Module 4: Spotify Metric

We developed a highly informative metric that is both easily scalable and computationally efficient. In this report we will detail the construction process of our metric and provide justification for its design. Finally, we present the results of the clustering analysis.

1 Word2vec embeddings

In this program, our design is based on word2vec embedding [2]. The effectiveness of word2vec embeddings has been demonstrated in various applications. We embed each word into 300-dimension space and use the mean of words' embedding to denote the description we derived from spotify API.

2 Construction of different style metrics

To ensure the metric is highly informative, we constructed it based on the styles of content, such as motivating, thrilling, or poetic. Specifically, we began by selecting representative texts that exemplify these styles and computed their average embedding vectors using a pre-trained Word2Vec model (word2vec-google-news-300). For each episode, we then calculated its similarity to these representative vectors using cosine similarity, which quantifies the alignment of semantic content. To normalize the distribution of scores to approximate a Gaussian, we applied the Box-Cox transformation to the similarity values, yielding our final metric. The parameters for this transformation were pre-determined.

We selected three metrics to focus on: Natural Science, Finance and Market, and Thrilling. For example, the episode "SC EP:489 Two Strange Encounters" has a Natural Science score of 0.5704, a Finance and Market score of 0.4692, and a Thrilling score of 0.9416.

3 Justification of our metric

This approach ensures that the metric captures the thematic essence of the content while being interpretable and meaningful for categorization. It is worth noting that although Spotify provides a rough classification for each podcast, it does not classify individual episodes. Moreover, our metrics not only fill this gap but are also more detailed than Spotify's classification. By quantitatively analyzing the content styles of each episode, we offer a more precise categorization capability. Since this construction is quite straightforward, the style of a specific episode could be directly obtained by the score.

Compared to PCA-based metrics, our metric offers significantly enhanced interpretability, greater informativeness, improved scalability, and higher computational efficiency.Additionally, since PCA heavily relies on the selected sample data, its performance may be prone to bias in practical applications. In contrast, our metric relies only on representative samples, making it highly adaptable and easy to adjust. Below are the K-Means [1]clustering results derived from our metric.

(a) Cluster based on finance and thrilling

(b) Cluster based on science and finance

(c) Cluster based on science and thrilling

Figure 1: K-Means Cluster Outcome

Contributions

	Xiangsen Dong	**Xupeng Tang**
Report	Main body of report	Reviewed/edited and provided feedback
Code	Construction of metrics and code realization	Data cleaning and preprocessing
Shiny	Reviewed/edited and provided feedback	Shiny App Implementation

Table 1: Team Contributions

References

[1] Abiodun M. Ikotun, Absalom E. Ezugwu, Laith Abualigah, Belal Abuhaija, and Jia Heming. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622:178–210, 2023.

[2] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.