Spotify Cluster Prediction

Zhaoqing Wu, Yunze Wang

1 Introduction

Spotify is the largest platform for music and podcasts. This project aims to enhance user experience by analyzing the podcast section using statistical and machine learning models. We propose 2 metrics to cluster users by preferences and recommend similar podcasts.

2 Data Preprocessing

Using Spotify's API, we collected podcasts across 16 categories, each capped at 20 episodes, totaling 14,665 entries. Data cleaning removed irrelevant elements (e.g., emojis, URLs). We used NLTK [1] for tokenization and stop-word removal, retaining only adjectives and nouns for higher data quality.

3 Model: BTM, PCA, and KMeans

The cleaned dataset was processed using the BTM [2], a model specifically designed for topic modeling in short texts. BTM successfully identified 600 highly relevant keywords for each category. These keywords were then utilized to construct a word frequency matrix for each podcast episode, which is shown in Fig. 1.

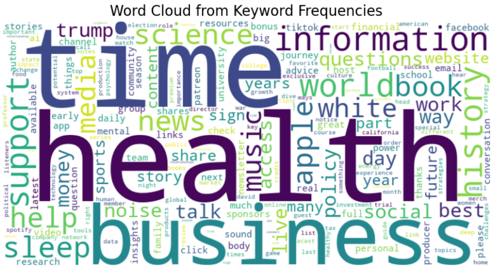

Figure 1: Word cloud of the Top 600 Keywords.

Using the 600 keywords, a word frequency matrix was generated for all episodes. To ensure compatibility with PCA [4], a logical column was added to the matrix to avoid rows containing only zeros. This resulted in a $14,665 \times 601$ matrix. The matrix was standardized using the StandardScaler and subsequently reduced to two dimensions via PCA, yielding two principal components: PC1 and PC2. For newly selected episodes, we processed the data similarly and applied KMeans clustering [3] to group episodes based on similarity with the existing dataset, providing personalized recommendations.

4 Model Interpretation and Analysis

The 600 keywords were ranked based on their contributions to PC1 and PC2. The results are summarized in Table 1. The top-ranked keywords provide meaningful insights. For instance, the keyword 'un' likely refers to the United Nations, which aligns with the 'news' category. This demonstrates that the proposed metrics are consistent with real-world semantics. Moreover, PCA effectively reduced the dataset's dimensionality while retaining variance, producing orthogonal components without multicollinearity issues.

Table 1: Top Parameters for PC1 and PC2

PC1		PC2	
Keyword	Score	Keyword	Score
un	0.160	appropriateness	0.164
noise	0.159	illustrative	0.164
white	0.156	coltivar	0.164

5 Strengths and Weaknesses

Strengths: The BTM model ensures representative data and meaningful results.
Weaknesses: BTM is computationally intensive, limiting its use in lightweight applications.

6 Contribution

- **Zhaoqing Wu**: Model building, coding, Shiny App development
- **Yunze Wang**: Web scraping, data cleaning, model building, coding, summary writing, GitHub maintenance

References

[1] Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, 2006.

[2] Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):2928–2941, 2014.

[3] J MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*, 1967.

[4] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.