CA4009 Search Technologies Lab #2
Date: 14/11/2019
Name(s): Maksims Kompanijecs - 15306971, Matthew Farelly - 15366246

Topics file: 310 - 450.xml - **150 TREC format topic statements**
Sample output file: trec678.res - **Sample output for the 150 TREC topics in the standard TREC output format**
Relevance assessment: qrels.trec678.adhoc - **Manual relevance assessments for the 150 TREC topics**

Question 3 - Manual examination of a TREC test collection and sample search results.

Three topics that we chose for this exercise: 336 - Black Bear Attacks, 370 - food/drugs laws, 412 - airport security.

And then we compared the ranking of these documents using their title, description and narrative. The topics that we chose were at random and for BM25 we chose values k=1.2 and b=0.75

Topic 1 - 336 - Black Bear Attacks

	Title		Description		Narrative	
Ranking	**BM-25**	**TF-IDF**	**BM-25**	**TF-IDF**	**BM-25**	**TF-IDF**
1st article	1	1	**N/A**	**N/A**	**N/A**	**N/A**
2nd article	2	2	222	**N/A**	336	**N/A**
3rd article	4	5	**N/A**	**N/A**	**N/A**	**N/A**

Topic 2 - 370 - food/drug laws

	Title		**Description**		**Narrative**	
Ranking	BM-25	TF-IDF	BM-25	TF-IDF	BM-25	TF-IDF
1st article	1	1	282	282	**N/A**	644
2nd article	3	2	4	150	**N/A**	**N/A**
3rd article	5	3	501	4	**N/A**	**N/A**

Topic 3 - 412 - airport security

	Title		Description		Narrative	
Ranking	**BM-25**	**TF-IDF**	**BM-25**	**TF-IDF**	**BM-25**	**TF-IDF**
1st article	3	1	291	**N/A**	132	132
2nd article	6	5	**N/A**	**N/A**	28	**N/A**
3rd article	8	42	**N/A**	**N/A**	**N/A**	28

From carrying out this exercise it was very clear to see that most effective search query is the title.

Question 4 - Exploring Evaluation Metrics

The goal of this exercise was to examine output file 'trec678.res' with the standard information retrieval evaluation application *trec_eval*

When we compile and run trec_eval program with the sample retrieval file 'trec678.res' and corresponding qrel file using '**./trec_eval qrels.trec678.adhoc trec678.res'**.What we recieve is a set standard information retrieval metric results for the TREC which was 500,000 news articles mainly taken from the LA Times newspaper. The results here are averaged based on **ALL** of the topics.

```
maksimk2@l101-29:~/Lab 2/trec_eval_latest/trec_eval-9.0.7 $ ./trec_eval qrels.trec678.adhoc trec678.res
runid                   all     lm
num_q                   all     150
num_ret                 all     142395
num_rel                 all     14010
num_rel_ret             all     7282
map                     all     0.2145
gm_map                  all     0.0991
Rprec                   all     0.2644
bpref                   all     0.2368
recip_rank              all     0.5778
iprec_at_recall_0.00    all     0.6393
iprec_at_recall_0.10    all     0.4549
iprec_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
maksimk2@l101-29:~/Lab 2/trec_eval_latest/trec_eval-9.0.7 $
```

Examine the values of Precision at rank cutoffs, Mean Average Precision (MAP) and Recall.

By my understanding of Precision at rank cutoffs - the most relevant results will be displayed in the first few articles.

Question 5 - Exploring Consistency of Relevance Assessment for Topics

Three topics selected from the topic set

1. 425 - counterfeiting money
2. 395 - tourism
3. 363 - transportation tunnel disasters

For this exercise I selected 3 different topics. On purpose I have selected topics that we are different size in length. I think this will show that the shorter the title search queries will return results more accurately matched to the file ranking document.

I will use BM-25 ranking function to estimate the relevance in the first 10 documents by a given topic. My values will be the same as the first exercise k=1.2 and b=0.75

Topic 1 - 425 - counterfeiting money

Article #	*Relevant?*			
	BM-25	**qrel**	**Article ID**	**Summary**
1	1	1	FBIS4-26260	Counterfeit money to buy goods in China
2	1	1	FBIS3-54773	Production and sale of counterfeit money in Russia
3	1	1	FBIS3-54773	Counterfeiting money is a widespread crime in Russia
4	1	1	LA091590-0091	Counterfeit money found in California
5	1	1	FBIS3-58171	Head of gang jailed for counterfeiting money in Vienna
6	1	1	FBIS4-58263	Counterfeit bills found on Jordanian market
7	1	1	FBIS4-47199	How to recognize counterfeit money - Moscow
8	1	1	LA102189-0077	Man accused of counterfeiting money in Manteca, USA
9	1	1	LA022289-0114	Counterfeit money found in a bag in California
10	0	0	LA010390-0055	Company had been selling counterfeit goods

Topic 2 - 395 - tourism

Article #	*Relevant?*			
	BM-25	**qrel**	**Article ID**	**Summary**
1	1	0	FBIS4-66770	Someone appointed as minister of tourism

2	0	0	FBIS3-21296	Privatization of tourism sector in Egypt
3	1	0	FBIS4-51171	Article on tourism development in China
4	0	1	FBIS4-56176	Cuban tourism ministry created
5	1	0	FT944-7133	Carribean tourism season
6	0	0	LA022290-0039	Tourism dropped in Tijuana and Rosarito
7	1	0	FBIS3-60947	Article about optimism of tourism in Cairo
8	0	0	FR940926-2-00017	Travel and Tourism advisory board in US
9	1	0	FBIS4-10441	New structure for Cuban tourism was announced
10	1	0	FBIS4-35578	Minister of tourism for China

Topic 2 - 363 - transportation tunnel disasters

Article #	*Relevant?*			
	BM-25	**qrel**	**Article ID**	**Summary**
1	1	0	FT922-12816	Flooding in underground tunnel system
2	0	0	FT944-11817	New ‘Austrian’ tunneling method
3	0	0	FT923-10334	Corporate crisis management software
4	1	0	LA071490-0103	Fire in a railway tunnel
5	1	0	FT922-12325	Underground tunnel burst in chicago
6	1	1	LA070890-0174	Muslim pilgrims suffocated in tunnel in Mecca
7	0	0	FT941-5434	Article on Channel tunnel
8	0	1	FT934-16621	Development of channel tunnel
9	0	0	FT932-2651	Development of two high speed tunnels through Alps
10	0	0	FT941-6061	Tunnel development projects planned through Alps

From carrying out this exercise it was strange to see that search query with the lenght of two words returned the most accurate results. This really came as a suprise as I was expecting one worded search requests to return the most accurate results.