Project Proposal for Data 620: Web Analytics

Said Naqwe
Eddie Xu
Mohamed Hassan-El Serafi

Final Project Proposal: Twitter Sentiment Analysis of Airlines and Network Analysis of Spotify Songs

Instructions

Your proposal should describe at a high level what you're seeking to accomplish and your motivation for performing this analysis. A guiding question or hypothesis to test is one good way to start. If you are going to work in a small group (encouraged!), You should also list your partners' names.
You should briefly describe your data sources, plan for doing the work, and upfront concerns. If you are working in a group, please describe the roles and responsibilities of each group member. We'll treat this proposal as a planning document, not a blueprint containing "firm, fixed requirements."

Overview

For our final project, we will be using two separate datasets to create a sentiment analysis and network analysis. For our Sentiment Analysis, we will explore the different types of sentiments from Twitter users about airlines. For our exploratory data analysis, we will evaluate the amount of sentiment each airline received (positive, negative, neutral) and the type of reasons each airline received. We will then create word clouds for positive and negative tweet sentiments, as well as word clouds for each airline. Finally, we will predict sentiments from the tweet text data, turning the airline sentiment into binary values (0 for negative sentiment, 1 for positive). The classifiers we will use are Support Vector Machine, Decision Tree, Random Forest, Naïve Bayes, Logistic Regression, K-Nearest Neighbors, and Ada Boost. The dataset will be retrieved from Kaggle, where you can find the link to the data here: https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment. We will identify the best performing classifier by determining the accuracy of each model.

For our Network Analysis, we are considering using the same airline dataset. We would create a 2-node network using airline name and usernames. We would analyze the centrality measures of users, while also examining their relationship with the airlines based on the negative reasons column. However, we are unsure if the variables in the dataset can be appropriately used to conduct a network analysis. As an alternative, we are thinking to utilize a Spotify dataset that contains 30,000 songs. The variables in the dataset include track name, track artist, track album, playlist name, playlist genre, and track album release date. Because the dataset is large, we are planning to subset the dataset, focusing on songs from 2019. We will build a 2-node network using the name of the artist and the genre. We want to analyze the centrality measures for the artists, in particular which artists are most connected to other artists. In addition, we will examine which song genres are most connected with other genres, as well as their connections with the artists. This dataset was also retrieved from Kaggle, which you can find here: https://www.kaggle.com/datasets/joebeachcapital/30000-spotify-songs?select=spotify_songs.csv.

Based on the centrality measures, we hope to identify the artists and genre that are most influential, as well as how the artists and genre of their music are connected.