CA4009: Search Technologies
Laboratory Session 2
10th November 2016

Jordan Healy - 13379226
Tríona Barrow - 11319851

- *Manual examination of TREC data and search results*

After downloading and examining the files - we began making sense of all of the IDs within the three files. All were pretty straightforward to get to grips with, as the layout for all of them was quite straightforward.

The XML file was just a list of summaries of the documents within the TREC system. It included the document titles, document number, description, and narrative/specification.

doc-number title desc narr

The RES file was also straightforward, it used the layout:

doc-number query-number 0 document-id rank score exp

We compared ranked positions of three documents using their Title, Description and Narrative. We chose three documents that have a variance in title length - one with three words, one with two, and the last with one word. When searching using the BM25 IR Model we chose the values k=1.2 and b=0.5. These were our results:

- Document 301

Title: International Organized Crime

RES File	**Title**		**Description**		**Narrative**	
Rank	*tf-idf*	*BM25*	*tf-idf*	*BM25*	*tf-idf*	*BM25*
1	3	2	38	Not found	Not found	Not found
2	12	4	820	38	Not found	Not found
3	5	5	Not found	44	Not found	Not found
4	2	15	793	820	Not found	Not found
5	1	38	44	Not found	Not found	Not found

- Document 308

Title: Implant Dentistry

RES File	**Title**		**Description**		**Narrative**	
Rank	*tf-idf*	*BM25*	*tf-idf*	*BM25*	*tf-idf*	*BM25*
1	1	1	33	1	1	1
2	3	2	1	254	Not found	Not found
3	2	3	35	33	70	70
4	26	5	69	35	Not found	Not found
5	5	10	88	116	Not found	Not found

- Document 312

Title: Hydroponics

RES File	**Title**		**Description**		**Narrative**	
Rank	*tf-idf*	*BM25*	*tf-idf*	*BM25*	*tf-idf*	*BM25*
1	1	2	Not found	Not found	Not found	2
2	2	1	Not found	Not found	Not found	Not found
3	3	3	Not found	Not found	Not found	Not found
4	4	5	Not found	Not found	2	Not found
5	5	4	Not found	Not found	Not found	Not found

We found as a result of this that the most effective search queries were with title, and the shorter the query the more the results matched to the RES file ranking for the documents. Description was useful in some ways - however this seems to be more like a summary for the document - so will not be able to match strings to the document as effectively as with the title, as some terms that appear in the description may not appear in the document itself. Narrative was again worse off in comparison to the RES file - from looking at the XML file this seems to be more a specification for what is needed from the document, so generally will show as less relevant to the subject of the document. Both evaluation functions seemed to rank results relatively similarly for the most part.

- *Exploring Evaluation Metrics*

```
-  trec_eval qrels.test results.test
```

Looking at the output of running trec_eval to generate the sample retrieval on the file “trec678.res” we are given a set of standard information retrieval metric results. We are only concerned with the following metrics:

Output Type	**Description**	**Value**
num_ret	Total no. retrieved docs	142395
num_rel_ret	Total no. retrieved relevant docs (according to qrels file)	7282
P_5	Precision of first 5 docs	0.4240
P_10	Precision of first 10 docs	0.4027
P_15	Precision of first 15 docs	0.3738
P_20	Precision of first 20 docs	0.3487
P_30	Precision of first 30 docs	0.3111
P_100	Precision of first 100 docs	0.1940
P_200	Precision of first 200 docs	0.1369
P_500	Precision of first 500 docs	0.0785
P_1000	Precision of first 1000 docs	0.0485
map	Mean average precision (map)	0.2145
iprec_at_recall_0.00	Interpolated Recall - precision averages at 0.00 recall	0.6393
iprec_at_recall_0.10	IR - precision averages at 0.10 recall	0.4549
iprec_at_recall_0.20	IR - precision averages at 0.20 recall	0.3590
iprec_at_recall_0.30	IR - precision averages at 0.30 recall	0.3014
iprec_at_recall_0.40	IR - precision averages at 0.40 recall	0.2377
iprec_at_recall_0.50	IR - precision averages at 0.50 recall	0.1925
iprec_at_recall_0.60	IR - precision averages at 0.60 recall	0.1510
iprec_at_recall_0.70	IR - precision averages at 0.70 recall	0.1125
iprec_at_recall_0.80	IR - precision averages at 0.80 recall	0.0728
iprec_at_recall_0.90	IR - precision averages at 0.90 recall	0.0528
iprec_at_recall_1.00	IR - precision averages at 1.00 recall	0.0272

Looking at the number retrieved relevant documents compared to the number retrieved documents, we can see that only a little over 5% of the documents are deemed the most relevant.

Considering that the values for the precision of the first n documents decreases as n increases, it’s safe to say that the smaller the value for n, the more relevant these documents are to the user. There is always a decrease precision in the values above, but there is a sharp decrease in precision as n goes from 30 to 100. This means that the user is more likely to look at the first 5, 10 or maybe even 30 documents to satisfy their need, but it is highly unlikely that they will browse the first 100 documents or more.

The mean average precision for this is a calculation based on the recall and precision values for the documents retrieved. These are broken up into segments based on the ranking - and run between 0 and 1 for recall value. Similarly to the precision value - it decreases the further down the rankings the result is. However, it starts out as a higher number as it is based on the number of queries - so acts more as a measure of quality for relevance feedback.

Recall is also a good metric for the relevancy of the results - as it is calculated from the number of relevant documents retrieved. It only looks at the rate of correct/relevant results in regards to the overall number of relevant results. The interpolated recall rates, alongside the other ones - also follow the similar trend, suggesting that the most relevant documents are being ranked nearer the top, and that the best feedback is being given for those also.

- `trec_eval -q qrels.test results.test`

Looking at the output for running the same command only this time using the -q flag, we get the average values (just like before) for all documents but also new values for each document. We will focus only on the documents we used from part 1 (namely, 301, 308 and 312).

Output Type	**Document 301 Value**	**Document 308 Value**	**Document 312 Value**
num_ret	474	4	11
num_rel_ret	74	4	11
P_5	0.2000	0.4000	0.8000
P_10	0.2000	0.2000	0.7000
P_15	0.2667	0.1333	0.5333
P_20	0.2000	0.1000	0.5500
P_30	0.1667	0.0667	0.3667
P_100	0.1300	0.0300	0.1100
P_200	0.1200	0.0150	0.0550
P_500	0.0920	0.0080	0.0220
P_1000	0.0740	0.0040	0.0110
map	0.0179	0.5138	0.7433
iprec_at_recall_0.00	0.3077	1.0000	1.0000
iprec_at_recall_0.10	0.0967	1.0000	1.0000

iprec_at_recall_0.20	0.0000	1.0000	1.0000
iprec_at_recall_0.30	0.0000	1.0000	0.8000
iprec_at_recall_0.40	0.0000	1.0000	0.7143
iprec_at_recall_0.50	0.0000	1.0000	0.7000
iprec_at_recall_0.60	0.0000	0.0435	0.7000
iprec_at_recall_0.70	0.0000	0.0435	0.6111
iprec_at_recall_0.80	0.0000	0.0119	0.6111
iprec_at_recall_0.90	0.0000	0.0119	0.6111
iprec_at_recall_1.00	0.0000	0.0119	0.6111

For document 301, we can see that there is a considerably larger number of documents retrieved, which in turn gives us a larger number of relevant documents retrieved. Due to this, we obtain more accurate precision values. (This is because, Precision = No of Relevant docs retrieved / Total docs retrieved). Even with a small value of n, the proportion of retrieved documents that are relevant is only 20%. Within the retrieved documents, there is also a low number of relevant documents retrieved. (This is because, Recall = No of relevant docs retrieved / Total relevant docs in collection). We can see this as the documents recall values falls to zero at only 20%.

For document 308, there is an extremely low number of documents retrieved, but all of them are deemed to be relevant. Since these values are so low, its precision values will be less accurate. We can see this since the values for P(5) to P(20) jump considerably. Within the retrieved documents, there is also an extremely high number of relevant documents retrieved. We can see this as the documents recall values stay at one for the entire first half.

The same can be said for document 312, as its number of documents retrieved is also low. It’s precision values are too, inaccurate, but there is a slight improvement from document 308. Within the retrieved documents, there is also a high number of relevant documents retrieved, although not as high as document 308. We can see this as the documents recall values stays relatively high for the first 70%.

- *Exploring Relevance Assessment*

We decided to reuse the three topics we picked for the first part of this report - International Organized Crime, Implant Dentistry, and Hydroponics. We decided to only use tf-idf evaluation algorithm, to make this easier. We used the name notation as in the qrel file for relevancy - 0 for irrelevant, 1 for relevant.

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

- *Topic 1 - International Organised Crime*

1	Does not identify an organisation/type of illegal activity
2	Does not identify a type of illegal activity
3	Identifies organisation (internal affairs) with members conducting multiple named illegal activities
4	Does not identify an organisation
5	Does not identify an organisation
6	Does not identify an organisation/type of illegal activity
7	Identifies the mafia as the organisation and corruption as the type of activity
8	Does not identify a type of illegal activity
9	Does not identify an organisation
10	Identifies mafia and cocaine smuggling

- *Topic 2 - Implant Dentistry*

1	No discussion of advantages/disadvantages, comfort, etc - anecdote about implants
2	No discussion of advantages/disadvantages, comfort, etc - FDA reclassification doc
3	No discussion of advantages/disadvantages, comfort, etc - FDA reclassification doc
4	Discusses breast implants - not dental related
5	List of names and wages - not related
6	Discusses breast implants - not dental related
7	Discusses ions - not dental related
8	List of names and wages - not related
9	Press release for promotion - not dental related
10	Political statement about the NHS - not related

- *Topic 3 - Hydroponics*

1	Criminal ruling regarding marijuana - not related
2	Doesn't go into details for nutrients/experiments/substrates/etc
3	Report on discovery of marijuana - not related
4	Discusses preparing tomatoes for cooking - not related
5	Discusses residents of a specific area - not related
6	Segment of a contract for a tomato picker - not related
7	Report on drug discoveries by law enforcement - not related
8	Discussion of sources of renewable energy, no mention of the science of hydroponics, it's merely listed as an example - not related
9	No discussion of the science of hydroponics, topic is related to BIBs loan program - not related
10	First half of document describes an experiment with regards to growing plants under water (which is relevant), but then describes a method of cooking these plants with duck- is related

As a result of this, we noticed that the two queries with multiple words were more accurate to the information need than the single word query. Both the first and second query only had one incorrect entry in the qrel file. However the last query only generated one relevant result, and the qrel file was incorrect on 5 results - this could also be linked to how uncommon the term "hydroponics" is in everyday use, which allows for less relevance feedback.