Congressional Tweets Analysis Project

Outline

- Project background & proposal
- Data Information & ERD
- Exploratory Data Analysis (EDA)
- Results
- Conclusion & Recommandation
- Appendix

Project Background & Proposal

Project background & proposal

- Which client/dataset did you select and why?
  Client 2: Lobbyists4America (Congressional Tweets Dataset -- 2008-2017 data)

- This project is about analysing congressional tweets data from the USA to determine how tweets, politicians could affect the congress most, so as to understand legislative trends and develop better lobbying strategies.

- Audiences are companies like Lobbyist4America that try to help their customers who want to affect legislation by using effective lobbying strategy.

- This dataset contains congressional tweets data and user data from 2008-2017, and we can analyze this dataset to understand how different features of tweets (such as the number of words per tweet) affect the favourite count of the corresponding tweet and whether it is a key feature that affects the number of followers of the corresponding politician.

Project Background & Proposal

Approach

- I plan to look at following features at first: text, retweeted_count, favourites_count, followers_count, created_at.
- I want to explore whether the positiveness of a tweet affects the popularity of the corresponding tweet, in order to check this relationship, I plan to use NLTK library to analyse the sentiment class of a tweet.
- Followers_count of a politician and favourite_count of a tweet are metric that I plan to use.
- I plan to calculate the word_count and created hour of tweets as new metric columns.
- I will to calculate TF-IDF for text and apply LDA modelling for topic modelling.

Project Background & Proposal

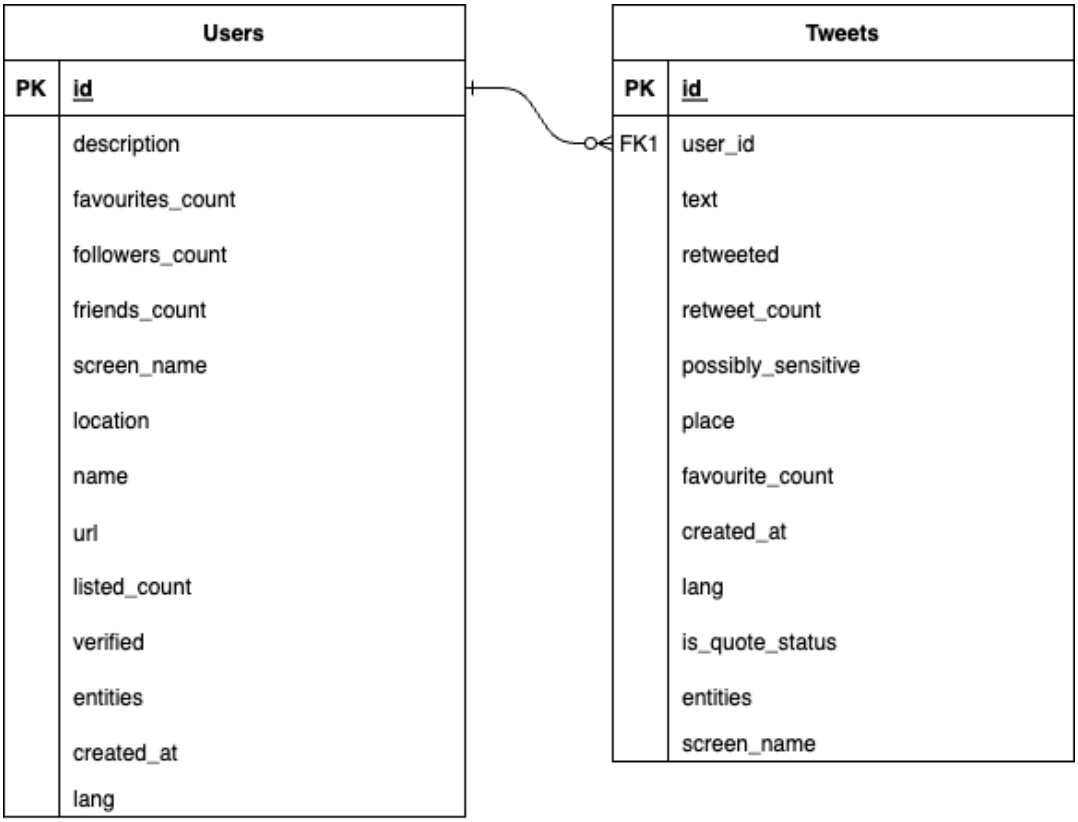
Initial questions & Hypothesis

- Initial questions
  - What is the average word length of a popular tweet?
  - Whether a positive tweet would be more popular?
  - What is the average tweet count of a popular politician?
- Hypothesis
  - The politician who has posted more tweets overall have more followers.
  - The politician who has more friends on twitter have more followers.
  - The longer political tweets (more words) are more popular.
  - Positive tweets get more favourites.
  - Retweeted count is positively correlated with favourite count of tweets

Data Information & ERD

- There are 2 tables in this dataset: Tweets, Users
- We have dropped some columns from two tables that not so relevant to this project.

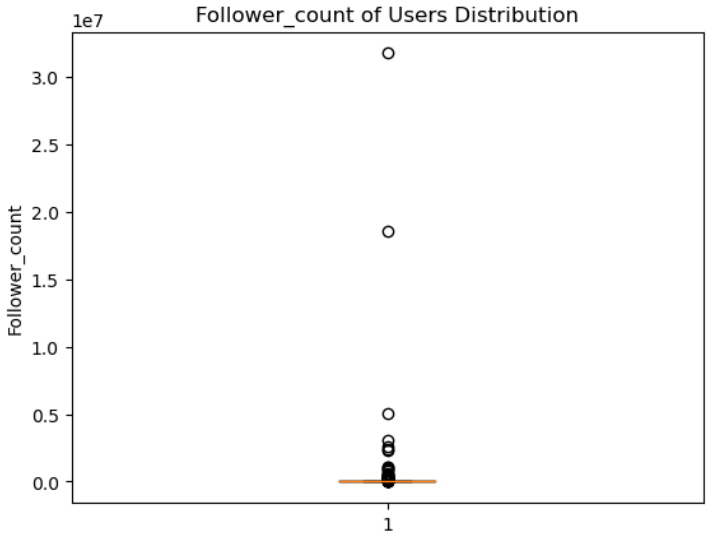
ERD of the congressional tweets dataset

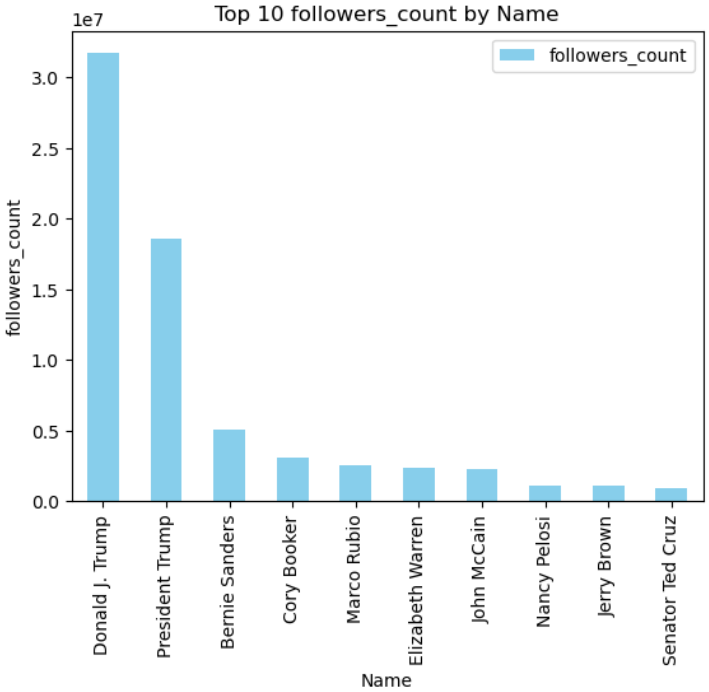


Exploratory Data Analysis (EDA)

Followers count of users

- The mean of Follower_count of users is very low, only a few outliers have much more followers.
- We have below listed 10 users with most followers.

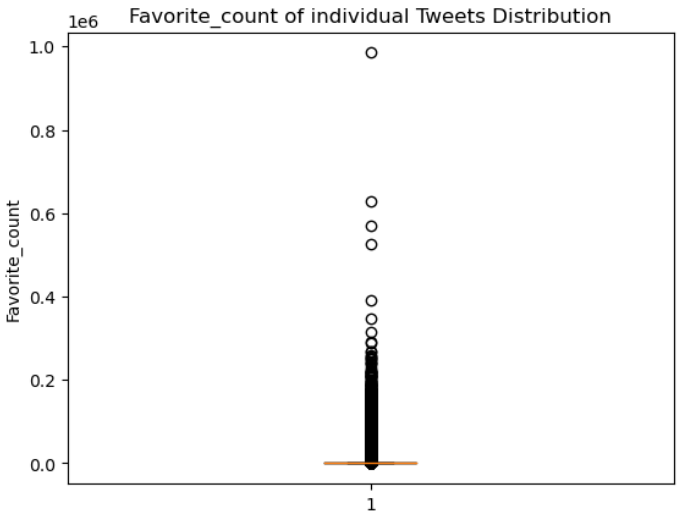


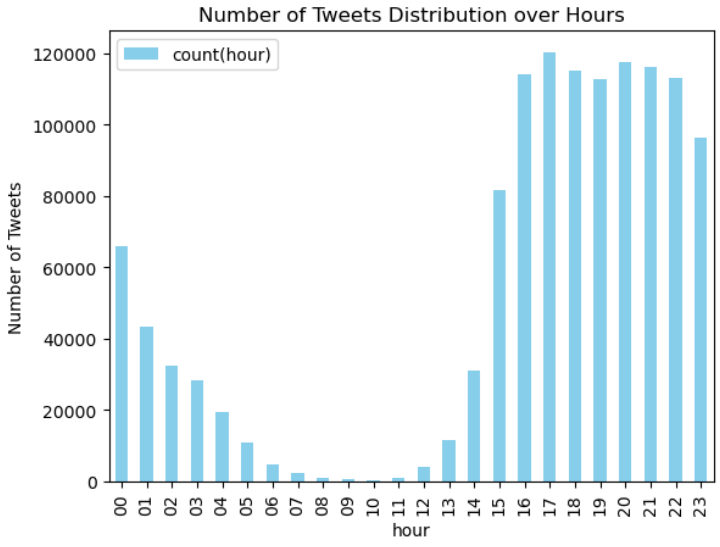


Exploratory Data Analysis (EDA)

Favourite count of tweets

- The mean of Favourite_count of tweets is very low, only a few outliers have much more Favourites.
- Most users post their tweet between 15:00 -24:00

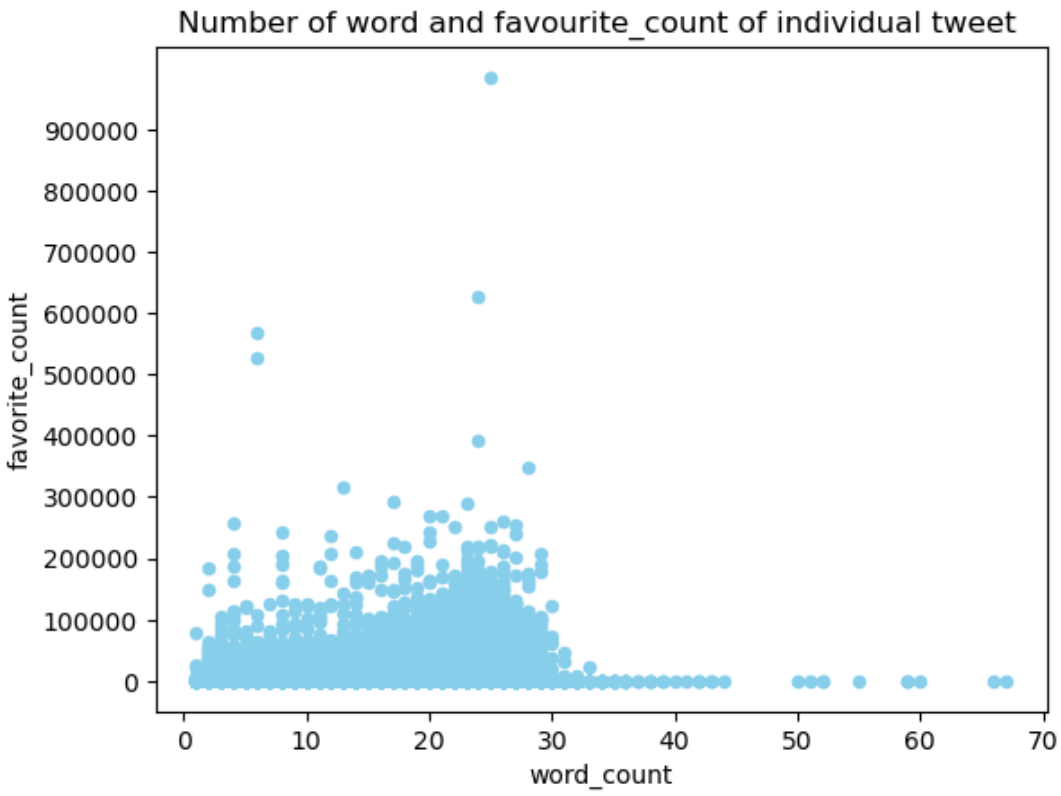




Results

Relationship between word count and favourite count of single tweet

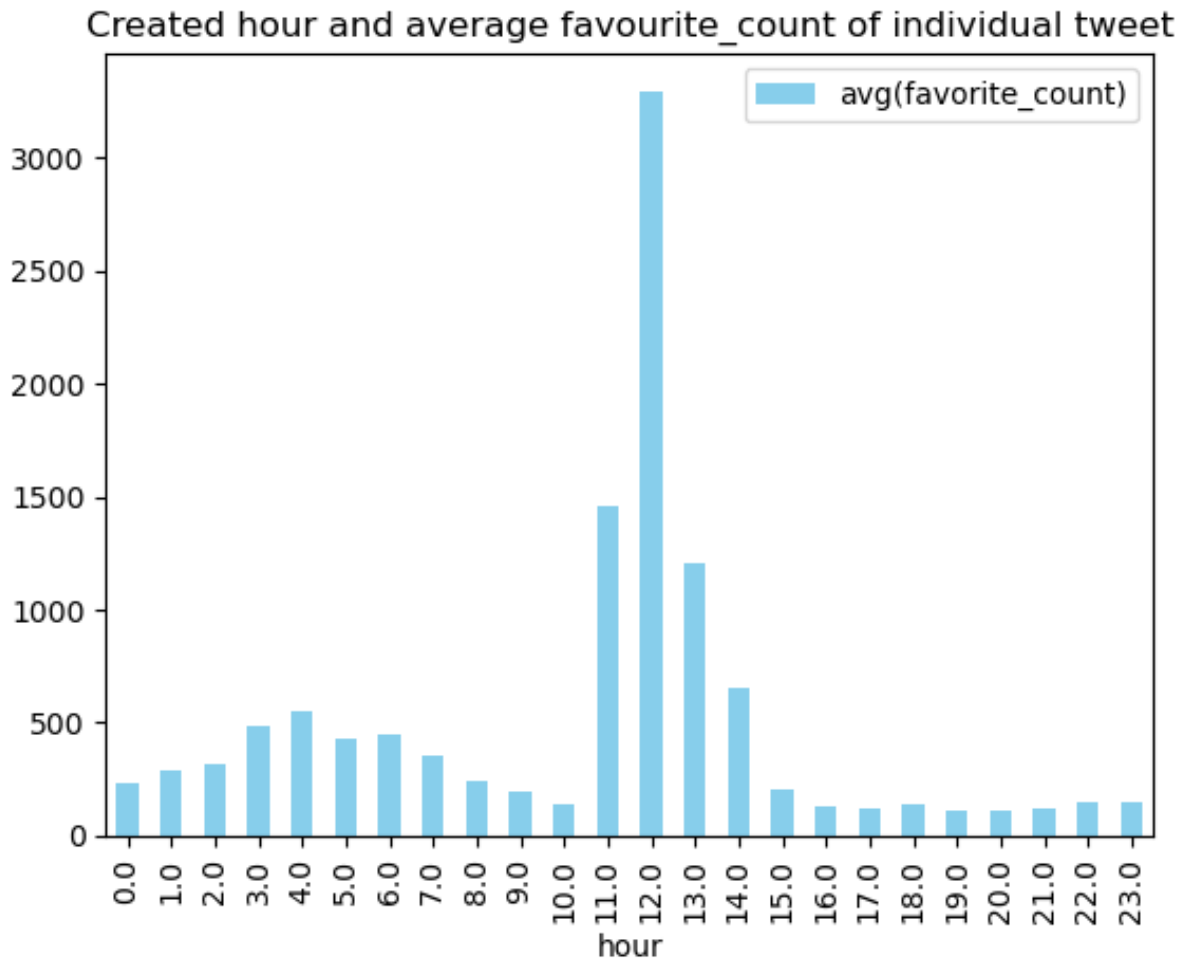
- Hypothesis : The longer political tweets (more words) are more popular.
- We have calculated the number of words of each tweet text and used it as a new metric word_count
- Here we used favourite count of a tweet to indicate the popularity of the tweet, from the scatter plot below, we can see that more words doesn't bring to a more favourite count.
- So our initial hypothesis is wrong, instead, the tweet should not be longer than 30 words.



Results

Relationship between created hour and average favourite count

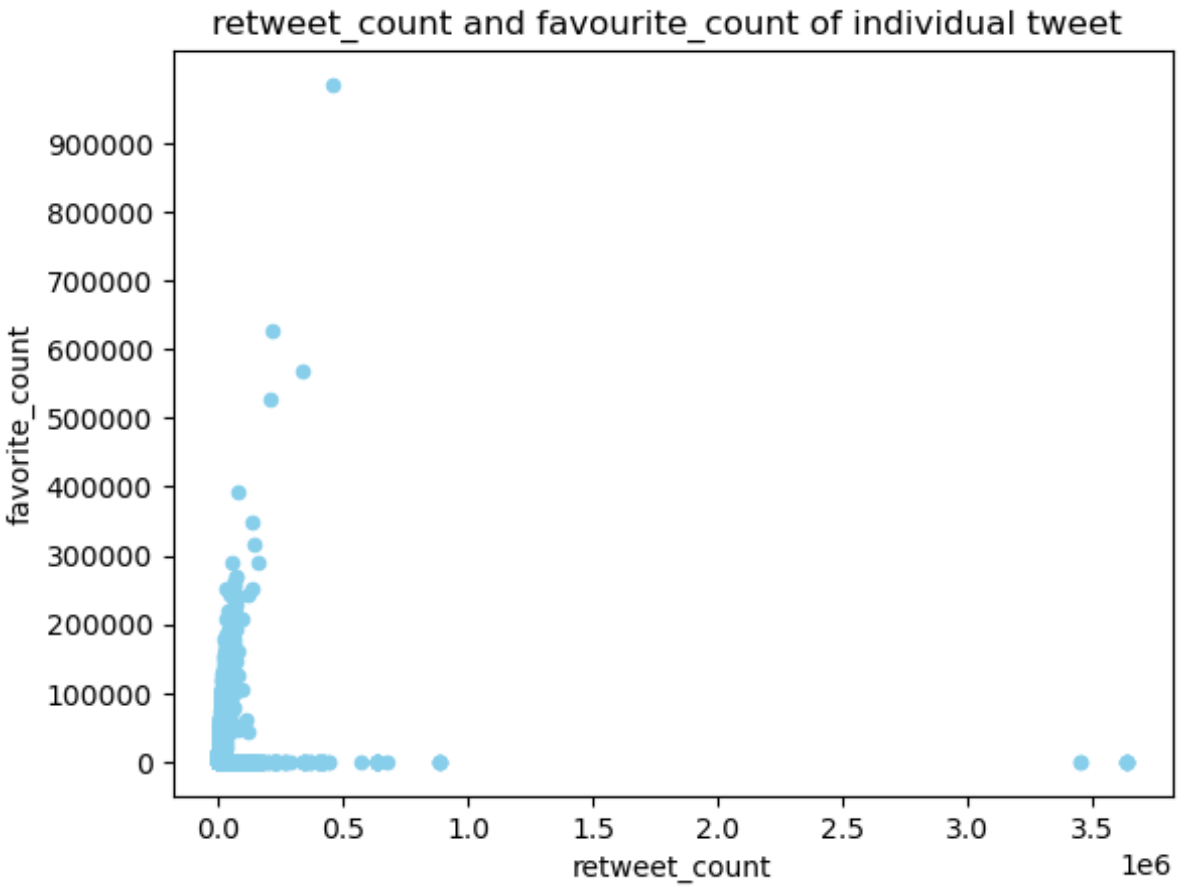
- We have created a hour metric based on the created column
- From the bar plot below, we can observe that the tweets posted between 11:00-13:00 have obvious higher favourite counts.



Results

Relationship between favourite_count and retweeted_count

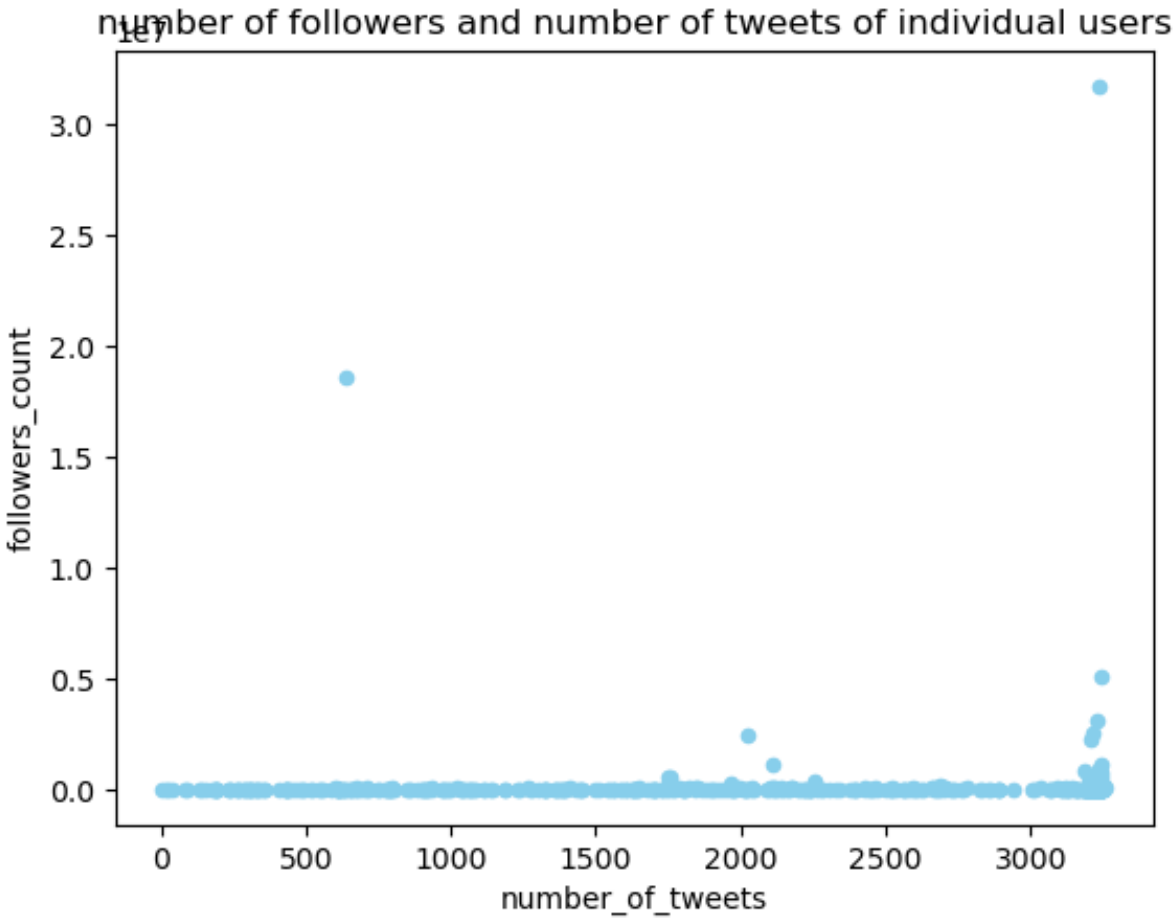
- Hypothesis: Retweeted count is positively correlated with favourite count of tweets.
- From the scatter plot below, we can learn that tweets with a very high retweeted count (more than 0.5*10^6) got a low favourite count.
- So our initial hypothesis is disproven, retweeted count is not positively correlated with favourite count of tweets.



Results

Relationship between tweet_count and follower_count

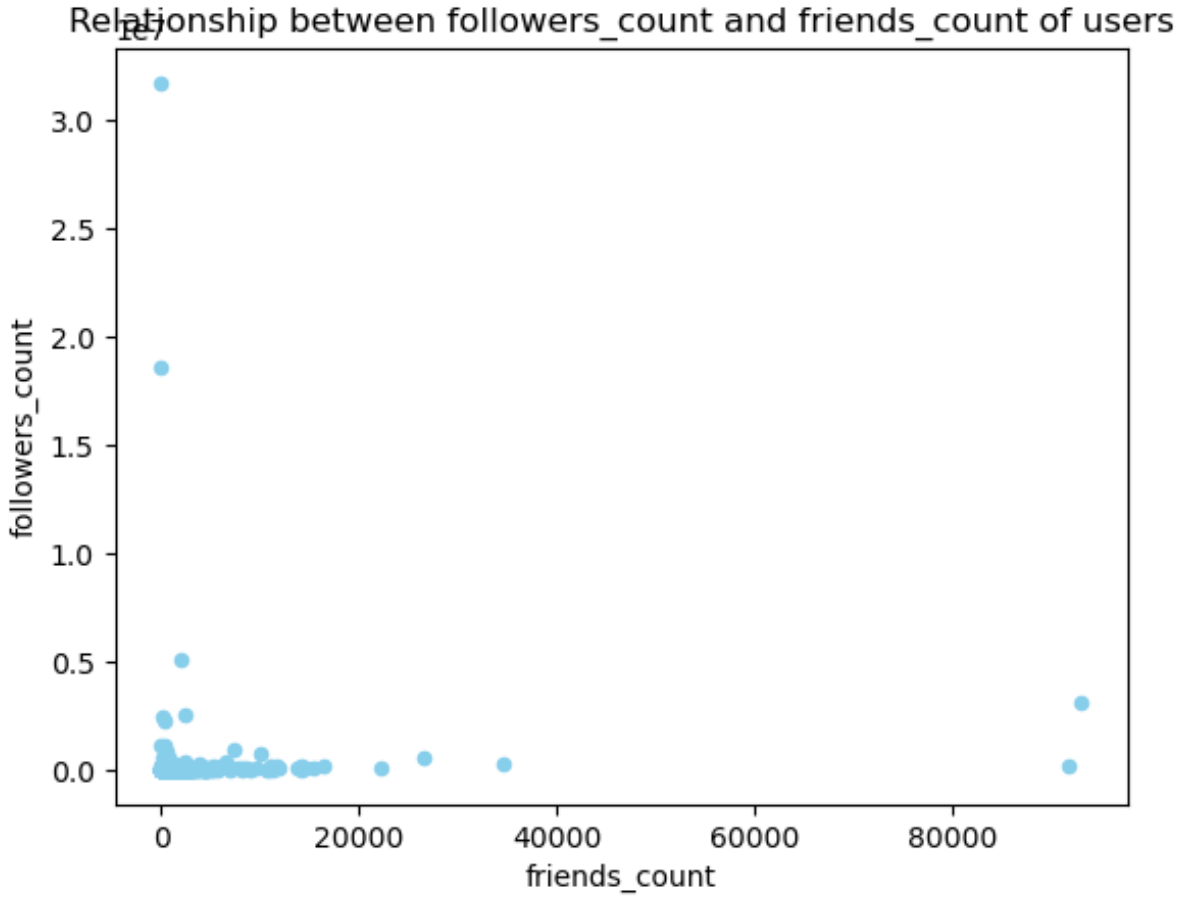
- Hypothesis: The politician who has posted more tweets overall have more followers
- From the scatter plot below, we can see that generally, the followers count of politicians is not related to the number of tweets from the politician.
- However, for several most popular politicians, they have posted a lot of tweets.
- So the initial hypothesis is not true, but in the meanwhile, if a politician want to be the most popular, he should try to post a large number of tweet.



Results

Relationship between followers_count and friends_count

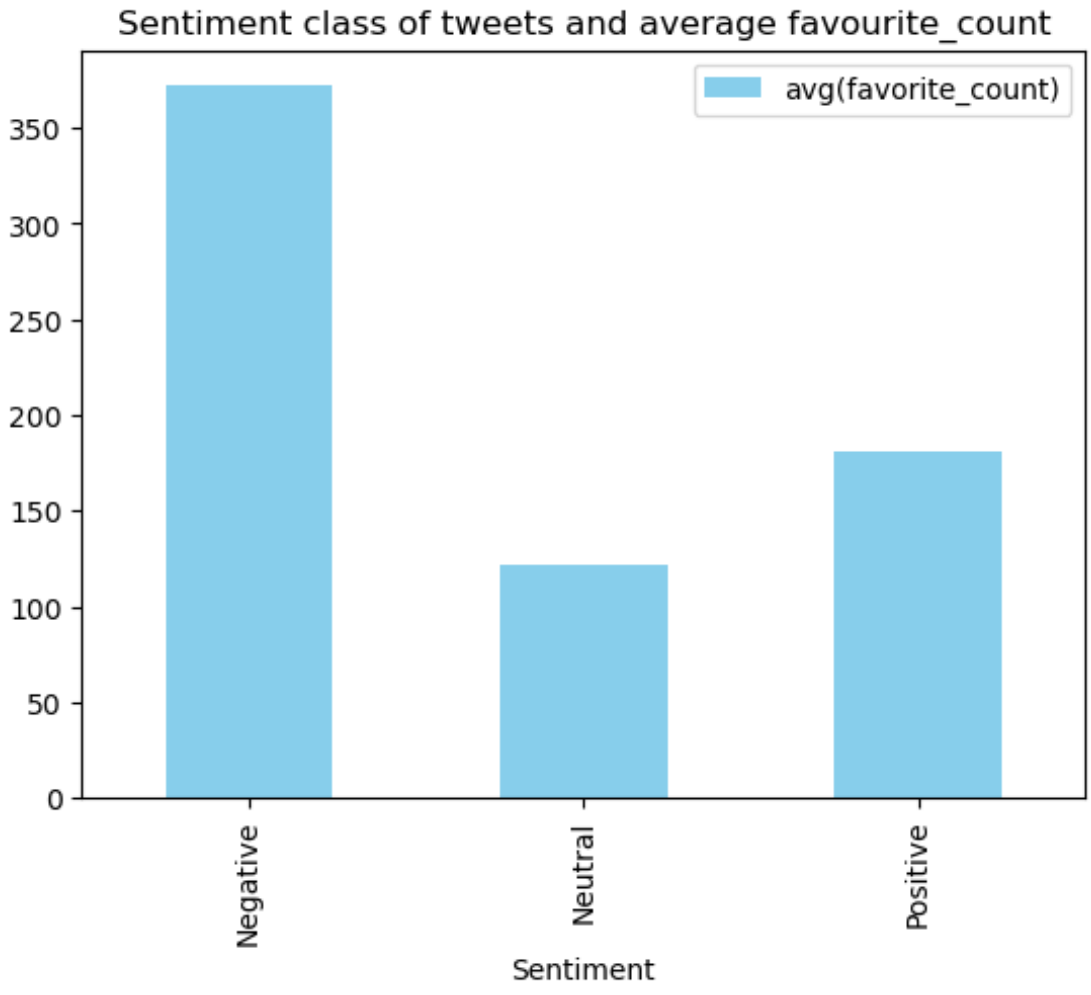
- Hypothesis: The politician who has more friends on twitter have more followers.
- There's no clear relationship between the follower_count and friend_count
- So our initial hypothesis is wrong.



Results

Relationship between sentiment class and average favourite_count

- Hypothesis: Positive tweets get more favourites.
- We have calculated a sentiment score for each tweet text and assign them to a sentiment class to a new sentiment class metric column.
- Based on the chart below, negative tweets have a clear higher favourite_count than other two classes, and positive tweets are still more popular than neutral tweets.
- So our initial hypothesis is wrong.

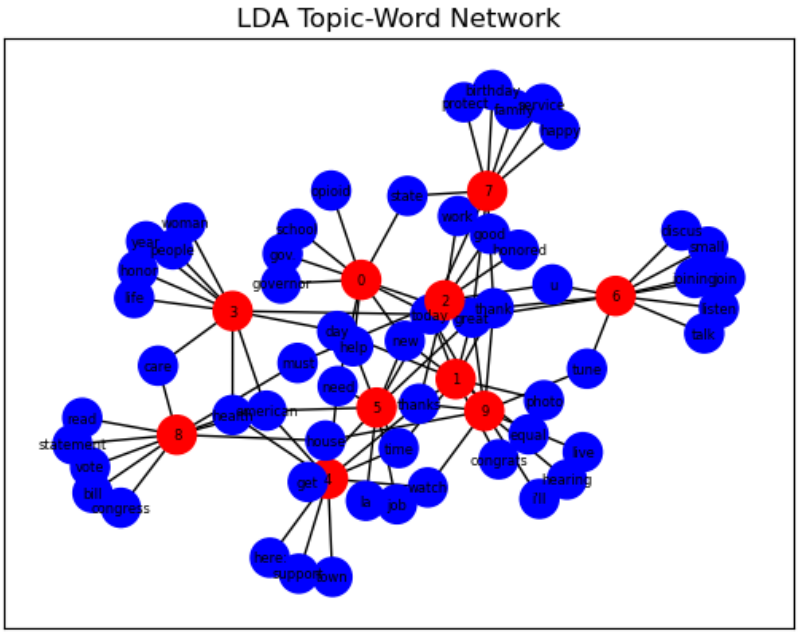


Results

TF-IDF calculation and LDA topic modelling

- Hypothesis: Positive tweets get more favourites.
- We have created a TF-IDF column based on tweet text and used LDA model to do topic modelling
- We have modelled in total 10 topic as list below and their graph plot.

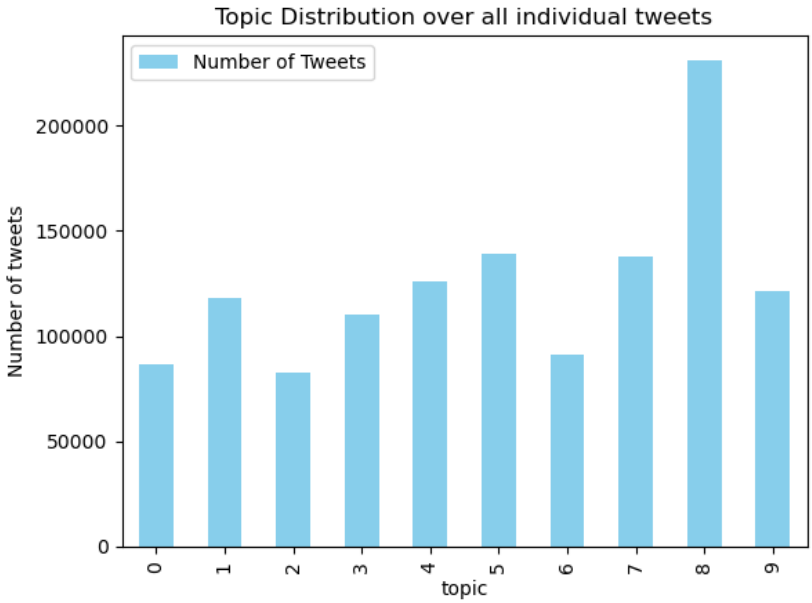
```
The topics described by their top-weighted terms:
+-----+-----------------------------------------------------------------------------+
|topic|topicWords                                                                   |
+-----+-----------------------------------------------------------------------------+
|0    |[great, new, governor, opioid, state, need, today, help, gov., school]       |
|1    |[great, new, day, thank, today, congrats, photo, equal, time, thanks]        |
|2    |[great, today, thank, u, thanks, honored, work, new, good, must]             |
|3    |[american, woman, today, day, people, health, year, life, honor, care]       |
|4    |[need, american, health, town, thanks, support, house, here:, watch, time]   |
|5    |[great, help, need, today, time, american, job, new, la, get]                |
|6    |[tune, u, great, join, discus, listen, talk, small, joining, today]          |
|7    |[happy, great, birthday, thank, today, work, family, state, protect, service]|
|8    |[house, bill, health, vote, american, must, care, statement, congress, read] |
|9    |[tune, watch, today, great, live, thank, thanks, house, hearing, i'll]       |
+-----+-----------------------------------------------------------------------------+
```

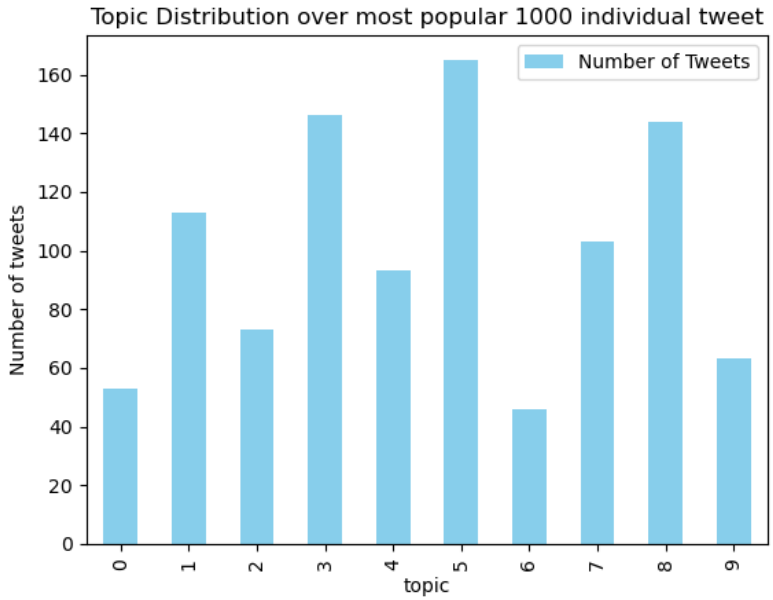


Results

Topic distribution of all tweets and 1000 most popular tweets

- Generally, the topic 8 is the most popular topic over all tweets.
- But for the most popular 1000 tweets, the topic 5 is the most popular and than follows topic 3 and 8

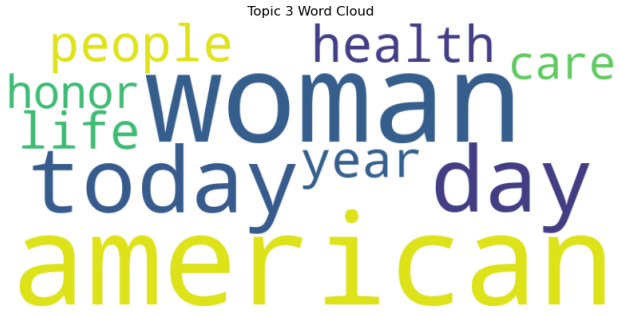


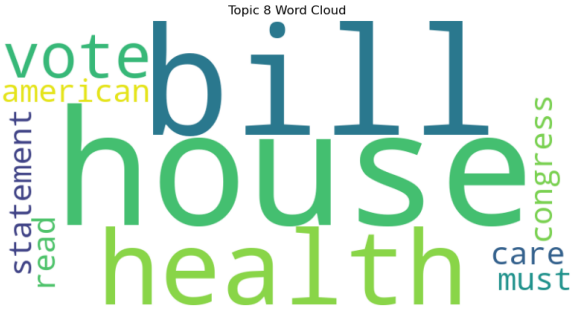


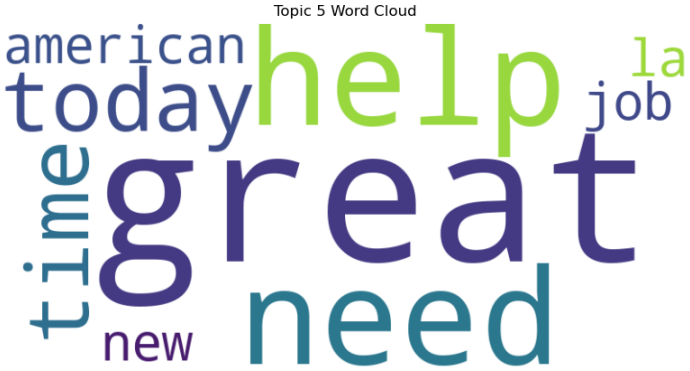
Results

Wordcloud of popular topics

- To have a better insight of popular topics we have found, we have provided some word cloud graph for these topics below.







Conclusion & Recommandation

Conclusion

- Top tweets should not exceed 30 words
- Tweets posted between 11:00-13:00 have significantly higher favourites
- Retweets are not related to favourites
- The number of tweets by a politician is not related to the number of followers they have
- There is no clear relationship between the number of followers and the number of friends of a politician
- Negative tweets get more favourites on average
- The most popular topic of all tweets is the topic 8: house, bill, health, vote, american, must, care, statement, congress, read
- The most popular topic of 1000 most popular tweets is the topic 5: great, help, need, today, time, american, job, new, la, get

Recommandation

- For companies that need an effective lobbying strategy, they need to focus on politicians who always tweet between 11:00-13:00, and their tweets should be short and clear (no more than 30 words)
- They also need to do more investigations to focus on politicians who like to post negative sentiment tweets, or at least positive sentiment tweets.
- They should focus on politicians who are interested in topics such as jobs, America, voting, houses, bills, etc.

Appendix

- All relevant assets like origin project proposal, Python code notebook, SQL queries, charts, Notebook outputs:
- https://github.com/schickwu/sql_final