Spotify Tracks Dataset Report

Student Nr. 311511857 – Eduardo Nunez (凡艾爾)

Overview

This visualization analyzes the effect of certain characteristics energy, danceability, speechiness, acousticness, instrumentalness and liveness with values between [0-1] and their importance towards placing a song on the current top rankings in Spotify.

For this analysis, the top-n tracks are compared directly with the rest of the data base (around 114k tracks!) in terms of the previously mentioned characteristics. The visualization is aided by a normalized histogram which reveals the distribution of the features along their domain. The correlation of a certain feature with the ranking can be derived from the histogram by observing how much area of the histogram overlaps. The bigger the overlapping area, the less relevant the feature is for the ranking. Features with less overlapping area mean that top songs differ in this aspect from their lower-ranked counterparts.

Secondly, a pie chart provides an easy-to-read overview of the importance of each feature, derived from the previously mentioned histogram overlap. Last, one can observe the statistical distribution of median and quartiles for each feature to obtain the final details of the visualization.

Usage Example and Analysis

In this section we illustrate how the provided graphs and information can be interpreted to draw conclusions.

As the instructions state, after selecting a specific number for the top-n songs, the histograms can be analyzed. In this example, we focus on three features (danceability, liveness and speechiness) and draw conclusions from them. Similar data analyses can be done for all other features.

By observing the danceability histogram, we can clearly observe that the top 116 tracks of the dataset are clearly located above the middle value of 0.5, with a significant peak around 0.8. In comparison, the remaining tracks on the dataset span the entire range but are slightly skewed towards the upper half of range (0.5 and above). From this we can already say that current top tracks have slightly higher danceability values in comparison to the rest. By judging the amount of overlap between the histograms, we can observe that danceability distribution clearly differs between the two sets, thus the correlation between the danceability value and the song ranking is expected to be somewhat higher.

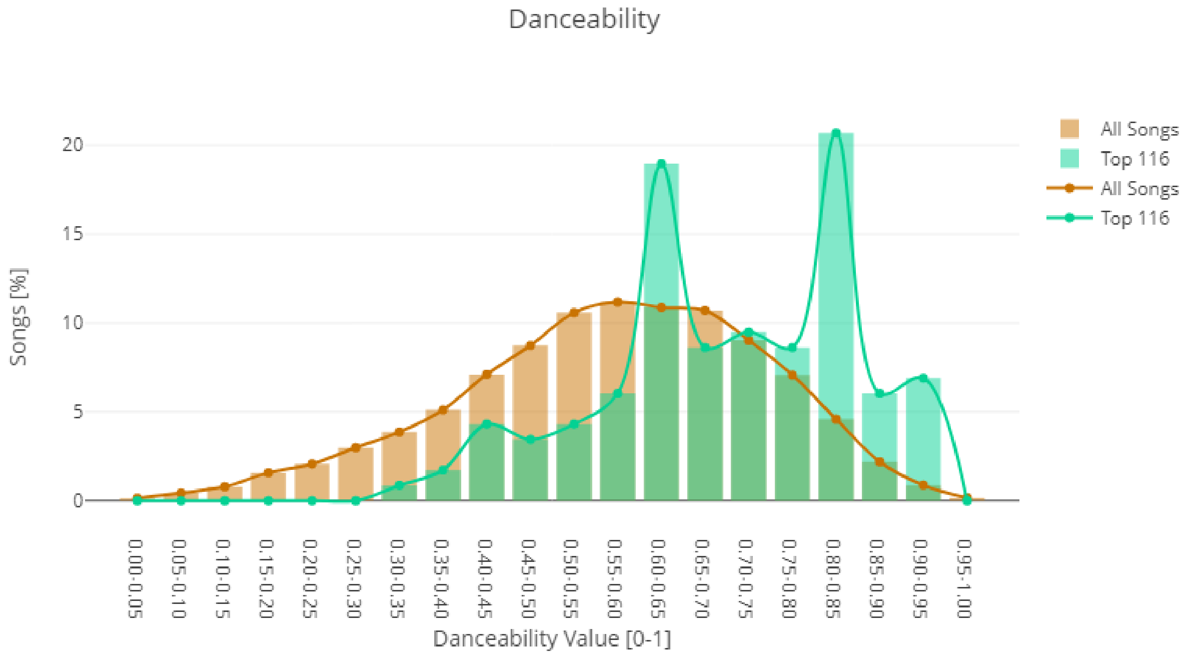

Figure 1: Normalized danceability histogram for the top 116 songs vs. all songs.

In contrast, if we observe the speechiness, we can observe that the two histograms almost overlap perfectly and a much higher total area is shared commonly in comparison to the danceability feature. This is a clear sign that the speechiness feature is not as strongly correlated to ranking in comparison to danceability.

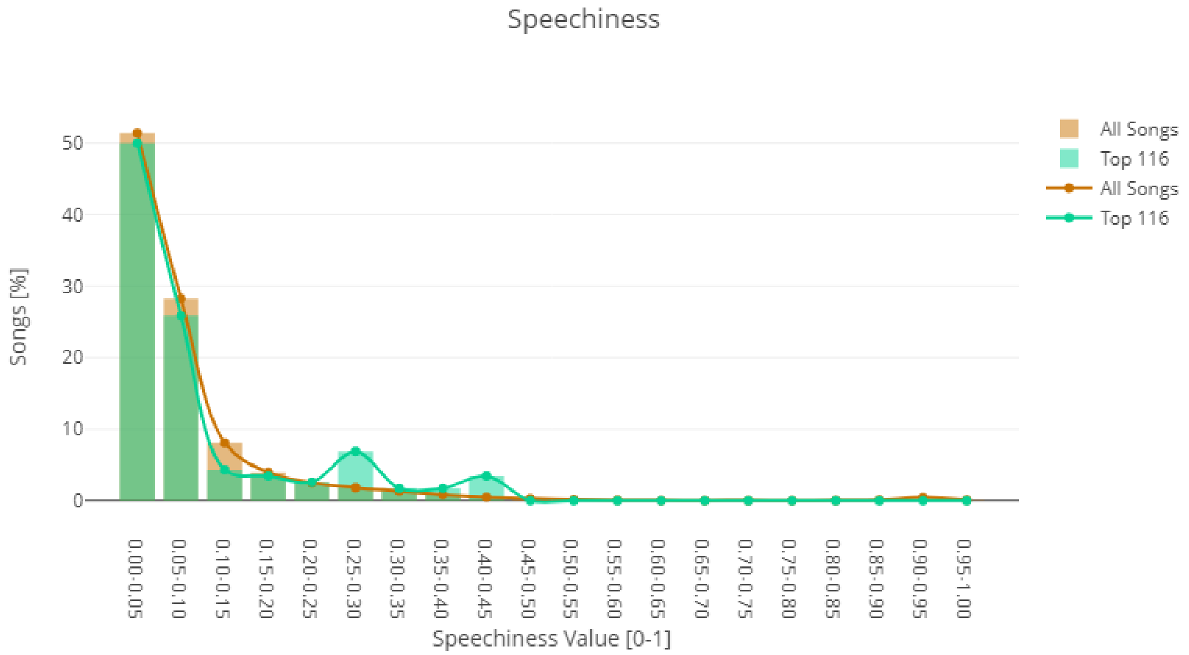

Figure 2: Normalized speechiness histogram for the top 116 songs vs. all songs.

Taking a look at the liveness we can observe the general tendency of top songs to have lower liveness values (a clear majority between 0.05 and 0.1) compared to all songs, which have slightly higher values of between 0.1 and 0.15.

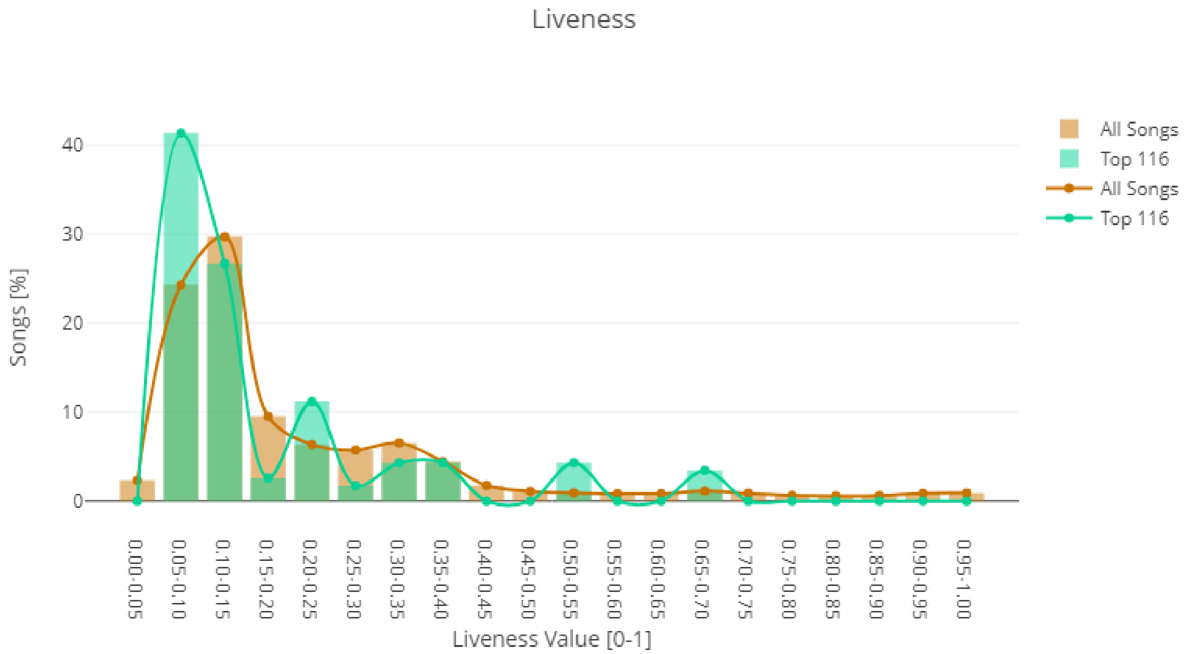

Figure 3: Normalized liveness histogram for the top 116 songs vs. all songs.

From these three histograms and our observations we can already conclude that liveness and danceability are more relevant features (stronger correlation) than the speechiness if we want to identify a top song from the rest. This can be clearly visualized on the pie graph provided on step 3, as shown below.

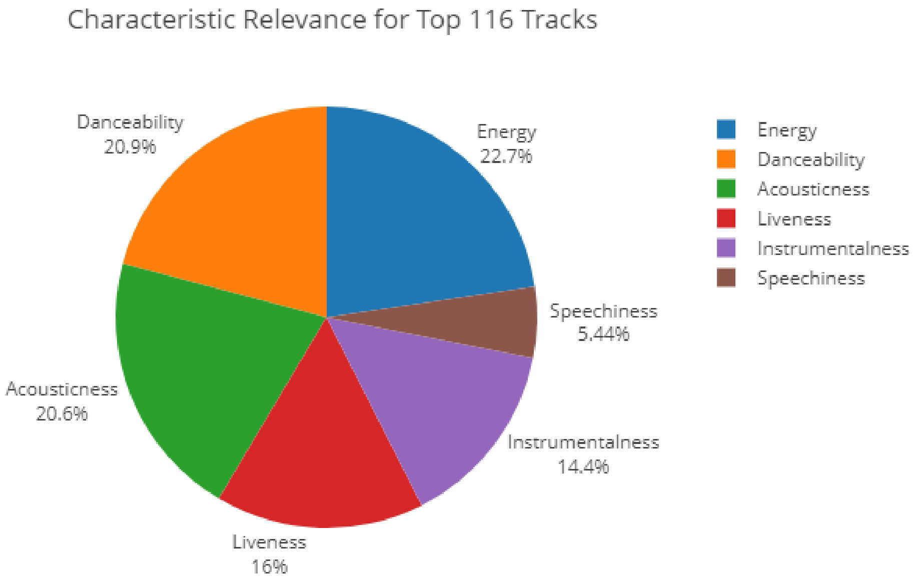

Figure 4: Pie graph showing the relevance of each feature relative to each other.

From the pie graph we can confirm our hypothesis and can observe that the correlation of the characteristics with the ranking can be read in a descending order by reading in counter-clockwise direction: Energy, Danceability, Acousticness, Liveness, Instrumentalness, Speechiness.

Finally, we can draw conclusions on which values concretely can be expected for each feature in step 4. The graph is shown below. Focusing on the evaluated features danceability, liveness and speechiness, one can observe following trends.

The danceability of top songs is much greater (median 0.704) than all the songs in the database (median 0.58). Additionally, danceability is the second most important factor related to the track's ranking, after energy. Secondly, The 1st and 3rd quartile distribution of the liveness feature is located lower in the top songs in comparison to all other songs, indicating that top songs tend to have less live-music characteristics. Additionally, liveness is more weakly correlated to the ranking than danceabilty but stronger than speechiness. Third, speechiness is the least relevant feature when it comes to ranking, and the top songs do not differ in any significant way from the rest of the dataset.

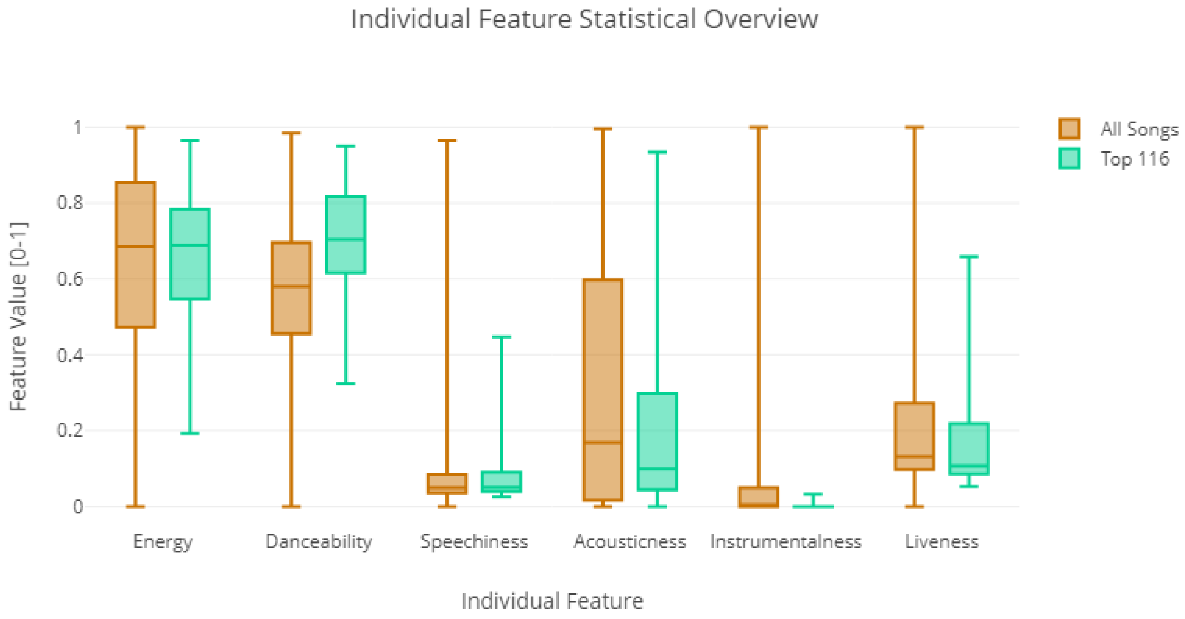

Figure 5: Statistical Overview of all features for both top-n songs and all songs.

Furthermore, additional and more complex insights can be gathered by comparing different top-n graphs. In this example, the top 5000 songs, which are briefly shown below.

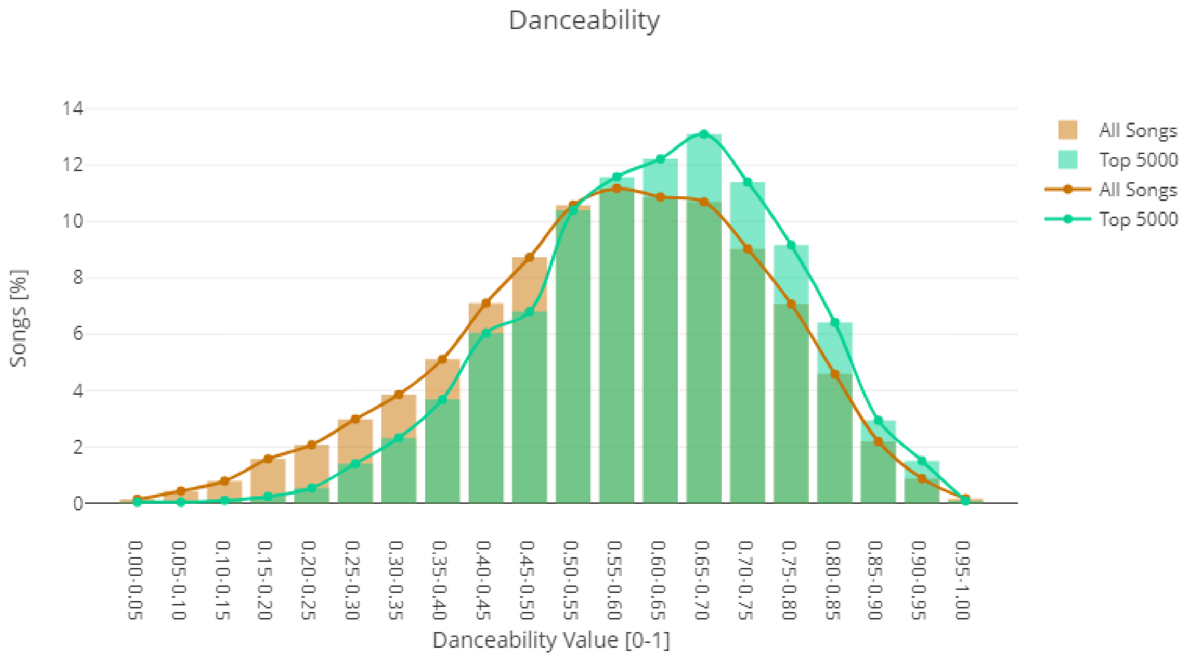

Figure 6: Normalized danceability histogram for the top 5000 songs vs. all songs.

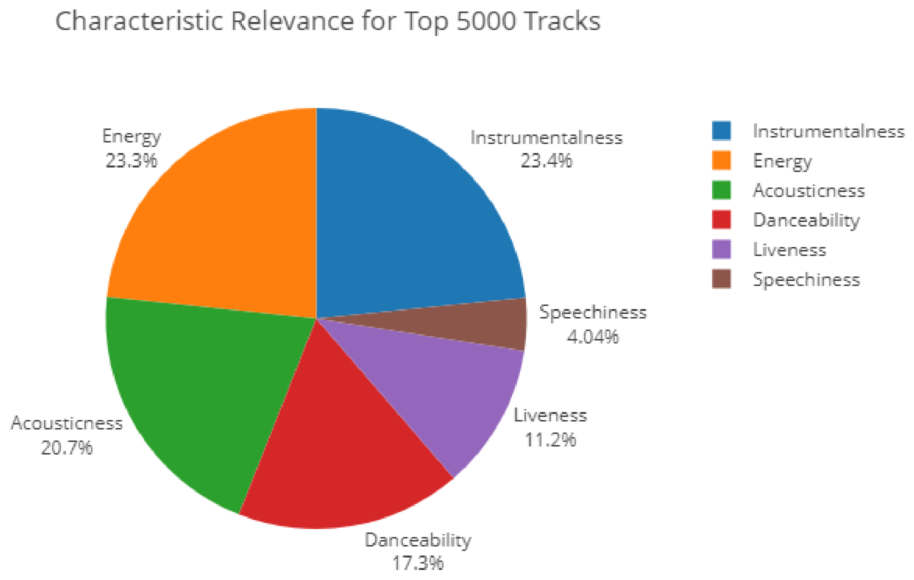

Figure 7: Pie graph showing the relevance of each feature relative to each other

By comparing these two classifications, we can see that instrumentalness (or lack thereof) takes the first place as the most important feature for identifying the top 5000 songs. This would underline the fact that top songs tend to have instrumentalness values very close to 0. By observing how the histograms changed of the top 116 vs top 5000 tracks, we can observe that the danceability resembles much more the histogram of all the songs. This would indicate that the songs located between the 116 and 5000 places are songs that have mostly a danceability value of around 0.65, because this explains the smoothening of the danceability histogram, and the shifting of the peak from 0.85 to around 0.65.