CA4009: Search Technologies
Laboratory Session 2
2nd November 2021

Joanna Talvo - 18342523
Chloe Ward - 18302716

4 Manual examination of a TREC test collection and sample search results

From our observation on the downloaded files, the XML file contains the summaries of each document which is in the format of `document_number title description narrative`. The qrels file tells you whether the document id is relevant or not through the relevance indicator (0 or 1). The res file is also similar but it has extra values such as the rank of the document, score, and the run_id.

We searched for the top 5 results from the IR Model generator for each ranking function (TF-IDF and BM25). For the BM25 function we used the values k=1.2 and b=0.75. We compared the rankings of three topics based on three different fields which are title, description, and narrative. We also compared the rankings of these documents in the qrels file as shown in the table. The following table below are the results:

Document ID: 301
International Organized Crime

RANK	**TITLE**		**DESCRIPTION**		**NARRATIVE**	
TF-IDF	RI	RRES	RI	RRES	RI	RRES
1	No results	3	0	38	0	No results
2	0	12	0	820	1	No results
3	1	5	0	No results	1	No results
4	0	2	0	793	1	No results
5	0	1	0	44	No results	No results

RANK	**TITLE**		**DESCRIPTION**		**NARRATIVE**	
BM25	RI	RRES	RI	RRES	RI	RRES
1	0	2	0	No results	No results	No results

2	0	4	0	No results	1	No results
3	1	5	0	820	1	No results
4	0	15	0	No results	1	No results
5	0	8	1	5	1	No results

Document ID: 302
Poliomyelitis and Post-Polio

RANK	**TITLE**		**DESCRIPTION**		**NARRATIVE**	
TF-IDF	RI	RRES	RI	RRES	RI	RRES
1	1	2	1	3	1	14
2	1	3	1	2	0	31
3	1	1	1	27	1	6
4	0	4	1	37	1	3
5	1	6	0	4	0	11

RANK	**TITLE**		**DESCRIPTION**		**NARRATIVE**	
BM25	RI	RRES	RI	RRES	RI	RRES
1	1	3	1	3	1	44
2	1	2	1	2	0	31
3	0	4	0	4	1	3
4	1	1	1	37	1	58
5	1	6	1	1	1	14

Document ID: 310
Radio Waves and Brain Cancer

RANK	**TITLE**		**DESCRIPTION**		**NARRATIVE**	
TF-IDF	RI	RRES	RI	RRES	RI	RRES
1	1	1	1	1	0	840
2	0	2	0	2	1	1
3	0	4	1	7	0	121
4	0	13	0	9	1	7

5	1	11	0	18	0	848

RANK	**TITLE**		**DESCRIPTION**		**NARRATIVE**	
BM25	RI	RRES	RI	RRES	RI	RRES
1	1	1	1	1	1	1
2	0	2	0	2	0	121
3	0	4	0	127	1	7
4	0	12	0	No results	0	31
5	1	11	0	No results	0	840

*RI = relevance indicator, RRES = rank in trec res file

From observing the results, we noticed that title is the most helpful and effective type of search query as the ranked positions for each ranking function are the closest and similar to the ranked positions from the TREC res file. The description can also be effective in some documents. Description contains more information or words so some of the words or strings in the description might not be in the document. Therefore, it doesn't seem to be not as effective as the title as a search query. From the topics, we can see that some of the descriptions are not even found in the res file using the BM25 function. For the narrative, most of the rankings from each ranking function are far off from the results in the res file and some are not even found. We noticed that the narrative field contains specifications of what relevant documents for a certain topic should have. BM25 and TF-IDF seem to return similar ranking positions / results for the most part.

5 Exploring Evaluation Metrics

By running the command below we get the results:
trec_eval qrels.trec678.adhoc trec678.res

Output type	**All documents**	**Value**
runid	all	1m
num_q	all	150
num_ret	all	142395
num_rel	all	14013
num_rel_ret	all	7282
map	all	0.2145

gm_map	all	0.0991
Rprec	all	0.2644
bpref	all	0.2367
recip_rank	all	0.5778
iprec_at_recall_0.00	all	0.6393
iprec_at_recall_0.10	all	0.4549
prec_at_recall_0.20	all	0.3590
iprec_at_recall_0.30	all	0.3014
iprec_at_recall_0.40	all	0.2377
iprec_at_recall_0.50	all	0.1925
iprec_at_recall_0.60	all	0.1510
iprec_at_recall_0.70	all	0.1125
iprec_at_recall_0.80	all	0.0728
iprec_at_recall_0.90	all	0.0528
iprec_at_recall_1.00	all	0.0272
P_5	all	0.4240
P_10	all	0.4027
P_15	all	0.3738
P_20	all	0.3487
P_30	all	0.3111
P_100	all	0.1940
P_200	all	0.1369

P_500	all	0.0785
P_1000	all	0.0485

Based on the values of the output result, we compared the values of the num_ret (number of retrieved documents) to the num_rel_ret (number of relevant documents retrieved) and 5.11% of the documents are the most relevant (num_ret_rel / num_ret * 100)

Looking at Precision, the precision values decrease as the X number of documents increases. We conclude that the smaller the value of X, the most significant these documents are based on the query. We also noticed that there is a sharp decline in the precision values as X goes up from 30 to 100 documents. This is because the first 5 documents are most relevant to the search query. The higher the X number of documents, the less significant it is.

For the Interpolated Recall-Precision Averages, they are broken down into different recall levels to measure precision. The values are similar to the Precision values, as it declines as well as the recall level increases. The values started in a higher value than the Precision, due to the fact that it is based on the number of queries. Therefore we can say that this gives a more accurate measure of quality for relevance.

The results for the command are below:
trec_eval -q qrels.trec678.adhoc trec678.res

Output Type	**Document 301**	**Document 302**	**Document 310**
num_ret	1000	1000	1000
num_rel	474	77	13
num_rel_ret	74	64	6
map	0.0179	0.4737	0.1424
Rprec	0.0949	0.4935	0.2308
bpref	0.1230	0.5109	0.1716
recip_rank	0.2000	1.0000	1.0000
iprec_at_recall_0.00	0.3077	1.0000	1.0000
iprec_at_recall_0.10	0.0967	0.7059	0.2857
iprec_at_recall_0.20	0.0000	0.6667	0.2727
iprec_at_recall_0.30	0.0000	0.6316	0.2500
iprec_at_recall_0.40	0.0000	0.5500	0.0120

iprec_at_recall_0.50	0.0000	0.5185	0.0000
iprec_at_recall_0.60	0.0000	0.4860	0.0000
iprec_at_recall_0.70	0.0000	0.4480	0.0000
iprec_at_recall_0.80	0.0000	0.3875	0.0000
iprec_at_recall_0.90	0.0000	0.0000	0.0000
iprec_at_recall_1.00	0.0000	0.0000	0.0000
P_5	0.2000	0.6000	0.2000
P_10	0.2000	0.7000	0.2000
P_15	0.2667	0.6667	0.2000
P_20	0.2000	0.7000	0.2000
P_30	0.1667	0.6333	0.1333
P_100	0.1300	0.4700	0.0400
P_200	0.1200	0.3200	0.0250
P_500	0.0920	0.1280	0.0120
P_1000	0.0740	0.0640	0.0060

For this part, we picked the three topics we used from the previous part of the lab (301, 302 and 310).

For the topic 301, there is a large number of retrieved documents which is the reason why there is a good large amount of relevant documents retrieved. Talking about the precision values, we can see that even if the X number of documents increases, the decline of the precision values is not considered to be much. For the IPREC values, we can see that the values from levels .20 to 1.00 are zero. It also started out in a low number of values so it is deemed to not be relevant.

For the topic 302, there is a small amount of retrieved documents but the number of retrieved relevant documents is close to the number of retrieved documents (77 - 64). This could also be the reason why the precision values for X number of documents are very high. Same with the IPREC values for each recall value which started out in high numbers.

For the topic 310, the number of retrieved documents is very low which results in a considerably low number of relevant documents retrieved. It is also obvious that the precision values for each recall level are very low and for level .50 onwards, it has 0.0000 precision values. For each number of X documents, the precision values are also low.

We think that the closer the number of retrieved documents to the relevant documents retrieved, the higher the precision levels are which is the case for topic 2. In the case of topic

3, the retrieved relevant documents are very low which reflects the low precision values.

6 Exploring Consistency of Relevance Assessment for Topics

The three topics that we have chosen are Radio Waves and Brain Cancer, International Art Crime and Cult Lifestyles. For the purpose of the relevance assessment, we decided to use the TF-IDF IR Model with the evaluation method. To denote relevance, we are following the format in the qrel file, with the relevance indicator which is set to 0 if it's relevant and 1 if it is not relevant.

Doc No.	**Topic 1**		**Topic 2**		**Topic 3**	
	Team	*qrel*	*Team*	*qrel*	*Team*	*qrel*
1	1	1	0	0	0	0
2	1	1	0	0	0	1
3	1	0	0	0	1	0
4	0	0	1	No results	0	0
5	1	1	0	No results	0	0
6	0	0	1	0	0	0
7	No results	No results	0	No results	1	0
8	0	0	1	No results	0	0
9	0	0	0	No results	0	0
10	1	1	1	No results	0	0

- *Topic 1* - Radio Waves and Brain Cancer

1	Includes statistical study
2	Includes news items which report on the incidence of brain cancer
3	Does not include any relevant information.
4	Does not include any relevant information
5	Includes statistical study

6	Contains no relevant identifications
7	No results
8	No relevant identifications
9	No relevant identifications
10	News items which report on the incidence of brain cancer being higher/lower/same

- Topic 2 - International Art Crime

1	Contains no relevant identifications
2	Contains no relevant identifications
3	No results
4	No results
5	Contains no relevant identifications
6	No results
7	No results
8	No results
9	No results
10	No results

- Topic 2 - Cult Lifestyle

1	Does not contain information about members lifestyles or goals
2	Gives information about cult members lifestyles and the ultimate goal
3	Contains no relevant identifications
4	Contains no information about members' lifestyles or how they dress etc.
5	Contains no information about how they contribute
6	Contains no information about what they do to attain the ultimate goal of the organisation

7	Does not include how they contribute to the cult
8	Does not include information about the members lifestyles
9	Does not include information about lifestyle or what they eat and the name of the cult.
10	Does not offer information about members' lifestyles.

By looking at the results, we see that topic 2 which was the Cult Lifestyle has no results in regards to the qrel file this was because in the qrel file for a topic it was not judged for relevance when the qrel file was created for TREC as it was already considered to be non-relevant by the TREC organisers.

When we read the documents it seemed like they were sometimes accurate but no results for the document ID appeared in the qrel file. We see from topic 1 that it generated 4 relevant results whereas topic 2 generated all non relevant results with no results found in the qrel file and from topic 3 we see that it only generated 1 relevant result. Both topic 1 and topic 2 generated more relevant results than topic 2 which highlighted to us that the words cult and lifestyles may be used to describe non relevant stuff in documents all the time, hence the relevance of only 1 document in the top 10 documents.