Report - Group 7

Yuchen Xu, Yudi Wang

I. **Background Information/Data Cleaning**:

This project analyzes podcast episodes using descriptive text data. The dataset originally contained 11 columns, including key fields such as show_name, name, description, category, and show_id. Key cleaning steps included:Verifying and addressing duplicates in the dataset, Tokenizing description text, removing stop words for meaningful analysis. The cleaned dataset comprises 20 categories and 368,955 episodes.

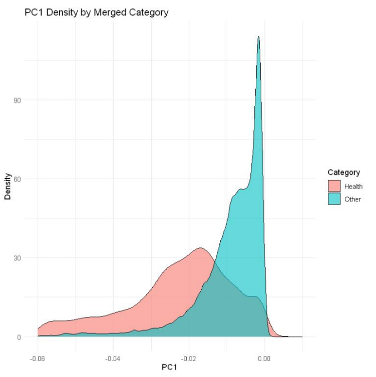
II. **Metrics**

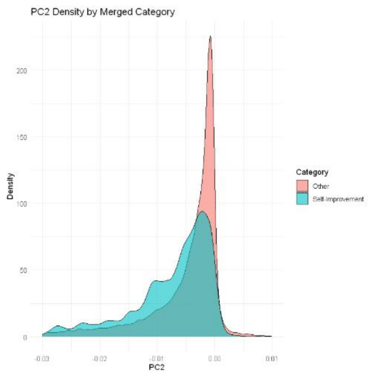
Metrics are derived using Principal Component Analysis (PCA) on the TF-IDF weighted document-term matrix (DTM) of episode descriptions. The DTM captures each episode as a vector of its 10,000 most frequently occurring words. PCA reduces the dimensionality to 10 principal components, each representing a distinct dimension of variance in the dataset. The first two principal components together explain approximately 50% of the variance in the dataset, indicating that they capture a significant portion of the information in the original high-dimensional space.

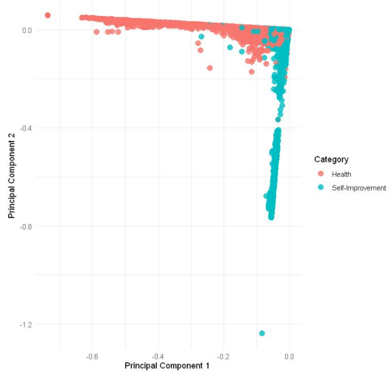
PC1 Score: Reflects the primary theme of episodes. Higher PC1 scores align with general topics, while lower scores correspond to focused themes like health, as shown in the density plot distinguishing "Health" and "Other".

PC2 Score: Highlights secondary thematic variations. Episodes with lower PC2 scores align with specific subcategories like "Self-Improvement," while "Other" categories exhibit broader content diversity.

This scatter plot visualizes Health and Self-Improvement episodes using PC1 and PC2, showing a clear separation in overall distribution, though some overlap remains in the central region.


PC1 Density by Merged Category


PC2 Density by Merged Category



III. **Strengths & Weaknesses, Conclusion**

Using PCA as a basis for metrics simplifies analysis by reducing thousands of features into interpretable dimensions. Metrics like PC1 offer insights into both thematic focus and diversity, making them suitable for clustering, recommendations, and content analysis. However, PCA assumes linear relationships among words and may lack semantic clarity due to the abstract nature of principal components. Also issues like overlapping names between show_name and episode_name during tokenization may introduce ambiguities affecting metric accuracy. Despite these limitations, PCA-based metrics are scalable and robust, ensuring applicability to large and diverse podcast datasets..

Contributions:

	Yuchen Xu	**Yudi Wang**
Code	Responsible for data cleaning code and building metrics.	Responsible for geting raw data; Reviewed metrics.
Summary	Introduction, data cleaning sections, metrics	Strengths&weaknesses, conclusion sections
Shiny App	Reviewed and Beautification	Frame building and basic function realization

References:

[1] C.L. Clayman, S.M. Srinivasan, R.S. Sangwan, K-means clustering and principal components analysis of microarray data of L1000 landmark genes, Procedia Comput. Sci., 168 (2020), pp. 97-104

[2] N.F. Jansson, R.L. Allen, G. Skogsmo, S. Tavakoli, Principal component analysis and K-means clustering as tools during exploration for Zn skarn deposits and industrial carbonates, sala area, Sweden, J. Geochem. Explor., 233 (2022), Article 106909