**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

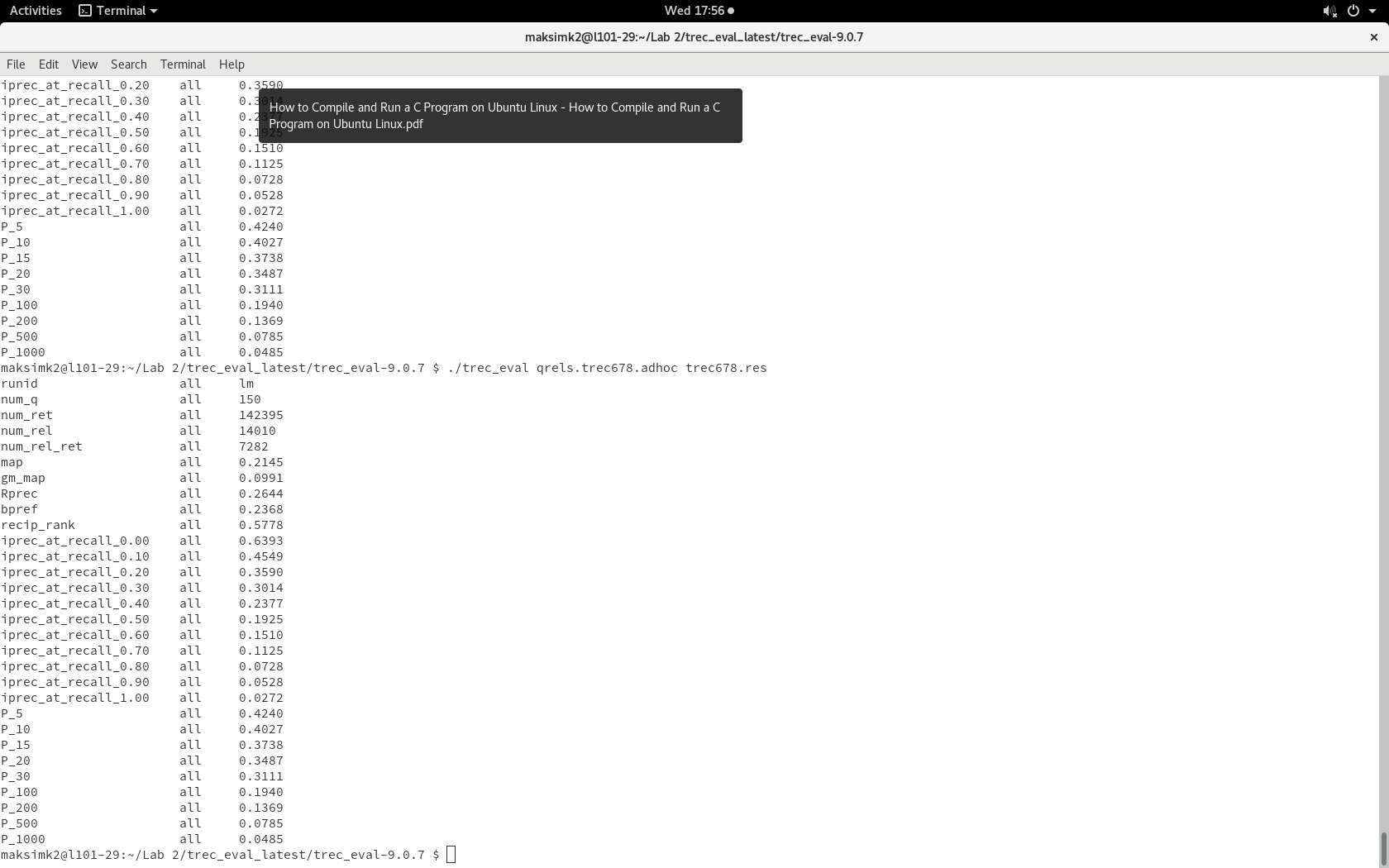
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Title | | Description | | Narrative | |
| **Ranking** | **BM-25** | **TF-IDF** | **BM-25** | **TF-IDF** | **BM-25** | **TF-IDF** |
| **1st article** | 3 | 1 | 291 | **N/A** | 132 | 132 |
| **2nd article** | 6 | 5 | **N/A** | **N/A** | 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.



**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](http://136.206.48.37:8084/IRModelGenerator/DocumentViewer?id=440279) | Counterfeit money found in California |
| **5** | 1 | 1 | [FBIS3-58171](http://136.206.48.37:8084/IRModelGenerator/DocumentViewer?id=314838) | Head of gang jailed for counterfeiting money in Vienna |
| **6** | 1 | 1 | [FBIS4-58263](http://136.206.48.37:8084/IRModelGenerator/DocumentViewer?id=294971) | Counterfeit bills found on Jordanian market |
| **7** | 1 | 1 | [FBIS4-47199](http://136.206.48.37:8084/IRModelGenerator/DocumentViewer?id=338465) | How to recognize counterfeit money - Moscow |
| **8** | 1 | 1 | [LA102189-0077](http://136.206.48.37:8084/IRModelGenerator/DocumentViewer?id=453507) | Man accused of counterfeiting money in Manteca, USA |
| **9** | 1 | 1 | [LA022289-0114](http://136.206.48.37:8084/IRModelGenerator/DocumentViewer?id=474500) | Counterfeit money found in a bag in California |
| **10** | 0 | 0 | [LA010390-0055](http://136.206.48.37:8084/IRModelGenerator/DocumentViewer?id=396918) | 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.