HomeWork-1 Report Template

Score of Top Relevant File of a Sample Query for each Retrieval Model

Model	Score
ES (built-in)	11.86
Okapi TF	2.18
TF-IDF	4.81
Okapi BM-25	11.79
Unigram LM with Laplace smoothing	-36.59
Unigram LM with Jelinek-Mercer smoothing	-11.54

Inference on the above results

It was mentioned that the ElasticSearch built in algorithm would have similar results to the Okapi BM-25 algorithm so it made sense to see that the top outputs in the file for sample query 85 were very close in score. Okapi TF and TF-IDF were also similar, with the difference that TF-IDF featured another factor which was the log of corpus length divided by the term frequency so it made sense that these scores were also greater. The BM-25 is built on the shortcomings of TF-IDF and features additional terms to dampen the effect of term frequency so it makes sense that scores would also be greater. The language models use log likelihoods which lead to negative scores.

Retrieval Model Performance

[Highlight the scores more than 0.28]

Model	Average Precision	Precision at 10	Precision at 30
ES (built-in)	0.2829	0.44	0.3747
Okapi TF	0.236	0.396	0.32
TF-IDF	0.2722	0.42	0.36
Okapi BM-25	0.2802	0.436	0.3707
Unigram LM with Laplace smoothing	0.2251	0.4080	0.3147
Unigram LM with Jelinek-Mercer smoothing	0.2365	0.4	0.3160

Inference on above retrieval model results

The results that were obtained from doing the evaluations were very understandable based on knowledge about the various retrieval models. Of the first three models the Okapi TF had the lowest score. The TF-IDF is an improvement on the first model as it features the IDF term which measures how important or rare a term is. Incorporating the IDF makes the model more precise which is reflected in the evaluation. The BM-25 is a further improvement as it dampens the effect of term frequency which leads to more precision. Between the language models the Jelinek-Mercer smoothing had better precision. This may be because it was more optimized with its one parameter vs the language model with Laplace smoothing.

Table showing the Query used for Evaluation

Query number	93	89	57	77	61
Original Query	Document must describe or identify supporters of the National Rifle Association (NRA), or its assets.	Document must identify an existing or pending investment by an OPEC member state in any "downstream" operation.	Document will discuss how MCI has been doing since the Bell System breakup.	Document will report a poaching method used against a certain type of wildlife.	Document will discuss the role of Israel in the Iran-Contra Affair.
Processed Query	nation rifl associ nra	invest opec downstream oper	mci bell	poach	israel iran contra
Processed Query - Pseudo RF **(Only MS students)**					