- **Constructed metric**

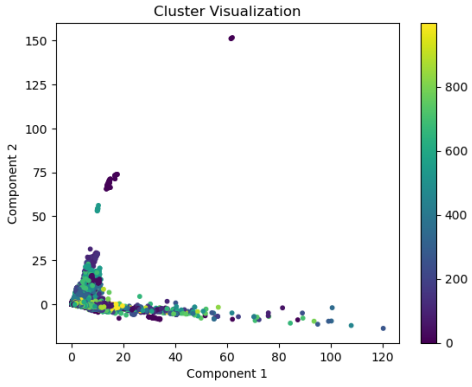
We extracted 514,590 episodes' descriptions from 4,538 shows. After word segmentation, we removed words that appeared fewer than ten times across all descriptions, and common terms (e.g., "shows" and "episodes"). After that, we got 34,223 unique words. Using these words, we constructed a word-count feature matrix with 34,223 features for each description.
Next, we applied Randomized SVD to this word-count feature matrix and selected the top 87 principal components (PCs) as our new metrics to evaluate episode features. However, here we chose to focus on the top 2 PCs as an example. Based on their factor loadings, we identified the first PC (top loadings included "get," "us," "visit," "free," "new") as "**Marketing**." The second PC (top loadings included "sleep", "noise", "sounds", "white") was interpreted as "**Sleeping**."

- **Explanation of the metric with example**

Here are two examples illustrating how the two metrics are applied. 'Marketing' is evaluating whether an episode is trying to commercialize some products. 'Sleeping' metrics is evaluating whether it's a sleep aid episode. Higher value for 'Marketing' means episode contained more marketing information in the episode. And higher value for 'Sleeping' represents episode mainly about sleeping topics, might be sleeping-aid episode. First example is: ***"Don't forget to check out Dax Shepard's own podcast Armchair Expert available wherever you get your podcasts! … Bert and Dax talk all about Dax's podcast Armchair Expert and how he made a splash in the podcast world, even though he entered it somewhat late in the game…"*** For this episode, the PC1 value is 4.38, while mean PC1 value among all description data of 1.85, PC2 value is -1.05, while the mean PC2 value is -0.13. According to our metric, high value of PC1 suggests that this episode may contain many marketing-related content. Another example is: ***"There are songs you want to groove to, and songs you want to pump yourself up with. But sometimes, you just need something for winding down at the end of a long day. If you're having trouble sleeping in silence or are just looking for new tracks to add to your bedtime rotation, we've got you covered."*** The PC1 and PC2 values for this episode are 0.92 and 0.11, respectively. Based on high PC2 value compared with the mean PC2 value, we infer that this is a "help-sleep" episode to help people relax.

- **Cluster to Identify the Nearest Episodes**

We applied K-means clustering with 1,000 clusters to group the episodes. For any given episode, we first identify the nearest cluster center and then find the closest episodes within that cluster. The plot shown here uses only the top two PCs. So many clusters appear concentrated near the origin. Because we are using a total of 87 PCs, and most clusters are better separated when considering the additional PCs.



- **Strengths and weaknesses of the metrics**

Our metrics have certain limitations. First, we were unable to remove all marketing text, which resulted in several PCs being dominated by marketing-related terms. Then, our feature set, consisting of around 30,000 words, may have reduced the effectiveness. Another limitation is for some PCs (like PC2) the negative loadings focus on some meaningless words count, making it hard to explain meaning of negative PC values. However, our approach still offers strengths. The matrices derived successfully capture the main features of each episode and given meaningful interpretation metrics. By using the top 87 PCs, we can explain 30% variance in the data, and the top 2 PCs can explain about 5%, compare with other PCs, top PCs can explain more information. Overall, our metrics are informative and provide a useful, interpretable representation of the underlying patterns in the episodes.

Contributions	Meiyi Yan	Xinrui Zhong
Summary	Responsible for four parts.	Reviewed and provided feedback. Made plot.
Code	Extracted data. Cleaned data. Responsible for feedback.	Extract data. Cleaned data. Responsible for word count. Responsible for PCA. Responsible for cluster.
Shiny App	Responsible for Shiny app. Reviewed/edited and provided feedback on Shiny App.	Responsible for Shiny app. Reviewed/edited and provided feedback on Shiny App.