# **Milestone 1: Project Proposal and Data Selection/preparation**

## Step 1: Preparing for Your Proposal

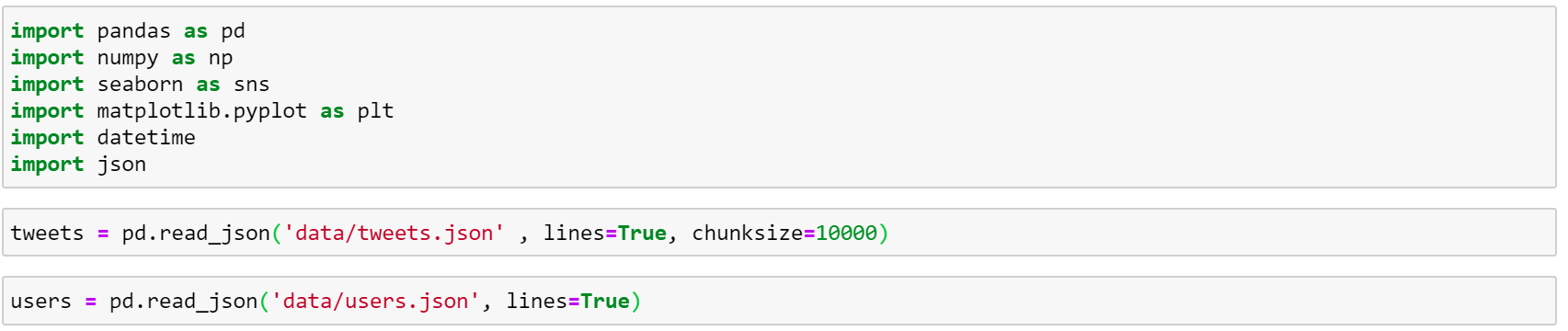
1. Which client/dataset did you select and why?

I chose **Lobbyst4America** as a client. I am very interested in being able to extract data from tweets and find useful insights.

1. Describe the steps you took to import and clean the data.

There are two files in the Lobbyst4America database, both in JSON format. “tweets.json” and “users.json”.

I have used pandas to read the json files and turn them into a dataframe.



I have created a general function to check overall information of the dataset.



From this one can find that the following columns are almost in their entirety empty and should be removed from the analysis.

|  |
| --- |
| --------Missing value Columns--------- |
| contributors |
| coordinates |
| geo |
| in\_reply\_to\_screen\_name |
| in\_reply\_to\_status\_id |
| in\_reply\_to\_status\_id\_str |
| in\_reply\_to\_user\_id |
| in\_reply\_to\_user\_id\_str |
| place |
| possibly\_sensitive |
| extended\_entities |
| quoted\_status\_id |
| quoted\_status\_id\_str |

1. Perform initial exploration of data and provide some screenshots or display some stats of the data you are looking at.

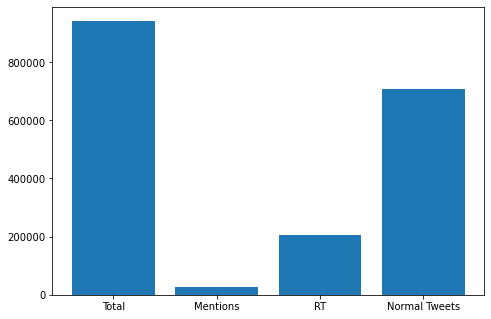
Calculated the number of tweets listed and the number of mentions and retweets.

Total tweets: 943370

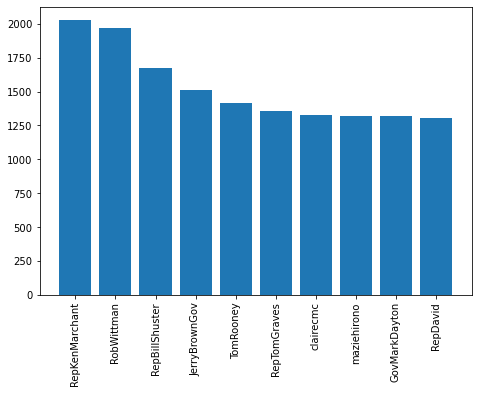
Total RT: 206679

Total mentions: 27482

Total normal tweets: 709209



Top 10 people writing the most tweets in the dataset.



1. Create an ERD or proposed ERD to show the relationships of the data you are exploring.

Below you can find the ERD proposed showing the relationship of the data. As one can see there is a lot of columns specially in the users database that are not really useful. s



## Step 2: Develop Project Proposal

### **Description**

The goal is to analyse the congressional tweets in order to understand key topics, members, and relationships within Congress. These insights will help the company focus their efforts.

### Questions

Create 2-3 questions that you want to answer with the data:

1. What are the main topics discussed over Twitter?
2. What are the relationships between congress people? Who is the most influential on Twitter?

### Hypothesis

1. The initial hypotheses is that using the hashtags used by the tweets one can find the main topics discussed and used them to identify main topics.
2. Looking at the retweets of people, we can find whether the different congressman are connected. In addition, we can look at the number of friends they have.

### Approach

1. I will be checking the text column in the tweet database to extract the hashtags. The metric used will be the relationship between hashtags and congressmen.
2. Under users database we can find friends\_count which can be used to our benefit. The best metric for this is number of mentions each congressman has and how many times their messages have been retweeted.