Spotify Podcast New Metric Project

Introduction

This project aims to use Spotify's podcast data to develop new metrics for each podcast episode. New metrics are constructed through the analysis of important textual information about podcasts, such as the name and description of each episode of the podcast, and these metrics will provide users with more personalized recommendations. The goal is to enhance the user experience by personalized recommendations, ultimately providing a more engaging and satisfying listening experience.

Data Cleaning Process

The raw data contains 7 variables, the 7 variables are id, name, description, duration_ms, release_date, show_id, duration_min, description_length.Since we mainly want to create new metrics for each podcast episode based on the content, we only focus on name and description that can provide us with information about the podcast content, and edit these two variables collectively into a "text" variable to facilitate subsequent cluster analysis. The “text” variable is turn into corpus, and all its contents, lowercase, remove punctuation, remove numbers, remove spaces, and then generate a Document-Term Matrix, and apply TF-IDF (Word Frequency-Inverse Document Frequency) weights to it to extract features for the subsequent clustering analysis.

Creating two new metrics

First, using the K-means algorithm in cluster analysis, our TF-IDF matrix is divided into five clusters; the reason for the five clusters is to both allow each cluster to better and more accurately show similar content, as well as to allow for a sufficient sample size to generalize to more similar podcasts.After completing the cluster analysis, I proposed two clusters, cluster “1” and cluster “5”, for which I created corpus respectively again to facilitate the subsequent deletion of some unrepresentative words.The document-word matrix is then generated again to summarize the high-frequency words and TF-IDF weights are applied to them to extract features as a measure of the importance of a word for clustering. Then we created data visualizations to show the top 20 high-frequency words and the top ten words with the highest importance for clustering to summarize a metric that could describe this type of podcast in terms of those words. I created two new metrics, metric 1 should be Technology shareability, and metric 2 should be Scientific innovativeness. Since most of our dataset is using tech podcast content, our new metrics can be better for users who like tech podcasts to do more segmentation filtering, such as if users like technology applied to life, such as how chatgpt should be used in our work life, the advantages and disadvantages of some use of robot dogs, how AI can better help ordinary people, then it can be focused on the Focus on metric 1 which is Technology shareability.On the contrary, if users are more interested in information about breakthroughs and innovations in the scientific community, such as current AI innovations in convolutional neural networks, which cutting-edge algorithms have been optimized, and whether brain-computer interface projects are getting new breakthroughs, then they can focus on the metric 2, which is Scientific innovativeness.

Strengths and weaknesses of the metrics

Since my metrics are summarized from the high-frequency words for each cluster and the few words with the highest importance to the cluster, they can be a good generalization of similar podcasts. The advantage is simple, strong usability, according to some high-frequency and important words can be easily classified, and allows for a good generalization of similar content. The disadvantage is that the accuracy could be improved.

Conclusion

This project extracted content features from Spotify podcast data, developed two new metrics, and summarized podcast clusters through clustering analysis. This can be very good to help our users go to better enhance their personalized choices and tap into the business value of different podcast topics.

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