**Step 1: Preparing for Your Proposal**

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

I selected the Lobbyists4America dataset. This dataset includes all tweets relevant to the lobbying for new laws by the general public. Having an avid interest in American politics and also intrigued by the unique concept of lobbying, this data appealed to me the most.

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

The tweets were available in a huge (1.6Gb) json file. All tweet parameters were in separate dictionaries, so basically 1 dictionary for every tweet, separated by a ‘\n’, i.e a new dictionary began in a new line. Python does not parse multiple dictionaries using inbuilt json functions.

Separating these dictionaries according to ‘\n’, I tried to make a list of all these dictionaries, which I could then parse into a pandas df. But the problem with the json was its size. Such a large json could not be handled by Python and it showed a memory error. Therefore I researched on storeable variables and dataframes.

I found a module named ‘pickle’ which stores pandas dataframe object into the local memory and can be easily read or written to using Spyder. So that is what I used. I currently am running continuous loops to load all data into the df. It takes around 15mins for 10k records and takes longer as I go deeper into the data. As yet, I have stored 1lakh entries successfully and do not see the end of the data yet. I do not know how many tweets are there in that file but I believe 1 lakh tweets is a sufficient database to work on.

Another problem was, some rows has a list as its value, i.e the ‘key’ in the original dictionary has a list of dictionaries as its ‘value’! I found a way to parse this and add these buried values and add them to the df column.

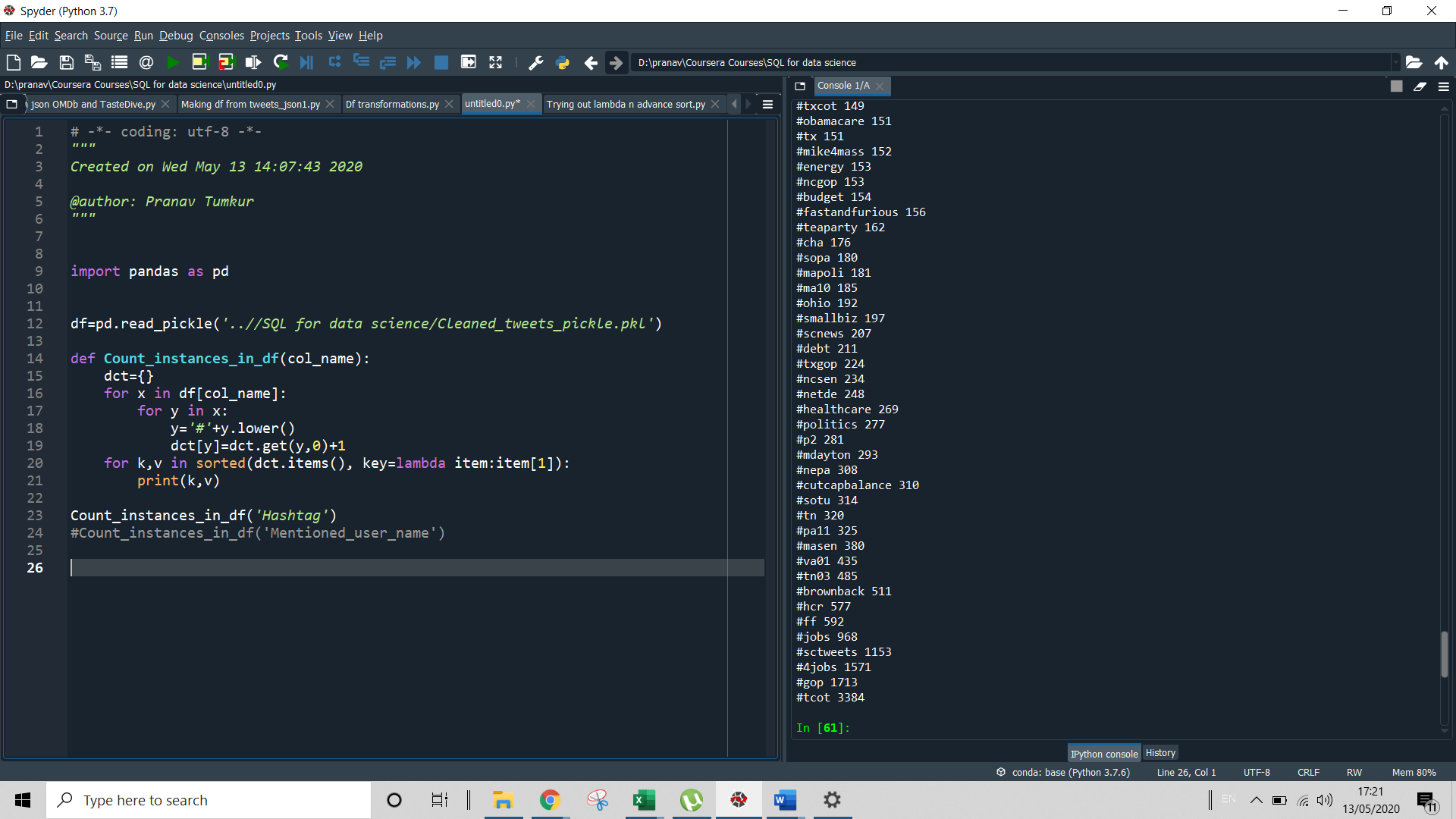
The json file could not be opened (notepad says the file is too large, Chrome crashes and no online json editor takes that big a file) and I was not aware of even how it looked. So I kept writing to a .csv file at every step to check how the data looked.

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

There were some parameters like the mentioned users, followers count, hashtags used, official site which I thought needed to be explored further in addition to the actual tweet text of course.

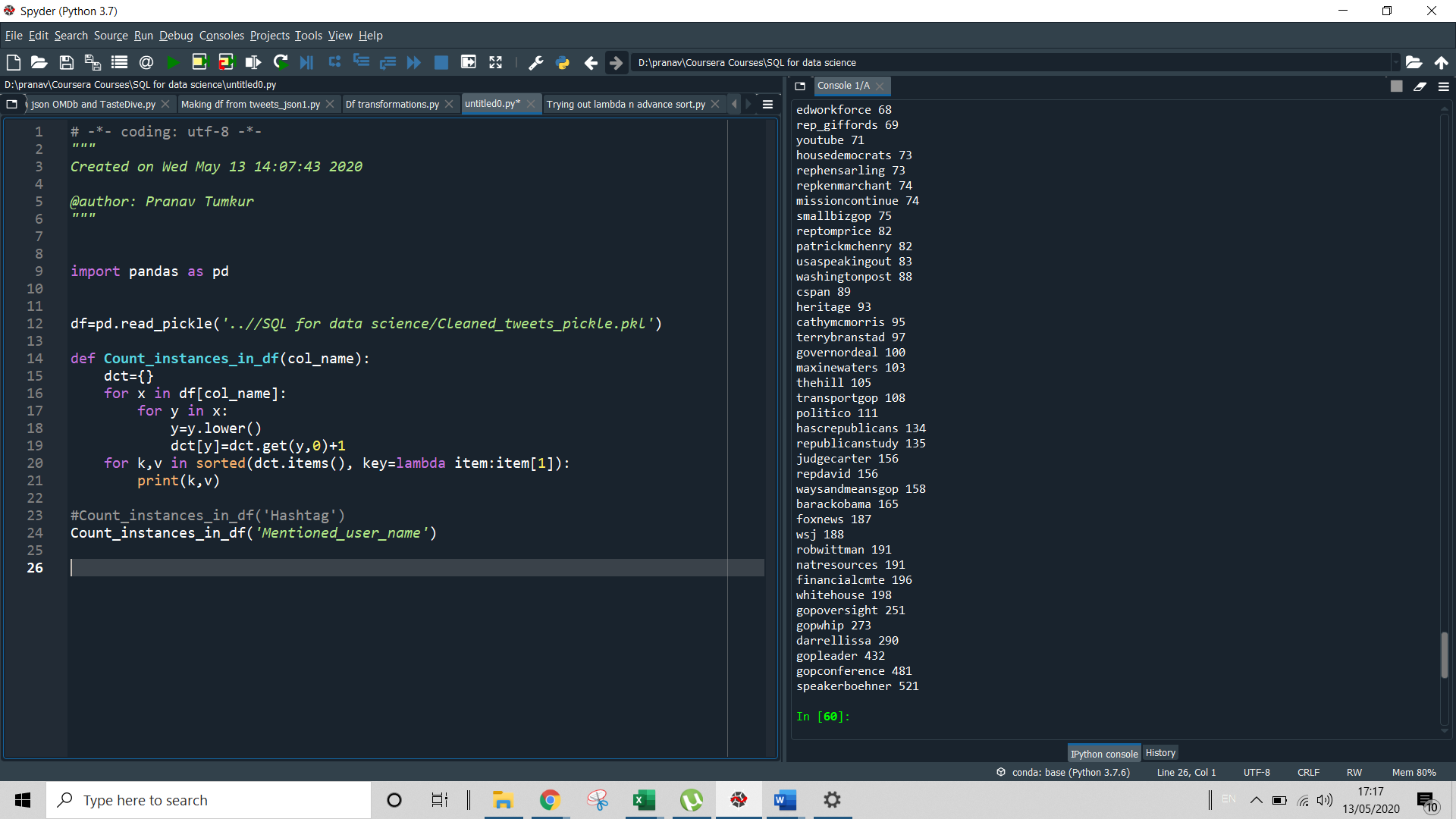
* Over the years a few hashtags like #tcot, #gob, #4jobs etc. have been used a large no. of times. Though as of now the meaning/full-form of these hashtags is unclear, these are subject for further analysis.

There’s also #fastandfurious mentioned 156 times over the year. But that seems to be unrelated!!

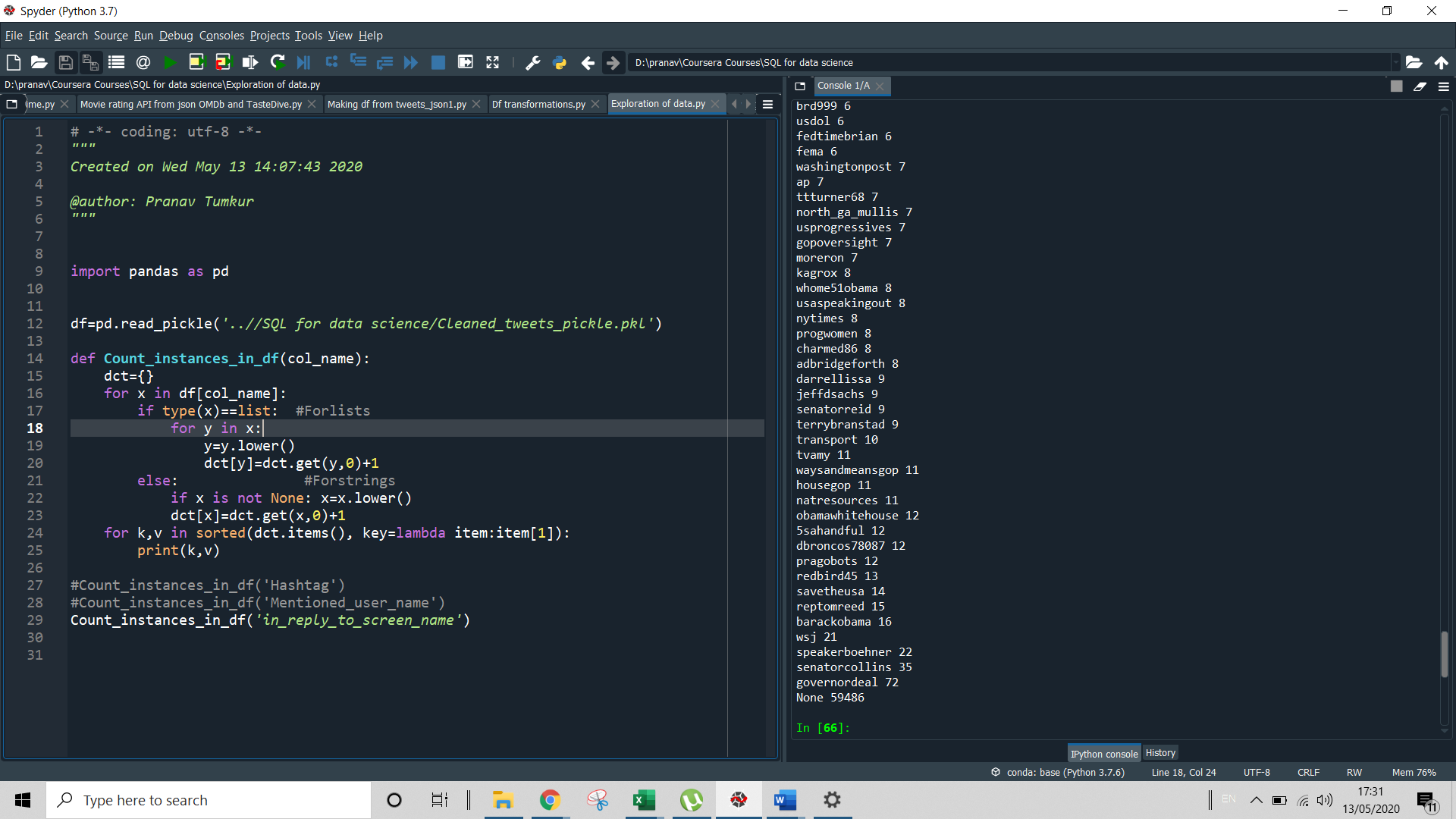


* There are also several mentions and replies to specific people quite frequently on the topic:

Mentions: speakeboehner, gopconference, gopleader etc.

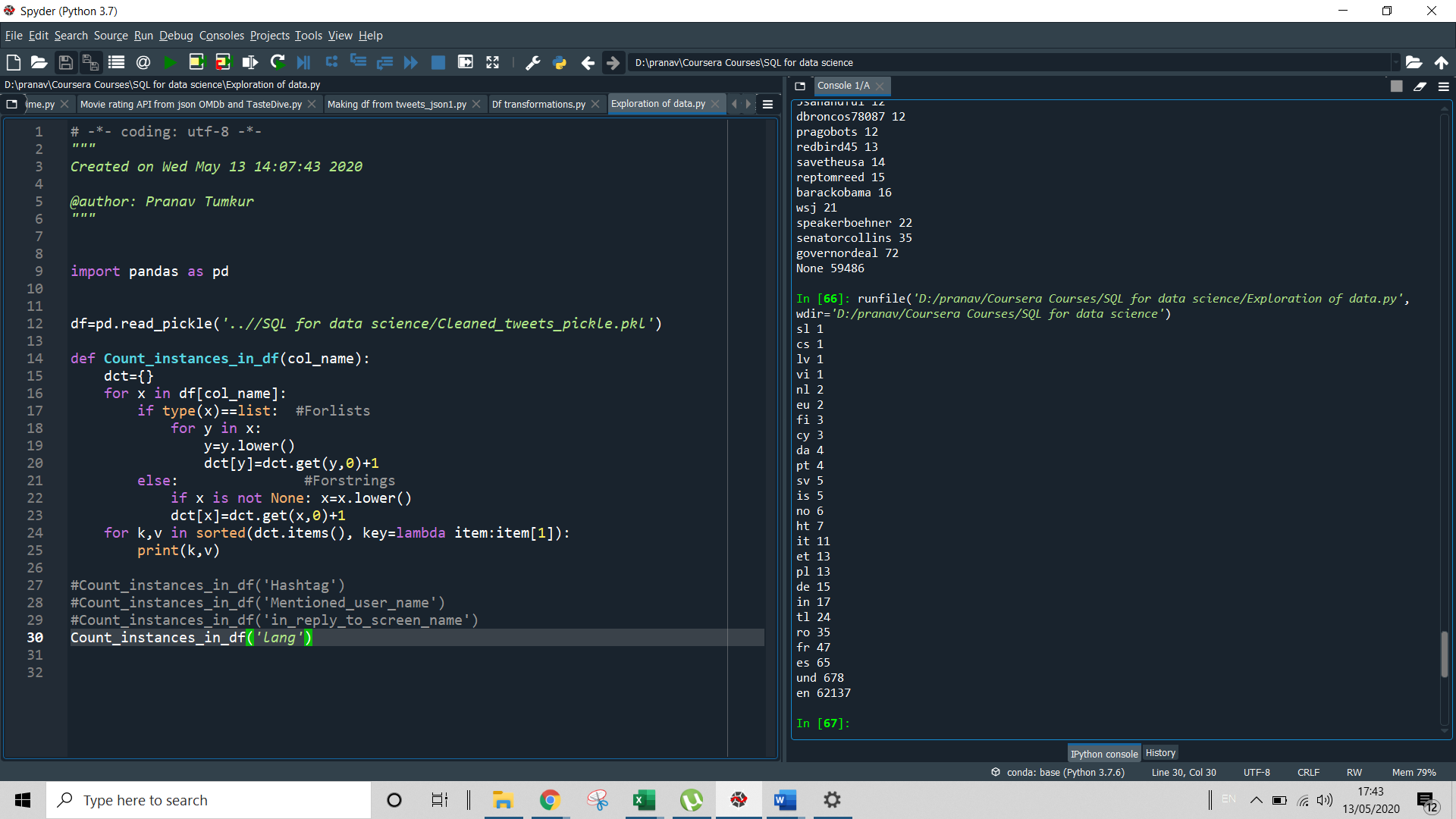


* Replies:

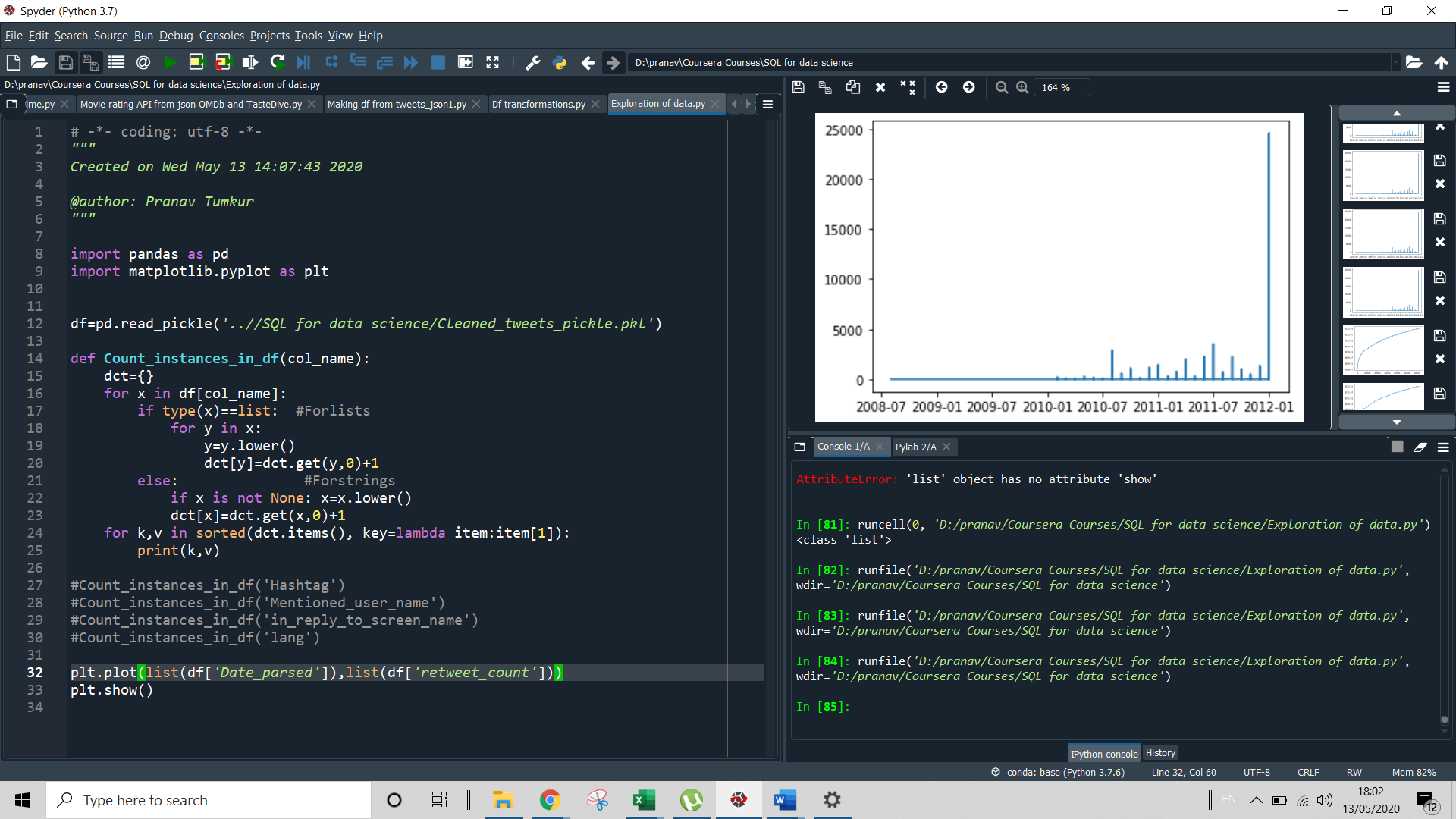


We see that replies do not possess a significant quantum, since over 4 years the highest no. of replied to a single id has been just 72, i.e to governordeal. We can conclude that this is not very significant data.

* The language also seems to be predominantly English by a large majority, undefined being a disant second and Spanish being 3rd and still having only 65 tweets in 4 years:

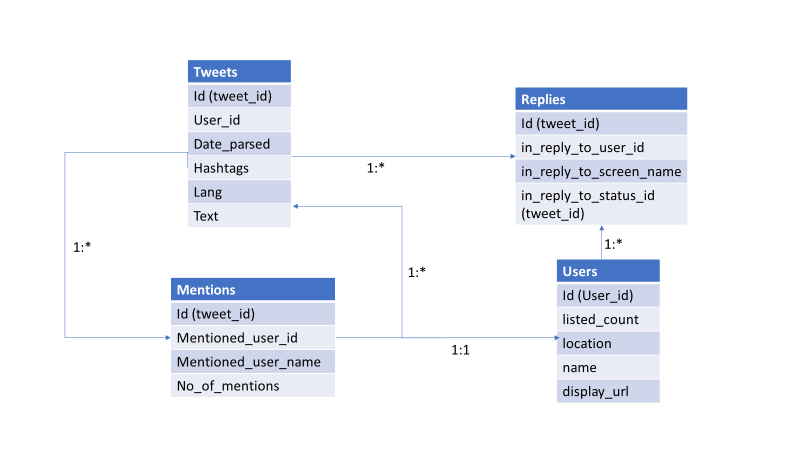


* Lastly, let us see the trend of the retweets. This will show us how if people have started caring any more for the cause of lobbying.



And we see a HUGE increase in 2012. As many as 24k retweets have been received which means that people are starting to care for the cause. We may try to investigate why and how this has happened. Has a famous personality joined the battle? Are official twitter channels getting more invested in the cause?

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

PFB the proposed ERD (made in MS Powerpoint):

Logics followed are as below for relationships:

* One tweet can have multiple mentions/replies
* A mention will be for only one user
* One user can have many tweets/replies

**Step 2: Develop Project Proposal**

Description

Lobbyists4America has a unique concept of lobbying by the pubic for a common cause and contributing to this particular cause. This project involves analysing twitter data to see if there has been any increase in response from government officials to tweets as also to see if more people have become committed to the cause. The audience could be the CEO of Lobbyists4America or the Congress representatives (Senators) who’d be interested in knowing how many people have lobbied for a particular cause.

Questions

1. Has there been any increase in response from government officials to tweets?
2. Has the number of lobbyists gone up?

Hypothesis

Since there is an election scheduled in 2012, we expect politicians to be more active on Twitter, responding to people and we would also see a surge in the tweets traffic.

W​rite 2-3 assumptions about the data that you'll want to go back to prove or disprove. You will want to keep them in front of you as you look at the data to keep them or change them. You may see relationships that you want to explore and will develop a "belief" about the data.

1. I have assumed that all tweets are only related to Lobbyists4America. If any other data seeps in, it would skew our observations.
2. I have also assumed that the tweets are solely due to the efforts of the institution- Lobbyists4America. If there was some other group which conducted similar initiatives and therefore the interest in the topic increased, it would not reflect purely the progress of Lobbyists4America.

Approach

Describe in 5-6 sentences w​hat approach you are going to take in order to prove (or disprove) your hypotheses. Think about the following in your answer:

1. Check the trend of retweets, replies, mentions to adjudge the public response
2. Find out how many government officials replied to the public or retweeted them
3. Find out if the no. of users who have shown support to the cause has increased
4. Keywords, hashtags mentioned most frequently and what they mean
5. A general sentiment analysis of people’s tweets to judge if they’re angry, depressed, rude and to determine the tone of the tweet.

W​hat features (fields/columns) are you going to look at first?

1. Hashtags
2. Mentions
3. Replies
4. Count of tweets trend

I​s there a relationship that exists that you want to explore?

Wish to see trend wrt time

W​hat metric/ evaluation measure will you use?

I will mostly use the Count and try plotting the same against time period. Will also make a word cloud for hashtags and sentiment analysis.