American Songbook VS Disney

Motivation

Being a huge fan of music, especially jazz, I was interested to learn more about “The Great American Songbook”. The reason I chose “The Great American Songbook” was due to the fact it is a large portion of the jazz repertoire. A large portion of the songs that were performed during the 30’, 40’s, and 50’s came from a handful of composers. Many jazz musicians of that era sang these lyrics including well known vocalists like Ella Fitzgerald, Frank Sinatra, and Billie Holiday. Not only were these songs sung by jazz musicians, but many country artists sang these songs as well. This great musical content was the popular music of their time and many artists had their own versions of these standards. I was interested to see how the lyrics that were composed during this time period compared to other genres. The original idea was going to compare it to contemporary pop music, but once I thought a little harder about the source material of “The Great American Songbook”, that wouldn’t make too much sense. “The Great American Songbook” was music that was mostly written for musical theatre and musical films. So it would make more sense to compare it to contemporary music that is written for musicals or musical films. Then came the idea of comparing it to Disney films. The Walt Disney Company has some of the most popular and well known songs in the world today. So then came the idea of seeing if I could create a model that could predict if the songs were written during “The Great American Songbook” era or for Disney.

Problem Statement

Lyrics have a lot of similarities in the content they talk about from love to heartbreak. “The Great American Songbook” has hundreds of songs that revolve mostly around these topics and so does Disney. If there are a lot of similar words like ‘love’ in musical lyrics can I create a model that is able to predict if the lyrical content is from “The Great American Songbook” or from Disney. The goal is to see what differences there are between the two genres of music.

Approach

Since the lyrics are not readily available for me to do any analysis or any modeling I first have to gather the data. Once the data has been gathered I will then need to clean and remove any unnecessary information or characters that shouldn’t be in the dataset. After finishing with the data cleaning. I will then be able to do some initial data analysis to learn a little more about the data. Then I will try as many different models to see which models perform the best. Then finally I can assess how the model did and how I could improve it from there.

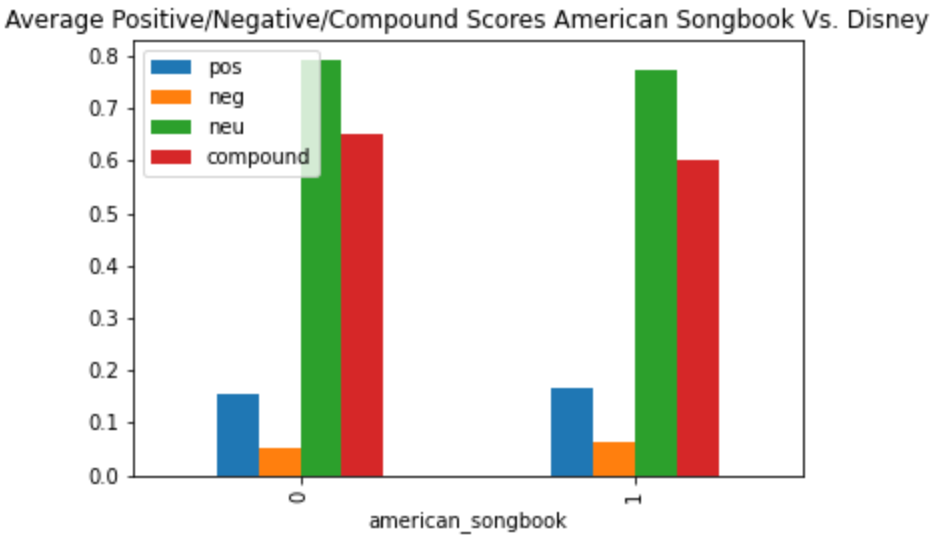
Data Collection

There were not many sources that had the data that I wanted. After doing some research I found there were some datasets that already existed for lyrics, but most of those lyrics were of popular music from the current day and going as far back as 30 years ago. So the best opportunity I had for gathering the lyrics was to scrape it off of lyrics websites. Most of my lyrics were scraped from lyricsfreak.com. I was able to gather all the lyrics I needed for "The Great American Songbook." Some of the composers I chose from this time period were Cole Porter, George Gershwin, and Irving Berlin. The largest issue I found scraping from this website was the limit of lyrics they had. Some of these songwriters had written hundreds of songs and only a fraction of the lyrics made it to this website. I looked around to see if other sites had more lyrics, but they all seemed to have the same lyrics on their website. I then had to gather the lyrics for the Disney songs. It seems as though 'lyricsfreak' wouldn't be the best source since they had a lot of missing lyrics. Not only did they have a lot of missing lyrics, but they also had poorly organized the lyrics. It was difficult to make sure I was getting the right lyrics. The best source I found for Disney lyrics was this website disneyclips.com. I used the libraries BeautifulSoup and requests to gather all the lyrics. I also used the Time function so I could put a sleep time on my requests. I wanted to be respectful to the websites I was gathering the data from.

Data Cleaning and Analysis

Once I had the data in hand I had to clean it. There were many html characters in the lyrics and there were some spacing issues that had to be remedied. The lyrics also had to be combined into one dataframe so I could easily manipulate the data. Before combining the lyrics I had to mark them with what category they belonged to. For example for "The Great American Songbook" I gave it the number "1" because that was going to be my goal for the model to be able to predict. All other songs were given a "0." I also wanted to add an extra layer to my data to differentiate between the Disney Renaissance songs and the songs that didn't belong to that period. I was interested to see how the Disney Renaissance songs performed in the model versus the other Disney Songs. Disney Renaissance songs are marked by a "1" and all other songs by a "0."

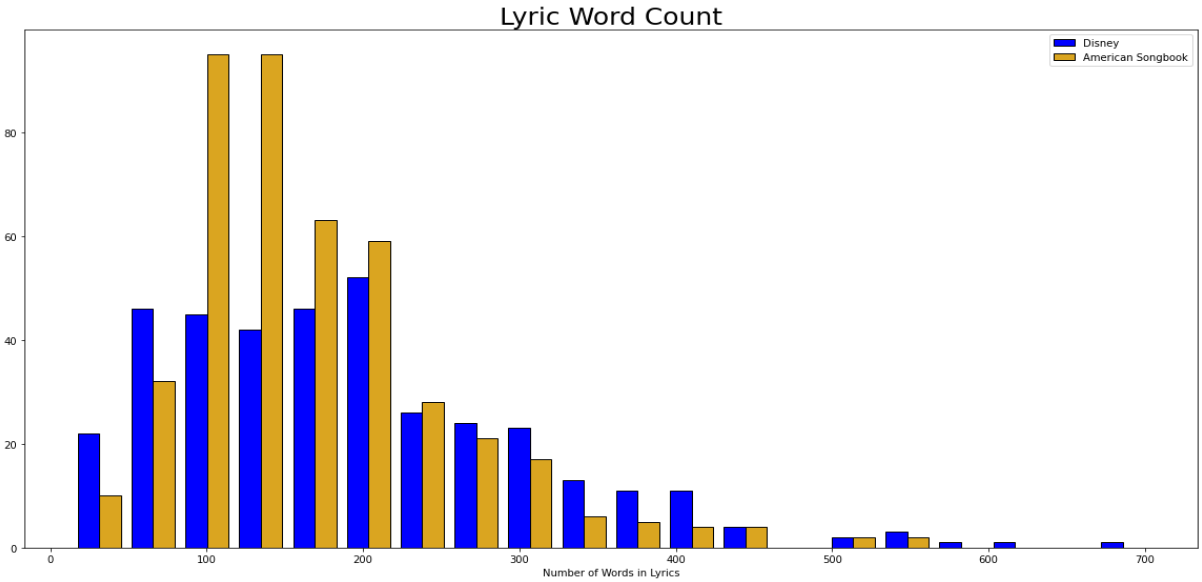
What was really interesting to learn about the differences between the two sets of lyrics was the sentiment analysis. I used the Vader Sentiment analysis as my library of choice to see how the lyrical content differed. What was amazing to see was how eerily similar the sentiment analysis was. As you can see in the image below there is not much of a difference.



0 = Disney **1 = American Songbook**

Also what was interesting was that at the top of each category (‘neg’, ‘pos’, ‘nue’, ‘compound’) from the sentiment analysis most of the lyrics were from Disney. One reason I think this happened is that I was able to retrieve all the lyrics from the Disney movies and so there was a full range of emotions that could be expressed. The lyrics I was able to grab for “The Great American Songbook” were songs that mostly were the more popular sung lyrics and sung by many musicians. The fact that only a handful of songs were used from all these composers could affect some of the outcomes of the sentiment analysis.

Looking at the word count you can also see that “The Great American Songbook” had many songs around the same number of words. Where Disney had a larger spread and most of the outliers as well.



Modeling

When it came to modeling I wanted to try as many different models I felt would be appropriate. Since I was trying to use a model to figure out between two sets of lyrics it was a classification problem. Before I could use any modeling I had to turn the words into tokens or some form that my model could use. To make the process a little easier I used a Pipeline so this process could be done in one motion. I then also used GridSearch so I could try as many different hyperparameters to see what parameters created the best model to predict between the two categories. The models I used were Logistic Regression, Random Forest Classifier, Extra Tree Classifier, AdaBoost Classifier, GradientBoost Classifier, and SVC. I also used both TFIDFVectorizer and CountVectorizer on the lyrics to see if one performed better over the other. The model that I ended up choosing because it performed the best was SVC using TFIDFVectorizer as the word tokenizer. It performed with an accuracy of 83%.

Findings

What I found to be very interesting was the percentage of what the model guessed incorrectly. 70.6% of the incorrect predictions were Disney songs falsely predicted as “The Great American Songbook”. Of those Disney songs that were predicted to be “The Great American Songbook” 50% of them were from the Disney Renaissance. This was very interesting to find out seeing that only about 30% of the songs were from the Disney Renaissance. Looking at some of the coefficients you can see why some songs were predicted incorrectly. For example the third word in the coefficients is the word ‘man’. One of the songs that was predicted incorrectly was the Mulan song “Make a Man Out of You.” The lyrics are littered with the word ‘man’ and it would be very difficult for the model to differentiate between the two lyrical styles.

Since this was a classification model I wanted to see some of the classification metrics. What was nice to see was the model had a pretty high sensitivity rate at 90.9%. It was pretty good at guessing the songs that were from “The Great American Songbook.” The specificity score was a lot lower at 74.2%. Looking at what the model guessed incorrectly this makes sense because most of the songs it predicted incorrectly were Disney songs.

Risks/limitations/assumptions

Some of the risks in this project is trusting the lyrics are all correct from the websites. It looks as though the lyrics are submitted by people online who want to add lyrics and the content might not be checked to see if it is reliable. One example is when trying to scrape lyrics for Mary Poppins on ‘lyricsfreak’ and one of the lyrics posted was actually an Avril Lavigne titled “S8ter Boi.” The lyrics could also include dialog or character names to signify who is singing the part.

The limitations for this project is that the amount of lyrics out there are limited. You are limited to the lyrics that have been written and are readily available. Especially in the sense of Disney lyrics you can only work off the lyrics from the movies that have been made. The other limitation is that not all the lyrics from “The Great American Songbook” are freely available. Only a fraction of all the lyrics that were written during this time period are available online. The best option would be to purchase books with the full lyrical discography or buy all the separate sheet music and manually type them all in.

Summary

This was a very interesting project to take on. It was interesting to learn more about two sources of music that I enjoy very much. “The Great American Songbook” is a great source of some of the most well-known and widely used music in jazz and pop culture. Disney is a behemoth producing some of the most popular songs in the world. Most recently Disney had a huge hit with Frozen’s “Let It Go.” It would be interesting to see how much better the model could have done with more source material. This also could be an interesting start to see if you could create an AI that could try to create lyrics in the style of “The Great American Songbook.” I was only able to pull a fraction of the amount of music that was produced during that time period. So if more lyrics could be pulled then the model would have more information to train on. There is a big difference between the two musical genres, but it was also interesting to see how similar some characteristics were to each other.

Sources:

“The Great American Songbook”:

- https://thesongbook.org/
- https://archive.org/details/americanpopulars00alec/page/23
- https://en.wikipedia.org/wiki/Great_American_Songbook#Songwriters_and_songs
- https://www.udiscovermusic.com/in-depth-features/cover-to-cover-the-story-of-the-great-american-songbook/

Disney and the Disney Renaissance:

- https://en.wikipedia.org/wiki/Disney_Renaissance
- https://www.youtube.com/watch?v=JX0gZY9VKlM