Podcast Clustering: Exploring Duration and Thematic Insights

Introduction

This project aims to construct two new, distinct, and informative metrics for podcast episodes (Spotify for Developers, n.d.) to enable effective clustering and comparison. Overall, the first metric categorizes episodes by duration into five groups using KMeans clustering, while the second metric identifies thematic clusters using LDA (Latent Dirichlet Allocation). The Shiny app built for this project allows users to select one or more podcasts, visualize their metrics, and compare them to identify similarities between episodes or podcasts.

Time-based metric

To construct the time-based metric, I applied KMeans clustering on podcast episode durations and categorized them into five groups: short, short-medium, medium, medium-long, and long. The KMeans model identified the following cluster boundaries: 39.68, 80.04, 142.19, and 375.81 minutes. These values reflect natural separations in episode lengths and align with intuitive expectations about podcast durations.

This metric is informative because it creates clear, data-driven distinctions in duration, enabling more granular analysis compared to simple thresholds like "under 1 hour." However, KMeans assumes uniformity within clusters and may oversimplify subtle differences in episode lengths.

Latent Dirichlet Allocation

We use Latent Dirichlet Allocation (LDA) to construct topic-based metric. LDA is particularly well-suited for analyzing unstructured text, as it assigns each document (in this case, a podcast episode) a distribution over a set of predefined topics, where each topic is represented by a group of highly probable words.

In this project, we applied LDA to the descriptions of podcast episodes, using it to classify episodes into distinct thematic categories. The meaning of each category was defined by its representative “hot words.” For example, a category with words like “sleep”, “health”, and “relax” was labeled as Health. This process ensures that each topic has a clear and interpretable semantic meaning. In total, we classified the podcasts into 8 categories based on their dominant themes. These categories include:

Category 1: Story category, represented by hot words such as *stories*.

Category 2: Politics category, represented by hot words such as *trump*.

Category 3: Crime category, represented by hot words such as *murder* and *crime*.

Category 4: Health category, represented by hot words such as *sleep*.

Category 5: Non-English category, represented by hot words such as *vida* and *redes*.

Category 6: Business category, represented by hot words such as *business*.

Category 7: Sport category, represented by hot words such as *football* and *nba*.

Category 8: Comedy category, represented by hot words such as *comedy*.

LDA offers a structured and data-driven approach to classify podcast episodes by uncovering latent patterns in large text datasets. It effectively handles diverse topics, reduces manual bias, and ensures intuitive, actionable categories using representative words for each topic. This topic-metric organizes episodes into meaningful groups, supporting thematic analysis, content organization, and comparative insights across podcasts.

Spotify for Developers. (n.d.). *Get multiple shows*. Spotify. Retrieved December 7, 2024, from https://developer.spotify.com/documentation/web-api/reference/get-multiple-shows

Jelodar H, Wang Y, Yuan C, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey[J]. Multimedia tools and applications, 2019, 78: 15169-15211.

Maier D, Waldherr A, Miltner P, et al. Applying LDA topic modeling in communication research: Toward a valid and reliable methodology[M]//Computational methods for communication science. Routledge, 2021: 13-38.

Contributions	Minyuan Zhao	Minliang Yu
Summary	Topic-based metric	Introduction Time-based metric
Code	Topic-based metric	Get data from api Time-based dimension
Shiny App	Shiny app Review	Shiny app code
Percentage	50%	50%