# Sentiment Analysis Using Python Part 2

By: Zach Adair

11/4/2018

MSDS 600

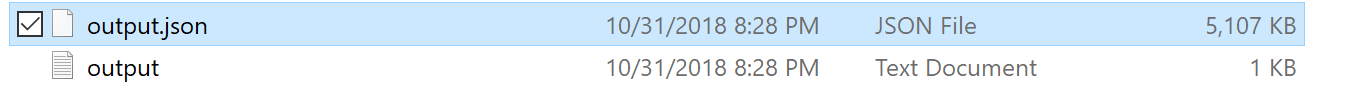
[zadair@regis.edu](mailto:zadair@regis.edu)

### What Part 2 entails

This report is done in continuation of part 1 which was finished and turned in last week. The objective of Part 2 is to take the output.json file created in Python last week and run a Sentiment Analysis on the tweets within that file. A sentiment analysis is a textual analysis which puts weights on specific words that can be used in speech (or in this case tweets) and created a value which determines if the speech was either positive or negative and how strong that tweet was towards either positive or negative speech. Our word rankings were provided for us and is titled AFINN-111.txt and has over 2000 words with values applied to them between -5 and 5. In this project we will download the necessary files and packages into python and run the analysis. At the end the mean sentiment value will be computed, and will also be between -5 and 5, and we will be able to see how positive or negative the tweets collected last week were.

### Steps in the Process

1. Leaving off from last week’s part 1 of the Sentiment Analysis, an output file had been made for what needs to be used for this portion of the project.
2. Start by taking the output file and making sure it is a JSON file.

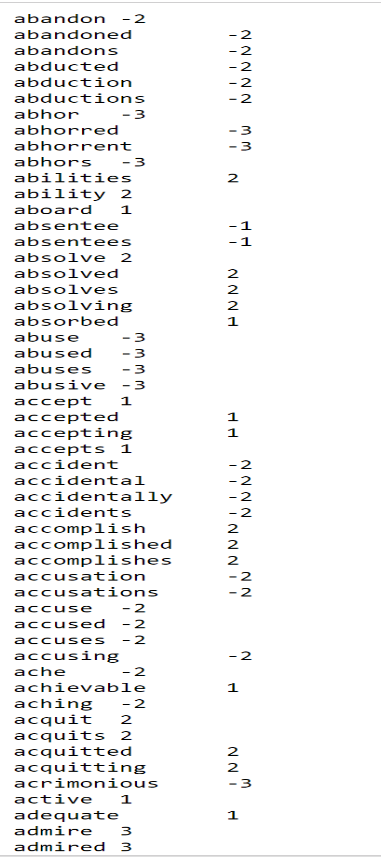


1. From the previous week I saved an output file as a .txt and a .json file
2. Next, take the AFINN-111.txt file and the tweetsentiment.py file from the assignment page and download them into my folder.

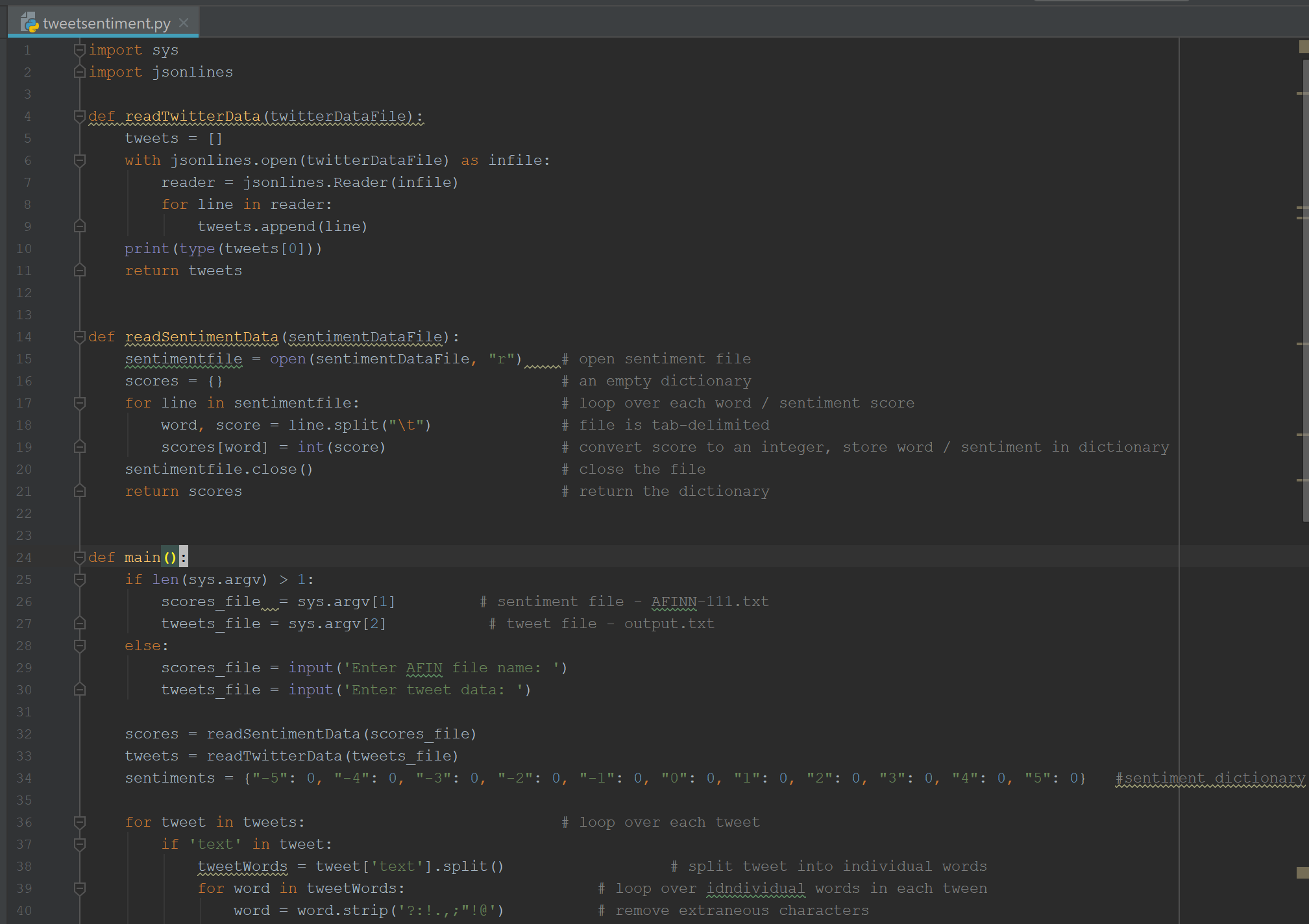




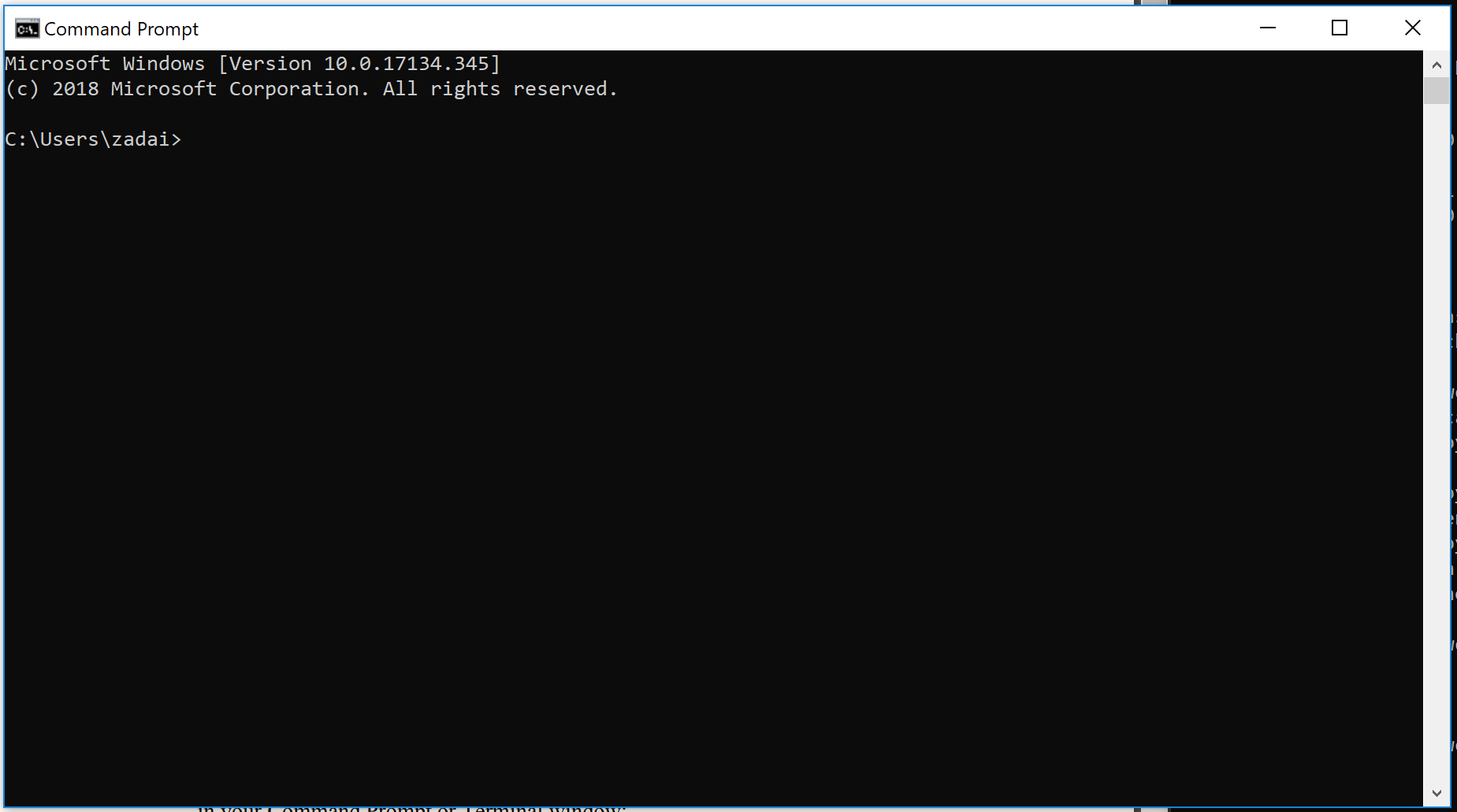
1. These two files will be used together along with the output file from last week to create our Sentiment Analysis.



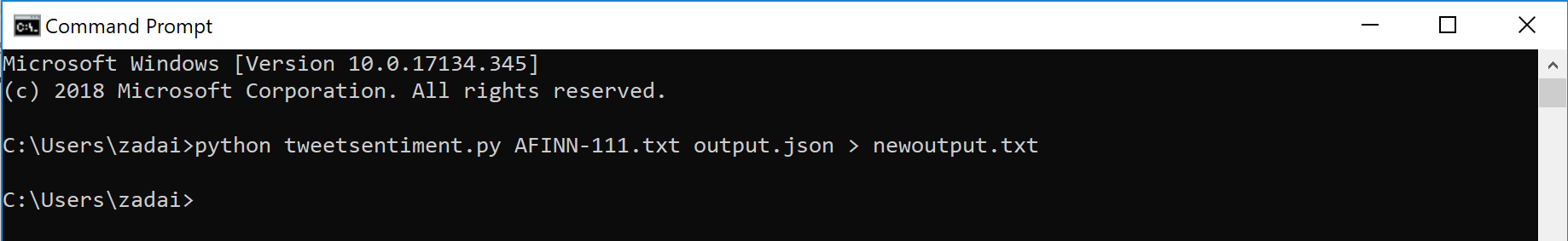
1. First is the AFINN-111.txt file, which assigns a value from -5 to 5 of each word in the list.



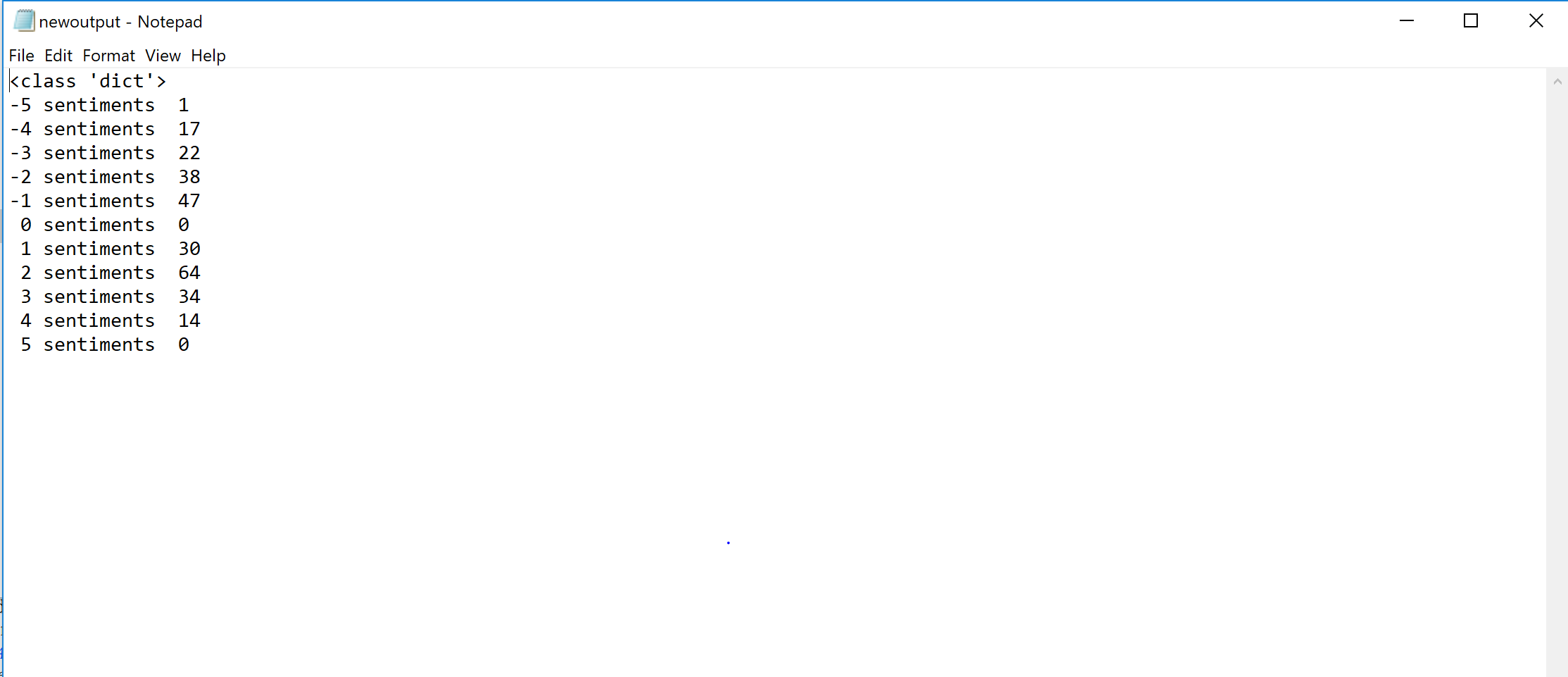
1. The other file downloaded was the tweetsentiment.py file, seen above in Pycharm and will be run in the next step.
2. Open the Command Prompt or an IDE, I decided to open the Command Prompt



1. Run the code *python tweetsentiment.py AFINN-111.txt output.json > newoutput.txt*



1. Hit *Ctl+c* and the newoutput.txt is created.
2. Go to the file location of the newoutput.txt file and open it up.



1. These are the results I got for my Sentiment Analysis for values -5 to 5.
2. Now that the python file has been run and the text file has been created it is time to answer the questions asked at the end of the analysis.

### Questions for the end of the Analysis

* What type of data are the sentiment values i.e., {-5, -4, … , 4, 5}?

Sentiment values also measure the emotional intensity of language, words and phrases within text. Allowing for there to be a measure on if a reaction is positive or negative based on the analysis of the response. Sentiment values are similar to the Likert Scale which we discussed last week in our discussion posts. Like Likert Scale data, Sentiment values are ordinal data where the numbered options are categories to which specific words from the AFINN-111.txt file slotted values, between -5 and 5 to base on the words in tweets collected from last week. The spread of my sentiment values mean, based on the tweets collected last week and the AFINN-111 file that the words assigned values in that file appeared however many times for their specific number category. Even though sentiment values are ordinal data we can still use them to find measures of central tendency like mean, median and mode.

* Compute the mean for the sentiments.

The mean value for sentiment analysis is based on a weighted mean, taking all the entries per number category and multiplying them by the categorical value and then dividing by the total amount, the result should lead to a value between -5 and 5.

Mean = (1(-5) + 17(-4) + 22(-3) + 38(-2) + 47(-1) + 0(0) + 30(1) + 64(2) + 34(3) + 14(4) + 5(0)) / 267

Mean = (-5 + -68 + -66 + -76 + -47 + 0 + 30 + 128 + 102 +56 + 0) / 267

Mean = 54 / 267 = 0.202

Thus the mean of my sentiment analysis is approximately 0.202

* What does this mean value mean?

What my mean value of 0.202 means is that based on the AFINN-111.txt my tweets are not strongly positive or negative but very close to neutral because my mean is so close to zero. What this means for an analysis is that from my group of tweets they didn’t lean strong one way or the other, at the time of collection, the tweets were almost balanced evenly between positive and negative posts. I’m sure some tweets were scaled heavily on one side or the other but the data shows that it was fairly balanced on both sides.

* In calculating the mean, what additional assumptions about the type of the sentiment data have been made?

There are plenty of additional assumptions one can make about the sentiment data made. First, is how we have designated the words will accurately tell us how positive or negative someone’s tweet is towards whatever topic they are referring to. Another, is that we are assuming that people are using specific words as either positive or negative and cannot be used for another reasoning, so in the case of some words being used in the correct context or being used incorrectly either as slang or something else the regardless result will be the same for the use of the given word. The same can be used for any other type of contextual misrepresentation of the English language like if someone were to use sarcasm. An additional assumption we are making is there was no human error, that none of the words being analyzed were by mistake, so no one meant to spell one word or make a different point about a topic and just ended up using the wrong word. My final assumption that I want to highlight is that are results can differ depending on the collection and what events going on in the world. If there is a huge national event occurring like an election or a giant sporting event, our tweets collected will pertain more to that “in the moment” topic than specific words associated with that event would be captured at any other time. For example, last Sunday night the Green Bay Packers and the New England Patriots played an NFL game, and a lot of attention was centered around the two quarterbacks, Aaron Rodgers and Tom Brady because they are two commonly debated players for who is truly the Greatest of all Time, or G.O.A.T. This could affect a sentiment analysis because words like goat, greatest, and championships will appear a lot more frequently within our collection of data than it would if we decided to do our tweet collection on a random boring Wednesday. This causes a subjective element to put into our analysis because of the heavy influence an event can have on what our collection of words from tweets may be.

* Suppose that in AFINN-111.txt that there were different numbers of words that have been rated. What effect would this have on your analysis?
  + What might be a way to compensate for this problem?

The effect different numbers on our analysis would do is it could change how positive or how negative our analysis is, or change the balance of our analysis entirely. It just depends on how those words are assigned and the distribution of those words throughout our collection of tweets.

The effect different numbers on the rated words would be most likely a shift either positively or negative in the mean. For example if a majority of the words were now rated as strongly positive or strongly negative, then depending sample, the result might be closer to a strong reaction in either way.

A way to possibly compensate for this problem is if you were to change around the numbers on the words that you make it as balanced a change as possible in terms of how positive and negative numbers are changed. That way if there is a change it wouldn’t have a giant effect on the mean of the results.

* Visit this site and discuss how sarcasm and other factors will effect your above results. <http://darenr.github.io/afinn/>

Sentiment analysis can be an extremely difficult form of analysis because of how different individuals use human language, slang for instance in the English languages opens the door for a slew of words to be slapped with double-meanings which can make it difficult measuring for sentiment analysis. A word can be weighted in one regard but could be used in a different way than originally weighted on a sentiment scale. For example, if we took the word bad, that word could mean either something positive or negative just based on how the word is used; and there’s seemingly 100s of words like it! Slang is one of the biggest causes of contextual misunderstandings and when those occur it can really throw off a sentiment analysis.

Sarcasm is another factor which can effect sentiment analysis because it can misconstrue what a person is saying. For example, someone can tweet, “Dang, I love Taco Bell cause it is sooooooo good for you!” With the long drawn out o’s you can deduce that whoever tweeted this is probably being sarcastic, but if in your sentiment analysis you have words like love and good positively weighted than someone using them sarcastically to actually dismiss something can cause for the tweet to be misrepresented within the analysis.