Lab 03 – MATH 240 – Computational Statistics

Brendan Mariano
Colgate
Mathematics
bmariano@colgate.edu

Abstract

In this lab, we used the tidyverse to group each artist and summarize their data for each song measuring feature. This ultimately allowed us to compare each artists similarity to Allentown for select features and determine who was the biggest contributor.

Keywords: Cleaning; Merging; Summarizing

1 Introduction

The overarching goal for this entire project is to determine whether The Front Bottom or the Manchester Orchestra contributed the most to their joint song Allentown. In this specific lab, we sought to determine how to effectively clean our data to put it into form for statistical analysis. One of the difficulties is that we used several sources to obtain the data, which means that they were initially stored in different places. Some features came from the essentia music extractor, others came from the `Essentia` models and the data analyzing lyrics came from the LIWC file. In order to have workable data, we extracted important information from each of them and merged it all into one data frame. This gave us the ability to analyze each artist and determine whose lyrics and music features most closely align with Allentown. For the end of the lab, we created 1-3 statistical plots based on a feature in order to reach such a conclusion.

1.1 Intro Subsection

2 Methods

In order to answer the given question, we were given a step by step process. The first entailed extracting desired features for each json file created by the music extractor and putting them into a data frame. Since the music extractor didn't contain all of the features that we needed, we loaded the `Essentia Model Output.csv`(into a different data frame), which contained data from two other essentia feature extractors. The two extractors had their own value for many of the features, so we averaged them and created additional columns to store the newly calculated values. At this point, we just needed the data from the lyrics so we loaded that file into another data frame. Since we now had three separate data frames, we merged all them together using the `merge()` function. Although some people had issues when merging with creating additional rows, I was able to avoid that by using the `artist, album, and track` as reference when merging.

2.1 Methods Subsection

Our final step was to make a conclusion about the overall question of artist contribution to Allentown using 1-3 plots. I used the Shiny app for this portion because the plots looked better and were easier to create. I settled upon using the avg.valence, spectral energy and instrumental because their values are unique to each artist and because they represented different things– avg.valence is about the emotional feeling of the song, spectral energy is about the how energy is distributed and instrumental is about the amount of instrument use.

3 Results

For every feature we chose, Manchester Orchestra was closer to the Allentown data compared to The Front Bottoms. For avg valence, Allentown had a value of 4.046 while The Manchester Orchestra has a median of 4.28, which is twice as close as The Front Bottoms with a median of 4.51.
The Allentown value for spectral energy is 0.179. In comparison, The Manchester Orchestra has a median of .207 and The Front Bottoms has a median of .294, making The Manchester Orchestra more than 80 percent closer to the Allentown value.
The Allentown value for instrumental is 0.029. In comparison, The Manchester Orchestra has a median of .0281 and The Front Bottoms has a median of .0367, which makes the Manchester Orchestra about 70 percent closer to Allentown.
Since the Manchester Orchestra more closely resembled Allentown with all three features, we concluded that they had a greater responsibility in creating Allentown.

3.1 Results Subsection

4 Discussion

It was clear that the data from the Manchester Orchestra most closely resembled the data from Allentown, which suggests that they had a greater contribution. While I am confident in our findings, the majority of our effort in this lab went towards cleaning the data rather than analyzing it. By only using 3 features, we didn't greatly diversify the features that we actually analyzed. All in all, the conclusion that we were able to make after extracting the proper data was promising, but we need to analyze more features in order to strengthen it.

Bibliography: Note that when you add citations to your bib.bib file *and* you cite them in your document, the bibliography section will automatically populate here.

5 Appendix