

Project Final Report

Quote Retrieval

GitHub Repo: <https://github.com/VaishnaviHire/Quote-Retrieval>

Maxwell King, Vaishnavi Hire, Tushar Goel

CS 6200 IR Fall 2019

king.ma@husky.neu.edu

hire.v@husky.neu.edu

tushar.1173goel@gmail.com

Task Definition

The retrieval task we have undertaken deals with getting quotes and opinions of a person or event as attributed to them in any media form news, editorials, interviews, books etc. We have also extended this task to take into consideration documents that talk about these quotes/opinions. The aim of this task is to identify how well the existing models perform to retrieve the quotes when given a wide range of queries and how the retrieval task can be improved based on the results that are achieved with these queries.

The motivation behind undertaking this task was to try and capture the most relevant document collection given the complexity of the queries. The results for queries vary widely for lesser-known personalities, most influential personalities, for people that are widely covered by the media sources. Opinions of a person can also be captured in indirect form like editorials, articles discussing the interviews or speeches, etc. Hence it makes it a challenge to capture desired quotes.

Our approach to getting a good evaluation set was to add diverse queries. We made sure to expand our queries out to include a large variety of things that would have quotes about them or in them that users would like to see. There were 30 queries in total that were used in this project covering a wide selection of potential interests.

Query Description

All 30 of the queries that were used in this project had the same format. The queries contained the query itself (title). And then a description of how the user evaluated what they were looking for in the query. Below is a sample query that breaks down what the 30 queries looked like.

SAMPLE QUERY:

Title (Query):

'Abraham Lincoln Gettysburg'

Description (evaluation Standards):

Relevant documents must include direct quotes by Abraham Lincoln from his Gettysburg Address. The relevance of these documents will be judged on how much of the speech they contain and how accessible those quotes are in the document.

Result

We tested our queries across search engines and retrieved a set of documents, typically 20 per query. These documents were then evaluated by the member who created the query for relevance. Our relevance scaled from 0 to 4, where 0 is a non-relevant document according to the evaluation standards and 4 is the most relevant document. These relevance scores were then used in the evaluation of how a baseline model would perform. Below is a look at what a sample result looks like for the query that was show in the sample query above.

<https://www.nationalaffairs.com/publications/detail/lincoln-at-gettysburg> - A speech transcript will have a relevance of 4.

<https://usa.usembassy.de/etexts/democrac/25.htm> - A document describing and pulling out quotes about the address will have the relevance of 3.

<https://www.wbalTV.com/article/this-day-in-history-abraham-lincoln-delivers-gettysburg-address/29837521> - Description of the event, without the speech transcripts or quotes will be considered as an irrelevant document with a score 0.

Human Performance

To evaluate the performance of manual retrieval, we computed count and averages of relevant documents for each set of queries. Here a query set is a collection of all queries created by a single user. Count of relevant documents are determined taking into consideration the relevance score.

We have three query sets, each containing 20 queries. Following tables describe the relevance distribution and Rankings across set.

Breakdown of Rankings per set

Row Labels	0	1	2	3	4	Grand Total
1	13	24	40	45	75	197
2	37	26	45	28	58	194
3	48	9	27	41	42	167
Grand Total	98	59	112	114	175	558

Human Set One Averages

Row Labels	Average of Ranking
Battles fought by Guru Gobind Singh	1.666666667
Elon musk interview	3.1
elon musk interview with nyt	3.181818182
elon musk latest	2.5
Elon musk news	3.4
elon musk on clean energy	3.1
elon musk on self driving cars	3.444444444
Elon musk speech	3
elon musk twitter	2.666666667
Facts about Sushruta	1.7
Jack Ma's and Elon Musk's views on AI	3.1
Jeff Bezos portrayed by top business magazines like Forbes	2.7
Kim Jong-un relationship with China	3.111111111
Leonardo da Vinci Quotes	3.4
Narendra Modi's view on Indian Economy	2.111111111
quotes and opinions of elon musk, as attributed to them in news articles, social media, books, etc.	2.9
quotes by Elon Musk	3.7
Sundar pichai's views on H-1B visa policy by the Trump Government	2
Warren Buffet views and media interactions in regard to start-ups	1.5
World Media views on Imran Khan	2.363636364
Grand Total	2.736040609

Human Set One Count Distribution

Row Labels	0	1	2	3	4	Grand Total
Battles fought by Guru Gobind Singh	1	4	2	1	1	9
Elon musk interview		1	2	2	5	10
elon musk interview with nyt			1	7	3	11
elon musk latest			7	1	2	10
Elon musk news			2	2	6	10
elon musk on clean energy		1	2	2	5	10
elon musk on self driving cars			1	3	5	9
Elon musk speech		3		1	6	10
elon musk twitter	1	1	2	1	4	9
Facts about Sushruta	1	3	4	2		10
Jack Ma's and Elon Musk's views on AI	1		1	3	5	10
Jeff Bezos portrayed by top business magazines like Forbes	1	1	2	2	4	10
Kim Jong-un relationship with China			3	2	4	9
Leonardo da Vinci Quotes		1		3	6	10
Narendra Modi's view on Indian Economy	2	2		3	2	9
quotes and opinions of elon musk, as attributed to them in news articles, social media, books, etc.		1	2	4	3	10
quotes by Elon Musk				1	8	10
Sundar pichai's views on H-1B visa policy by the Trump Government	2	1	4	1	2	10
Warren Buffet views and media interactions in regard to start-ups	2	3	3	2		10
World Media views on Imran Khan	2	2	1	2	4	11
Grand Total	13	24	40	45	75	197

Human Set Two Averages

Row Labels	Average of Ranking
Abraham Lincoln Gettysburg	3.444444444
AJ Green's Health	2.555555556
Battles fought by Guru Gobind Singh	1.6
Bill Simmons Roger Goodell	3
Celtics 11/20/2019	0.4
Facts about Sushruta	2.3
Gandhi's views on truth	1.777777778
Gladiator Russell Crowe Quotes	1.875
Jack Ma's and Elon Musk's views on AI	3.333333333
Jeff Bezos portrayed by top business magazines like Forbes	2.333333333
JFK space race	1.8
Kim Jong-un relationship with China	2.363636364
Leonardo da Vinci Quotes	3.1
Mohammad Ali	1.1
Mohammad Ali Sting like a bee	2.888888889
Narendra Modi's view on Indian Economy	2.8
President Obama Inauguration	1.1
Sundar pichai's views on H-1B visa policy by the Trump Government	2
Warren Buffet views and media interactions in regard to start-ups	1.8
World Media views on Imran Khan	3.090909091
Grand Total	2.226804124

Human Set Two Count Distribution

Row Labels	0	1	2	3	4	Grand Total
Abraham Lincoln Gettysburg		1		2	6	9
AJ Green's Health	2		2	1	4	9
Battles fought by Guru Gobind Singh	2	2	5		1	10
Bill Simmons Roger Goodell		2	1	2	5	10
Celtics 11/20/2019	9				1	10
Facts about Sushruta		2	5	1	2	10
Gandhi's views on truth	4		1	2	2	9
Gladiator Russell Crowe Quotes	1	2	3	1	1	8
Jack Ma's and Elon Musk's views on AI		1		3	5	9
Jeff Bezos portrayed by top business magazines like Forbes	1	1	4		3	9
JFK space race	3	2	1	2	2	10
Kim Jong-un relationship with China			8	2	1	11
Leonardo da Vinci Quotes		1	1	4	4	10
Mohammad Ali	4	3	2		1	10
Mohammad Ali Sting like a bee	1	1	1	1	5	9
Narendra Modi's view on Indian Economy	1		3	2	4	10
President Obama Inauguration	6	1	1		2	10
Sundar pichai's views on H-1B visa policy by the Trump Government		4	3	2	1	10
Warren Buffet views and media interactions in regard to start-ups	2	3	2	1	2	10
World Media views on Imran Khan	1		2	2	6	11
Grand Total	37	26	45	28	58	194

Human Set Three Averages

Row Labels	Average of Ranking
Abraham Lincoln Gettysburg	1.222222222
AJ Green's Health	0.5
Bill Simmons Roger Goodell	2.7
Celtics 11/20/2019	0.25
Elon musk interview	3.666666667
elon musk interview with nyt	2.555555556
elon musk latest	2.5
Elon musk news	2.5
elon musk on clean energy	1.833333333
elon musk on self driving cars	2.857142857
Elon musk speech	2.555555556
elon musk twitter	2.5
Gandhi's views on truth	1.75
Gladiator Russell Crowe Quotes	3.444444444
JFK space race	1.777777778
Mohammad Ali	1
Mohammad Ali Sting like a bee	1.714285714
President Obama Inauguration	1.444444444
quotes and opinions of elon musk, as attributed to them in news articles, social media, books, etc.	2
quotes by Elon Musk	3.555555556
Grand Total	2.119760479

Human Set Three Count Distribution

Row Labels	0	1	2	3	4	Grand Total
Abraham Lincoln Gettysburg	6		1	2		9
AJ Green's Health	5	2	1			8
Bill Simmons Roger Goodell	3		1	6		10
Celtics 11/20/2019	7	1				8
Elon musk interview			2	4		6
elon musk interview with nyt	1	1	2	2	3	9
elon musk latest		1	5	2	2	10
Elon musk news	1	3	5	1		10
elon musk on clean energy	1	1	3		1	6
elon musk on self driving cars			2	4	1	7
Elon musk speech		2	1	5	1	9
elon musk twitter	1	3	5	1		10
Gandhi's views on truth	2	2	1	2	1	8
Gladiator Russell Crowe Quotes				5	4	9
JFK space race	4	1	2	2		9
Mohammad Ali	6	2	2			10
Mohammad Ali Sting like a bee	4			3		7
President Obama Inauguration	5	1	1	2		9
quotes and opinions of elon musk, as attributed to them in news articles, social media, books, etc.	2			2		4
quotes by Elon Musk			1	2	6	9
Grand Total	48	9	27	41	42	167

As seen, the ranking average is similar across all three of the sets, with set 1 having the highest accuracy with 2.7 and sets 2 and 3 with 2.2 and 2.1 respectively. Most of the queries across the sets provided relevant results. However, less relevant (relevance score : 1,2) and irrelevant documents had major share in all the 3 sets, 269 documents. This shows that, there is a discrepancy in what was expected from the queries and search results from the search engines.

Though the human results show that the retrieval methods are working in consistently there is still the issue that we are working with human retrieval here and 11 of these sets were acquired differently by different people. So though these results indicate that the retrieval system is not functioning at a high level with the queries it was given it is not enough to stop here. The team decided to use some standard retrieval algorithms to build out a search engine of their own and some evaluation techniques.

Baseline Model:

The Baseline model that the team used in this project was a simple bag of words approach that used a query likelihood algorithm to rank the documents. The bag of words for this baseline model was built out by only using the terms that appeared in the queries. The ranking query likelihood algorithm was just the generic algorithm with some slight smoothing coefficient to account for missing terms.

$$QL = \sum_{i \in q=0}^q \log\left(\frac{f_{id}+1}{|D|+|V|}\right)$$

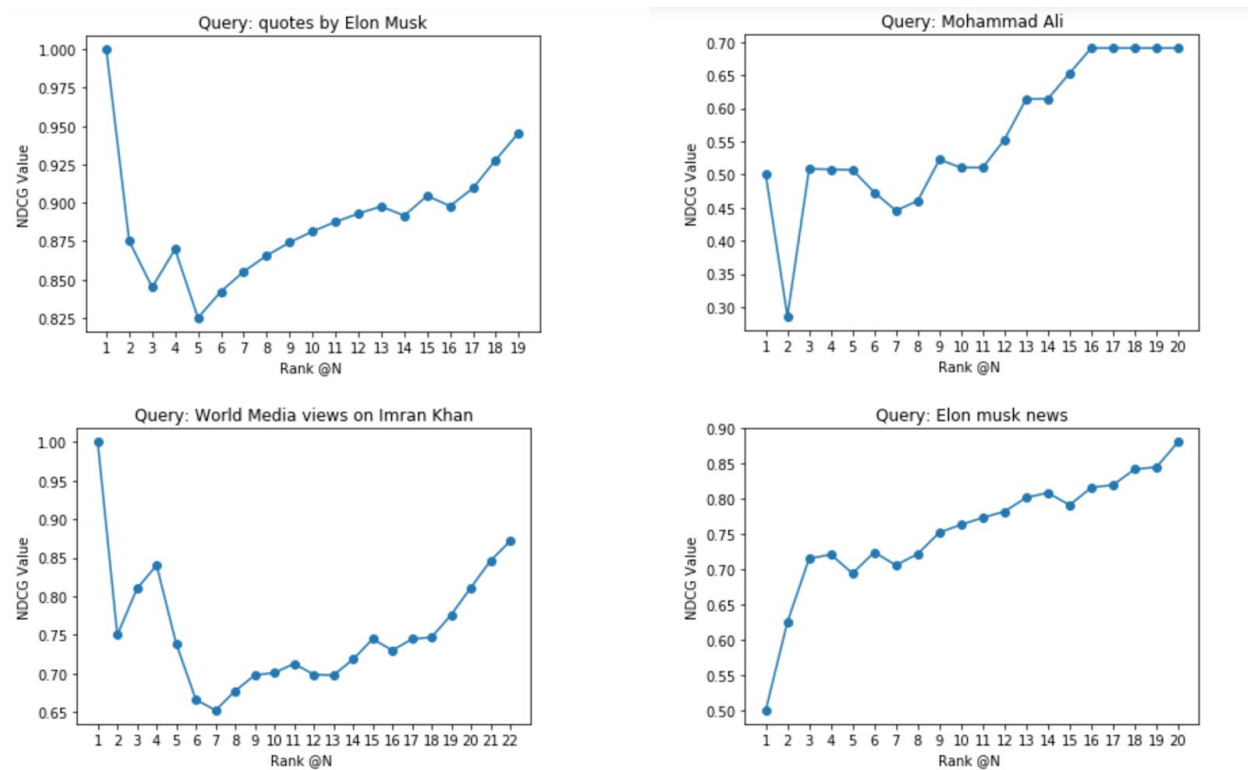
Evaluation Methods:

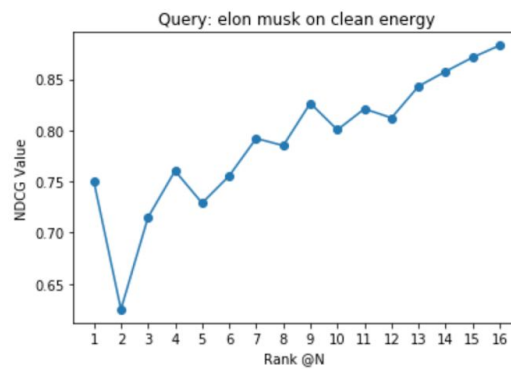
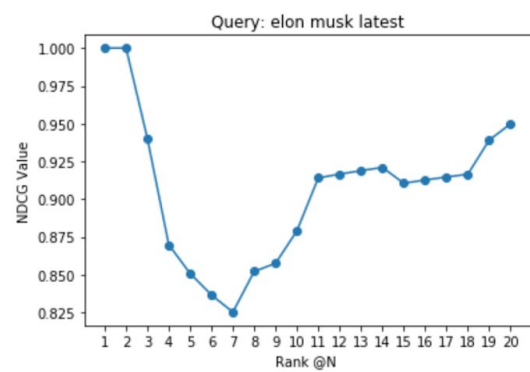
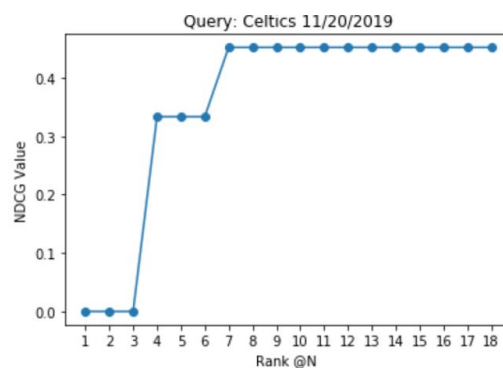
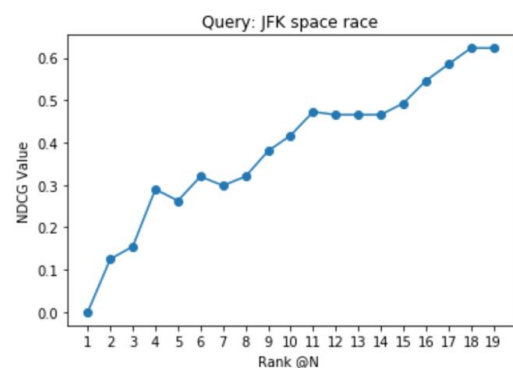
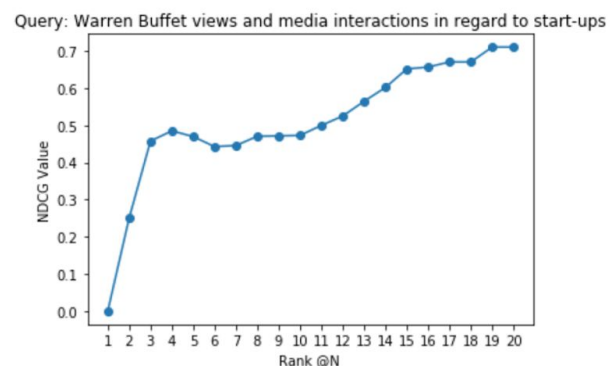
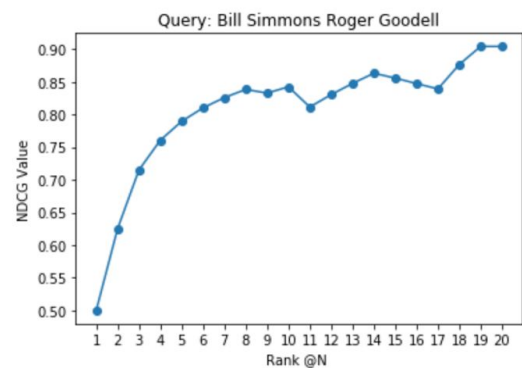
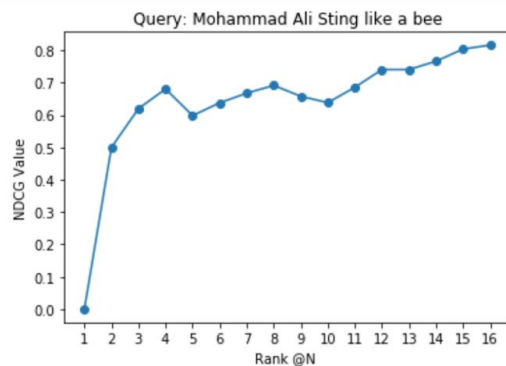
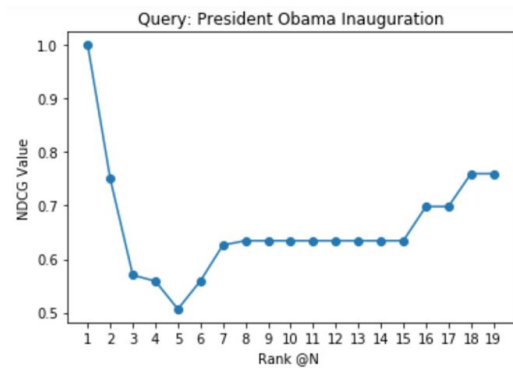
To evaluate the results we used Normalized Discounted Cumulative Rate (NDCG). NDCG was chosen due to its ability to evaluate non-binary rankings. The baseline model was first used to rank the documents and then using the ratings given by the team to the queries (0-4) Discount Cumulative (DCG) was calculated. These scores were then normalized by taking the ideal DCG that the query could have had and dividing each of the ranked documents DCG by this ideal DCG.

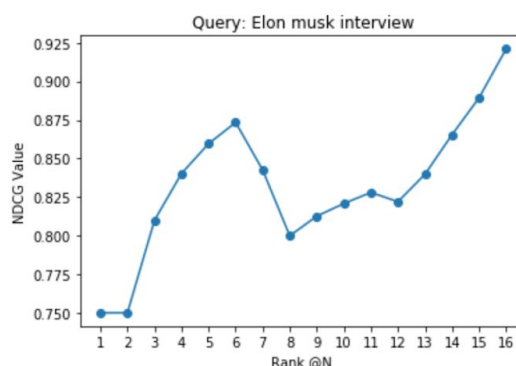
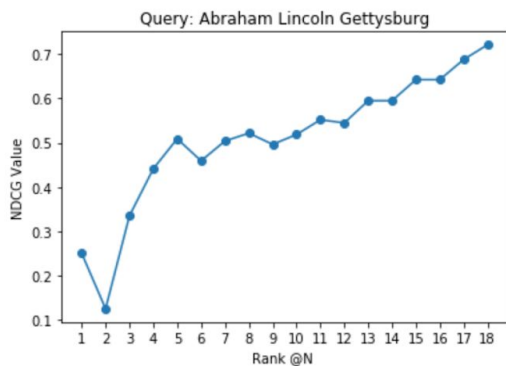
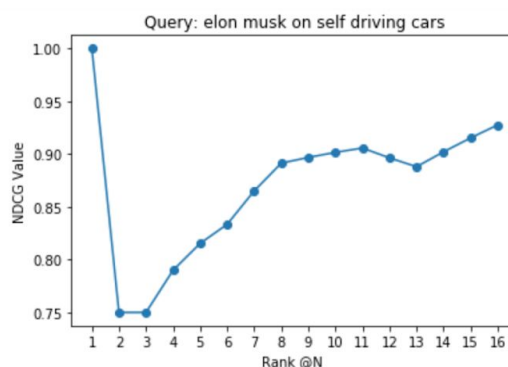
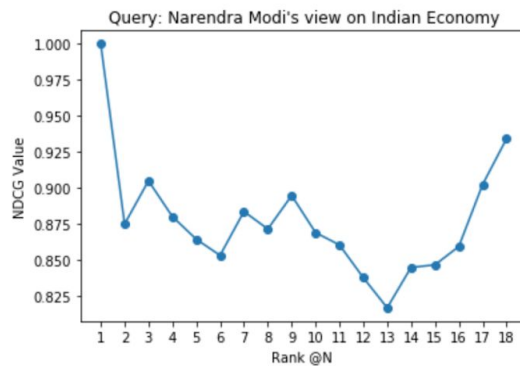
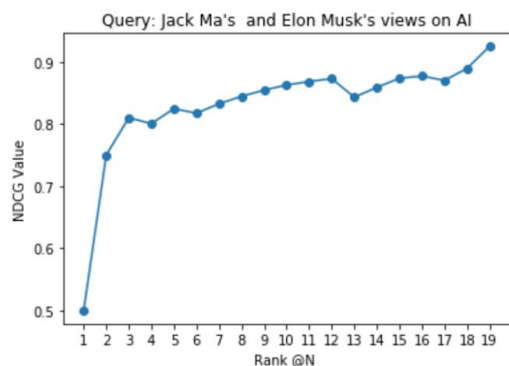
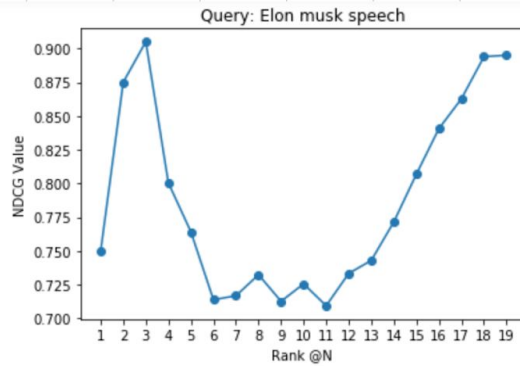
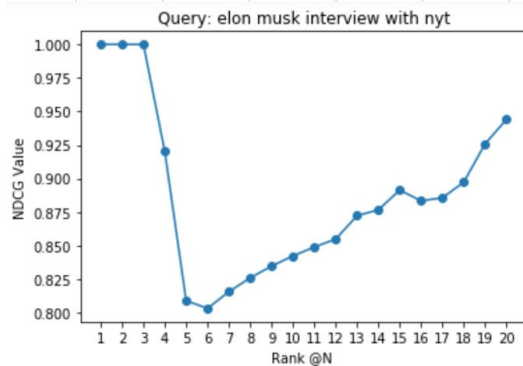
$$DCG_p = rel_1 + \sum_{i=2}^p \frac{rel_i}{\log_2 i}$$

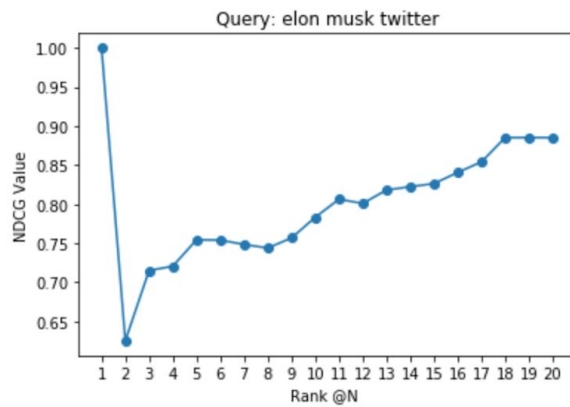
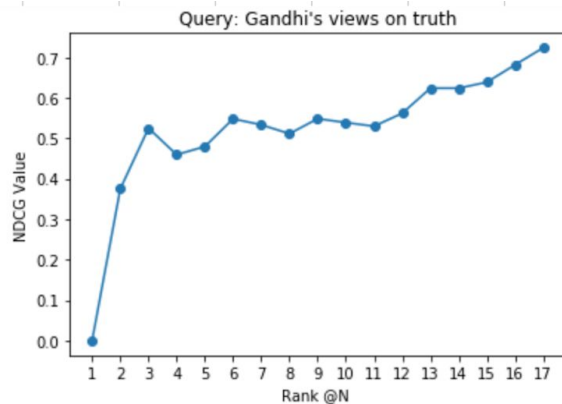
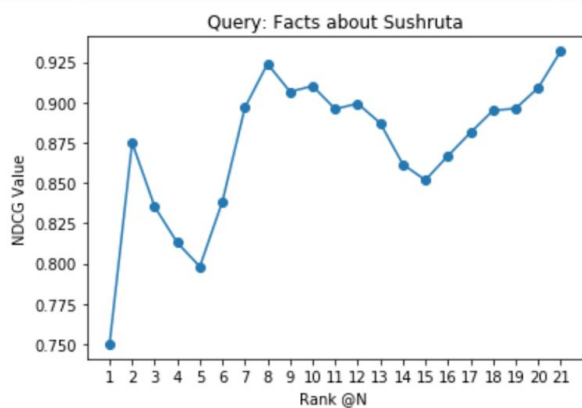
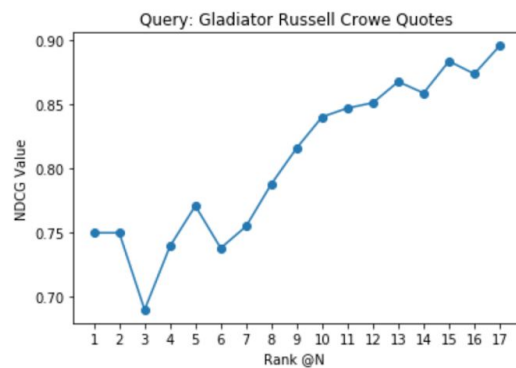
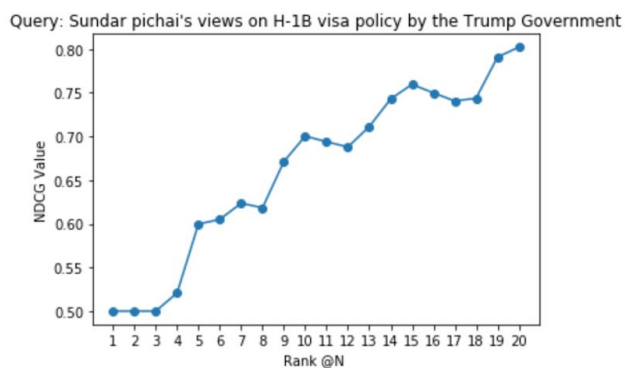
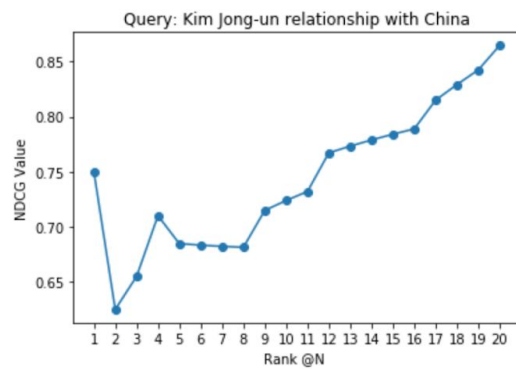
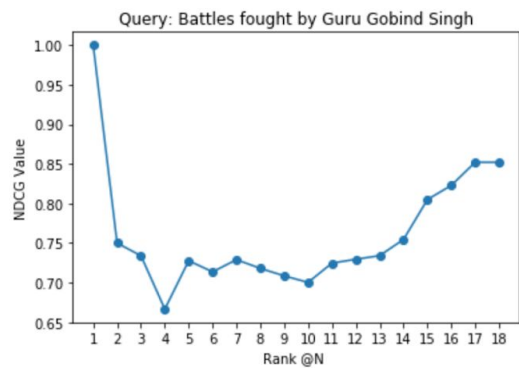
Baseline Results:

The results from the baseline were saved into separate csv files and can be viewed in the Github repository. On top of that, the results were also charted to see how well the baseline model performed on the given queries. Below are the results from those charts showing the NDCG value at the specific ranking. Note that the closer to 1 the value is the better the baseline model did at indexing the most relevant documents towards the top of the list.

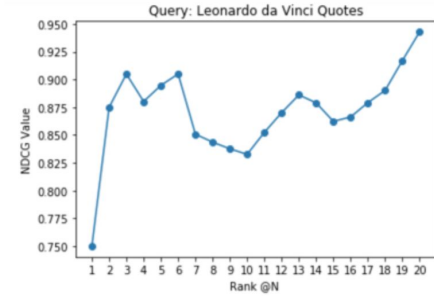
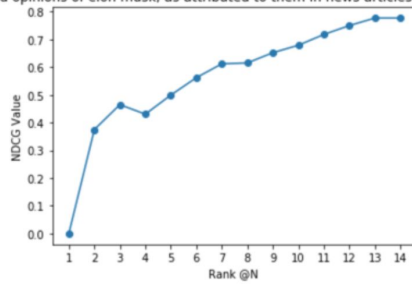






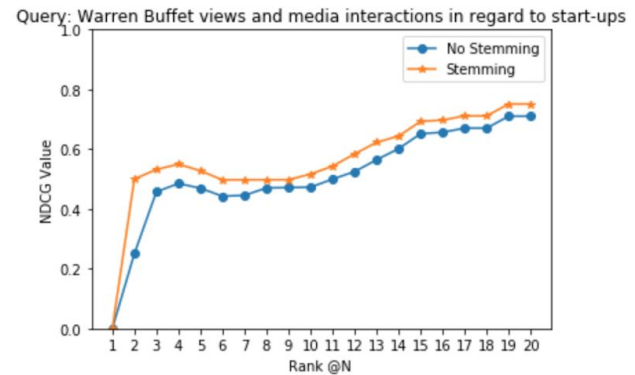
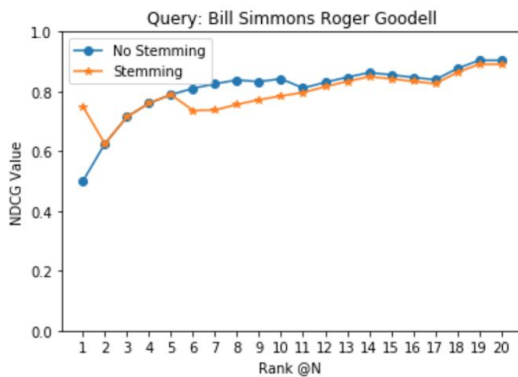
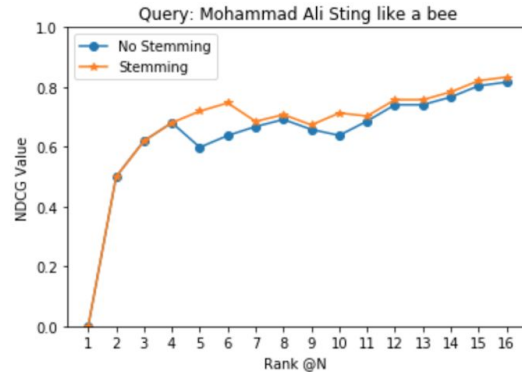
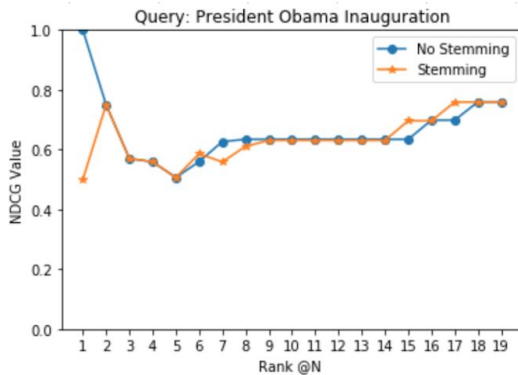


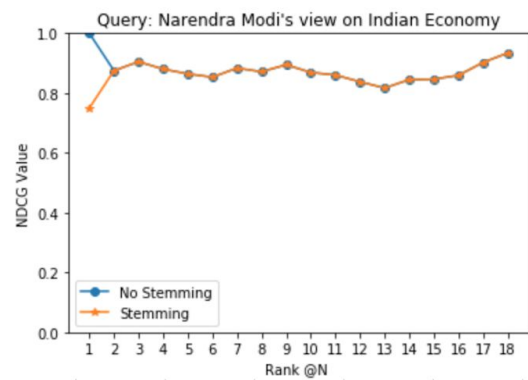
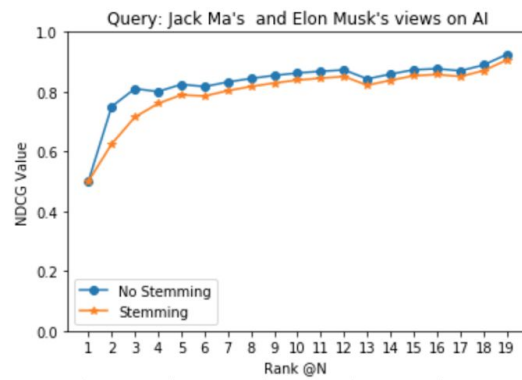
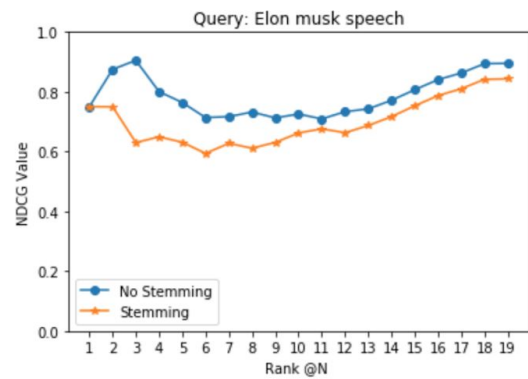
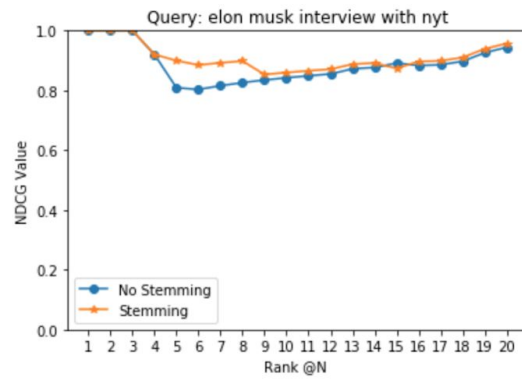
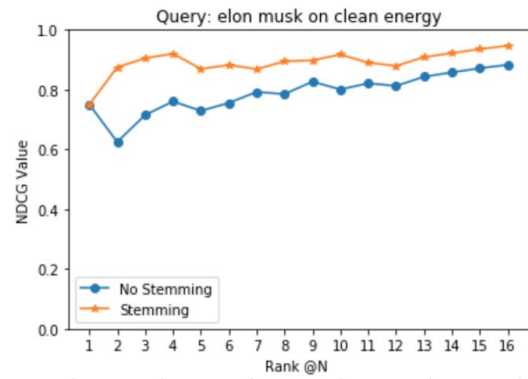
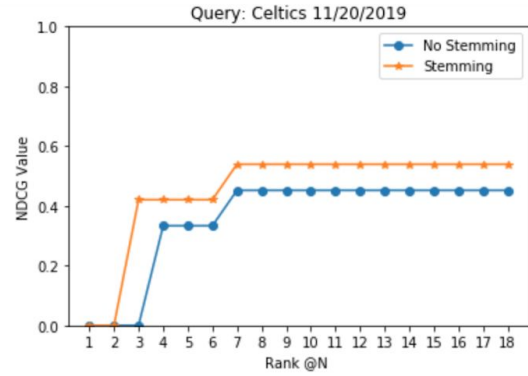
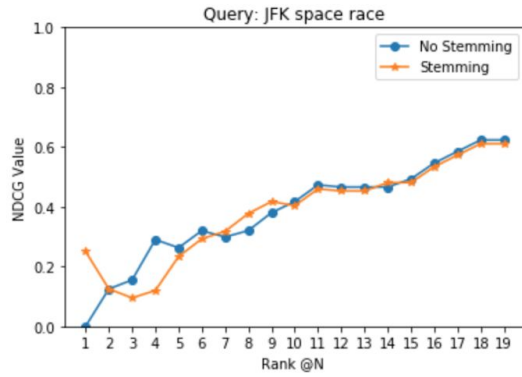
Query: quotes and opinions of elon musk, as attributed to them in news articles, social media, books, etc.

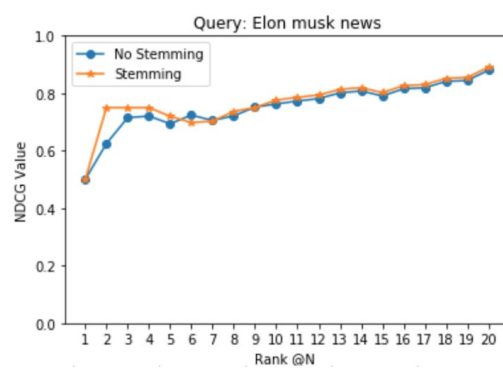
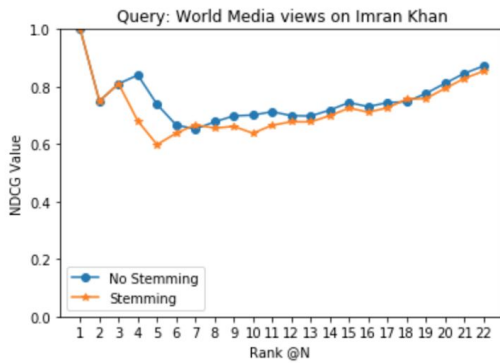
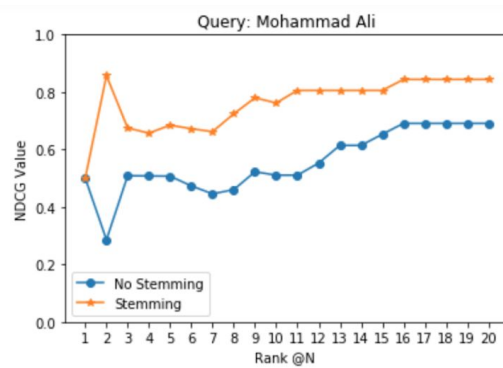
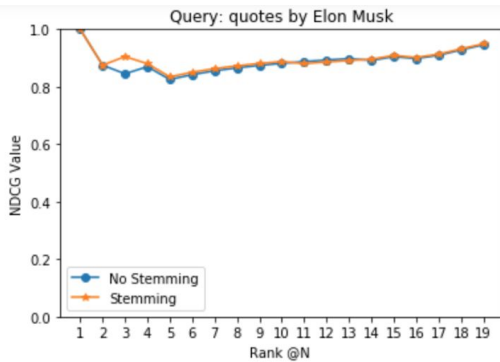
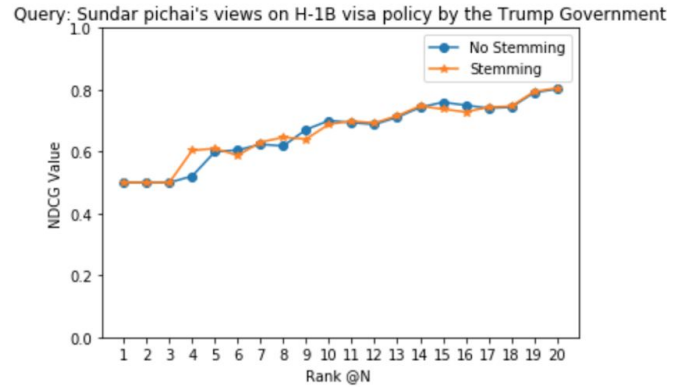
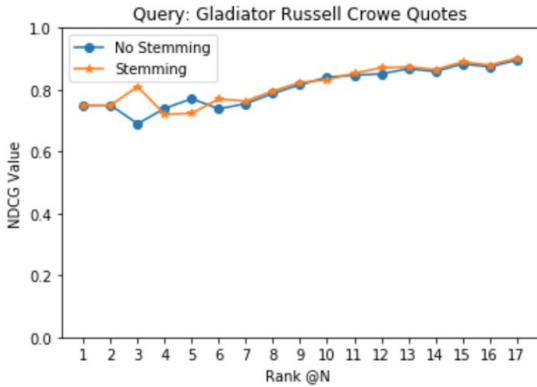
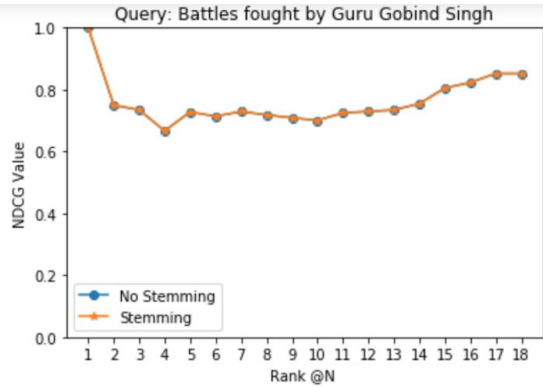
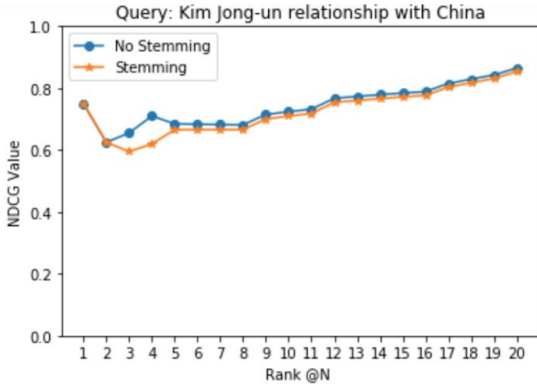


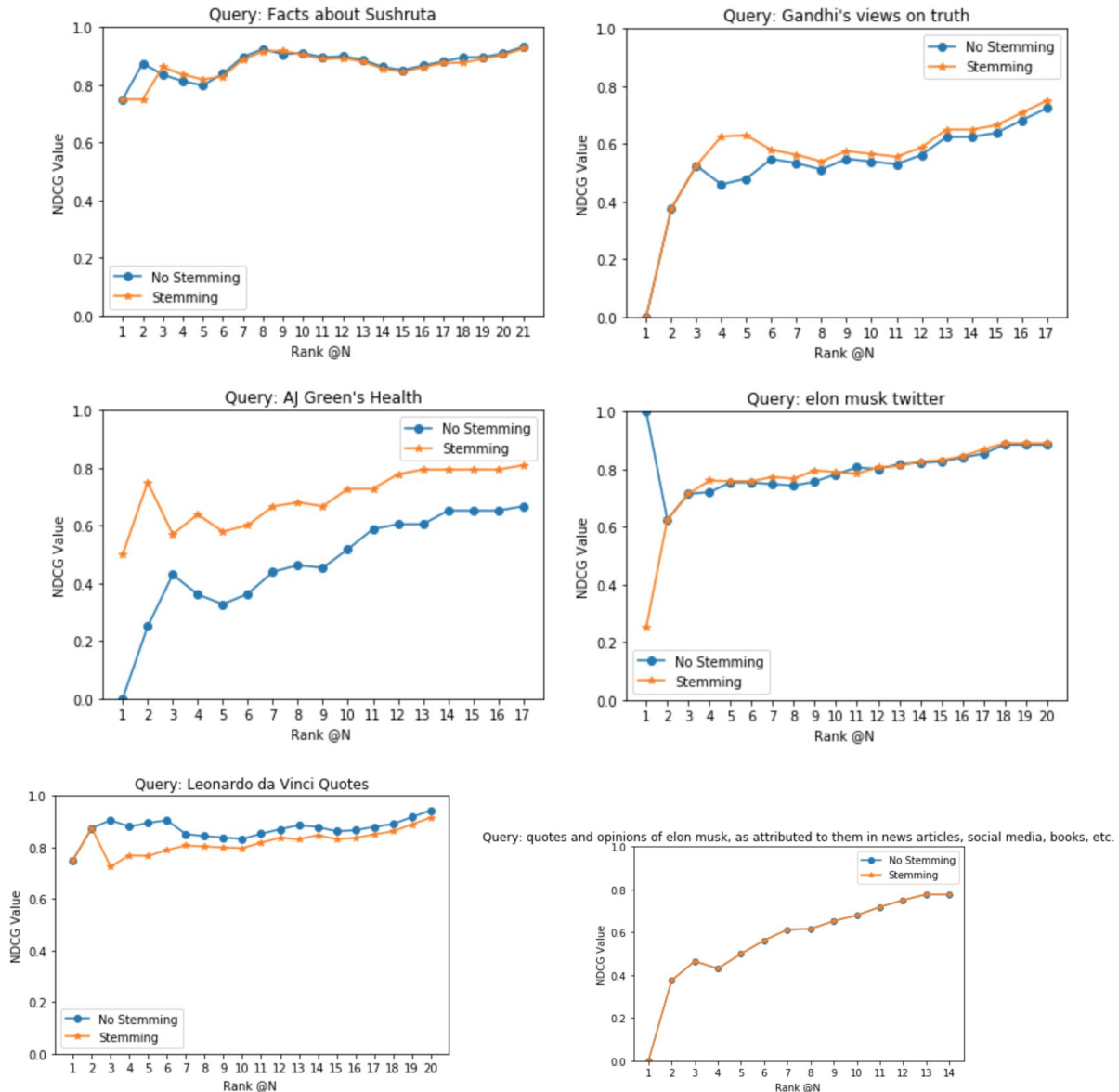
Stemming Changes:

Due to the poor results of the baseline model and the correlation to key query terms such as 'quote' and 'speech' performing better in the baseline model. Stemming was added to the system to see if it couldn't improve the retrieval system. The terms ['said', 'spoke', 'interview', 'speech', 'quote'] were added to all of the queries that were given to the baseline model. The results from adding in this stemming can be seen below. In almost all cases the addition of these words increased the likelihood that the retrieval model would give you the more relevant documents first









Conclusion

The results from this project showed that basic retrieval systems, in this case, query likelihood, do indeed struggle in retrieving such non-trivial queries like quotes. By looking at our resulting graphs and numbers it is clear that the baseline model was unsuccessful in giving the user relevant information in the correct rank. Queries that include such keywords such as 'speech', 'quote' and a few more did better at this task. So in an attempt to improve the model, these words along with a few other words were added to all of the queries. This model while overall better than the baseline model still could use some improvement.

We also observed that human performance for this task also retrieved less relevant documents. For this task, relevance information can improve search engine performance. However, this information is not usually available with the search engine. Using specific keywords like 'views', 'quotes' and 'opinion' also resulted in more relevant document retrieval.

Based on what we saw in our evaluations of current retrieval models when looking for quotes. Our team has come up with some suggestions to improve the retrieval models beyond just adding in stemming. The first suggestion would be to look at adding higher weights to the stemming words and characters.

Especially the characters, quotation marks, indentions, and italicized text are all indications that quotes are being made in the document. Another suggestion would be to look at the instances of pronouns in the document. Many of the more important documents that our team came across had dialog that started with the name of the person and then what they said. So if that could be captured and weighted higher in the retrieval then the model then it could increase the ability for the model to return relevant documents.

Lastly, our team only used the query likelihood algorithm as a retrieval model. So we as a team think that it might be prudent to look into other retrieval models such as BM25.