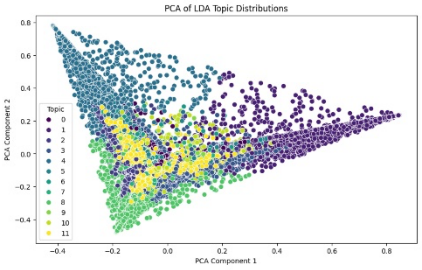
Summary for 628 Module 4—Spotify Podcast
Group 16 Tianyu Yao/ Kangxin Zheng

Data Collection: We used the URL, client ID, and client secret to retrieve an authentication token. Upon inspecting the podcast website, we identified 13 categories, and we pulled 50 shows for each category. A function was then created to retrieve episode details, such as episode ID, name, duration, and release date, based on the podcast data obtained in the previous step.

Challenges : We encountered a 429 error, indicating that we had exceeded the rate limit by making too many requests within a 30s window. To address this, I checked the retry policy and implemented a time.sleep() function to prevent this issue in subsequent requests.

Data Cleaning: We removed URLs, email addresses, and punctuation. All text was converted to lowercase.Stop words and manually selected ones were excluded using the NLTK package.



Modeling: We employed Latent Dirichlet Allocation (LDA) to assign topics to podcast descriptions.

Topic Selection: To determine the optimal number of topics for the LDA model, we calculated the perplexity scores for various topic counts. We selected topics=12 based on it.We used CountVectorizer to convert the podcast descriptions into a document-term matrix. Following this, we trained the LDA model with 12 topics, which represented the following themes: Politics & Election, Business & Success, Sports & NFL, Health & Wellness, etc.

Dimensionality Reduction with PCA: Next, we applied PCA to the topic distribution data from LDA. We chose n=9, as the explained variance ratio reached 90% when n=9.

Metric Construction: We used the PCA-transformed values to build metrics. For example, for a given podcast, we might have topic probabilities such as Topic_0_Prob = 0.1, Topic_1_Prob = 0.3, ..., Topic_11_Prob = 0.2. The PCA transformation maps these probabilities into nine new values that represent the podcast's position along the principal axes of topic variation. The high explained variance ratio (90%) demonstrates that the PCA reduction is efficient and informative.

Strength and Weakness: The resulting metrics are straightforward to generate and are highly representative of the underlying data, facilitating both analysis and visualization. Interpreting the individual components of the PCA can be challenging, as they often combine several topics, leading to less interpretable results. Also, PCA may reduce the direct clarity of insights.

Dashboard Using Shiny: https://tianyuyty.shinyapps.io/628-mod4-SpotifyPodcast/

The interactive Shiny application provides the following key functionalities:

1. **Word Cloud**: Highlights prominent keywords in a selected episode's description.
2. **Topic-Based Clustering**: Visualizes podcast episodes using PCA-transformed values, with clustering results and clear episode highlights.
3. **Nearest Episodes**: Displays the top 10 nearest episodes based on PCA proximity, offering insights into content similarity.
4. **Topic Information**: Displays the dominant topic of the selected episode for better understanding.

Conclusion

This dashboard combines **topic modeling**, **PCA**, and **clustering** to deliver an insightful and user-friendly platform for analyzing podcasts. It enables both **quantitative analysis** and **visual exploration** of podcast episodes while highlighting content similarities and topic distribution.

References:

[1] Ben Marwick. (n.d.). Interactive PCA Explorer. Retrieved from https://github.com/benmarwick/Interactive_PCA_Explorer

[2] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.

[3] TidyTextMining. (n.d.). Topic Modeling. Retrieved from https://www.tidytextmining.com/topicmodeling.html

Contribution:

Kangxin Zheng: Responsible for data collection, data cleaning and modeling, as well as these parts in summary.

Tianyu Yao: Responsible for whole shiny app part, as well as shiny app part in summary.