The truth behind the lyrics

Author: Jinxu Xiang

This report is prepared with the following environmental settings.

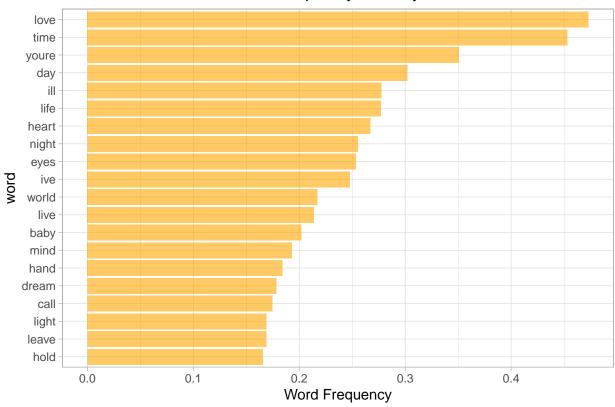
```
print(R.version)
```

```
##
                  x86_64-w64-mingw32
## platform
                  x86_64
## arch
                  mingw32
## os
                  x86_64, mingw32
## system
## status
## major
                  3
## minor
                  6.2
                  2019
## year
## month
                  12
## day
                  12
## svn rev
                  77560
## language
## version.string R version 3.6.2 (2019-12-12)
## nickname
                  Dark and Stormy Night
```

First,I processed the raw textual data 'lyrics.RData' saved in 'data' file by cleaning data, removing stopwords and creating a tidy version of texts which is saved in 'output' file.

Then,I combined the processed text with artist information 'artists.csv' and saved the joint data in 'output' file. The 'Origin' column of joint data contains the name of city and country (or state in America). So I extracted the names of each reigon and saved it as 'Precessed_country'.

Word Frequency of All Lyrics





This is the wordcloud of all lyrics. As you can see, the word 'love', 'light', 'heartlife', 'dream' etc. appear many times. Most of words in this graph are positive. However, words like 'ill', 'die' and 'cry' also appear frequently. But compared to the previous words, the number and frequency of negative words are significantly lower. Although at this time we will think that most music is positive, is it true that most of music is positive?

Question 1 - Is it true that most of music is positive?

To quickly find answers, first I made a dynamic shinyapp to search keywords under different condition. I designed the drop-down tab to select different genres and regions, and the slider to select the year interval. If you want to select all genres or regions, you can select 'All genres' or 'All Region'. (However, this cannot be displayed after being stored as html)

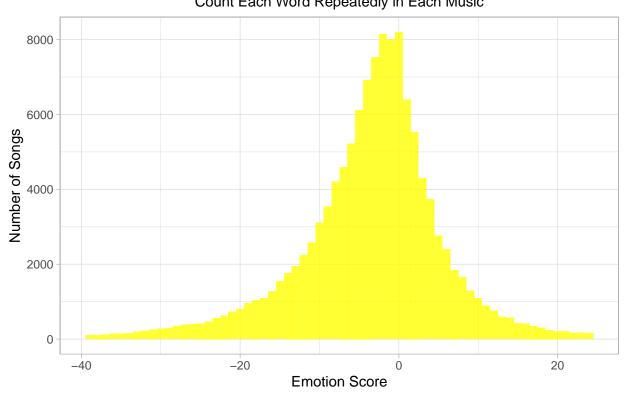
shinyApp(ui, server)

Shiny applications not supported in static R Markdown documents

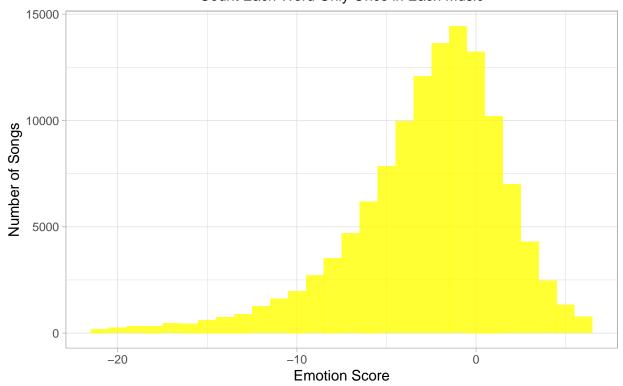
After using shinyapp to control variables, I have a preliminary understanding of the question I want to do. Next I use dataset 'bing' (from Bing Liu and collaborators, https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html) to distinguish positive and negative words.

First,I gave positive and negative words +1 score and -1 score respectively. Then, I scored and summed all stemmed words in each piece of lyrics. For the total score obtained, if the total score is positive, it will be classified as positive music, and if the total score is negative, it will be classified as negative music. I used two scoring methods. The first one counts the repeated words in each stemmed lyrics, and the second one counts the repeated words only once in each stemmed lyrics. The two types of score statistics are as follows.

Music Positive and Negative Emotion Distribution Graph Count Each Word Repeatedly in Each Music



Music Positive and Negative Emotion Distribution Graph Count Each Word Only Once in Each Music

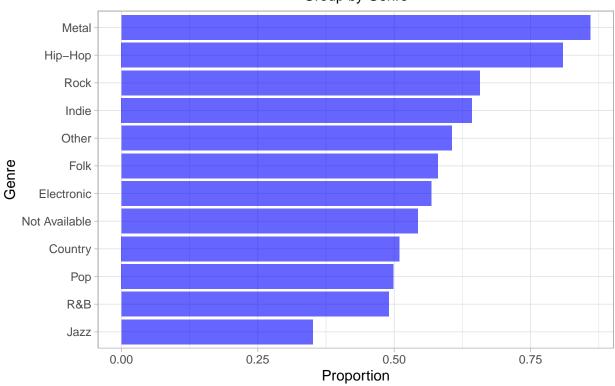


These two graphs above are the opposite of what I said before. Even though positive words occupy most of the wordcloud, about 70% music use more negative words. In addition, the emotion score which each word only count once looks like Poisson distribution. This property may be used in more places.

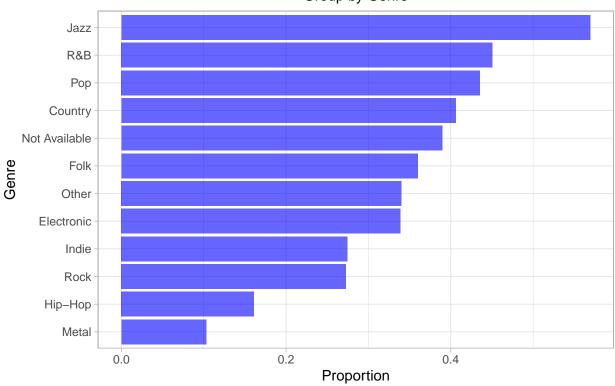
The statistical model of this data is too good, it is difficult to imagine that the 120,000 data can get a distribution that looks a lot like the real distribution. In addition, considering that there are many strange author names, song names and lyrics in this dataset, it will increase people's doubt about the reality of this dataset.

Since more negative words are shown below the appearance of positive words, so where, when and what kind of music use negative words more? Where, when and what kind of music use positive words more? The following graphs make statistics of different categories of lyrics.

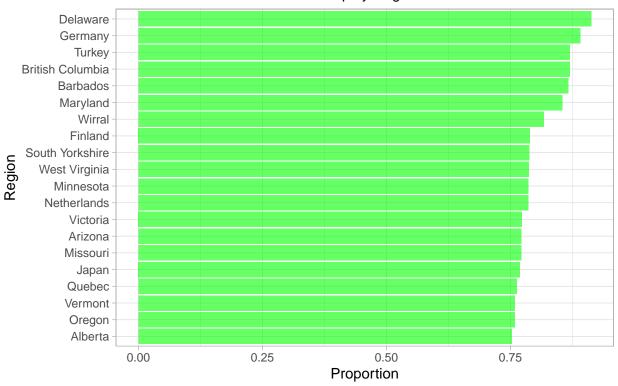
Proportion of negative song Group by Genre



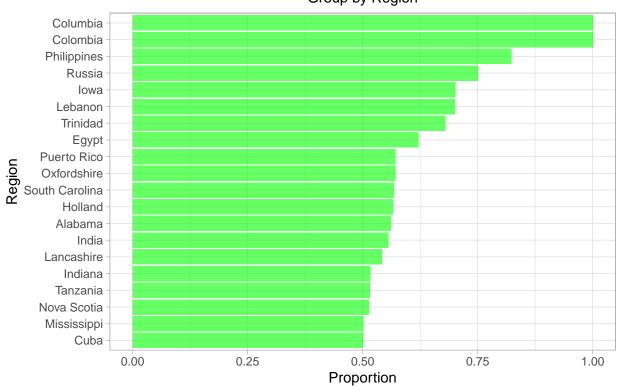
Proportion of positive song Group by Genre



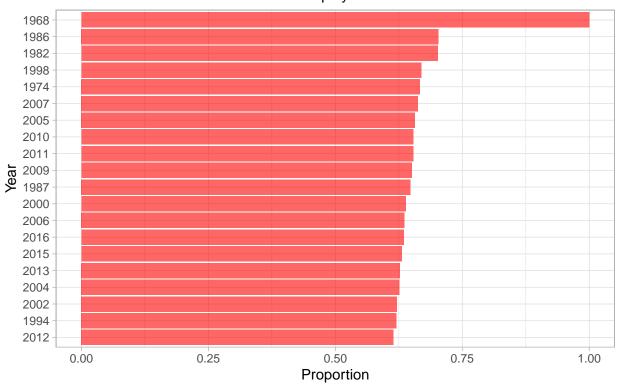
Proportion of negative song Group by Region



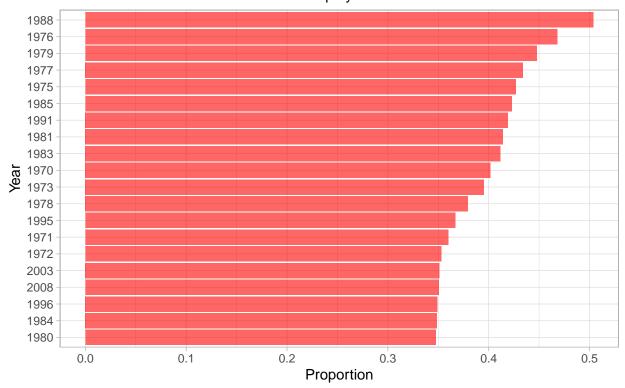
Proportion of positive song Group by Region



Proportion of negative song Group by Year



Proportion of positive song Group by Year

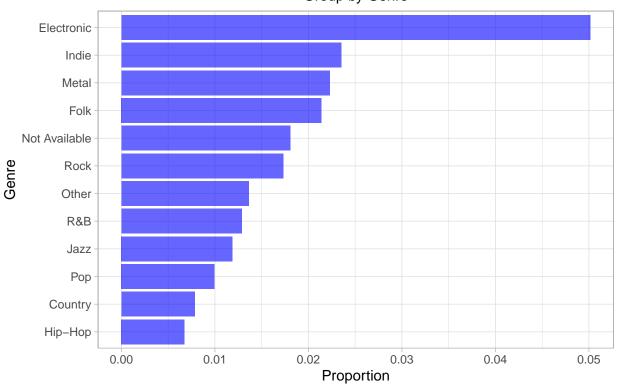


Metal and Hip-hop are the music genres with the most negative words and Jazz is with the least. Columbia, Phillippines and Russia are the regional using positive words while most of others use negative words more. Since there was only one song in 1968, the proportion reached 100%. In addition, the negative words were used the most in 1986 and 1982 and the least used in 1988.

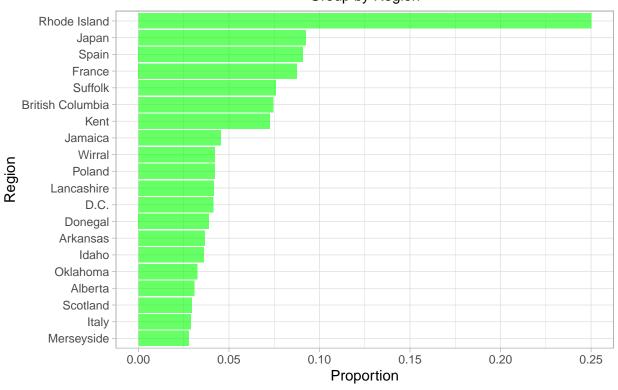
Question 2 - What kind of music is out of the ordinary?

From the shinyapp and statistical chart above, we can see that most of music uses words like 'love' and 'time'. we will be curious about such a question, what kind of music doesn't contain the most frequent words. After selecting the music that doesn't contain the most frequent 30 words, we can get music statistics in different categories as follows.

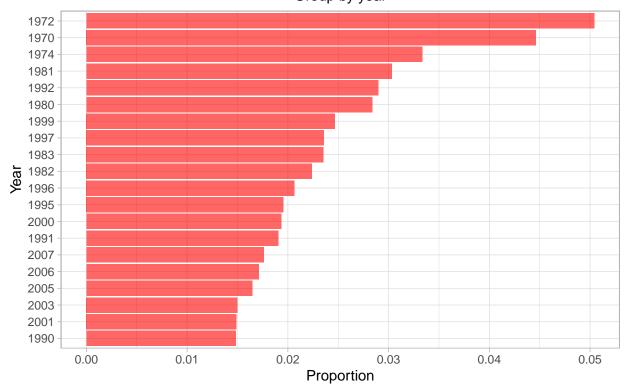
Proportion of music without frequent words Group by Genre



Proportion of music without frequent words Group by Region



Proportion of music without frequent words Group by year



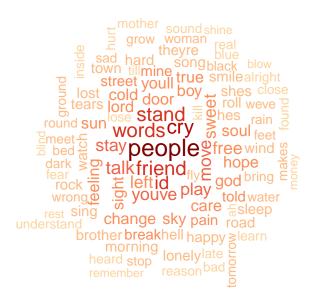
These statistical graphs show that electronic music uses the least high-frequency words when classifying music genres. Rhode Island, Japan, Spain, France, Suffolk and British Columbia use the least frequent words when classifying by country and people used the least frequent words in 1972, 1970 and 1974. Then, what are the musical keywords for these genres, places and times?

```
moment chorus

truth rainline gotta god shes
bring listen minetrue forget

polose hardwords pain kiss
youve id change forever
youve id change forever
state of the sun stop stands of the sun stands of t
```





The hue of the graphs above are consistent with the previous graphs. We can see that 'feeling', 'play', 'inside' are more used by Electronic music. These words also match the entertaining nature of this kind of music. 'burn', 'soul', 'lost' are more used in those regions above, many negative words appear in wordcloud. It it related to the instability in these areas. In 1970s, 'people', 'friend' and 'stand' used a lot. It can be seen that this is an age of peaceful development.

Frequent words are related to their characteristics, and they often appear in any lyrics. However, words other than them can better reflect the real thoughts of people in a certain region, era, and character. Statistics on these words can better reflect the story behind the lyrics.

Conclusions

- 1. Even though most of the music uses mostly positive words, negative words are actually used more frequently around the world.
- 2. When the emotion of the music is scored, the resulting distribution is similar to the Poisson distribution. This distribution fits well, may make people doubt about the reality of the data.
- 3. Frequent words appear frequently in any lyrics. After removing them, you can see the words that discribe the era and region.