HomeWork-1 Report Template

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

|  |  |
| --- | --- |
| Model | Score |
| ES (built-in) | 0.2517 |
| Okapi TF | 0.1901 |
| TF-IDF | 0.2635 |
| Okapi BM-25 | 0.1601 |
| Unigram LM with Laplace smoothing | 0.1752 |
| Unigram LM with Jelinek-Mercer smoothing | 0.1640 |

**Inference on the above results**

The above results clearly infer that the ES Built and TF\_IDF results in the highest score from querying the data, while the language models score comparitively lesser.

BM25 model is supposed to be on par with the ES built in model, but lacks comparitively due to data/query preprocessing reasons.

# Retrieval Model Performance

# [ Highlight the scores more than 0.28]

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Average Precision | Precision at 10 | Precision at 30 |
| ES (built-in) | 0.2517 | 0.3240 | 0.2853 |
| Okapi TF | 0.1901 | 0.2960 | 0.2360 |
| TF-IDF | 0.2635 | 0.3440 | 0.3027 |
| Okapi BM-25 | 0.1601 | 0.2120 | 0.1533 |
| Unigram LM with Laplace smoothing | 0.1752 | 0.2800 | 0.2080 |
| Unigram LM with Jelinek-Mercer smoothing | 0.1640 | 0.2480 | 0.1920 |

**Inference on above retrieval model results**

The precision of the above retreival models starts with a higher score, and lacks comparitively as the documents increase and the scores go down.

The unigram models show a similar trend as the score stays up at the initial 10 documents, but the score is then pulled down at the 30th document state.

# Pseudo-relevance Feedback Improvements[ ONLY MS STUDENTS]

[The highlighted scores that indicate an improvement in the average precision score of the model]

1. Result after adding the top 5 distinctive terms to each query.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Average Precision | Precision at 10 | Precision at 30 |
| ES (built-in) | 0.2434 | 0.3480 | 0.2480 |
| Okapi TF | 0.1783 | 0.2920 | 0.2280 |
| TF-IDF | 0.2544 | 0.3600 | 0.2973 |
| Okapi BM-25 | 0.1002 | 0.2120 | 0.1493 |
| Unigram LM with Laplace smoothing | 0.1620 | 0.2640 | 0.1760 |
| Unigram LM with Jelinek-Mercer smoothing | 0.1533 | 0.2400 | 0.1787 |

1. Results after adding top 5 significant terms from Elasticsearch aggs to each query.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Average Precision | Precision at 10 | Precision at 30 |
| ES (built-in) | 0.2434 | 0.3480 | 0.2480 |
| Okapi TF | 0.1783 | 0.2920 | 0.2280 |
| TF-IDF | 0.2544 | 0.3600 | 0.2973 |
| Okapi BM-25 | 0.1002 | 0.2120 | 0.1493 |
| Unigram LM with Laplace smoothing | 0.1620 | 0.2640 | 0.1760 |
| Unigram LM with Jelinek-Mercer smoothing | 0.1533 | 0.2400 | 0.1787 |

**Inference on the above pseudo-relevance results**

The relevant significant and distinctive terms which are fetched from the corpus, are introduced into the queries, however no observable improvement was concluded. This clearly shows maybe the terms fetched were not relevant to the document and hence the downfall in the score.

**Table showing the Query used for Evaluation**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Query number | 95 | 68 | 57 | 97 | 98 |
| Original Query | describe a computer application to crime solving | report actual studies, or even unsubstantiated concerns about the safety to manufacturing employees and installation workers of fine-diameter fibers used in insulation and other products. | discuss how MCI has been doing since the Bell System breakup. | identify instances of fiber optics technology actually in use. | identify individuals or organizations which produce fiber optics equipment. |
| Processed Query | ['comput', 'applic', 'crime', 'solv'] | ['actual', 'studi', 'unsubstanti', 'concern', 'safeti', 'manufactur', 'employe', 'instal', 'worker', 'finediamet', 'fiber', 'insul', 'product'] | ['mci', 'bell', 'system', 'breakup'] | ['instanc', 'fiber', 'optic', 'technolog', 'actual'] | ['individu', 'organ', 'produc', 'fiber', 'optic', 'equip'] |
| Processed Query - Pseudo RF **( Only MS students)** | ['comput', 'applic', 'crime', 'solv', ‘supercomput’,’faster’,’model’,’,machin’,’million’] | ['actual', 'studi', 'unsubstanti', 'concern', 'safeti', 'manufactur', 'employe', 'instal', 'worker', 'finediamet', 'fiber', 'insul', 'product', ‘fiber’,’reinforc’,’technolog’,’compani’, ‘lead’] | ['mci', 'bell', 'system', 'breakup', ‘local’, ‘charge’, ‘longdist’, ‘breakup’, ‘reduct’] | ['instanc', 'fiber', 'optic', 'technolog', 'actual', ‘glass’, ‘work’, ‘manufactur’,’swirbul’, ‘technolog’] | ['individu', 'organ', 'produc', 'fiber', 'optic', 'equip', ‘satellit’,’technolog’, ‘manufactur’, ‘opticalwaveguid’,’compani’] |