany?
7       How was the Renaissance influenced italian art?
8       Where was Hitler hiding, minutes before he was about to die?
9       Who was Leonardo Da Vinci?
10      Who was Michelangelo?
```

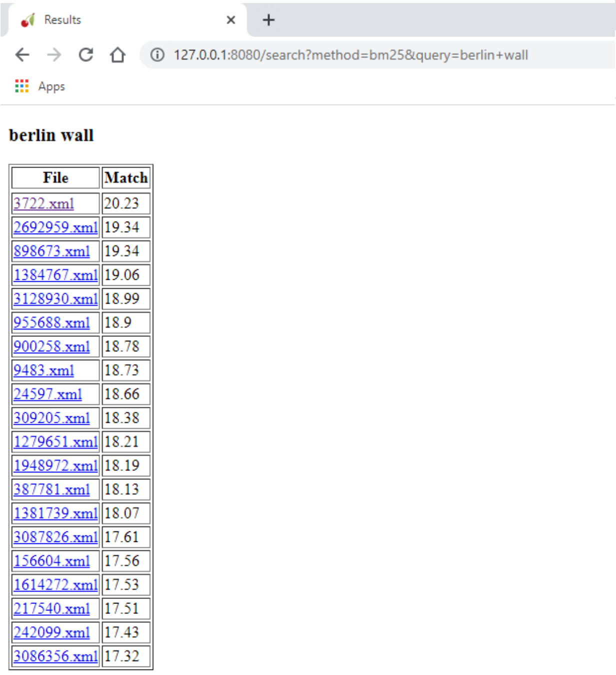
Figure 1 - Question created for the database in Notepad

The way I processed the questions to get to the final queries is to sift all "unnecessary" words. Another useful When I say, unnecessary words, are words like the five W and the H (Who, When, Where, Why, What and How), any pronouns (I, me, he, she, herself, it, we, you (singular), you (plural), that, they, each, few, many, who, whoever, whose, someone, everybody), and any other words that don't have any value if they are not put in context. However, this should take into context that some of the questions need the mentioned i.e. the five W and the H to get an understanding, and then decide what is correct for the final query.

So, the resulting queries are:

1. “Berlin wall fall”
2. “Adolf Hitler die” or “Adolf Hitler death”
3. “French Revolution start AND length”
4. “Benito Mussolini die” or “Benito Mussolini death”
5. “Peace Treaty signing place”
6. “Germany enemies”
7. “Renaissance influence Italian art”
8. “Hitler hiding before death”
9. “Leonardo Da Vinci”
10. “Michelangelo closest”

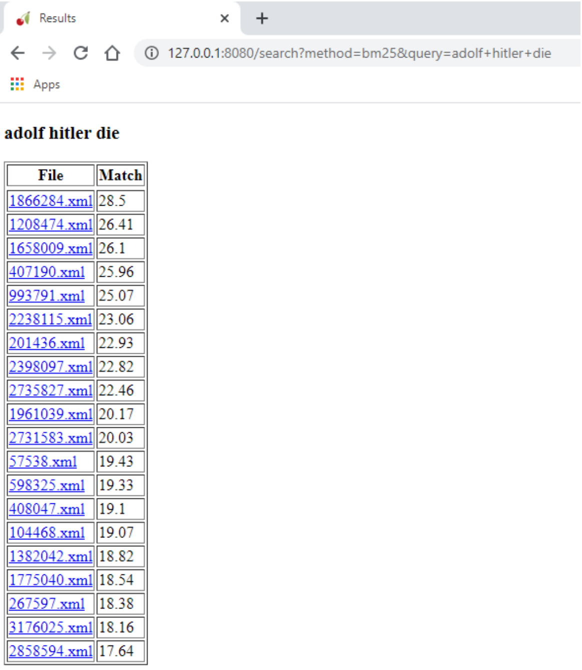
The generated files for the bm25 method:



127.0.0.1:8080/search?method=bm25&query=berlin+wall

berlin wall

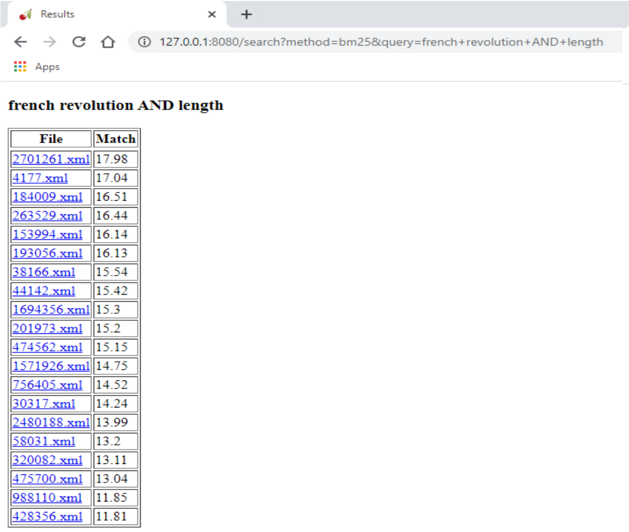
File	Match
3722.xml	20.23
2692959.xml	19.34
898673.xml	19.34
1384767.xml	19.06
3128930.xml	18.99
955688.xml	18.9
900258.xml	18.78
9483.xml	18.73
24597.xml	18.66
309205.xml	18.38
1279651.xml	18.21
1948972.xml	18.19
387781.xml	18.13
1381739.xml	18.07
3087826.xml	17.61
156604.xml	17.56
1614272.xml	17.53
217540.xml	17.51
242099.xml	17.43
3086356.xml	17.32



127.0.0.1:8080/search?method=bm25&query=adolf+hitler+die

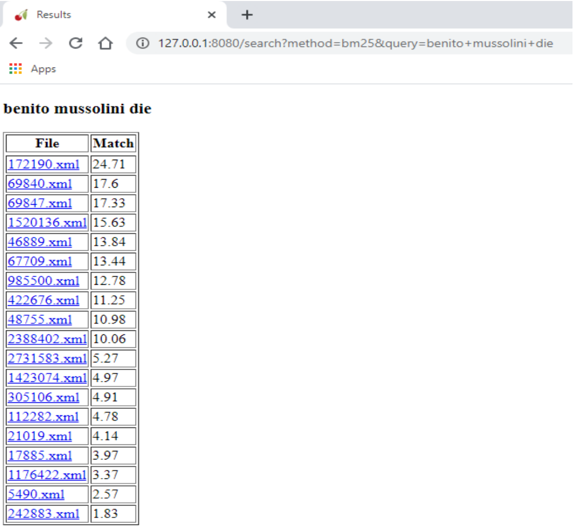
adolf hitler die

File	Match
1866284.xml	28.5
1208474.xml	26.41
1658009.xml	26.1
407190.xml	25.96
993791.xml	25.07
2238115.xml	23.06
201436.xml	22.93
2398097.xml	22.82
2735827.xml	22.46
1961039.xml	20.17
2731583.xml	20.03
57538.xml	19.43
598325.xml	19.33
408047.xml	19.1
104468.xml	19.07
1382042.xml	18.82
1775040.xml	18.54
267597.xml	18.38
3176025.xml	18.16
2858594.xml	17.64


Results
127.0.0.1:8080/search?method=bm25&query=french+revolution+AND+length
Apps

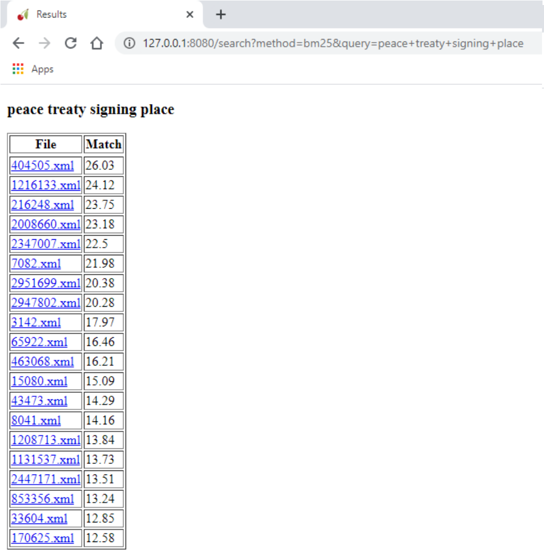
french revolution AND length

File	Match
2701261.xml	17.98
4177.xml	17.04
184009.xml	16.51
263529.xml	16.44
153994.xml	16.14
193056.xml	16.13
38166.xml	15.54
44142.xml	15.42
1694356.xml	15.3
201973.xml	15.2
474562.xml	15.15
1571926.xml	14.75
756405.xml	14.52
30317.xml	14.24
2480188.xml	13.99
58031.xml	13.2
320082.xml	13.11
475700.xml	13.04
988110.xml	11.85
428356.xml	11.81


Results
127.0.0.1:8080/search?method=bm25&query=benito+mussolini+die
Apps

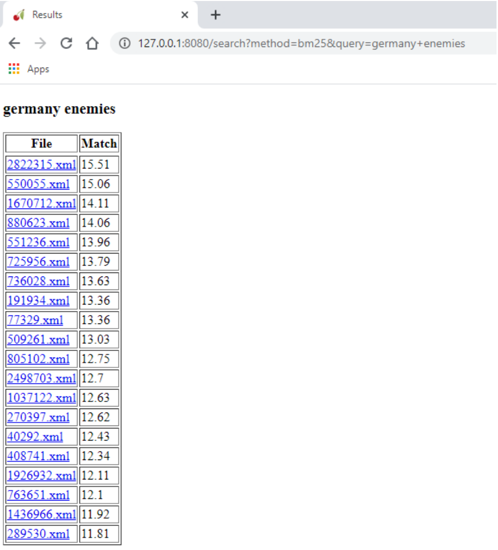
benito mussolini die

File	Match
172190.xml	24.71
69840.xml	17.6
69847.xml	17.33
1520136.xml	15.63
46889.xml	13.84
67709.xml	13.44
985500.xml	12.78
422676.xml	11.25
48755.xml	10.98
2388402.xml	10.06
2731583.xml	5.27
1423074.xml	4.97
305106.xml	4.91
112282.xml	4.78
21019.xml	4.14
17885.xml	3.97
1176422.xml	3.37
5490.xml	2.57
242883.xml	1.83


Results
127.0.0.1:8080/search?method=bm25&query=peace+treaty+signing+place
Apps

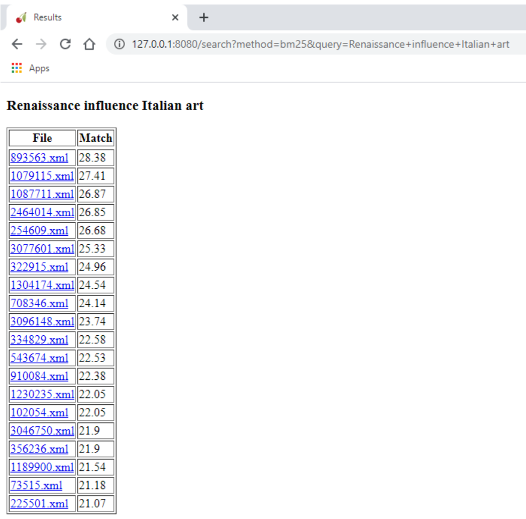
peace treaty signing place

File	Match
404505.xml	26.03
1216133.xml	24.12
216248.xml	23.75
2008660.xml	23.18
2347007.xml	22.5
7082.xml	21.98
2951699.xml	20.38
2947802.xml	20.28
3142.xml	17.97
65922.xml	16.46
463068.xml	16.21
15080.xml	15.09
43473.xml	14.29
8041.xml	14.16
1208713.xml	13.84
1131537.xml	13.73
2447171.xml	13.51
853356.xml	13.24
33604.xml	12.85
170625.xml	12.58



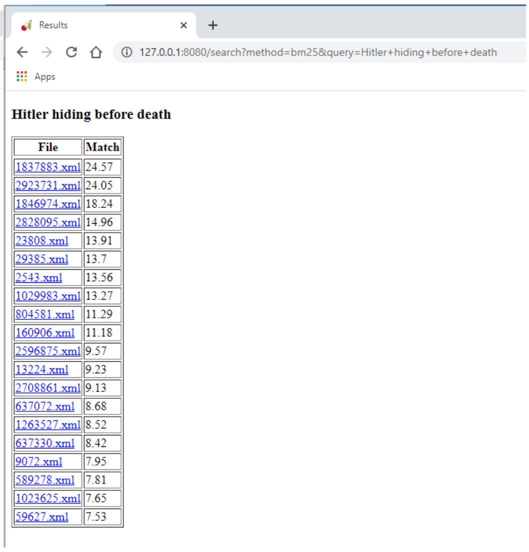
germany enemies

File	Match
2822315.xml	15.51
550055.xml	15.06
1670712.xml	14.11
880623.xml	14.06
551236.xml	13.96
725956.xml	13.79
736028.xml	13.63
191934.xml	13.36
77329.xml	13.36
509261.xml	13.03
805102.xml	12.75
2498703.xml	12.7
1037122.xml	12.63
270397.xml	12.62
40292.xml	12.43
408741.xml	12.34
1926932.xml	12.11
763651.xml	12.1
1436966.xml	11.92
289530.xml	11.81


Results
127.0.0.1:8080/search?method=bm25&query=Renaissance+influence+Italian+art
Apps

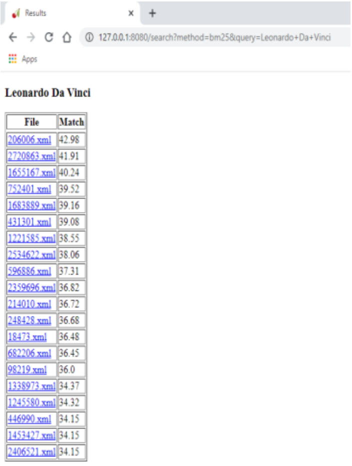
Renaissance influence Italian art

File	Match
893563.xml	28.38
1079115.xml	27.41
1087711.xml	26.87
2464014.xml	26.85
254609.xml	26.68
3077601.xml	25.33
322915.xml	24.96
1304174.xml	24.54
708346.xml	24.14
3096148.xml	23.74
334829.xml	22.58
543674.xml	22.53
910084.xml	22.38
1230235.xml	22.05
102054.xml	22.05
3046750.xml	21.9
356236.xml	21.9
1189900.xml	21.54
73515.xml	21.18
225501.xml	21.07


Results
127.0.0.1:8080/search?method=bm25&query=Hitler+hiding+before+death
Apps

Hitler hiding before death

File	Match
1837883.xml	24.57
2923731.xml	24.05
1846974.xml	18.24
2828095.xml	14.96
23808.xml	13.91
29385.xml	13.7
2543.xml	13.56
1029983.xml	13.27
804581.xml	11.29
160906.xml	11.18
2596875.xml	9.57
13224.xml	9.23
2708861.xml	9.13
637072.xml	8.68
1263527.xml	8.52
637330.xml	8.42
9072.xml	7.95
589278.xml	7.81
1023625.xml	7.65
59627.xml	7.53



Leonardo Da Vinci

File	Match
206006.xml	42.98
2720863.xml	41.91
1655167.xml	40.24
752401.xml	39.52
1683889.xml	39.16
431301.xml	39.08
1221585.xml	38.55
2534622.xml	38.06
596886.xml	37.31
2359696.xml	36.82
214010.xml	36.72
248428.xml	36.68
18473.xml	36.48
682206.xml	36.45
98219.xml	36.0
1338973.xml	34.37
1245580.xml	34.32
446990.xml	34.15
1453427.xml	34.15
2406521.xml	34.15

2. Relevance judgments

How was this done?

The way I have dealt with how to get the queries relevant, was to first answer the questions, in a clinical way. What I mean is that there is no scope for the answer to be ambiguous. So, the files which had to have the answer in the truest form. If it was, but not clearly stated then depending on how the answer can be perceived, the judgement will be done based off the content of the file.

What problems were encountered and how were they solved?

In terms of problems, some of the queries were yielding a very small set of results, so the query had to be altered to something more generic. So the following queries were changed to:

1. "Hitler hiding before death AND alone" to "Hitler hiding before death"
2. "Michelangelo closest" to "Michelangelo"
3. "Benito Mussolini die" to "Benito Mussolini"

3. Evaluation

3.1 BM25 Method

Outline what you did.

Include the following table, duly completed with your results. Numbers to two decimal places exactly as shown in the table below. The last line is for the averages - examples are shown.

Num	**Query**	**P** **(n=5)**	**P** **(n=10)**	**R** **(n=5)**	**R** **(n=10)**
0	tigers "commonly occur"				
1	Berlin wall fall	0.60	0.40	0.45	0.3..
2	Adolf Hitler die	0.40	0.20	0.6…	0.6…
3	French Revolution start AND length	0.20	0.10	0.20	0.20
4	Benito Mussolini die	0.20	0.10	0.20	0.20
5	Peace Treaty signing place	0.60	0.50	0.3…	0.5…
6	Germany enemies	0.40	0.40	0.43...	0.71
7	Renaissance influence Italian art	0.80	0.70	0.31	0.53
8	Hitler hiding before death	0.20	0.20	0.33..	0.6…
9	Leonardo Da Vinci	0.40	0.40	0.20	0.40
10	Michelangelo	0.80	0.70	0.28	0.50
Avg		**0.46**	**0.37**	**0.27**	**0.42**

P refers to precision. The formula to calculate precision is:

$$P = \frac{No.\ of\ relevant\ documents\ returned}{No. of\ documents\ returned}$$

And recall is:

$$R = \frac{No. of\ relevant\ documents\ returned}{Total\ no. of\ relevant\ documents}$$

3.2. TF*IDF Method

Repeat the above steps using TF*IDF. You do not need to re-index the collection, just use a different query function as shown in the Python program. You can select between BM25 and TF*IDF for each query via the Web interface.

In this table, I have done what it as done before. The only difference is that I have used a different method, which is, in this case, TF-IDF.

TF-IDF is a method used for calculating the term weight in a document. Two of the factors that have the most value is:

- Term frequency
- Inverted Document Frequency

Include the following, table duly completed.

Num	**Query**	**P** **(n=5)**	**P** **(n=10)**	**R** **(n=5)**	**R** **(n=10)**
0	tigers "commonly occur"	0.25	0.20	0.40	0.60
1	Berlin wall fall	0.40	0.60	0.18	0.54
2	Adolf Hitler die	0.20	0.30	0.16	0.50
3	French Revolution start AND length	0	0.10	0	0.40

4	Benito Mussolini die	0.80	0.70	0.26	0.46
5	Peace Treaty signing place	1	0.70	0.45	0.63
6	Germany enemies	0.80	0.50	0.40	0.50
7	Renaissance influence Italian art	0.80	0.50	0.40	0.50
8	Hitler hiding place before death AND alone	0.20	0.10	0.50	0.50
9	Leonardo Da Vinci	0.60	0.70	0.20	0.46
10	Michelangelo closest	0.80	0.70	0.26	0.46
Avg		0.53	0.46	0.29	0.50

3.3. Additional Experiment: BM25 parameters

Include the following, table duly completed.

Num	**Query**	**P** **(n=5)**	**P** **(n=10)**	**R** **(n=5)**	**R** **(n=10)**
0	tigers "commonly occur"	0.25	0.20	0.40	0.60
1	Berlin wall fall	0.60	0.40	0.45	0.3..
2	Adolf Hitler die	0.40	0.20	0.6…	0.6…
3	French Revolution start AND length	0.20	0.10	0.20	0.20
4	Benito Mussolini die	0.20	0.10	0.20	0.20
5	Peace Treaty signing place	0.60	0.50	0.3…	0.5…
6	Germany enemies	0.40	0.40	0.43...	0.71
7	Renaissance influence Italian art	0.80	0.70	0.31	0.53
8	Hitler hiding before death	0.20	0.20	0.33..	0.6…
9	Leonardo Da Vinci	0.40	0.40	0.20	0.40
10	Michelangelo	0.80	0.70	0.28	0.50
Avg		**0.90**	**0.70**	**0.30**	**0.50**

4. Discussion

Compose a short description of what the results show, any interesting problem cases, any technical problems encountered and so on. Don't forget to reference the article(s) you looked at and include your rationale for the new settings as well as describing how the results differ.

The results don't change as much as in the ranking.

Reminder: When you submit your report, make sure you **convert to .pdf first!** .doc files cannot be accepted. Then submit to Faser, following the assignment instructions. Don't forget to also upload a zip file with the "query.txt" file, the "output.txt" files as per instructions.

[i] https://whoosh.readthedocs.io/en/latest/index.html