Report

Github repository: https://github.com/wenjiejin01/FinalProject_IR.git

Section 1 - Data collection

The data collection consists 114614 tweets about the topic Covid-19, with total Retweets: 79095, unique Tweets: 35519 and Unique Users: 87930.

The keywords = {"coronavirus", "covid", "#COVID", "#coronavirus", "#COVID19", "recovery cases", "suspected cases", "death cases", "vaccines", "disinfection"}

It takes approximately 40 minutes to collect 100000 tweets from Twitter. In my case, I'm using the subsample with 4600 tweets to do the search engine.

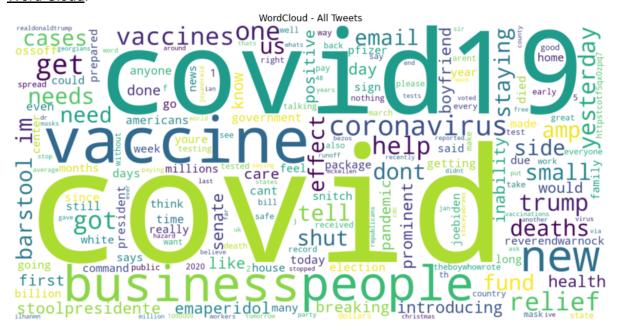
Section 2 - Search Engine

Pre-processing:

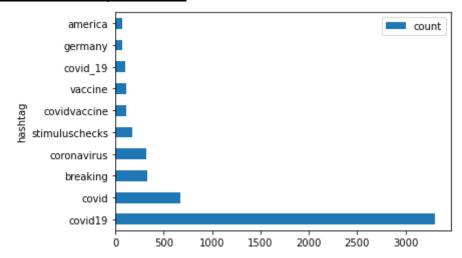
To build my search engine, the first thing I do is take only original tweets, that means we don't interest retweets. After that, I need to mention a new function implemented by myself, the Update function. Update is used to update the number of likes and retweets, because I'm scraping data with the function given by class, the "MyStreamListener" class. This function scrapes the status of the tweets that the user just sent, that is, when the number of likes or retweets is 0.

Finally, we exclude tokens from this data through the getTerms function and build the inverted index and compute tf, df and idf from the tokens. Need 43.68 seconds to build the index.

Word Cloud:

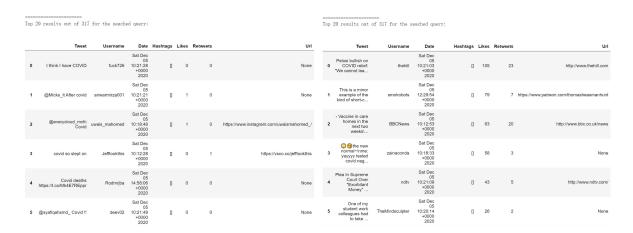


Bar plot of the 10 most frequent words:



My score:

My score is based on the idea G(d) + tf-idf. Where G(d) = numbers of likes + number of retweets. Because, In my opinion, likes and retweets is the most important factor to rank the results. The result of query "covid" is:



To show the difference between my score and TF-idf. Left hands we have searched with tf-idf ranking function, and Right is ranking with links and retweets. Both results have query terms which we search, but we are more interested in the right-hand result because they are ordered by numbers of likes and retweets.

Section 3 - RQs

• Subsection RQ1:

- List of 10 selected queries:
 - 1. covid
 - 2. covid19 america
 - 3. vaccine covid19
 - 4. breaking covid vaccine news
 - 5. death cases america
 - 6. suspected cases america
 - 7. recovery cases
 - 8. covid symptoms
 - 9. coronavirus death rate barcelona
 - 10. Wuhan covid19