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 3o 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.

SAMPLE RETRIEVED DOCUMENTS

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.

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.

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

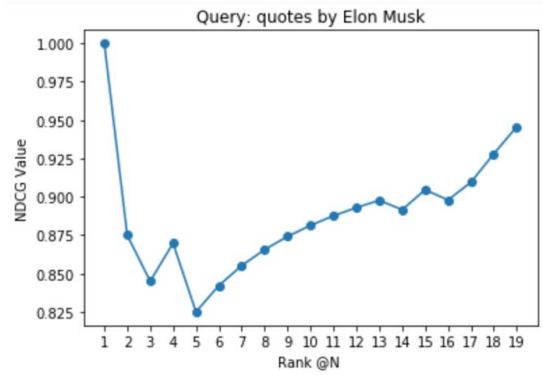
Baseline Model:

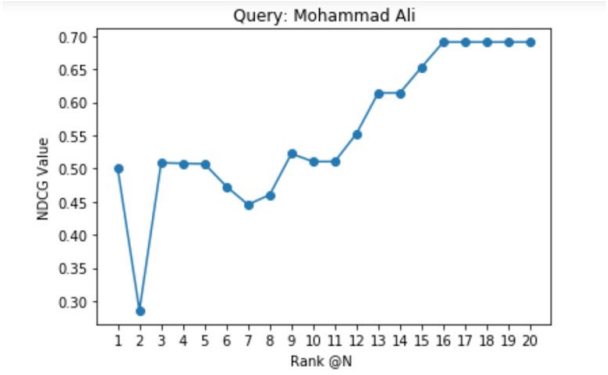
The Baseline model that the team used in this project was a simple bag of words approach that used a query likelihood algorithm to rank the documents. The bag of words for this baseline model was built out by only using the terms that appeared in the queries. The ranking query likelihood algorithm was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

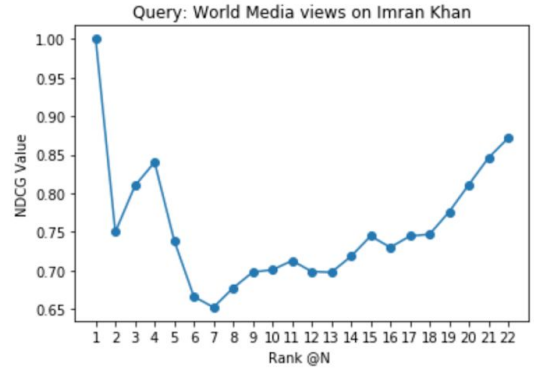
$$QL = \sum_{i \in q=0}^{q} log(\frac{f_{i,d}+1}{|D|+|V|})$$

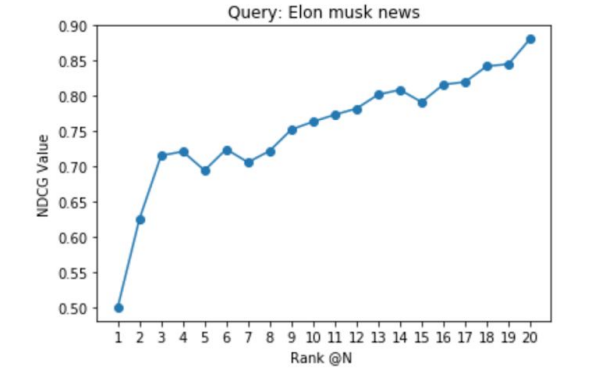
Baseline Results:

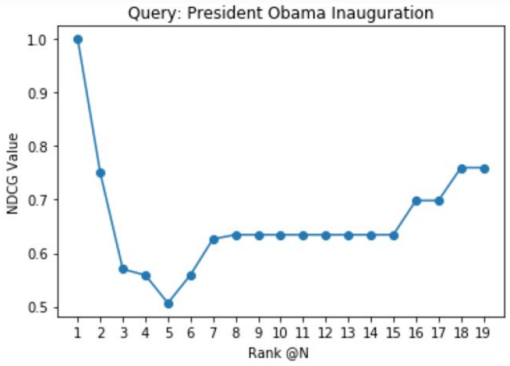
The results form the baseline were saved into separate csv files and can be viewed in the Github repository. On top of that, the results were also charted to see how well the baseline model performed on the given queries. Below are the results from those charts showing the NDCG value at the specific 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.

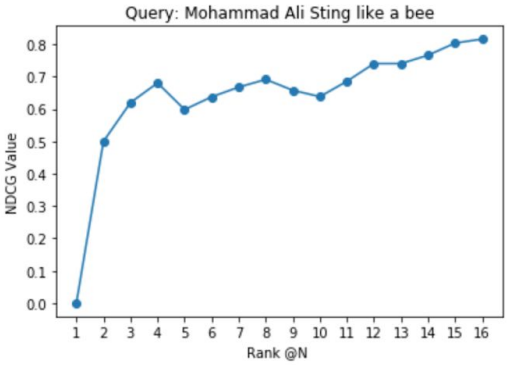


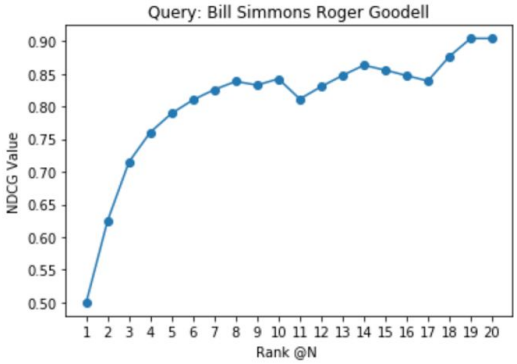


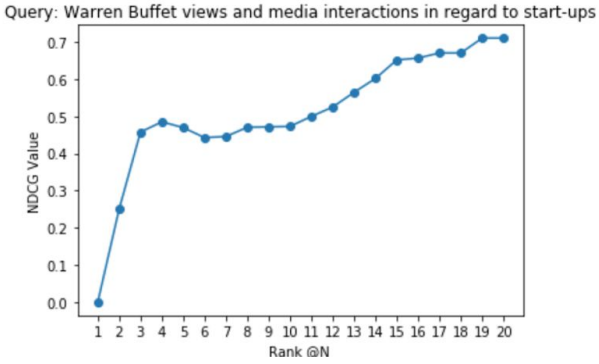


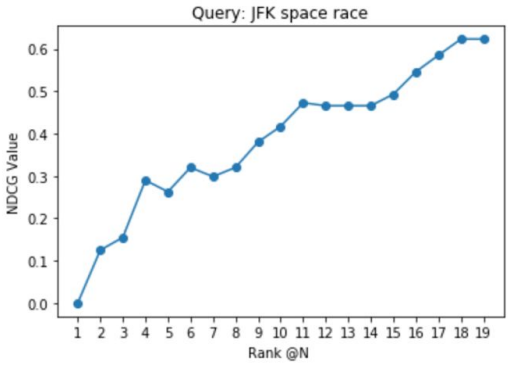


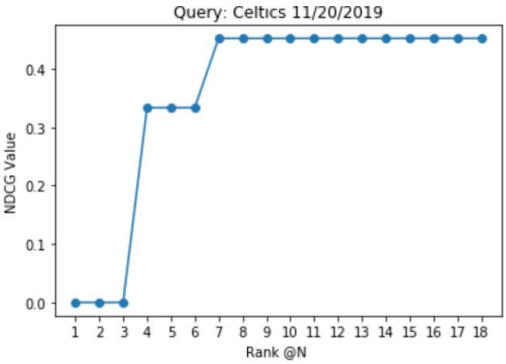

Query: President Obama Inauguration
NDCG Value
Rank @N

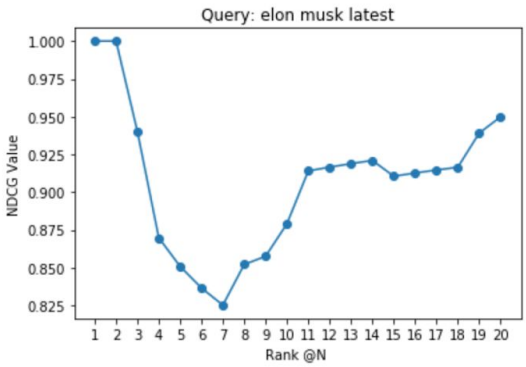

Query: Mohammad Ali Sting like a bee
NDCG Value
Rank @N

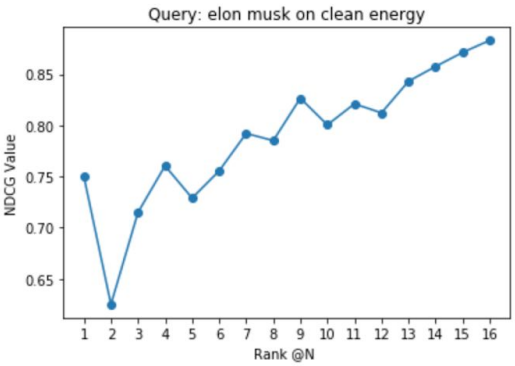

Query: Bill Simmons Roger Goodell
NDCG Value
Rank @N

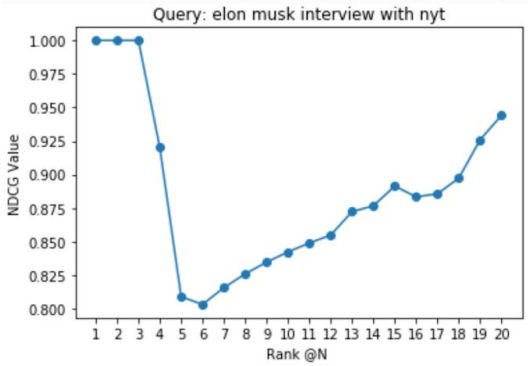

Query: Warren Buffet views and media interactions in regard to start-ups
NDCG Value
Rank @N

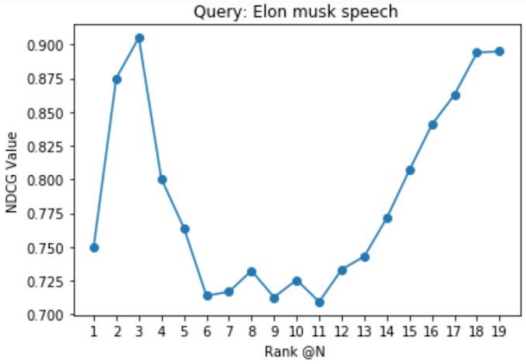

Query: JFK space race
NDCG Value
Rank @N

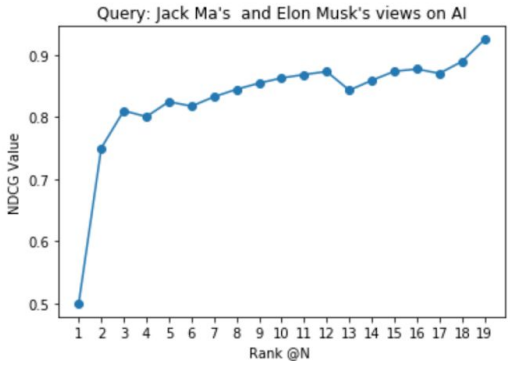

Query: Celtics 11/20/2019
NDCG Value
Rank @N

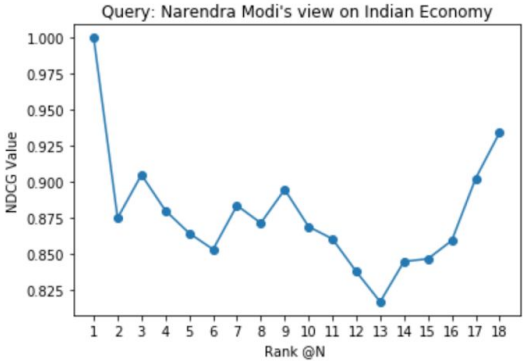

Query: elon musk latest
NDCG Value
Rank @N

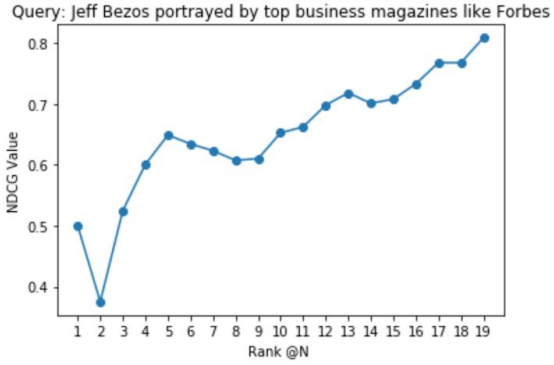

Query: elon musk on clean energy
NDCG Value
Rank @N

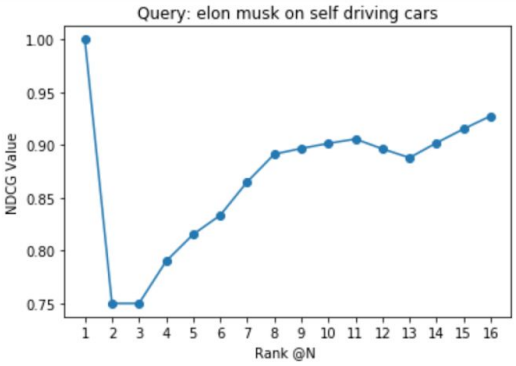

Query: elon musk interview with nyt
NDCG Value
Rank @N

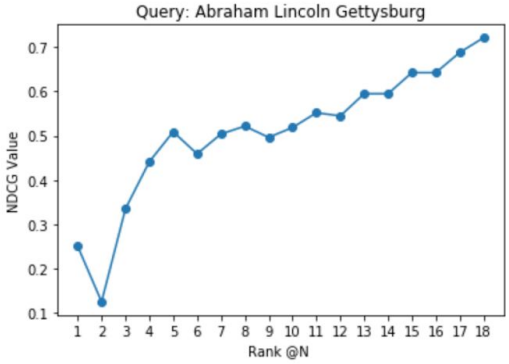

Query: Elon musk speech
NDCG Value
Rank @N

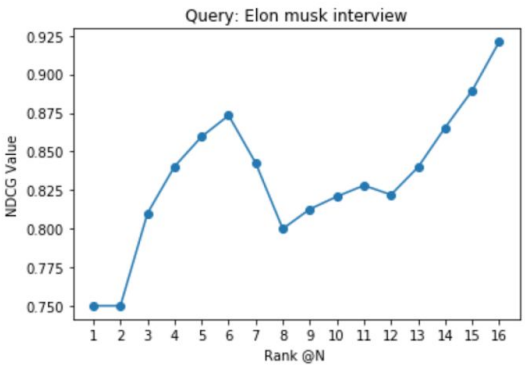

Query: Jack Ma's and Elon Musk's views on AI
NDCG Value
Rank @N

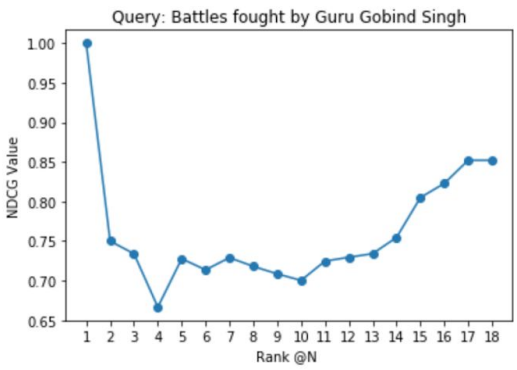

Query: Narendra Modi's view on Indian Economy
NDCG Value
Rank @N

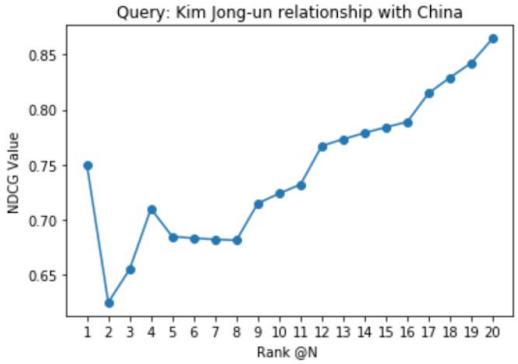

Query: Jeff Bezos portrayed by top business magazines like Forbes
NDCG Value
Rank @N

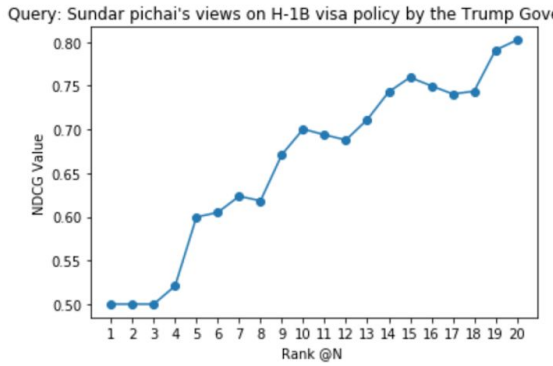

Query: elon musk on self driving cars
NDCG Value
Rank @N

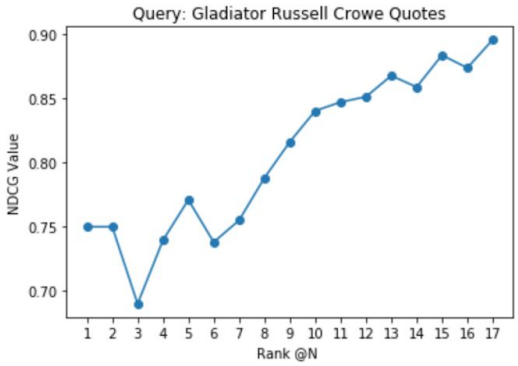

Query: Abraham Lincoln Gettysburg
NDCG Value
Rank @N

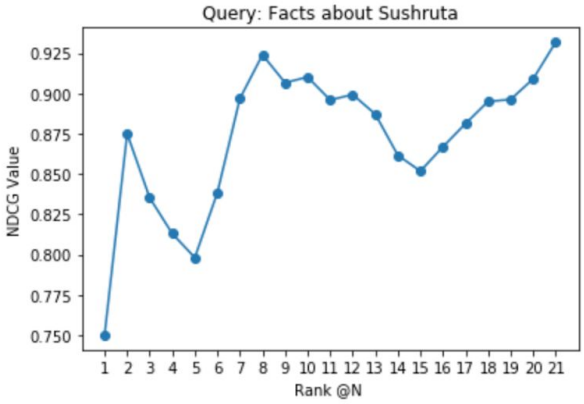

Query: Elon musk interview
NDCG Value
Rank @N

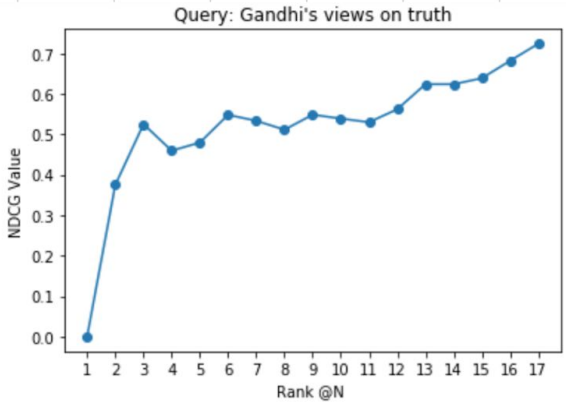

Query: Battles fought by Guru Gobind Singh
NDCG Value
Rank @N


Query: Kim Jong-un relationship with China
NDCG Value
Rank @N

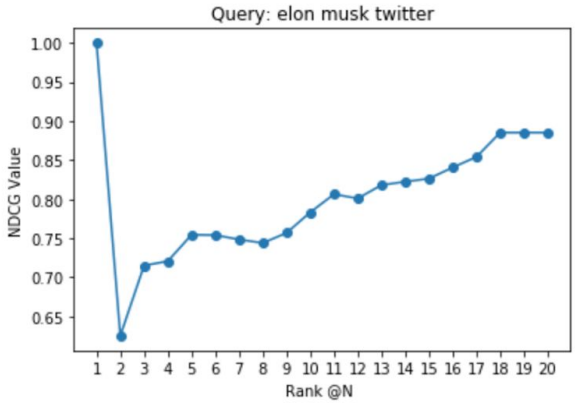

Query: Sundar pichai's views on H-1B visa policy by the Trump Government
NDCG Value
Rank @N

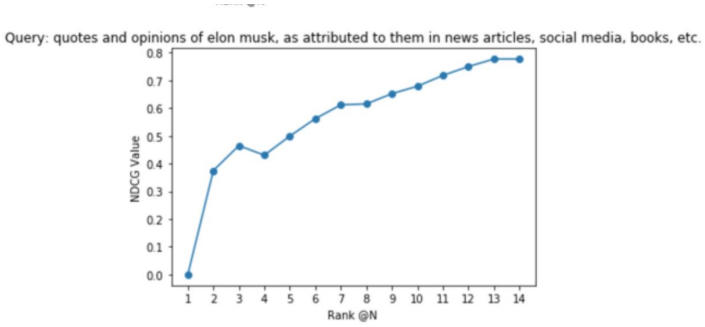

Query: Gladiator Russell Crowe Quotes
NDCG Value
Rank @N


Query: Facts about Sushruta
NDCG Value
Rank @N


Query: Gandhi's views on truth
NDCG Value
Rank @N

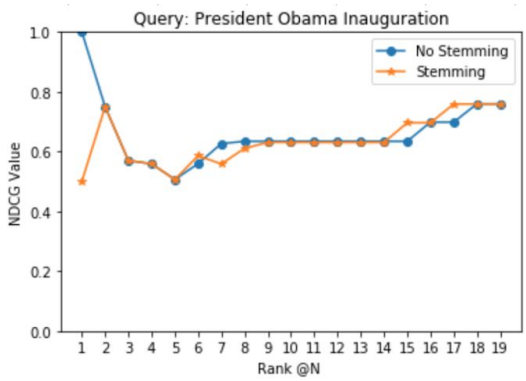

Query: AJ Green's Health
NDCG Value
Rank @N

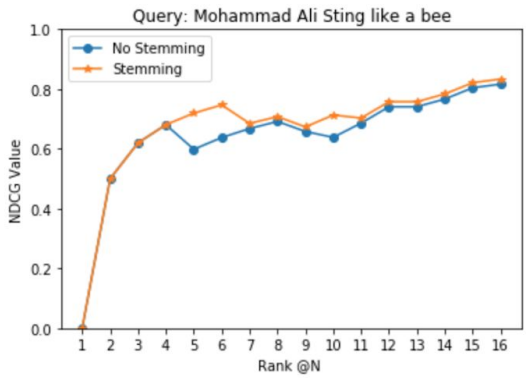

Query: elon musk twitter
NDCG Value
Rank @N

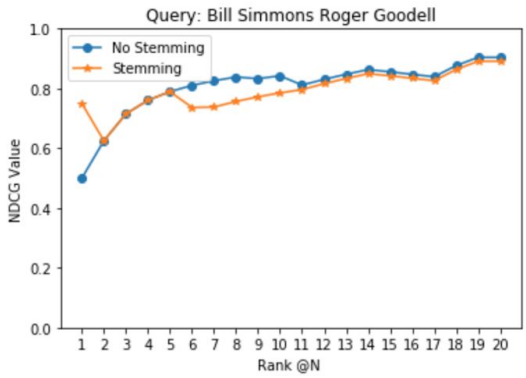


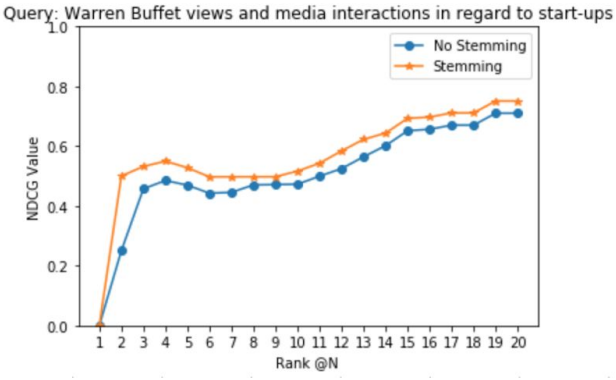
Stemming Changes:

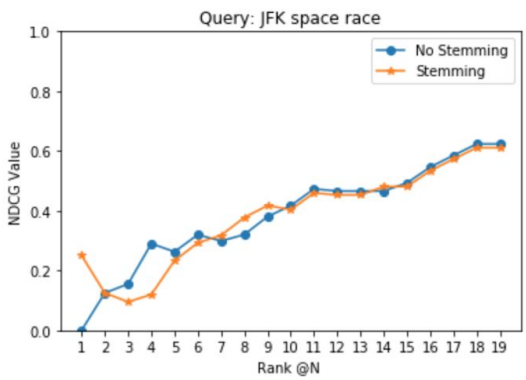
Due to the poor results of the baseline model and the correlation to key query terms such as ‘quote’ and ‘speech’ performing better in the baseline model. Stemming was added to the system to see if it couldn’t improve the retrieval system. The terms ['said', 'spoke', "", 'interview', 'speech', ‘quote’] 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

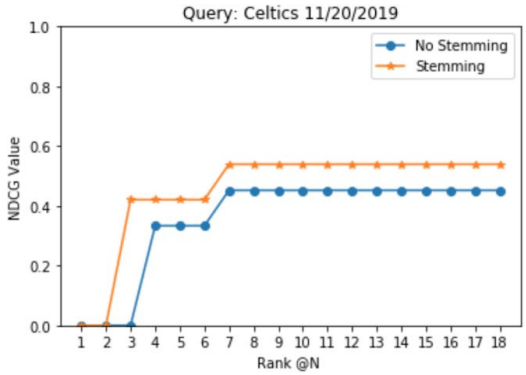


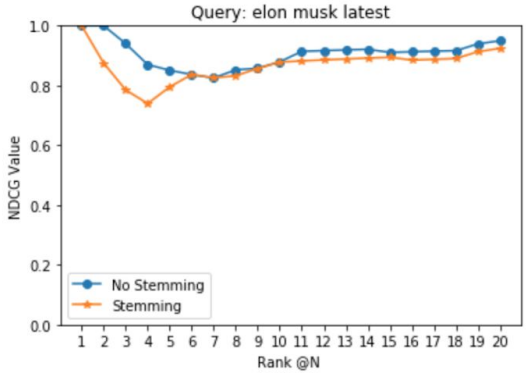


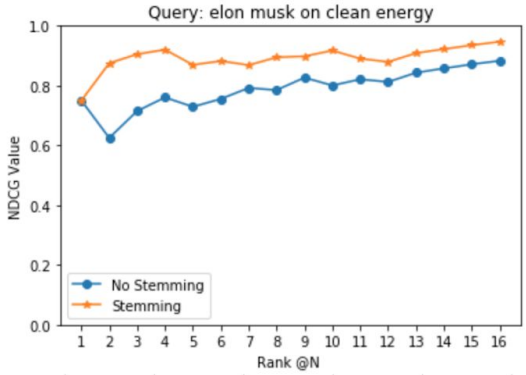


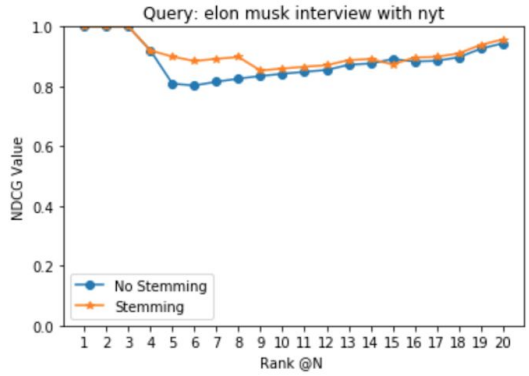


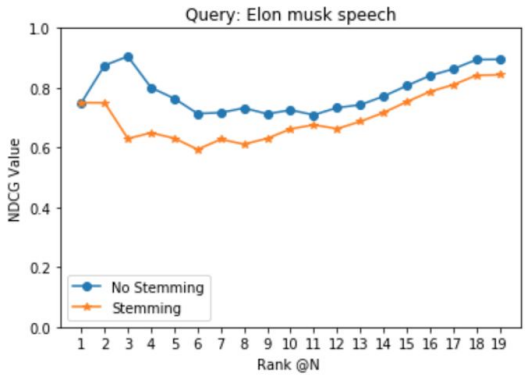

Query: JFK space race
No Stemming
Stemming
NDCG Value
Rank @N

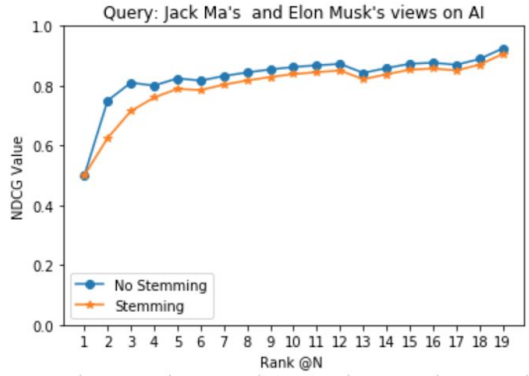

Query: Celtics 11/20/2019
No Stemming
Stemming
NDCG Value
Rank @N

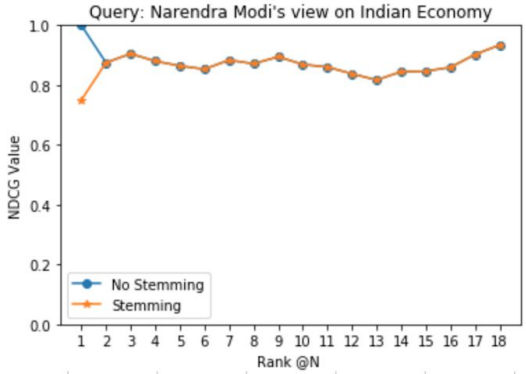

Query: elon musk latest
No Stemming
Stemming
NDCG Value
Rank @N

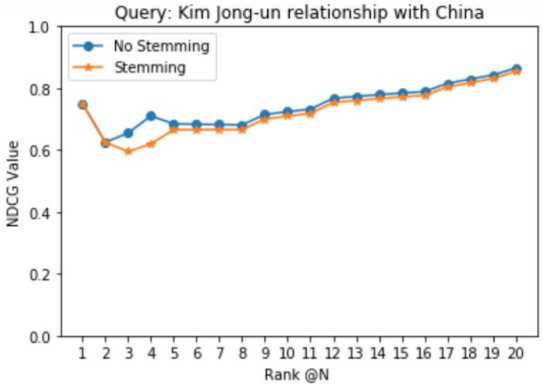

Query: elon musk on clean energy
No Stemming
Stemming
NDCG Value
Rank @N

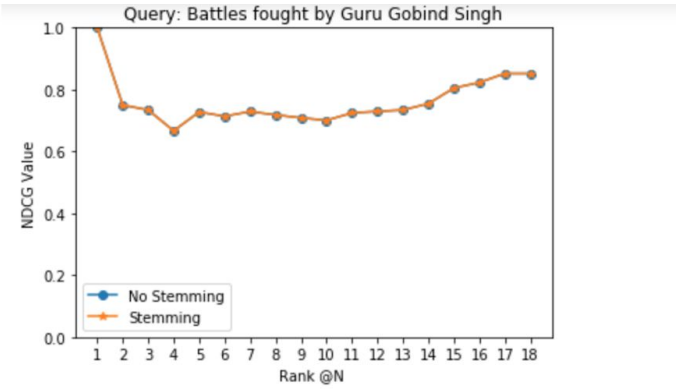

Query: elon musk interview with nyt
No Stemming
Stemming
NDCG Value
Rank @N

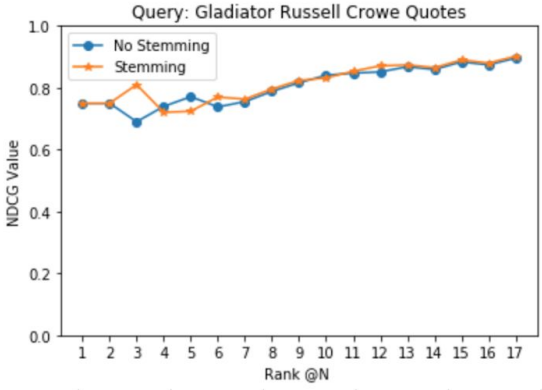

Query: Elon musk speech
No Stemming
Stemming
NDCG Value
Rank @N

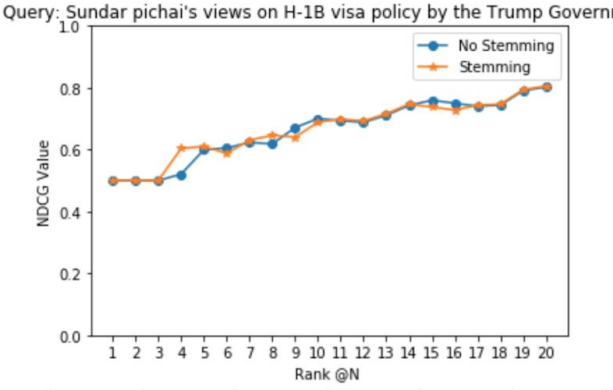

Query: Jack Ma's and Elon Musk's views on AI
No Stemming
Stemming
NDCG Value
Rank @N

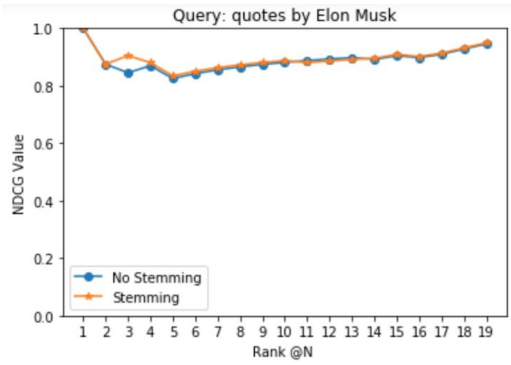

Query: Narendra Modi's view on Indian Economy
No Stemming
Stemming
NDCG Value
Rank @N

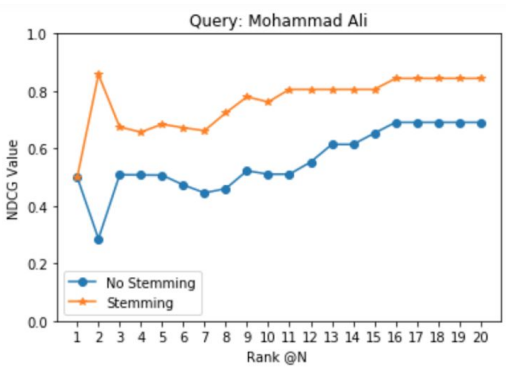

Query: Kim Jong-un relationship with China
No Stemming
Stemming
NDCG Value
Rank @N

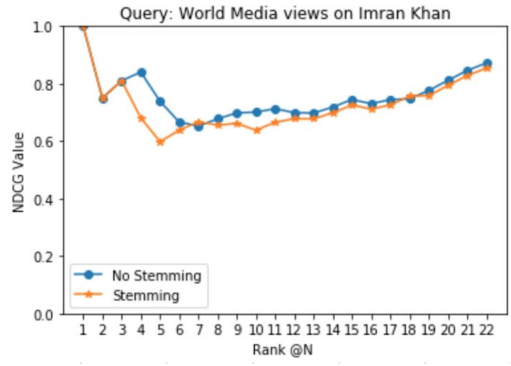

Query: Battles fought by Guru Gobind Singh
No Stemming
Stemming
NDCG Value
Rank @N

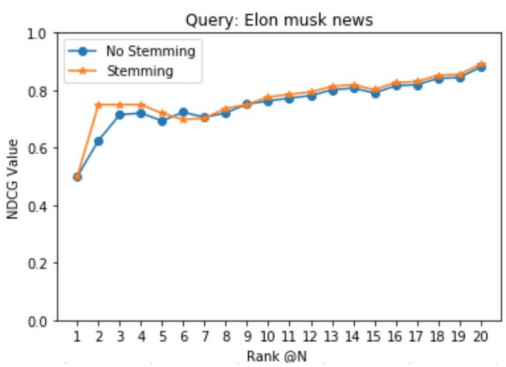

Query: Gladiator Russell Crowe Quotes
No Stemming
Stemming
NDCG Value
Rank @N

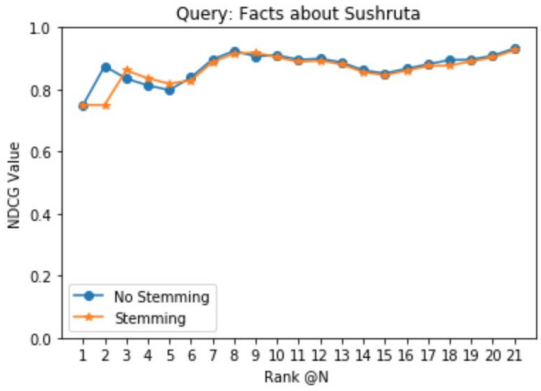

Query: Sundar pichai's views on H-1B visa policy by the Trump Government
No Stemming
Stemming
NDCG Value
Rank @N

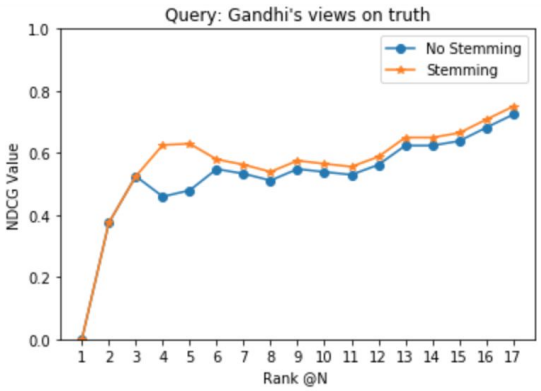

Query: quotes by Elon Musk
No Stemming
Stemming
NDCG Value
Rank @N

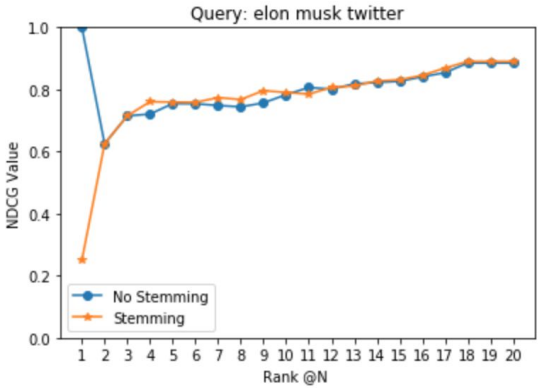

Query: Mohammad Ali
No Stemming
Stemming
NDCG Value
Rank @N

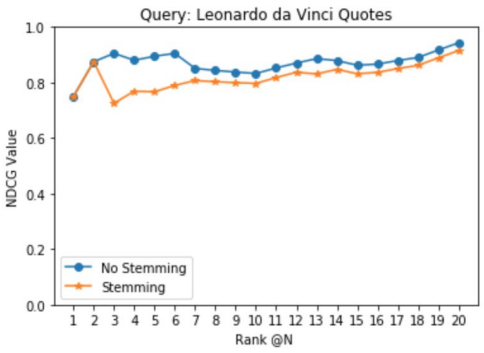

Query: World Media views on Imran Khan
No Stemming
Stemming
NDCG Value
Rank @N

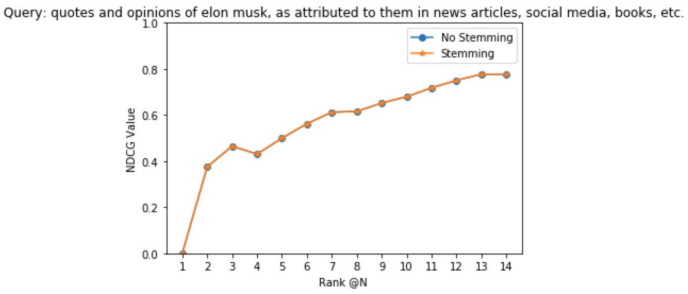

Query: Elon musk news
No Stemming
Stemming
NDCG Value
Rank @N











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. Querries 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.

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 looking 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 retrival model. So we as a team think that it might be prudent to look into other retrieval models such as BM25.