For looking at the Yelp data I thought that it would be a good idea to first look over some of the reviews within the JSON file. There were some obvious searches I could use (pronouns, etc.) but there were two searches that I thought would be more effective and creative. Is the use of numbers and star ratings correlated? In other words, do people who are happy tend to write in their reviews “this worked for 8 hours” and “it only cost $3”? Or do people tend to write, “it was 30 minutes late” and “I would never give this place 5 stars”? I imaged that those who used numbers tended to rate the company with higher stars.

My next question was whether or not punctuation and star ratings have a correlation. Are complainers good with punctuation? Do they use more commas and question marks? Or do they tend to use words with punctuation (i.e. contractions like ‘can’t’)? I hypothesize that punctuation is used more by those who give lower ratings.

At first I had to make sure that I was accurately getting the data. Initially my regular expression for numbers was printing out every word, whether it was a number or not. By printing out the data I was able to verify whether or not I was accurately getting the right information. My punctuation regex worked from the beginning.

I found that both features are negatively correlated. Stars and digits have a negative correlation of -.0168286. Similarly, stars and punctuation have a negative correlation of -.126945. I was surprised that numbers and star ratings were negatively correlated. Perhaps individuals quantify their sufferings more than those who are simply excited about a great product. I correctly hypothesized that punctuation would be negatively correlated especially mainly because words that are contracted (i.e. Like ‘can’ and ‘not’ = can’t) contain punctuation.