Information Retrieval Evaluation in the Field of <Language>

CE205 Assignment 2 2017-18

Cieran Almond (1604959)

1. Topic and Queries

Why I chose languages:

I decided to choose languages as my topic of choice because it's a topic that I find interesting, and also a topic I can device a diverse range of queries for, involving different languages, statistics, countries, words and phrases. I devised my queries originally from what I would want to find out about languages myself, and what I considered interesting. Towards the end of my original queries (around queries 15-20), I decided to read through the Wikipedia article on languages to take some inspiration from them, and devise some more complex queries.

My original queries:

1. "How many languages are there in the world?"
2. "What is a natural language?"
3. "What is a dialect?"
4. "Whats is a language encoded into secondary media?"
5. "Where did language originate?"
6. "What part of the brain processes language?"
7. "How long does it take to become fluent in a language from birth?"
8. "How many languages are spoken in China?"
9. "What are the top 3 most spoken languages in the world?"
10. "What are the Indo-European family languages?"
11. "What is the generative theory of grammar?"
12. "What is a syllable?"
13. "How many characters are in the traditional Chinese language?"
14. "How many writing systems are there in Europe?"
15. "How many words in the oldest language?"
16. "What is the International Phonetic Alphabet?"
17. "What is Morphology?"
18. "How many people in the United Kingdom are bilingual?"
19. "In which countries is English spoken?"
20. "Which language has the longest word?"

My queries in Whoosh querying language:

1. 'languages "in the world"'
2. 'language "natural"'
3. 'language "into secondary media"'
4. 'language "encoded secondary media"'
5. 'language "origins"'
6. 'language "brain processes"'

7. ‘language “time to be fluent”’
8. ‘languages “in china”’
9. ‘languages “most spoken”’
10. ‘languages “Indio-European”’
11. ‘grammar “generative theory”’
12. ‘what “is syllable”’
13. ‘languages “characters in Chinese”’
14. ‘europe “writing systems”’
15. ‘words “oldest language”’
16. ‘alphabet “international phonetic”’
17. **‘**morphology “what is”’
18. ‘united kingdom “number of bilingual people”’
19. ‘countries “speak english”’
20. ‘language “longest word”’

2. Indexing the Documents

I copy and pasted my ir directory file into my C:\Users\ca16873\ir, editied the python file “awf_analyse_wiki_files_34” to
files = glob.glob('c:\\wikipedia\\resources\\wikipedia\\processed**.xml'), then opened up my cmd/command prompt and typed: C: “Users\ca16873\ir”. Then, once I was in the directory, entered “python” and put in the supplied commands “from awf_analyse_wiki_files_34 import *” and “awf_create_index() “. The process took around 45 minutes to complete.

The only problem I encountered was that once the indexing was complete, my file size was 11.4gb in size, and I wanted to store the file on an 8gb memory stick to use at a later date. So I compressed the folder which took around 5 minutes and ended up being 1.50gb in size, small enough to store on my memory stick.

3. BM25 Performance

3.1 Method

I navigated to my C: "Users\ca16873\ir" directory in the command prompt and entered the following into the command line: "from awf_analyse_wiki_files_34 import *" , then entered awf_query_index_bm25('a query') for all of my relevant queries that I created above in task 1. I didn't need to do any previous setup from what I did in part 2.

After each query, I used the methods P = no.relevant documents returned / no.documents returned and R = no.relevant documents returned / total no.relevant documents to determine the values of P and R.

3.2 Results
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)**
1	'languages "in the world"'	0.40	0.20	0.22*	0.22*
2	'language "natural"'	0.80	0.70	0.26*	0.60
3	'language "into secondary media"'	0.00	0.00	0.00	0.00
4	'language "encoded secondary media"'	0.00	0.00	0.00	0.00
5	'language "origins"'	0.60	0.60	0.20	0.53
6	'language "brain processes"'	0.20	0.40	0.07	0.14
7	'language "time to be fluent"'	0.00	0.00	0.00	0.00
8	'languages "in china"'	0.60	0.70	0.25	0.58
9	'languages "most spoken"'	0.20	0.60	0.10	0.40
10	'languages "Indio-European"'	0.00	0.00	0.00	0.00
11	'grammar "generative theory"'	0.00	0.00	0.00	0.00
12	'what "is syllable"'	1.00	0.80	0.33	0.86
13	'languages "characters in Chinese"'	0.80	0.70	0.31	0.54
14	'europe "writing systems"'	0.20	0.50	0.13	0.63
15	'words "oldest language"'	0.60	0.30	1.00	1.00
16	'alphabet "international phonetic"'	0.20	0.40	0.08	0.33*
17	'morphology "what is"'	0.60	0.70	0.20	0.47
18	'united kingdom "number of bilingual people"'	0.00	0.00	0.00	0.00
19	'countries "speak english"'	0.40	0.30	0.20	0.30
20	'language "longest word"'	0.20	0.40	0.20	0.80
Avg		**0.34**	**0.37**	**0.18**	**0.37**

3.3 Discussion

With the returned results, I took the first 10 returned results, copied the file .xml directory and opened it in a notepad file. I then compared the information in the notepad file to my original question and used determination to consider if they are relevant or non relevant. For example, in my first query "languages in the world" any mention of specific languages mentioned in the context of them being applicable to a country, books written in a certain language or languages spoken by certain people groups, I considered relevant to the query. Another example is for language origins, I included anything as relevant that spoke about the origins of a specific language. I noted down every result I read with a 1 for relevant and a 0 for not relevant on a piece of paper. I couldn't find a corrorlation between the number given as a "relevancy score" and an actual relevant result upon inspecting the document; a document with a score of 15.27 was just as likely to contain relevant data as a document scoring 13.57 for my first query I deviced, 15.27 being the highest value returned and 13.57 being the lowest. As the pdf instructions instructed, for calculating the recall I continued past the first 10 results to find the relevancy of all 20 documents, counted the number of relevant responces and used them to work out the values of n=5 and

n=10. I repeated this process for every query, finding a similar correrlation from the results returned of their “relevancy score” not really determining the relevancy to the query in question.

For query 15, since it only returns 3 results, it only returns 0.60 and 0.30 due to there only being 3 results, however all of them are divided by 1 because I deamed all 3 results as revevant to the query.

My results show that, on average, BM25 comparing the first 10 queries for relevancy is more accurate than just comparing the first 5 results, and that its better to compare more documents to identify relevant documents then to compare a small sample size. The data presented however, may not be a 100% accurate interpretation, as there are several anomalies; 6 entries were entered as 0 as they did not return any results, and 1 entry only returned 3 results, which slightly skewed the relevance columns as they both contained “1.0” values.

4. TF*IDF Performance (not compulsory, see marking scheme)

4.1 Method

I used exactly the same method as previously mentioned for my BM25 queries, only this time entering awf_query_index_tf_idf('a query') where “a query” is the queries shown in the table below into the CMD (I did not use the web browser for the assignment).

4.2 Results

Include the following, table duly completed.

Num	Query	P (n=5)	P (n=10)	R (n=5)	R (n=10)
1	‘languages “in the world”’	0.60	0.50	0.33	0.55
2	‘language “natural”’	0.60	0.70	0.27*	0.63*
3	‘language “into secondary media”’	0.00	0.00	0.00	0.00
4	‘language “encoded secondary media”’	0.00	0.00	0.00	0.00
5	‘language “origins”’	0.60	0.60	0.30	0.60
6	‘language “brain processes”’	0.40	0.40	0.17	0.33*
7	‘language “time to be fluent”’	0.00	0.00	0.00	0.00
8	‘languages “in china”’	1.00	0.70	0.33*	0.46*
9	‘languages “most spoken”’	0.20	0.30	0.13	0.38
10	‘languages “Indio-European”’	0.00	0.00	0.00	0.00
11	‘grammar “generative theory”’	0.00	0.00	0.00	0.00
12	‘what “is syllable”’	0.80	0.70	0.25	0.44
13	‘languages “characters in Chinese”’	0.60	0.60	0.19	0.46
14	‘europe “writing systems”’	0.20	0.30	0.13	0.38
15	‘words “oldest language”’	0.00	0.00	0.00	0.00
16	‘alphabet “international phonetic”’	0.40	0.40	0.33*	0.66*
17	‘morphology “what is”’	0.20	0.60	0.11*	0.66*
18	‘united kingdom “number of bilingual people”’	0.00	0.00	0.00	0.00
19	‘countries “speak english”’	0.80	0.70	0.26*	0.46*
20	‘language “longest word”’	0.20	0.50	0.13	0.63
Avg		**0.33**	**0.35**	**0.15**	**0.33**

4.3 Discussion

Compose a short description of what the results show (was TF*IDF always better than BM25, always worse or sometimes better/worse?), any interesting problem cases, any technical problems encountered and so on.

An interesting problem case or observation is that BM25 was able to retrieve 3 results for query 15, where as TF*IDF was unable to retrieve any at all. My guess is there wasn’t a high concerntration of the term “oldest language” and the documents were very long, which since TF*IDF searches for keywords in concerntrations of text, may have not deemed it frequent enough to return a result.

Something to note is that the command prompt indicates that the relevancy value is much higher for TF*IDF than BM25, if we take my first query for example “languages in the world”, the highest value from this result is 368.6, the lowest being 155.91 which is significantly higher than what I indicated in my findings for the BM25 first result.In addition, the documents returned were a lot longer in length in comparision to the BM25 method.

In some aspects for some queries, however, TF*IDF did out perform BM25, for example in the first query it retrieved, what I deemed, more relevant data. On the whole, however, they both performed

quite similarly in terms of what the avg results returned and for each query had similar relevance in response.

5. Additional Experiment (BM25 Parameters or Named Entities)

5.1 Description

Explain what you did.

5.2 Results

Present the results in a table with a short analysis.

Appendix 1

Include the queries you used for your BM25 evaluation and list the IDs of up to four right answers found in the first ten responses returned by the system, listed in the order in which they are returned. Note that there might be as many as ten right answers found in the first ten, or as few as zero. Generally, only a minority of your queries should have no answers at all in the first ten. If more then four correct answers are returned, just list the first four here. Note that we can use this information to verify that these answers are really returned in response to the query, and that these answers are indeed 'correct' answers.

Num	Query	IDs of Answers
1	'languages "in the world"'	877389, 418385, 23940,
2	'language "natural"'	21173, 21652, 1661566. 301999
3	'language "into secondary media"'	0
4	'language "encoded secondary media"'	0
5	'language "origins"'	1424038, 616932, 69495, 2664500
6	'language "brain processes"'	160538, 5626, 1657256, 893696
7	'language "time to be fluent"'	0
8	'languages "in china"'	737850, 1458526, 3000009, 243875
9	'languages "most spoken"'	1844404, 195445, 1234492, 769881
10	'languages "Indio-European"'	0
11	'grammar "generative theory"'	0
12	'what "is syllable"'	316354, 1060562, 945490, 19930
13	'languages "characters in Chinese"'	225534, 187273, 61600, 261949
14	'europe "writing systems"'	2412305, 53682, 305738, 65373
15	'words "oldest language"'	37445, 202353, 589920
16	'alphabet "international phonetic"'	159470, 59045, 1268993, 533322
17	'morphology "what is"'	229130, 55402, 596255, 164089
18	'united kingdom "number of bilingual people"'	0
19	'countries "speak english"'	567017, 142525, 189109
20	'language "longest word"'	9467, 2201357, 387219, 1566678

Appendix 2

Include the same queries as in Appendix 1 plus the IDs of the right answers found when TF*IDF was used.

Num	Query	IDs of Answers
1	‘languages “in the world”’	59556, 904073, 626817, 445324
2	‘language “natural”’	1199964, 75359, 23032, 75599
3	‘language “into secondary media”’	0
4	‘language “encoded secondary media”’	0
5	‘language “origins”’	46279, 23032, 9249, 540382
6	‘language “brain processes”’	105714, 5626, 5664, 160538
7	‘language “time to be fluent”’	0
8	‘languages “in china’”	435358, 244908, 5405, 25734
9	‘languages “most spoken”’	626817, 9249, 25401
10	‘languages “Indio-European”’	0
11	‘grammar “generative theory”’	0
12	‘what “is syllable”’	1158155, 197367, 19930, 228026
13	‘languages “characters in Chinese”’	261949, 15606, 187273, 225534
14	‘europe “writing systems”’	49950, 2303, 53682
15	‘words “oldest language”’	0
16	‘alphabet “international phonetic”’	59045, 670, 2204, 248274
17	‘morphology “what is’”’	18916, 445324, 15227, 15105
18	‘united kingdom “number of bilingual people”’	0
19	‘countries “speak english”’	63881, 6340, 63892, 610214
20	‘language “longest word”’	9467, 9468, 387219, 156667

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 exactly.