STAT 628 Module 4 Summary

Introduction & Data Cleaning

The goal of this project was to develop 2 metrics for Spotify's podcasts, and provide recommendations of similar episodes based on these metrics. We first collected the data using Spotify's web API to find episode IDs for the top 50 podcasts in the US, according to Spotify Charts during the week of 11/25/24. For each show, episode data was gathered until 200 episodes were found, or until the API had been called 6 times, whichever occurred first. Data cleaning was done to change the names and descriptions to lowercase and remove URLs, punctuation, and stop words, and to combine the names and descriptions into one feature. Episodes from shows that do not have descriptions or only have irrelevant information (such as social media contacts) were removed manually, and episodes were assigned types based on Spotify's Podcast Genre Charts to help understand the value of the metrics developed. The final dataset had 10,970 podcast episodes.

Data Analysis & Results

Our plan was to use word counts in the description and name to create metrics. Initial attempts to use Principal Component Analysis to develop the metrics struggled due to the sheer size of the dataset - the prcomp() function ran for a very long time. So we reduced the word counts considered to only the top 200 most common words across all of the podcasts. This shrunk the sample set considered, allowing for quicker results, but likely had a negative impact on our results, since uncommon words may be the deciding factors in clustering more distinct genres (e.g., history or fiction).

After completing PCA, we chose the first two principal components as our metrics: PC1 and PC2. PC2 clusters podcasts intended for kids well, while PC1 separates out a segment of the news and politics podcasts (Fig. 1). Some words with high loading vectors for PC2 include 'read', 'aloud', 'book', and 'kids'. For PC1, key words were 'minnect', 'david', 'bet', and 'valuetainment'. Reviewing our dataset, only 1 podcast was classified as 'kids' - a podcast reading kids books aloud. This explains why PC2 is so effective at clustering these episodes. For PC1, Patrick Bet-David hosts a podcast called 'Valuetainment' and seems to often be a guest on news & politics podcasts.

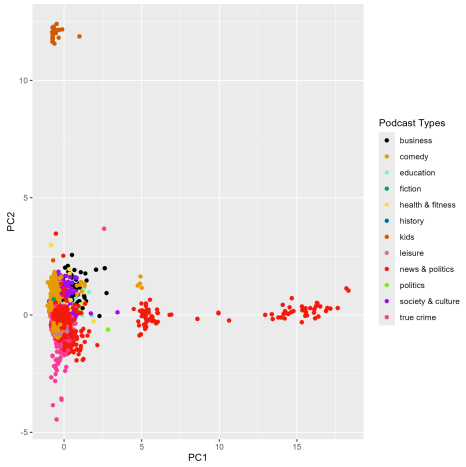

Fig. 1: Visualization of PCA with Podcast Types

Conclusion

Our metrics provide some insight into ways podcasts can be clustered and subsequently classified, but their effectiveness is limited by the lack of diverse data in our dataset. Many of the most popular podcasts in the US have similar topics, meaning genres are underrepresented in our dataset (e.g., kids podcasts). Additional cleaning and data could improve the metrics created.

STAT 628 Module 4 Summary

References & Contributions

Spotify. (n.d.). Web API | Spotify for Developers. Developer.spotify.com. https://developer.spotify.com/documentation/web-api

Spotify. (2024). The Podcast Charts - Spotify. The Podcast Charts - Spotify. https://podcastcharts.byspotify.com/

Contributions	Amy Merkelz	Yifan Zhang
Summary	Responsible for introduction, data cleaning, and analysis sections, and Figure 1.	Responsible for the conclusion section.
Code	Responsible for data pulling, cleaning, and analysis code.	Reviewed data pulling code, reviewed analysis code and approach.
Shiny App	Provided feedback on the Shiny app.	Responsible for the Shiny app.