Sam Bacon
Lab 17: Final Project Brainstorming
April 26, 2021

Favorite Labs/Assignments:

Homework 4: Missing Topic (Sentiment Analysis using NLTK)

Sentiment analysis is a process that I have wanted to learn more about for a while. I am fascinated by methods and procedures that allow us to use quantitative measurements to analyze something that is very organic and subjective, such as language. As I worked through the tutorial, I thought of a few things that I want to learn more about. Specifically, I was very interested in the TextBlob package from the challenge section I created. I was able to quantify the polarity and subjectivity of different texts. We scratched the surface of this during homework 4, and I definitely would like to dive deeper with my final project.

Homework 3: Twitter Analysis

This assignment got me thinking about what kind of data I will want to use for my project. As an aspiring statistician and data scientist, I have always been very interested in the data-collection process. The data set from this homework was my “favorite” data to work with because it seemed very current and real-world. I plan to look into ways that I could collect data on Twitter, Instagram, or other social media platforms because I feel that it is the most modern form of communication that exists. Also, I can see a lot of ways in which I could run sentiment analysis on this sort of data. I enjoyed homework 4 because it showed me different ways to sort and visualize Twitter data.

Lab 13: Clustering

Throughout this course, I have learned that I am very interested in unsupervised learning models. K-means clustering stood out to me because the algorithm is very simple to understand, and yet it can still provide powerful insights. In addition, we commonly use the Sum of Squared Error to evaluate the “precision” of the model. This is a value that I have learned a lot about in my statistics courses (typically with ANOVA), so it would be cool to use a measurement that I am very familiar with to perform an analysis that I am less familiar with.

Project Ideas:

1. *Sentiment Analysis: Comparing multiple languages*

When I was completing homework 5, I noticed that there were premade lists of stopwords in many different languages. This got me thinking about sentiment analysis across multiple languages. If I have an English sentence and translate it into Spanish, will its sentiment analysis scores remain the same, or will they change? I could also apply this to certain companies, celebrities, or any type of platform that releases data in multiple languages. For example, if Fox or CNN posts an article in English and Spanish, how do the sentiment scores compare? In theory, they should be pretty similar, but it would be interesting to examine this more closely. I have not decided on a specific data set yet, but this analysis could apply to any source that posts in multiple languages. This gives me a lot of flexibility.

2. *Sentiment Analysis: Comparing competing brands/companies*

Instead of looking at multiple languages, I could look at how the sentiments scores compare between two companies that are operating/competing in the same market. Some examples that come to mind would be 1) lyrics of top artists 2) articles from different news sources 3) ads of competing companies ...Again, there are a lot of different routes I could take on this. The hard part will be narrowing down a specific topic and data set that I am passionate about.

3. *Sentiment Analysis: Comparing TV shows*

In case you couldn't tell, I really want to do something with sentiment analysis. I saw on flowingdata.com that there are datasets with transcripts from shows such as *The Office*, *Friends*, and *Lost*. It could be really cool to compare the sentiments of different characters on the shows. I could also compare the sentiments of different seasons or episodes and see if there is some relationship between sentiment and episode/season popularity.

Final Idea

______I have decided to perform a sentiment analysis on lines from *The Office*. I will be using data from the 'Schrute' R package…
https://technistema.com/posts/introducing-the-schrute-package-the-entire-transcripts-from-the-office/

My analysis will incorporate multiple DMML tools and models that we have discussed. As of now, the two major topics that I will include are 1) Sentiment Analysis and 2) K-means clustering. I will perform sentiment analysis on the lines and obtain polarity and subjectivity values for each one. Then, I will be able to perform K-means clustering to identify any significant patterns from the sentiment values. In addition, I would like to try and train a model that could predict who spoke a given line. I have not narrowed down what type of model that would be yet.

I have looked through the data, and I feel that it is in a format that I can adapt to meet the needs of my project. I will obviously have to perform some preprocessing, but I am excited to see if this can teach me anything new about my favorite show