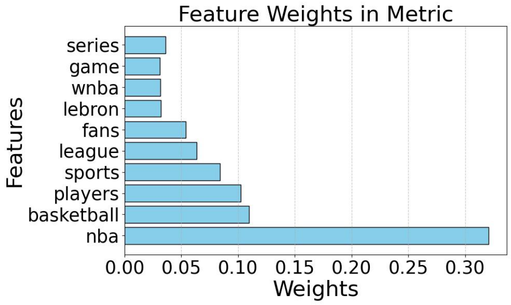
Spotify Metrics Report

Introduction/Motivation: This report aims to develop two informative metrics to analyze and interpret Spotify's podcasts. The approach involves tokenizing podcast descriptions to extract meaningful textual features and applying dimensionality reduction techniques, such as Principal Component Analysis (PCA) or Truncated Singular Value Decomposition (Truncated SVD), to identify patterns and insights within the data.

Data Cleaning: we pull the descriptions of podcast episodes by spotify api and clean it by removing punctuation, urls and repeated text.

Metric:

1. The metrics are constructed by tokenizing podcast descriptions to extract textual features and then applying dimensionality reduction techniques, such as PCA or Truncated SVD, to compute principal components or singular vectors that capture the most significant patterns in the data.
2. PCA and Truncated SVD create different metrics because PCA maximizes variance to capture dominant patterns, while Truncated SVD focuses on raw feature relationships, highlighting distinct data characteristics.



3. For example, in a podcast about basketball, terms like "NBA," "basketball," "WNBA," "player," and "LeBron" might have higher frequencies, indicating their significant influence on the metric.
4. The metric is informative as it captures key patterns in podcast descriptions, summarizing high-dimensional data effectively. Compared to Truncated SVD, PCA assigns higher weights to critical features, ensuring better focus on essential aspects, while Truncated SVD produces noisier, less interpretable results.

Assumption:

1. The features are standardized and scaled appropriately.
2. The most relevant information in the data is captured by the variance.
3. The relationships between features can be approximated linearly.

Advantages:

1. Dimensionality reduction: PCA effectively summarizes high-dimensional data into fewer components, capturing the most significant variance.
2. Interpretability: The principal components provide clear insights into feature contributions, making it easier to analyze patterns in podcast descriptions.
3. Focus on critical features: PCA emphasizes features with the highest variance, ensuring key information is retained.

Disadvantages:

1. Linearity assumption: PCA assumes linear relationships between features, which may not capture complex nonlinear patterns.
2. Sensitivity to scaling: The results depend on feature scaling, requiring careful preprocessing.
3. Loss of information: While focusing on variance, PCA may overlook smaller, yet meaningful, patterns.

Summary:

This project develops two informative metrics for analyzing Spotify podcasts by tokenizing descriptions and applying PCA and Truncated SVD for dimensionality reduction. PCA captures dominant patterns through variance maximization, while Truncated SVD highlights raw feature relationships. The PCA-based metric is preferred for its clarity and focus, offering actionable insights into podcast themes.