# **CS 6200: Information Retrieval**

# **Fall 2019**

# **Ad-hoc query retrieval and their ranking evaluation using BM25**

Team Members : Sai Divya Sangeetha Bhagavatula, Sayan Biswas, Sneha Agarwal, Varun Jagadeesh

# Problem Description

# The purpose of the project is to evaluate how Information Retrieval systems perform. There are many tasks which an Information Retrieval systems performs such as accepting query input from the user, transforming the query, finding relevant documents for the query, ranking the documents and finally displaying the documents in terms of relevance to the user who posted the query. With this project we aim to focus on the Ranking aspect of the Information Retrieval system and how it performs when fed with ad-hoc queries. We aim at using BM25 as the baseline model as it focuses on topical relevance and is often observed to achieve high accuracies and performs better than TF-IDF. We will be comparing the rankings obtained by BM25 to that derived by human judgements and compare it’s ranking and relevance to the one obtained using manual human judgement criteria. In the last we will be using query likelihood model and Jelinek-Mercer smoothing on top of query likelihood method to rank the documents and compare it’s ranking to the baseline model as well as with the one obtained using manual human judgement criteria.

# Ad-hoc queries for the task

A lot of queries of belonging to different types and expecting different kind of documents were tried. A few sample example are listed below:

1. Premier league stats for 2019 – The user of this query expects statistical and structured data for the specific year mentioned.
2. Pros and Cons of butterfly keyboard used by Apple – The user of this query expects a comparative analysis on the butterfly keyboard used by Apple.
3. Citizenship amendment bill – The user of this query expects details on this bill and people’s say on it.
4. How to evaluate a search engine? – The user of this query wants to know the various methodologies that can be used and applied on search engines to evaluate them

Relevant documents included for query “Premier league stats for 2019” in evaluation set:

1) <https://www.espn.com/soccer/stats/_/league/eng.1>

2) <https://www.premierleague.com/stats>

Evaluating Information Retrieval System based on Ranking

Finding relevant documents and getting metrics for estimating human performance:

Each team member took the queries posted by the other team members and used various sources of knowledge such as search engines(Google, Safari, Duck Duck Go, Bing, etc.), query specific websites (such as ‘GeekforGeeks’ for technical coding related queries, ‘espn’ for sports related queries), etc. to come up with a set of relevant documents for each query. In total 10 most relevant documents were found. These documents were shared with the team member who initially posted the query. Based on their expectations each document was re-ranked from 1 to 10 with 1 being the most relevant and 10 being the least relevant and also assigned a relevance score on a scale of 0-4, with 4 being the most relevant and 0 the least. On the basis of these relevance ratings, DCG@10(Discounted Cumulative Gain @10) was calculated. IDCG@10(Ideal Discounted Cumulative Gain @10) was also calculated. And then NDCG@10(Normalized Discounted Cumulative Gain @10) was calculated by dividing DCG@10 by IDCG@10.

These above 2 metrics – re-ranking on a scale of 1-10 and IDCG@10 were the metrics considered for estimating human performance on the task.

Evaluating queries on the baseline model:

We considered BM25 as our baseline model as it is a retrieval model based on the probabilistic retrieval framework. The main advantage of BM25 is its efficiency. It performs very well in many ad-hoc retrieval tasks. BM25 often achieves better performance compared to TF-IDF. We evaluated the baseline model on the evaluation set as obtained above.

All the documents for a specific query was passed to the BM25 model and the document scores obtained. These documents were re-ranked based on the document scores. These are the new rankings based on our baseline model BM25. These rankings were compared to the rankings obtained using human judgements. Again here on the basis of new ranking for the documents, DCG@10, IDCG@10 and NDCG@10 was calculated.

Comparison of results obtained using BM25 to the human judgements:

We found that human judgements performed better than the baseline model BM25. The human judgement performed better than BM25 when compared using the raw rankings assigned to the documents as well as on comparing the NDCG@10.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Query(e.g. “Premier league stats for 2019”) | Ranking(Human Judgement) | Graded relevance score | BM25 document score | Ranking(BM25) |
| Doc 1 | 1 | 4 | 4.23114584 | 5 |
| Doc 2 | 2 | 3 | 4.43129396 | 2 |
| Doc 3 | 3 | 3 | 4.56857724 | 1 |
| Doc 4 | 4 | 4 | 3.48422328 | 10 |
| Doc 5 | 5 | 0 | 3.69882664 | 6 |
| Doc 6 | 6 | 3 | 4.33734667 | 4 |
| Doc 7 | 7 | 3 | 4.37867215 | 3 |
| Doc 8 | 8 | 1 | 3.69399485 | 7 |
| Doc 9 | 9 | 4 | 3.52944513 | 8 |
| Doc 10 | 10 | 1 | 3.50723823 | 9 |

For this specific query, using, Human Judgement, NDCG@10 = 0.94 and using BM25, NDCG@10 = 0.78. This clearly shows that Human judgement superseded BM25.

In order to better the Information Retrieval system and bring its accuracy closer to human judgement we proceeded to implement query likelihood model for ranking the documents.

Evaluating queries on specialized model:

We initially simple query likelihood method for ranking the documents and saw some improvement in the Information retrieval system performance when using query likelihood as compared to BM25. We then modified the query likelihood model to include Jelinek-Mercer smoothing. On doing so we saw significant improvement and for most of the queries the rankings were similar to that given by human judgements. Also, the NDCG@10 score using query likelihood model and Jelinek-Mercer smoothing moved quite closer to the one obtained using human judgements.

Comparison of results obtained using query likelihood model with Jelinek-Mercer smoothing to BM25 and human judgements:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Query(e.g. “Premier league stats for 2019”) | Ranking(Human Judgement) | Graded relevance score | Ranking(BM25) | Query-likelihood document score | Ranking(query-likelihood) |
| Doc 1 | 1 | 4 | 5 | -34.928 | 1 |
| Doc 2 | 2 | 3 | 2 | -34.919 | 3 |
| Doc 3 | 3 | 3 | 1 | -34.942 | 6 |
| Doc 4 | 4 | 4 | 10 | -34.907 | 2 |
| Doc 5 | 5 | 0 | 6 | -34.881 | 5 |
| Doc 6 | 6 | 3 | 4 | -34.946 | 7 |
| Doc 7 | 7 | 3 | 3 | -34.960 | 9 |
| Doc 8 | 8 | 1 | 7 | -34.961 | 10 |
| Doc 9 | 9 | 4 | 8 | -34.948 | 8 |
| Doc 10 | 10 | 1 | 9 | -34.926 | 4 |

For this specific query, using, Human Judgement, NDCG@10 = 0.94 and using BM25, NDCG@10 = 0.78 as seen in the above comparison. NDCG@10 for query-likelihood using Jelinek-mercer smoothing = 0.84. This clearly performs far better than BM25 although it is still poor compared to human judgement.