Objective: The objective of this lab was to use SparkStreaming to track word usage across twitter. SparkStreaming gives us the ability to process data in real-time, as it is coming in (on selected time intervals). This can be very useful for things such as stock market analysis, or customer sentiment.

Data: The data consists of all tweets submitted to twitter that occur while the program is running. Unfortunately, in order to access twitters API and get access to the twitter stream, you need a developer account. I created a new twitter, and applied for a developer account, but the delay between application and approval can take a while. Without API access, it is impossible to access the twitter stream, making the actual runnning of this program impossible.

Evaluation: The actual wordCount part of the program was dead simple. We first have to pull in the tweets as Json objects. Then we convert them to strings, split along the white space. We can flatmap these split strings and create a ‘list’ of all words used. Then we can do a very simple map reduce, where we map each word to the value of 1 originally, then reduce them by key to combine the values of all same words. The results are then printed to console.

Conclusion: The actual mapReduce part of this assignment was quite simple at this point. But the connecting to twitter was somewhat more involved. I was able to find tons of good resources to help, the main one being <http://adilmoujahid.com/posts/2014/07/twitter-analytics/>.However, what I was completely unable to find was a way to connect to twitters API without first having access\_keys and consumer\_tokens. From the looks of older results, it used to be much easier to access the API, however with the rise of BOTS and the twitter events surrounding the 2016 election, I can understand twitter controlling api access much more closely now.