Lab 05 – MATH 240 – Computational Statistics

Cristian Palmer
Student
Mathematics
cpalmer@colgate.edu

Abstract

This lab is a continuation to the lab which we completed last week. Last week, we took preliminary steps to determine which of three bands, *The Front Bottoms's*, *Manchester Orchestra*, or *All Get Out* contributed the most to the collaboratory song *Allen Town* (Ross (2018)). For this week's lab, our goal was to continue our research into this question by acquiring and organizing important music data from different sources. In this lab, we utilized the `stringr` (Wickham (2023)) and `jsonlite` (Ooms (2014)) packages to extract data from the `Essentia streaming music extractor` (Alonso-Jiménez et al. (2020)), `Essentia models` (Bogdanov et al. (2013)), and `LIWC` (Pennebaker et al. (2015)), a lyric analysis tool.

Keywords: Data Frames : Loops : Merging

1 Introduction

This lab originated from the question of which band, *The Front Bottoms*, *Manchester Orchestra*, or *All Get Out* contributed the most to the song Allen Town. One way to approach this question is to use various programs to provide us data about each band's music catalog. For this the lab, we worked to acquire, and edit this data eventually combining all of our data into a final data frame housing all of the data. We also created two `csv` files, one housing every one of these band's songs besides *Allen Town* and one housing only *Allen Town*. For our next lab, these `csv` files along with the final data frame will prove extremely useful for answering our question. Additionally, for the end of this lab we took the first steps to analyzing our data by creating a couple graphs which could possibly lead to an answer of which band contributed most to the song *Allen Town*.

2 Methods

For this lab, we began by using the `stringr` and `jsonlite` packages to extract data from the `Essentia streaming music extractor` about the *The Front Bottoms's* song *Au Revoir (Audios)*. The purpose of doing this was to lay the ground work for the next step where we created a loop to complete this process for all 181 songs in the `EssentiaOutput` folder. We tasked our loop to save the data in an empty data frame. We then ran our loop, filling our data frame with `Essentia streaming music extractor` data for every song. This data frame is the first of three data frames which we eventually combined into one final inclusive data frame. Before we ran this loop however, it is important to note that we had to remove a `csv` file from the `EssentiaOutput` folder. If this file was not removed, then our loop would stop working part way through when it got to this file. Our next set of data came from the `EssentiaModelOutput.csv` file provided to us. For this data, we were tasked with manipulating the data to create new columns which would be of more interest to us for this lab. After creating, renaming, and deleting columns, we were able to finish creating our second data frame. For our third data frame, we were tasked with loading the `LIWCOutput.csv` file provided to us. `LIWC` is a software which analyses text, in our case lyrics and provides insightful data about said lyrics. Finally, we used the `merge()` function to combine all three data sets into one final data set housing data from all three of our sources. Since each of our data frames we were combining into one had columns, "artist", "track", and "album", we decided to merge on these three columns. In preparation for next lab, we also created two of our own `csv` files. One of these files housed every one of the multiple band's songs except for *Allen Town*, and the other housed only *Allen Town*. To complete this step, we simply utilized the `write.csv` function. As a final additional challenge, we also began to analyze this data by creating a couple graphs which could possibly lead us to an answer of which band contributed most to the song. For this step, we utilized the *Shiny App* (The Data Science Collaboratory at Colgate University (2024)) provided to us. To utilize this app, we simply used the `write.csv` again, this time to write our final data frame as a `csv` file. We then simply uploaded our new final data frame as a `csv` to the *Shiny App* and were able to use it to create various plots which could possibly end up being informative. However, the majority of data analysis will be done in our next lab.

3 Results

Our final data frame ended up consisting of 181 rows and 140 columns. This means that for all 181 songs, we have 140 different types of data about each song coming from our various data sources. We can verify that no data was lost when merging since our data frame of `Essentia streaming music extractor` data had 181 rows and 11 columns, our `LIWC` data had 181 rows and 121 columns, and our `EssentiaOutput` data had 181 rows and 14 columns. When merging these three data frames together we would expect to see 181 rows and

146 columns, since the number of rows will not change cause each row corresponds to one song and each data has the same set of songs. Our data only has 140 rows since we merged by "artist", "track", and "album". This means that these columns were not duplicated since they were the same in each data frame, therefor, it makes sense that we are 6 rows short of what we expected because two sets of these 3 columns were taking out. So, we correctly have 140 columns and 181 rows showing that no data was lost during our data altering or data merging steps. Viewing our final data frame there are also no visible issues concerning the naming of columns or data within the data frame itself. All of our data appears to be present with no "NA's", and no columns appear to be misnamed. Concerning the graphs we made, I will place the one which I believe has the highest possibility of being relevant in the appendix section below.

4 Discussion

The graph I chose to include is a Violin Plot created with the *Shiny App*. For each artist, this plot shows how the happiness of their catalog of songs can be distributed. Our data on happiness came from the `EssentiaOutput` data set. Looking at the graph, it appears as though the happiness level in *Manchester Orchestra's* catalog most closely alligns with the level of happiness in the song which *The Front Bottoms* and *Manchester Orchestra* created together. So, this graph provides some evidence that possibly *Manchester Orchestra* contributed most to the song. However, in our next lab we will go into much greater detail on analyzing and visualizing our data.

References

Alonso-Jiménez, P., Bogdanov, D., Pons, J., and Serra, X. (2020). Tensorflow audio models in Essentia. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 266–270. IEEE.

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., and Serra, X. (2013). Essentia: An audio analysis library for music information retrieval. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 493–498. ISMIR.

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

Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. LIWC.net.

Ross, A. R. (2018). Manchester orchestra and the front bottoms are finally together on "allentown".

The Data Science Collaboratory at Colgate University (2024). Collaboratory resources. Online Application. Retrieved from https://www.colgate.edu/about/campus-services-and-resources/data-science-collaboratory.

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

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

5 Appendix

	feature	artist	out.of.range	unusual	description
1	spectral_skewness	All Get Out	FALSE	TRUE	Outlying
2	spectral_skewness	Manchester Orchestra	FALSE	FALSE	Within Range
3	spectral_skewness	The Front Bottoms	TRUE	TRUE	Out of Range
4	spectral_rolloff	All Get Out	TRUE	TRUE	Out of Range
5	spectral_rolloff	Manchester Orchestra	FALSE	FALSE	Within Range
6	spectral_rolloff	The Front Bottoms	TRUE	TRUE	Out of Range
7	spectral_kurtosis	All Get Out	FALSE	TRUE	Outlying
8	spectral_kurtosis	Manchester Orchestra	FALSE	FALSE	Within Range
9	spectral_kurtosis	The Front Bottoms	TRUE	TRUE	Out of Range
10	spectral_entropy	All Get Out	FALSE	TRUE	Outlying
11	spectral_entropy	Manchester Orchestra	FALSE	FALSE	Within Range
12	spectral_entropy	The Front Bottoms	TRUE	TRUE	Out of Range
13	spectral_energyband_middle_high	All Get Out	TRUE	TRUE	Out of Range
14	spectral_energyband_middle_high	Manchester Orchestra	FALSE	FALSE	Within Range
15	spectral_energyband_middle_high	The Front Bottoms	TRUE	TRUE	Out of Range
16	spectral_complexity	All Get Out	TRUE	TRUE	Out of Range
17	spectral_complexity	Manchester Orchestra	FALSE	FALSE	Within Range
18	spectral_complexity	The Front Bottoms	TRUE	TRUE	Out of Range
19	spectral_centroid	All Get Out	TRUE	FALSE	Out of Range
20	spectral_centroid	Manchester Orchestra	FALSE	FALSE	Within Range
21	spectral_centroid	The Front Bottoms	TRUE	FALSE	Out of Range
22	erbbands_skewness	All Get Out	TRUE	TRUE	Out of Range
23	erbbands_skewness	Manchester Orchestra	FALSE	FALSE	Within Range
24	erbbands_skewness	The Front Bottoms	TRUE	TRUE	Out of Range
25	dissonance	All Get Out	FALSE	TRUE	Outlying
26	dissonance	Manchester Orchestra	FALSE	FALSE	Within Range
27	dissonance	The Front Bottoms	TRUE	TRUE	Out of Range
28	barkbands_skewness	All Get Out	TRUE	TRUE	Out of Range
29	barkbands_skewness	Manchester Orchestra	FALSE	FALSE	Within Range
30	barkbands_skewness	The Front Bottoms	TRUE	TRUE	Out of Range

Table 1: Summary of Selected Features