# Sentiment Analysis with Python Part 3

By: Zach Adair

MSDS 600

11/6/2018

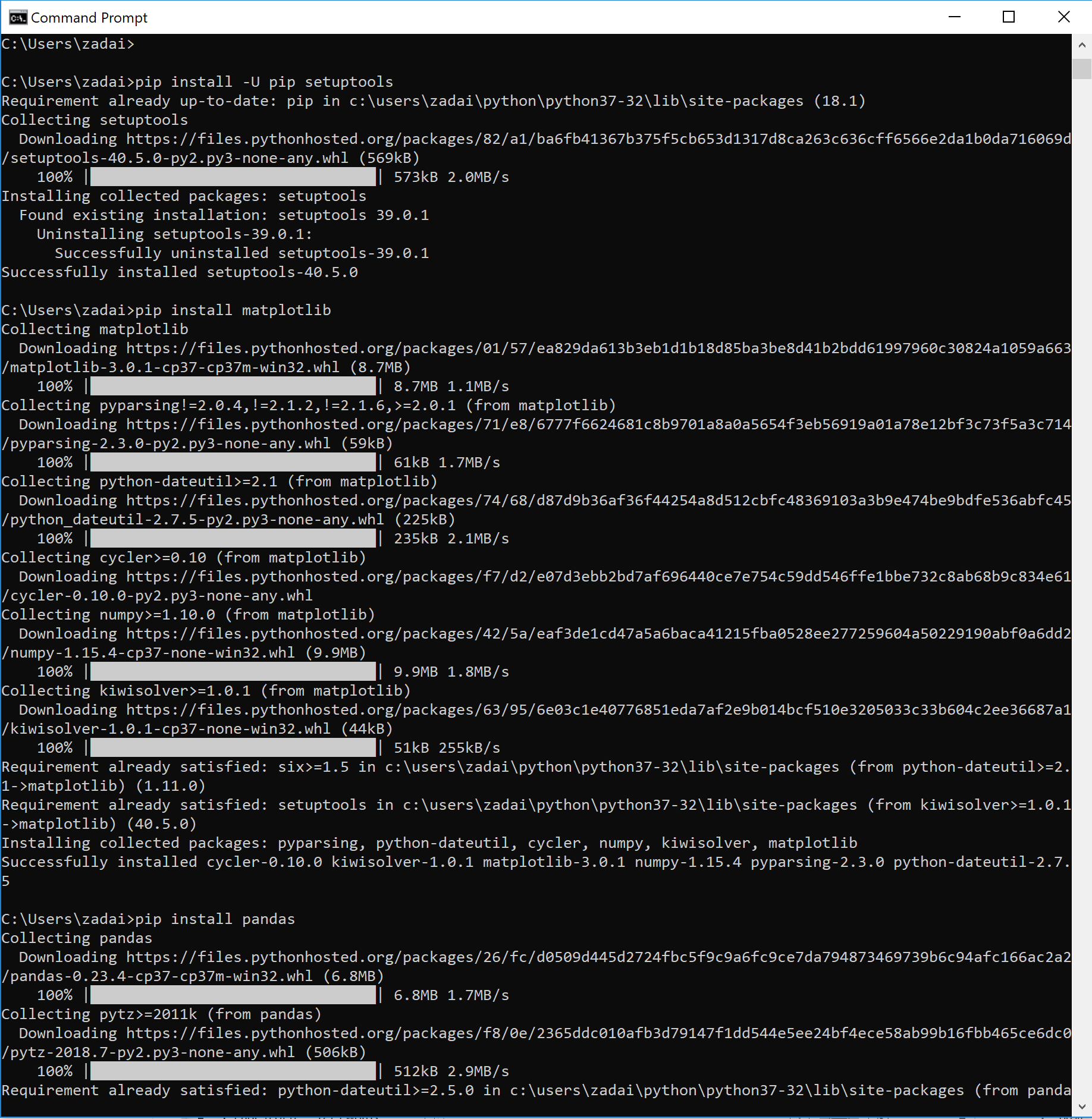
[zadair@regis.edu](mailto:zadair@regis.edu)

### What Part 3 Entails

Part 3 is final part of the Sentiment Analysis of our Python data which was created in output.json last week and was analyzed earlier in part 2 of this project. In part 3 we will be installing some more necessary tools in python to allow us to make lists and graphs within the application. With these new tools we will be able to analyze the top languages and locations of the tweets. The last section of this report is about the key words analysis that is possible with Python. I picked 5 different words and was able analyze how often those 5 words showed within our population of tweets.

### Steps in the Process

1. Open up the command prompt and install several packages
   1. Pip install -U pip setuptools
   2. Pip install matplotlib
   3. $ pip install pandas

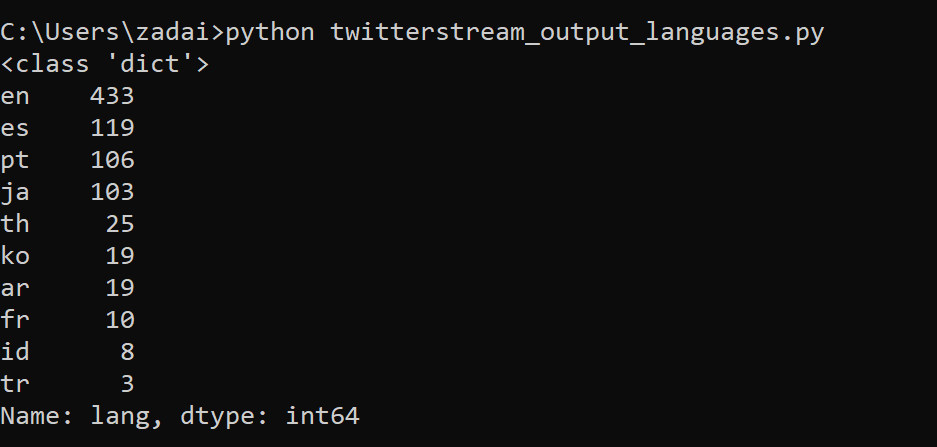


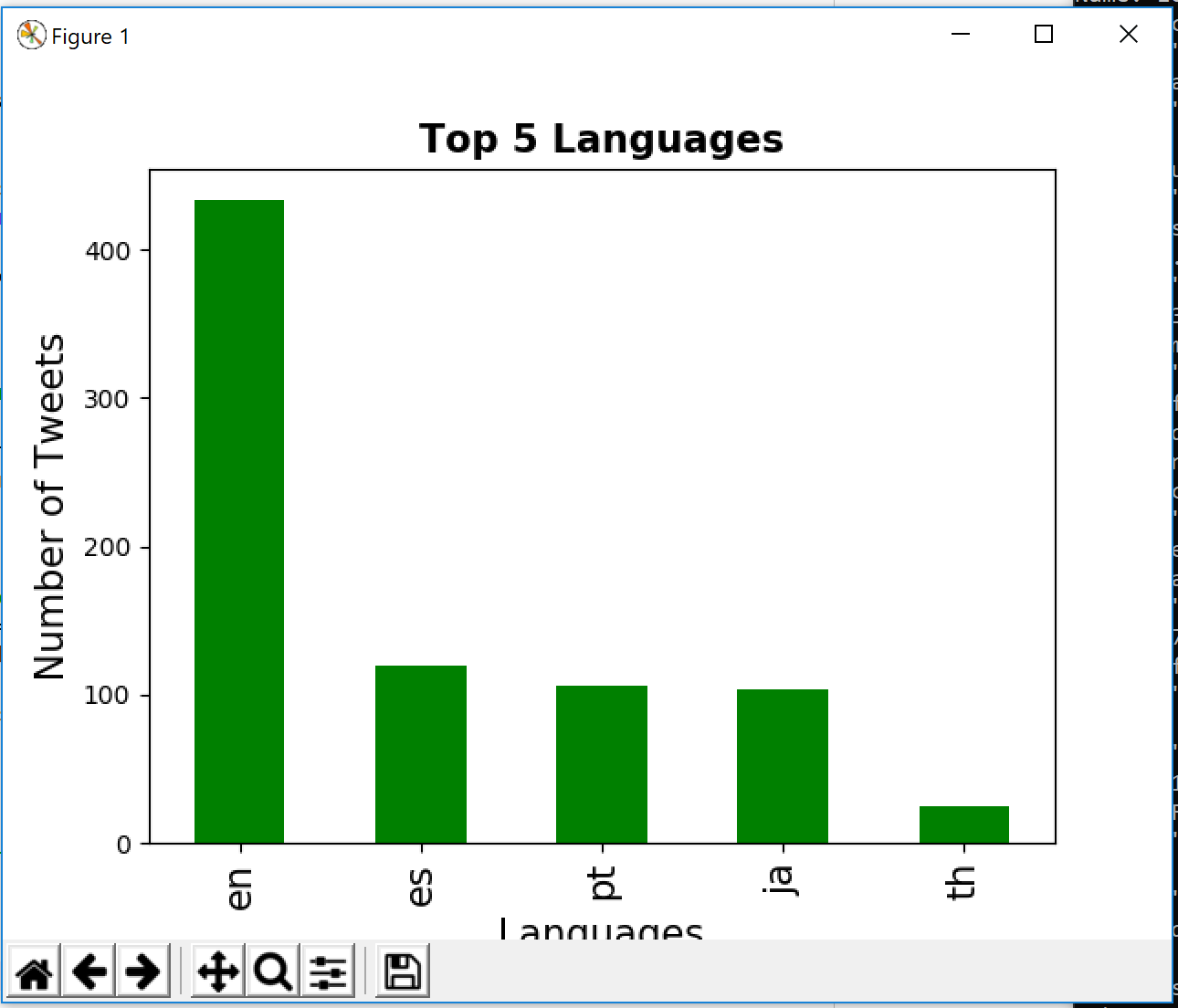
1. Once loaded in successfully, move into an IDE and open up the twitterstream\_output\_language.py file
   1. I am using the IDE Spyder, but from in there we are able to change how many different languages we will see in our output
   2. Also, we need to make sure the correct file is going to be grabbed by our code, my output from the previous portions of our analysis my output file is a JSON file so I need to make sure it grabs that.



* + 1. As you can see, I have made sure that my output file the code will be grabbing is my output.json file and I also switched up the number of lines my plot will distribute out from 10 to 5

1. Go back to the command prompt and run the line of code *python: twitterstream\_output\_languages.py*

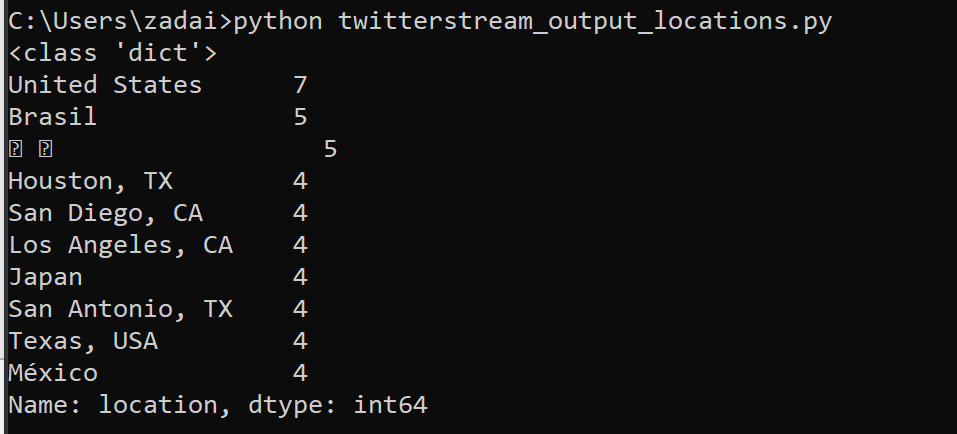


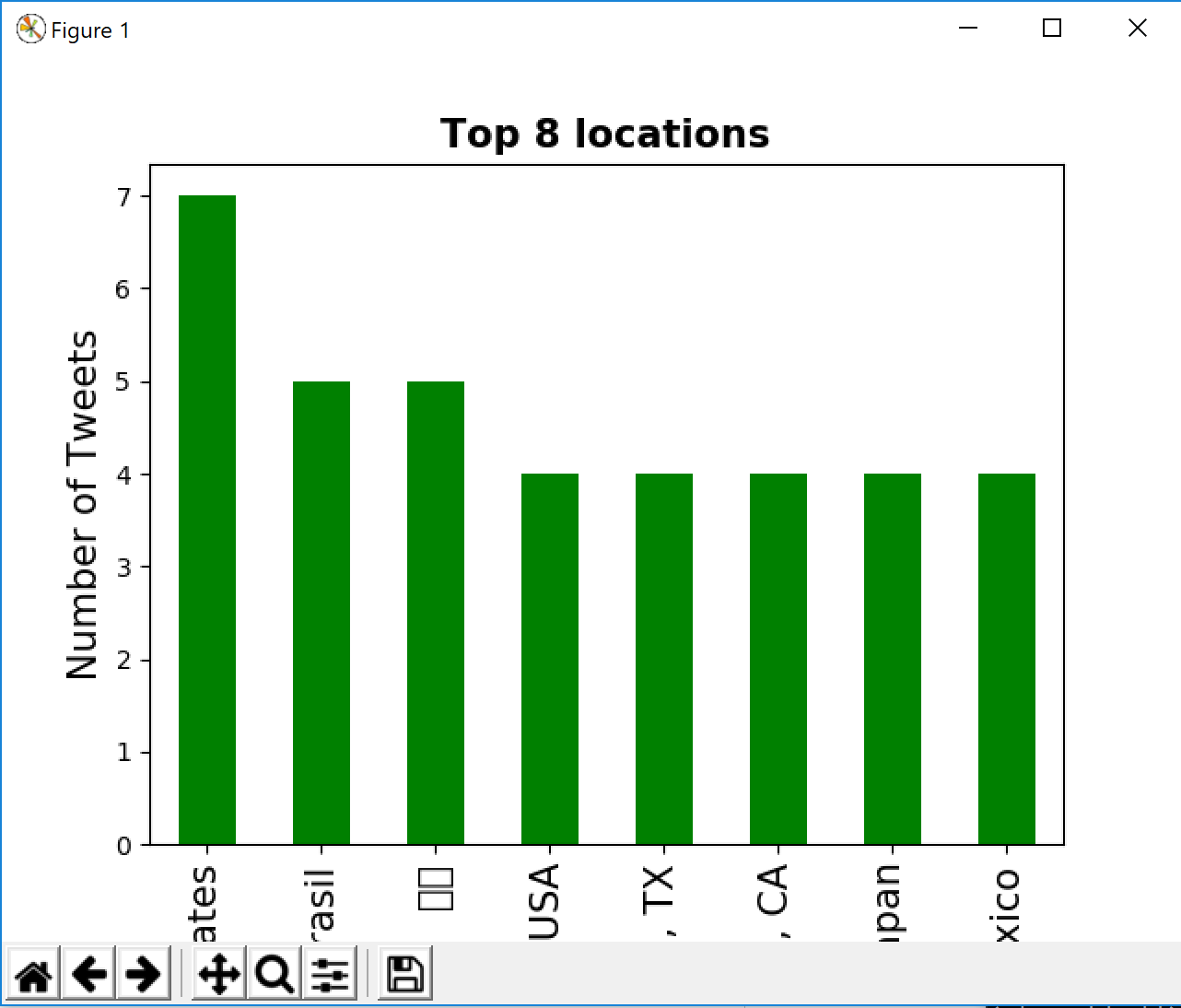


1. Above is the output from the code that was just run, a list of the top 10 languages by number of tweets in the prompt, than outside of the prompt a bar graph, in descending order of the top 5 languages by number of tweets, and from the data the top language by an overwhelming margin is English.
2. Now open twitterstream\_output\_locations.py in an IDE.
   1. Like our languages we too have the freedom to change the number of locations we will see as well.
   2. After checking and making sure Spyder would be getting my output.json file I changed the number of locations I’ll be graphing from 10 to 8

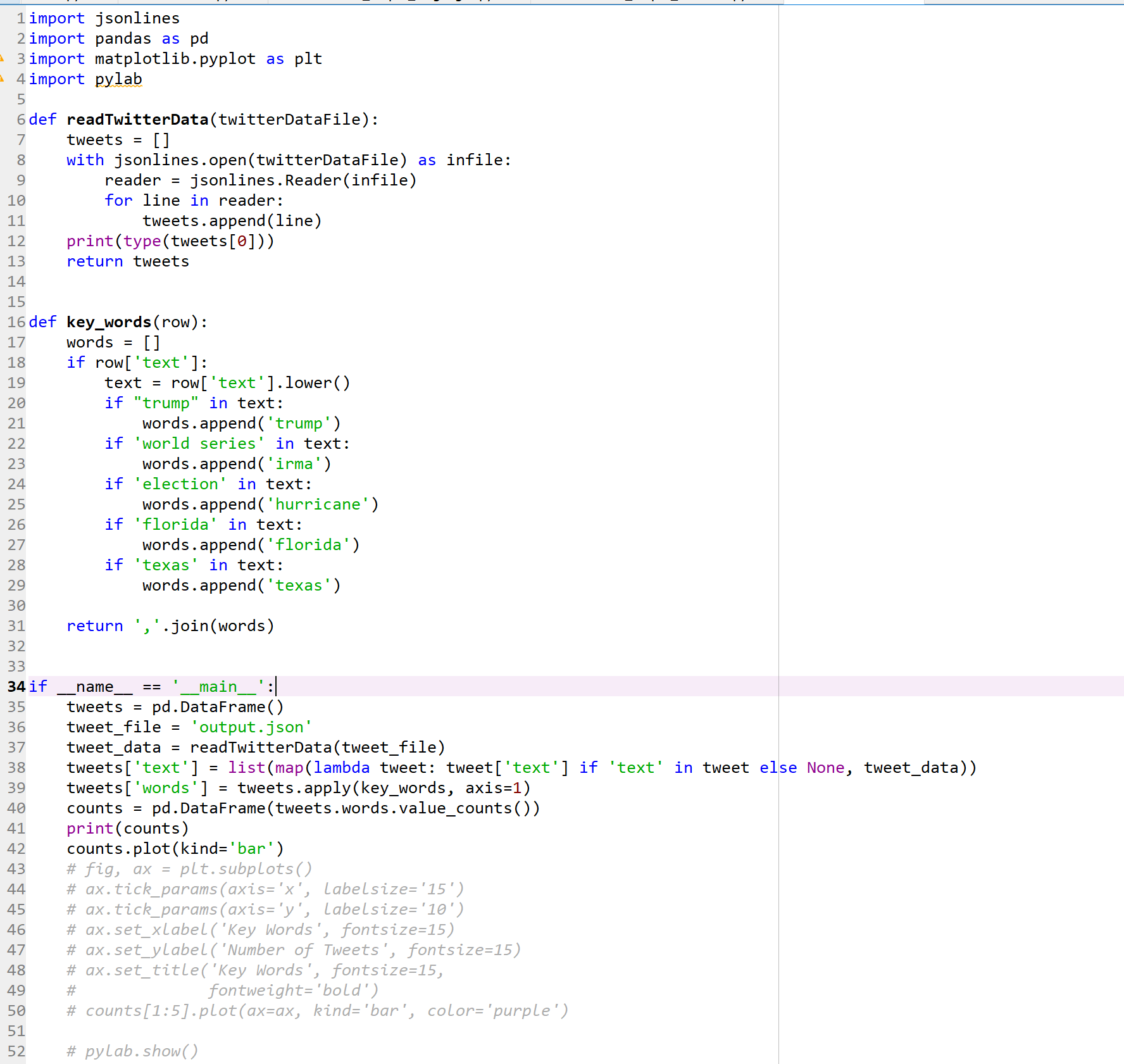


1. Save the changes made in twitterstream\_output\_locations.py and go back into the command prompt
2. Once back into the command prompt, run the code: *python twitterstream\_output\_locations.py*

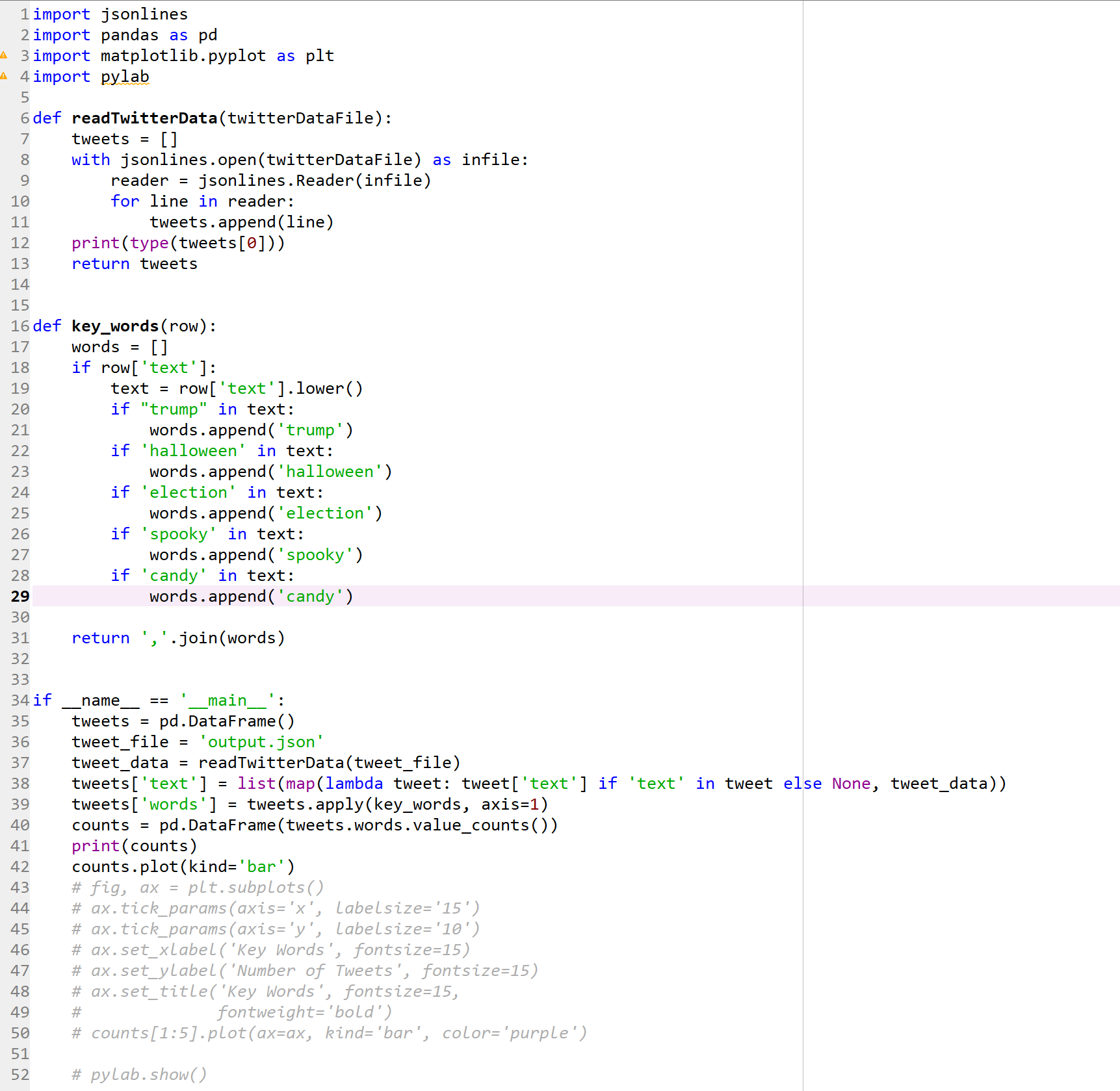




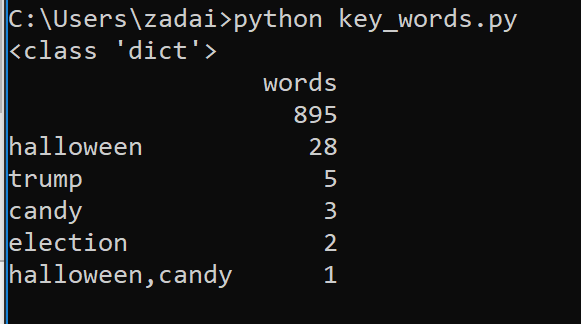
1. Above, is the list of the top ten locations of tweets from my sample of tweets and a bar graph displaying my top 8 locations of where my collected tweets were located. And as we can see the United States is my location where the majority of my tweets came from.
2. Now open up the key\_words.py file in the IDE and change the key\_words section to what you think will pick up the most tweets.



1. Since I ran my ouput.json file on 10/31 I decided to keep one of the words ‘trump’ and change the other 4.



* 1. The words I used were Trump, because he seems to always be in the news; election, because the midterm elections are coming very soon; and my last three words all pertain to Halloween so the actual word Halloween, spooky and candy.



* 1. I unfortunately could get a graph to pop up, but out of my 895 words my key words didn’t pop up as much as I’d hoped.
  2. 4 of my 5 words ended up showing up, all but spooky appeared at least once, Halloween and candy appeared together once, election was the minimum appearing twice and Halloween appeared 28 times throughout my total analysis.
* A few thoughts about my results from my analysis.

The reason I picked the words, Halloween, Trump, spooky, election and candy is because I wanted to compare two small sets of words that are about two different topics and find out which would get more results from my sample of tweets I collected. The tweets being collected on Halloween means that holiday would be a popular topic on twitter, and with the midterm elections only six days after Halloween, it appeared to be a good topic to put against Halloween. I decided on the words the five words I selected by taking the name of the two groups, Halloween and election, and taking one or two words for each that are closely associated with one of the two topics. So just going with what first came to my head I decided to use spooky and candy for Halloween and Trump for election. After I had my words I ran it through the python process and got the results that are seen above.

Looking over the results I came across a couple observations and conclusions. It is no surprise that Halloween end up being the highest tallied word that I was able to capture from my sample of words. Creating the output file on Halloween night made it almost a certainty that it would be the number one word in my opinion and the only way to prevent that is if some sort of news story broke and make Halloween go on the back-burner. I was disappointed in the words that I brought into my analysis that were associated with Halloween, I thought spooky and candy would yield more results than the 3 times with candy and not even a single time with spooky. If I were to choose again I think choosing words like haunted or trick-or-treat might have ben a better option at least for spooky. It is also possible that with a bigger sample my use of the word spooky could have yielded results.

My other two words are associated to what happened this week, Trump, who is always in the news or a talking point about something and election, which is held Tuesday. The week before the election didn’t exactly get me a heavy amount of results for those two words either. Thinking there might be some heavy talk less than a week before the election but unfortunately Trump only gave me 5 hits, and election only getting two. With all the buzz around the midterm election I just figured there would be campaigns or rallies and causing twitter to buzz more about it.

Pitting the two events, Halloween and the election were a good idea but it ended up have a flaw in its initial set up. Even though the word Halloween was the only word that is based on Halloween that generated a lot of results, It still did overwhelmingly more hits as a group compared to the two words I used for the election. I believe the biggest reason for this difference between the groups is that the output.json file was created on Halloween night, if created a day or two before or after Halloween would generate result sums which are much closer than what my result shows.

I think the words I chose could have easily had an effect on the mean of my sentiment analysis, which was 0.202, which is very slightly positive but would be classified as neutral word choice. The two topics and the words associated with them are also divisive in their own right. The election season brings a lot of strong feelings out of people in either a positive or negative way so if at the time my collection there was an influx of positive or negative tweets towards the president or the election it is possible it could effect my results. Halloween I believe is a holiday in which people look upon favorably, could effect my sentiment results by pushing them to lean more towards a strongly positive result.