Milestone 1: Project Proposal and Data Selection/preparation

Step 1: Preparing for Your Proposal

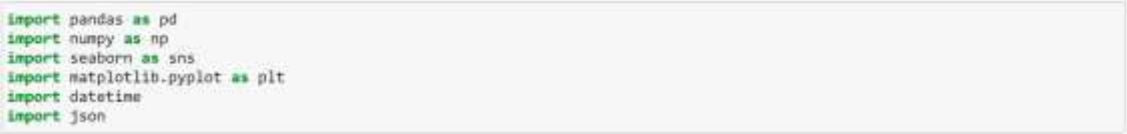
1. Which client/dataset did you select and why?

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

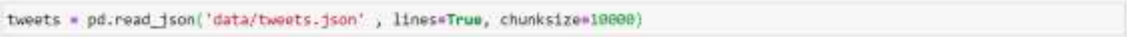
2. Describe the steps you took to import and clean the data.

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

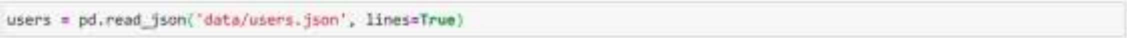
   I have used pandas to read the json files and turn them into a dataframe.



```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import datetime
import json
```

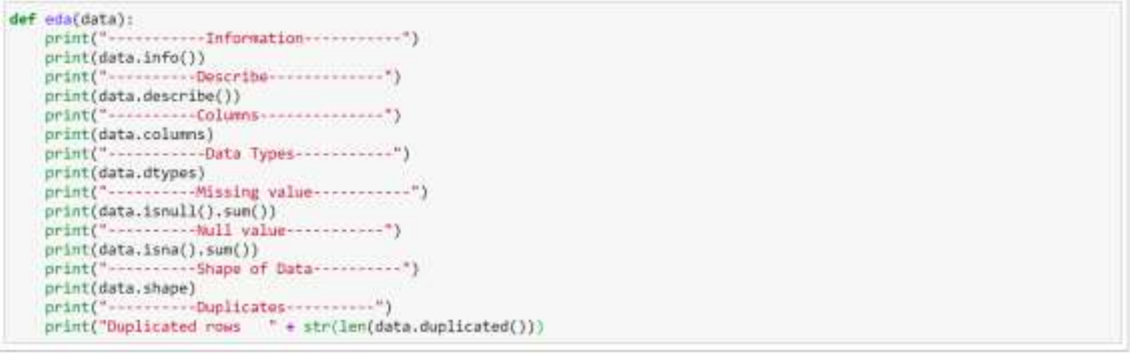


```
tweets = pd.read_json('data/tweets.json' , lines=True, chunksize=10000)
```



```
users = pd.read_json('data/users.json', lines=True)
```

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



```
def eda(data):
    print("-----------Information-----------")
    print(data.info())
    print("----------Describe-------------")
    print(data.describe())
    print("----------Columns-------------")
    print(data.columns)
    print("-----------Data Types----------")
    print(data.dtypes)
    print("----------Missing value----------")
    print(data.isnull().sum())
    print("----------Null value----------")
    print(data.isna().sum())
    print("----------Shape of Data----------")
    print(data.shape)
    print("----------Duplicates----------")
    print("Duplicated rows   " + str(len(data.duplicated())))
```

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

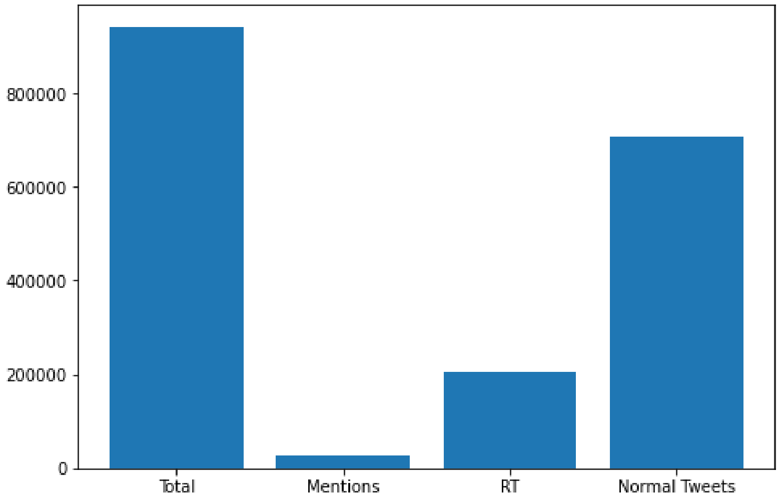
--------Missing value Columns---------
contributors
coordinates
geo
in_reply_to_screen_name
in_reply_to_status_id
in_reply_to_status_id_str
in_reply_to_user_id
in_reply_to_user_id_str
place
possibly_sensitive

extended_entities
quoted_status_id
quoted_status_id_str

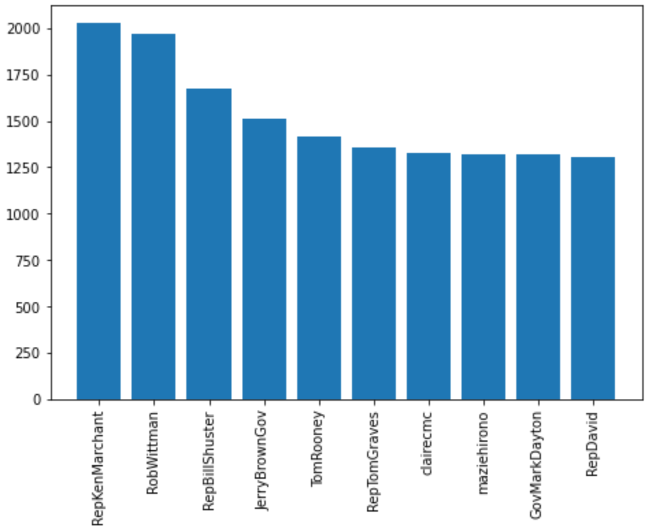
3. Perform initial exploration of data and provide some screenshots or display some stats of the data you are looking at.

   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

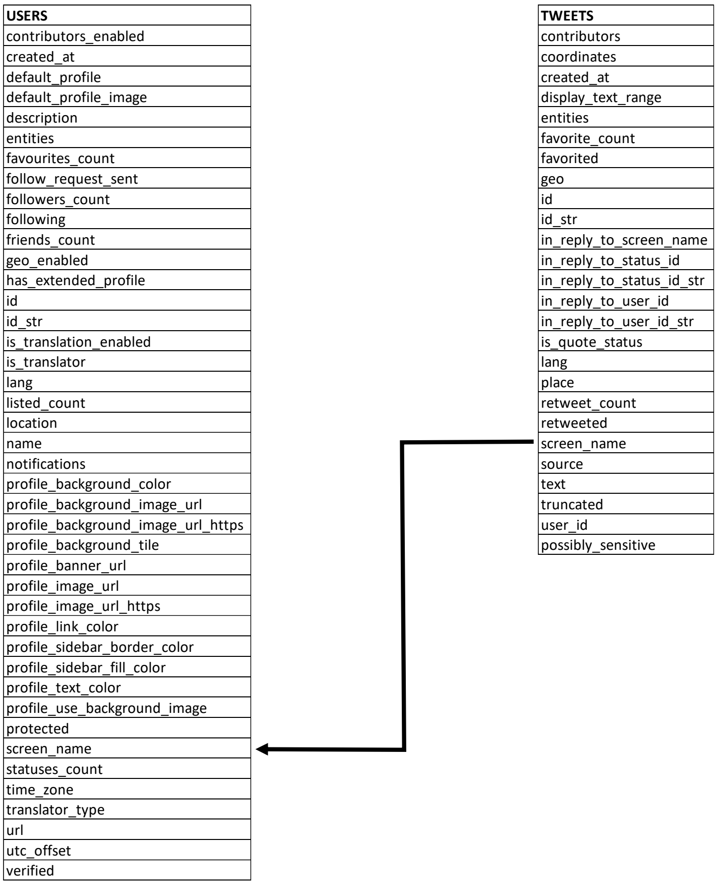


   Top 10 people writing the most tweets in the dataset.



4. 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



USERS
contributors_enabled
created_at
default_profile
default_profile_image
description
entities
favourites_count
follow_request_sent
followers_count
following
friends_count
geo_enabled
has_extended_profile
id
id_str
is_translation_enabled
is_translator
lang
listed_count
location
name
notifications
profile_background_color
profile_background_image_url
profile_background_image_url_https
profile_background_tile
profile_banner_url
profile_image_url
profile_image_url_https
profile_link_color
profile_sidebar_border_color
profile_sidebar_fill_color
profile_text_color
profile_use_background_image
protected
screen_name
statuses_count
time_zone
translator_type
url
utc_offset
verified

TWEETS
contributors
coordinates
created_at
display_text_range
entities
favorite_count
favorited
geo
id
id_str
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
is_quote_status
lang
place
retweet_count
retweeted
screen_name
source
text
truncated
user_id
possibly_sensitive

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.