Lab 3 – MATH 240 – Computational Statistics

Charles Hooey
Student at Colgate University
Mathematical Economics Major
chooey@colgate.edu

Abstract

This laboratory assignment involved the use of R to extract song data from three different bands in order to determine which band contributed most to a specific track. The song data was first sourced in the first part of the lab, cleaned within the second part, and finally summarized within the third part in order to come to a final conclusion that the band Manchester Orchestra had contributed most to the song "Allentown".

Keywords: For Loops; Packages; Lists; Data Frames

1 Introduction

This labratory assignment involves the examination of data extracted from songs using Essentia (Bogdanov et al., 2013) in order to determine the degree of contributions by each author made to a song that had been collaborated on by three different bands. Within the first part of this assignment, we seeked to use R in order to replicate a command line that would pull the Essentia data for an individual song for all songs within a specific directory. After sourcing this data, we cleaned and combined it with two other data sources during week two of the assignment. For our third and final week of this assignment, we attempted to make an inference of which band had contributed the most to the song "Allentown" which involved the comparison of Essentia and LIWC features from "Allentown" and other songs released by each of the bands.

2 Methods

As mentioned above, part one of this assignment involves the creation of a batch file in order to process every single song within our music directory through Essentia by generating command prompts that would run the Essentia command for each song, along with each song's file location, and json file format. To complete this task, we made frequent use of the `stringr` package for `R` (Wickham, 2023), which would assist us in breaking up the locations of each song file into a vector containing the artist, album, track of the song. This vector was established within two seperate loops, one which would sort through all albums and another that would create a new command line for each new song iterated through, and following such each song would have their Essentia command, file location, and json formatted file added to an empty vector that would then be written to a `.txt` file. Our created batch file would run the Essentia extractor command for each song within the MUSIC directory, which makes data collection much more efficient as had we not written this batch file, we would have had to write the Essentia command line for each individual song.

2.1 Collecting and Cleaning Data

Whereas week one involved the initial sourcing of song data by creating a batch file that would create an Essentia command line for all songs within the MUSIC subdirectory, during week two we seeked to pull data from various files provided within the assignment repository that was for a collection of 181 songs, data which will be helpful in summarizing the musical tendencies of each band who had collaborated on the song "Allentown". In order to complete this task, first within task two we use the `fromJSON()` function from the `jsonlite` package in `R` (Ooms, 2014) in order to pull Essentia model for an example track provided within the assignments repository. After pulling Essentia data for this song, we pulled this data for all files within the EssentiaOutput directory, and cleaned our resulting data by averaging data points that had been calculated using numerous data set into singular columns, and then removing the unaltered data from our data set. After compiling our cleaned data set, we combined this data with data sourced from a language analysis tool called LIWC that was also ran for each of the songs within our directory, as well as the data from our Essentia calls using the `merge()` function to create one aggregate dataframe.

2.2 Summarizing Data

Within the final week of the lab assignment, we used the cleaned song data we had previously sourced in order to collect specific statistical measurements for each band. These measurements included ranges that when compared against data from the song "Allentown" would allow us to determine whether or not each band was in range with the song. The results of this comparison are summarized within Figure 1, which provides data for 10 different musical features. Furthermore, I chose to represent my data visually with a bar chart using Shiny, as it provided a simple and easy to interpret graph which showed the relative frequency that each band had been in or out of range when compared with Allentown.

Figure 1: Summary of Features by Artist

artist	feature	description
All Get Out	spectral_skewness	Outlying
Manchester Orchestra	spectral_skewness	Within Range
The Front Bottoms	spectral_skewness	Out of Range
All Get Out	spectral_rolloff	Out of Range
Manchester Orchestra	spectral_rolloff	Within Range
The Front Bottoms	spectral_rolloff	Out of Range
All Get Out	spectral_energyband_middle_high	Out of Range
Manchester Orchestra	spectral_energyband_middle_high	Within Range
The Front Bottoms	spectral_energyband_middle_high	Out of Range
All Get Out	spectral_complexity	Out of Range
Manchester Orchestra	spectral_complexity	Within Range
The Front Bottoms	spectral_complexity	Out of Range
All Get Out	spectral_centroid	Out of Range
Manchester Orchestra	spectral_centroid	Within Range
The Front Bottoms	spectral_centroid	Out of Range
All Get Out	melbands_spread	Out of Range
Manchester Orchestra	melbands_spread	Within Range
The Front Bottoms	melbands_spread	Out of Range
All Get Out	melbands_flatness_db	Out of Range
Manchester Orchestra	melbands_flatness_db	Within Range
The Front Bottoms	melbands_flatness_db	Out of Range
All Get Out	erbbands_skewness	Out of Range
Manchester Orchestra	erbbands_skewness	Within Range
The Front Bottoms	erbbands_skewness	Out of Range
All Get Out	erbbands_flatness_db	Outlying
Manchester Orchestra	erbbands_flatness_db	Within Range
The Front Bottoms	erbbands_flatness_db	Out of Range
All Get Out	dissonance	Outlying
Manchester Orchestra	dissonance	Within Range
The Front Bottoms	dissonance	Out of Range

3 Results

After completing part one of this assignment, I was able to successfully create a batch file containing all songs within the MUSIC directory, which for each song included the command prompt to run Essentia, the file's name, as well as the file listed in json format. Furthermore, task two of this assignment lead me to source specific Essentia model data values for all songs within a specific directory, and further to merge these values of data with their Essentia call data and LIWC data. The cleaned and aggregated dataset that I had created within week two will provide me with a foundation on which I can make statistical inferences that seek to answer my overarching research question through summarizing data that will help us to better understand specific musical tendencies present within the distinct playstyles of each band. Additionally, I had attempted to solve the lab challenge by creating a boxplot which summarized the pitch salience values of all songs except "Allentown" in an attempt to better understand the vocal tendencies of one of the specific bands that had contributed to the previously mentioned collaboration track. While this graph alone did not provide me any signifigant insight toward my research question, it familiarized me with creating boxplots within R. Finally, the statistical measurements recorded within part three of this lab led me to make the conclusion that Manchester Orchestra had contributed the most to the song "Allentown". As seen within figure 2, among the 10 features included within my data sample, Manchester Orchestra was the only band consistently in range with "Allentown".

References

Bogdanov, D., Wack, N., Gómez Gutiérrez, E., Gulati, S., Boyer, H., Mayor, O., Roma Trepat, G., Salamon, J., Zapata González, J. R., and Serra, X. (2013). Essentia: An audio analysis library for music information retrieval. In Britto, A., Gouyon, F., and Dixon, S., editors, *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 493–498, Curitiba, Brazil. International Society for Music Information Retrieval (ISMIR).

Ooms, J. (2014). The jsonlite package: A practical and consistent mapping between json data and r objects. *arXiv:1403.2805 [stat.CO]*.

Wickham, H. (2023). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.5.1.

4 Appendix

Below, I have included my figures and plots for this assignment in order to avoid any formatting issues or complicatations:

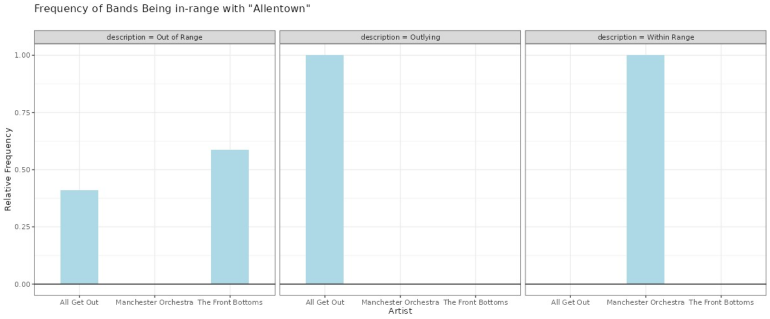

Figure 2: Relative Frequency of Each Band Being Within Range of Allentown