Spotify Podcast New Metric Project

1. Introduction

This project uses Spotify podcast data to create new metrics by analyzing episode titles and descriptions, providing personalized recommendations to users.

2. Data Cleaning

Podcast data retrieved from the Spotify API includes program names, descriptions, and related metadata. The following cleaning steps were performed to ensure the accuracy of the analysis:

Handling Missing Values: Podcasts with missing or malformed descriptions were excluded.

Text Preprocessing: Converted text to lowercase, removed punctuation, numbers, stopwords, and irrelevant phrases.

Feature Extraction: MUsed TF-IDF to generate a feature matrix and applied PCA to reduce its dimensionality, extracting the top 10 components for simplicity.

3. New Metrics

The following key metrics were developed to quantify the characteristics of podcast content:

Sentiment Polarity: Classified words in descriptions as positive or negative to help users understand the emotional tone of the podcast content.

Novelty: Highlights content creativity, enabling users to discover unique podcasts.

Hearability Complexity: Helps understand language style and complexity, catering to audiences with different reading preferences.

4. Clustering Methodology

Clustering analysis was performed using the K-means algorithm to group podcasts into distinct categoriesFeature **Selection**: Metrics such as sentiment polarity, novelty, and text complexity were used as clustering inputs.

Determining Cluster Count: Specified the number of clusters (K).

Clustering Computation: Podcasts were grouped into K clusters based on the Euclidean distance between feature values.

Visualization: The first two principal components from PCA were plotted, with different colors representing each cluster.

5. Clustering Results

K=3 Clusters:

Cluster 1: Simple language, neutral sentiment, average novelty – for straightforward content seekers.

Cluster 2: Positive sentiment, complex language, high creativity – for innovative content enthusiasts.

Cluster 3: Serious tone, intricate text, in-depth content – for deep discussion lovers.

Figure 1: K=3 Clusters

K=4 Clusters:

Cluster 1: Simple language, neutral sentiment, low novelty – for casual content listeners.

Cluster 2: Positive sentiment, rich language, complex structure – for energetic, dynamic content fans.

Cluster 3: High novelty, neutral sentiment, moderate complexity – for unique but simple content explorers.

Cluster 4: Negative sentiment, structured text, longer sentences – for academic or socially focused listeners.

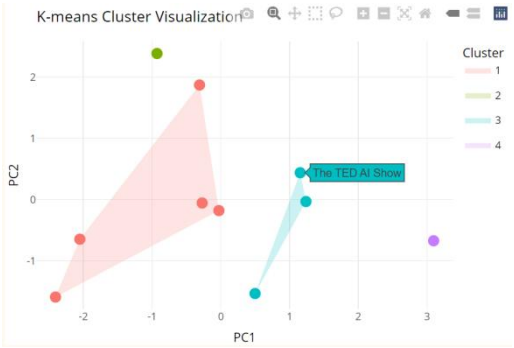

Figure 2: K=4 Clusters

6. Conclusion

This project analyzes and clusters Spotify podcasts based on sentiment, novelty, and listening complexity. It provides personalized recommendations, uncovers content patterns, and offers insights for users and creators via the Shiny web app: (https://andrewchanshiny.shinyapps.io/Spotify/).

Contributions

Parts	**Leyan Sun**	**Zhengyong Chen**
Summary	Responsible	Check and add some parts
Code	Responsible for cluster analysis	Responsible for data
Shiny App	Check and Format adjustment	Responsible

Table 1: Team Contributions