The truth behind the lyrics

Author: Jinxu Xiang

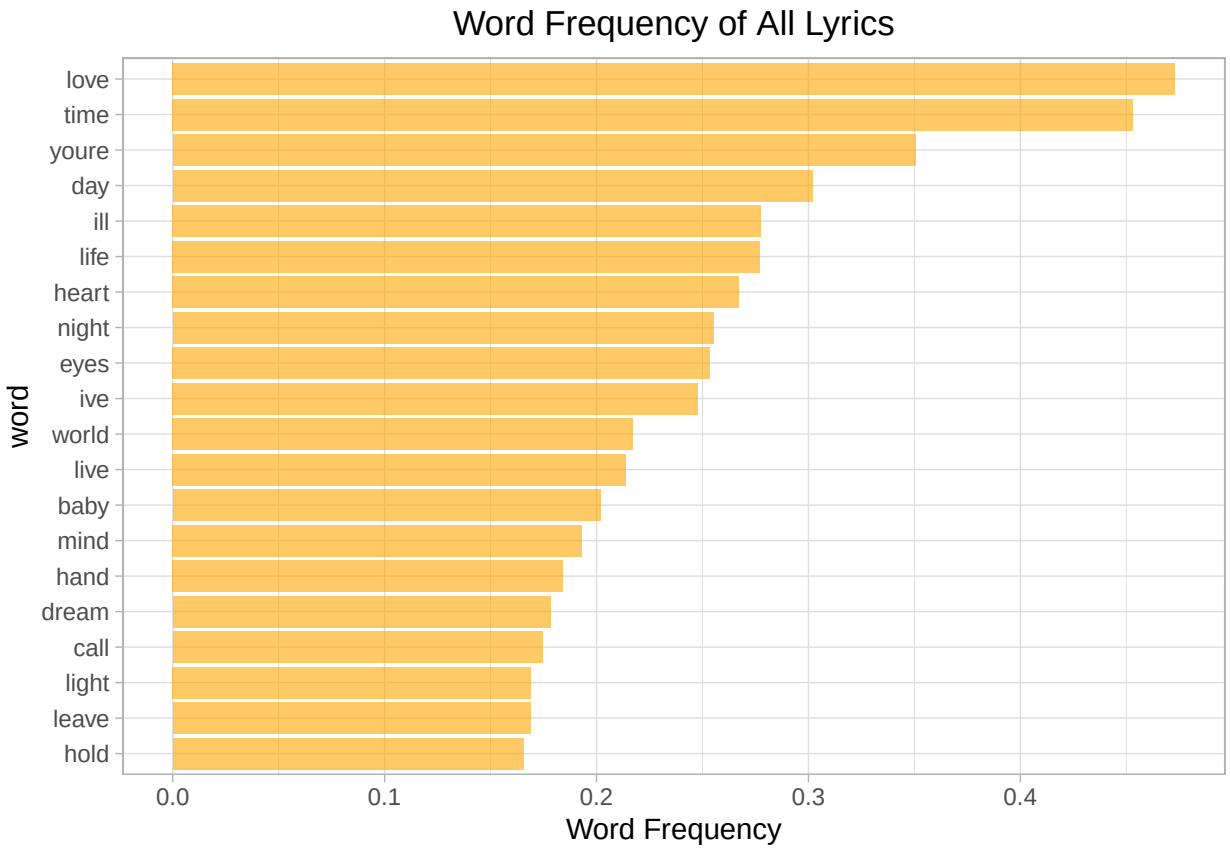
This report is prepared with the following environmental settings.

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

```
##                _
## platform       x86_64-w64-mingw32
## arch           x86_64
## os             mingw32
## system         x86_64, mingw32
## status
## major          3
## minor          6.2
## year           2019
## month          12
## day            12
## svn rev        77560
## language       R
## 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
love
time
youre
day
ill
life
heart
night
eyes
ive
world
live
baby
mind
hand
dream
call
light
leave
hold
word
0.0
0.1
0.2
0.3
0.4
Word Frequency

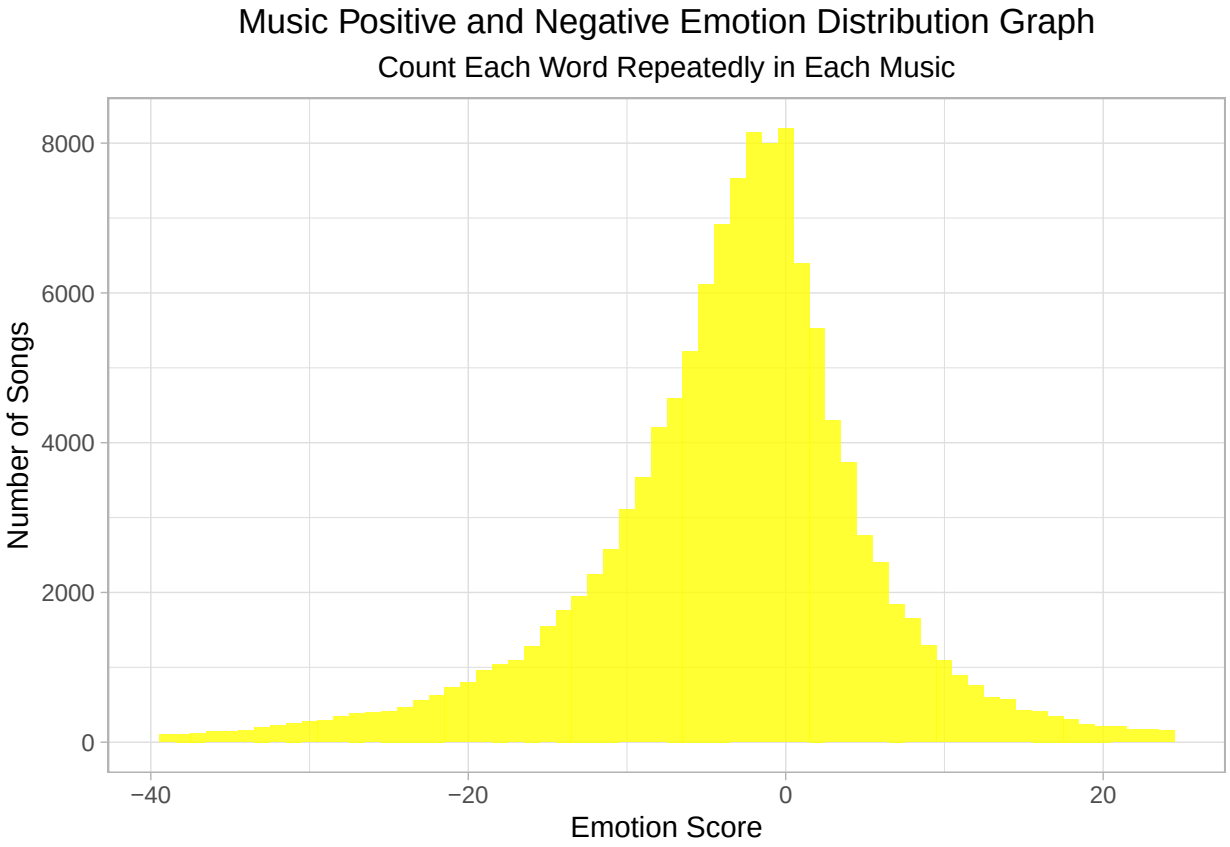
This is the wordcloud of all lyrics. As you can see, the word 'love', 'light', 'heartlife', 'dream' etc. appear many times. Most of words in this graph are positive. However, words like 'ill', 'die' and 'cry' also appear frenquently. But compared to the previous words, the number and frequency of negative 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?

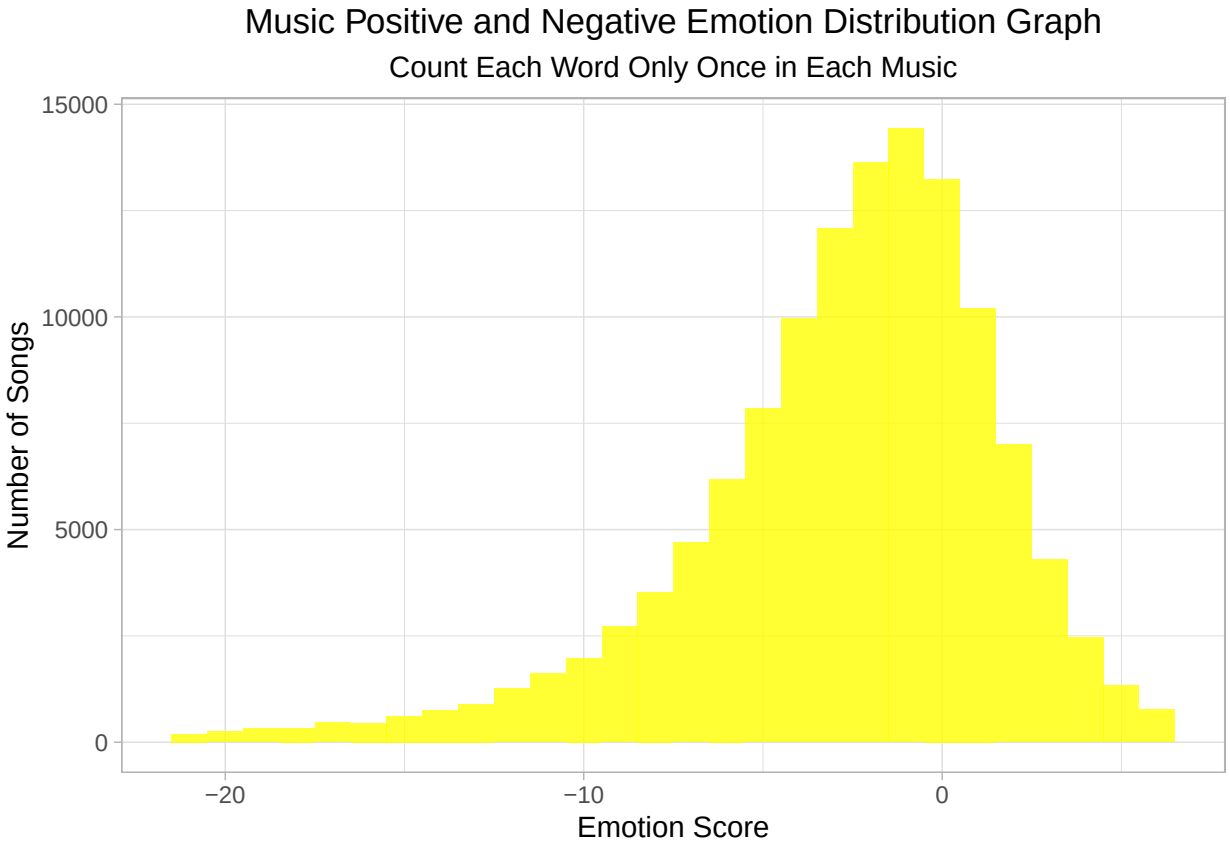
Question 1 - Is it true that most of music is positive?

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, so I put it in 'Project1-shinyapp' in 'doc' file.

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.

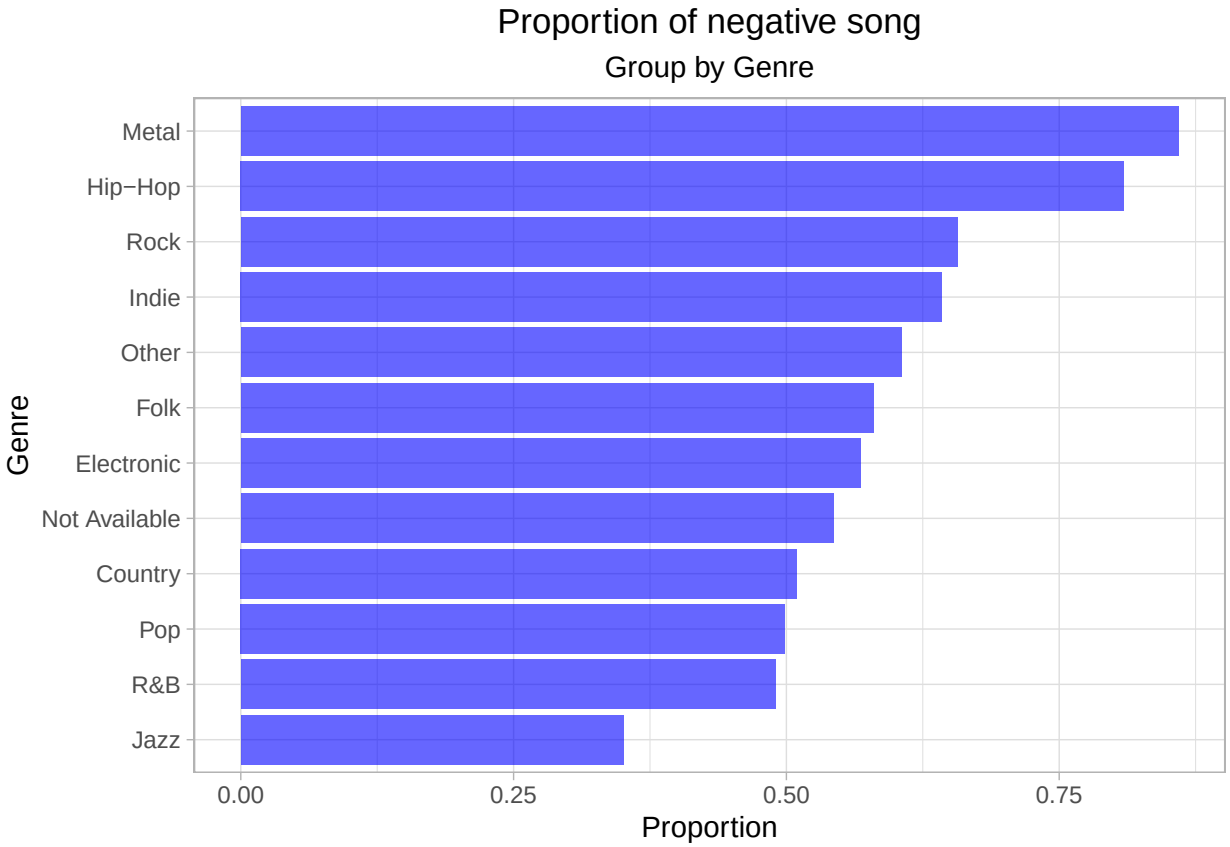

Music Positive and Negative Emotion Distribution Graph
Count Each Word Repeatedly in Each Music
Number of Songs
8000
6000
4000
2000
0
−40
−20
0
20
Emotion Score

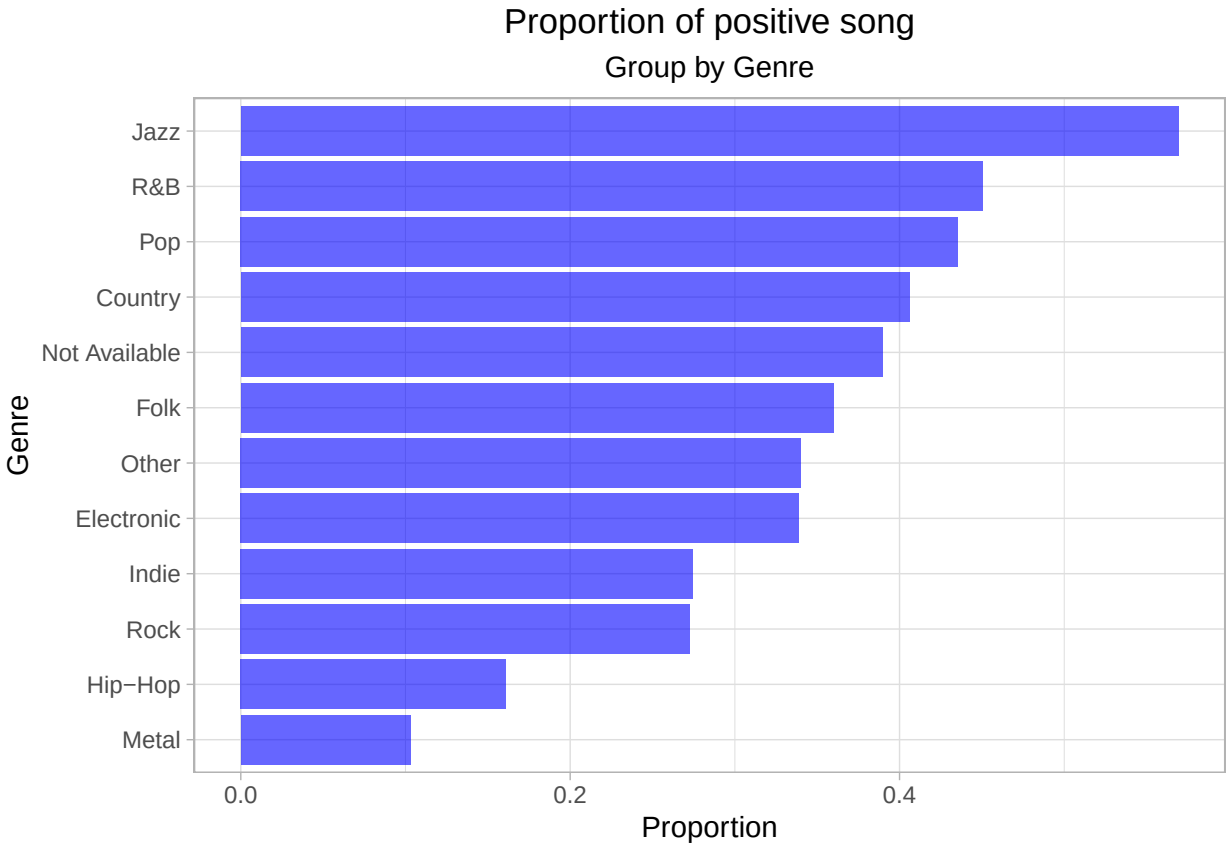


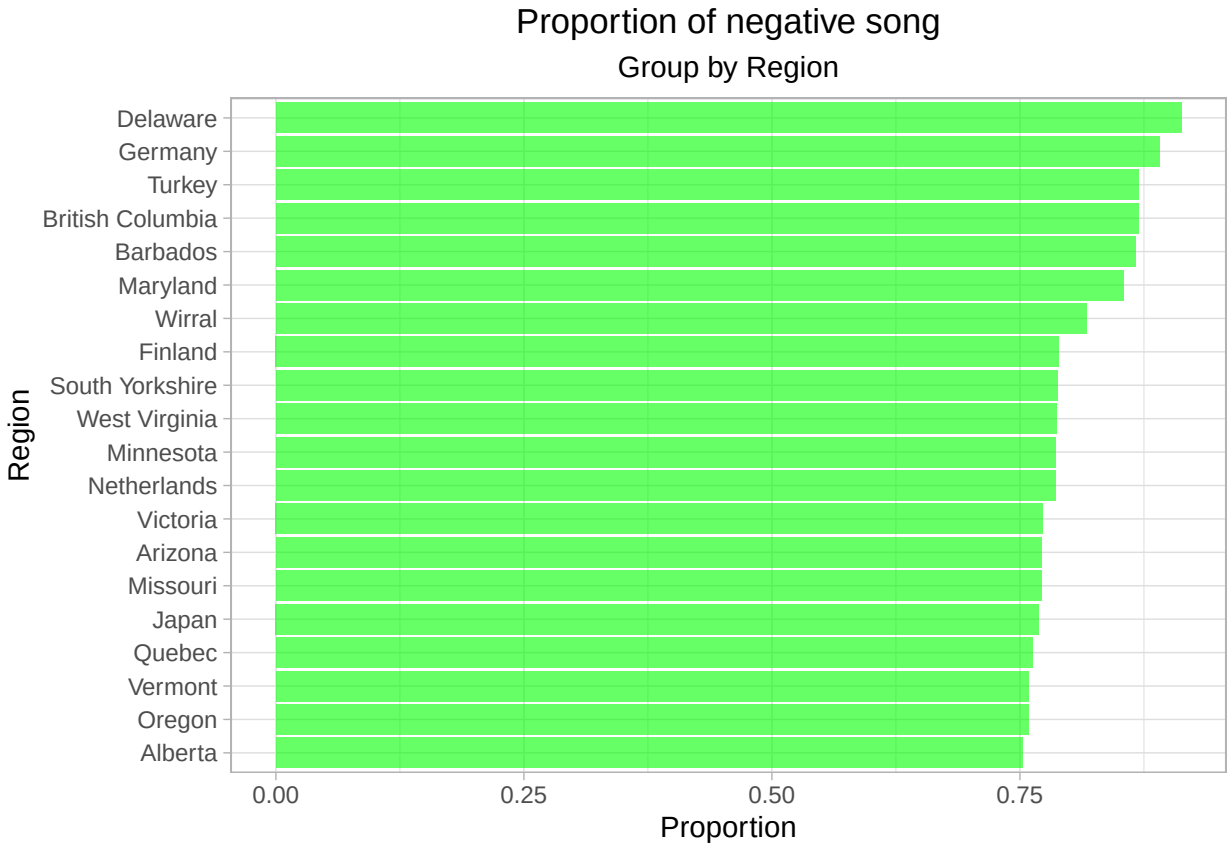
These two graphs above are the opposite of what I said before. Even though positive words occupy most of the wordcloud, about 70% music use more negative words. In addition, the emotion score which each word only count once looks like Poisson distribution. This property may be used in more places.

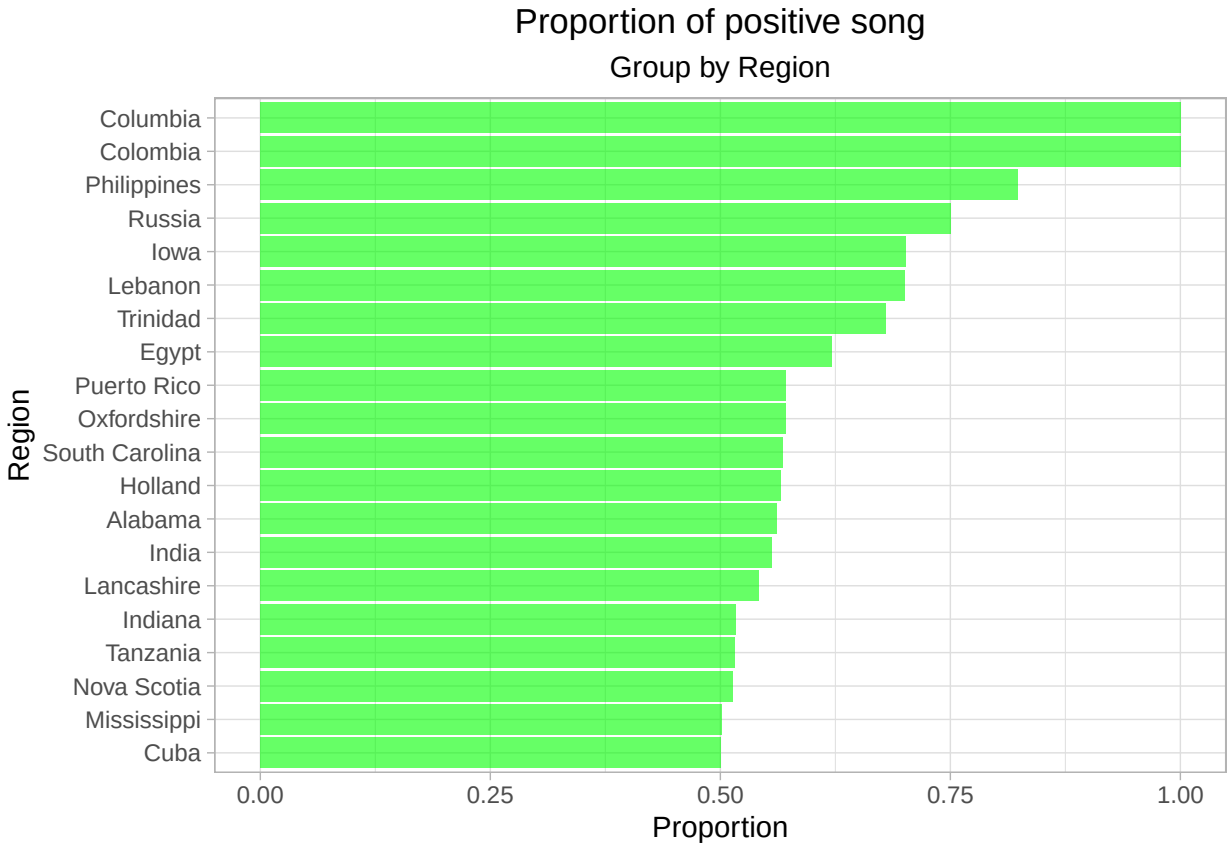
The statistical model of this data is too good, it is difficult to imagine that the 120,000 data can get a distribution that looks a lot like the real distribution. In addition, considering that there are many strange author names, song names and lyrics in this dataset, it will increase people's doubt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

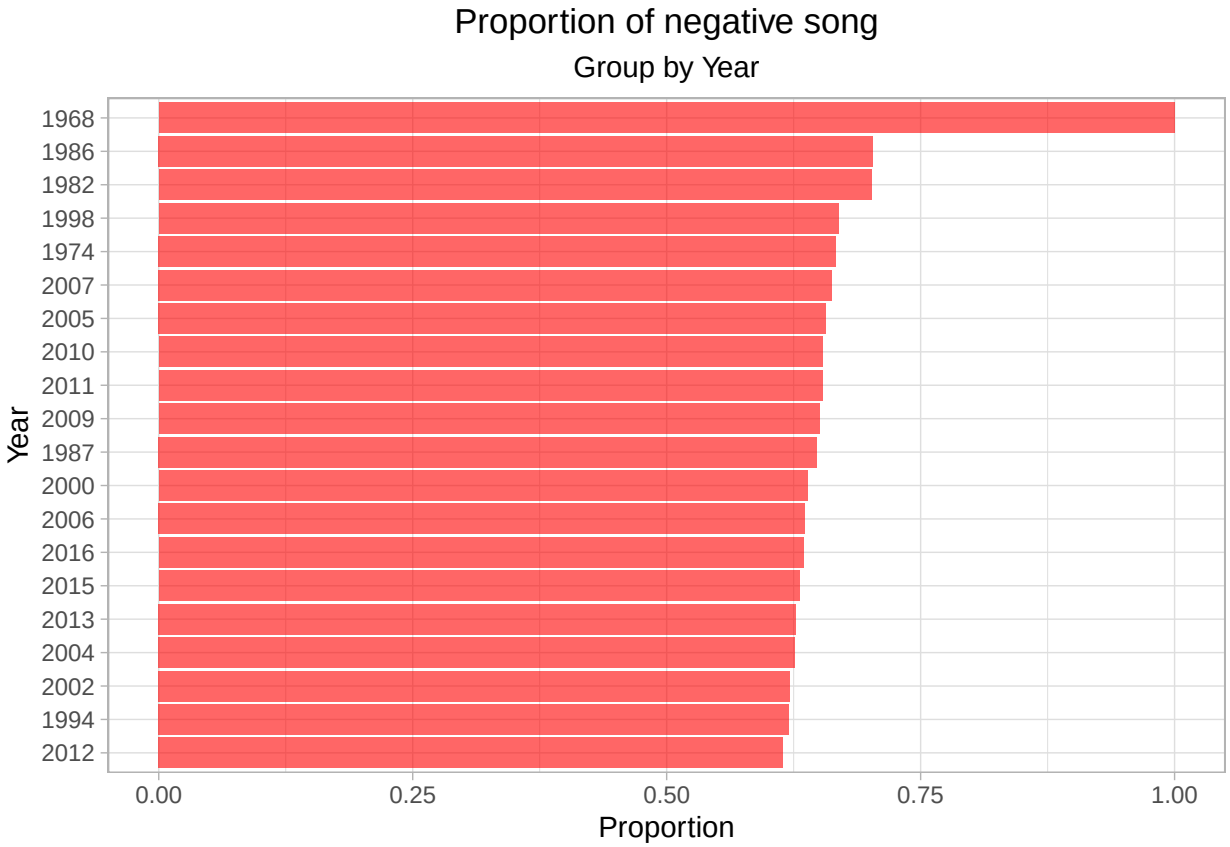
Since more negative words are shown below the appearance of positive words, so where, when and what kind of music use negative words more? Where, when and what kind of music use positive words more? The following graphs make statistics of different categories of lyrics.

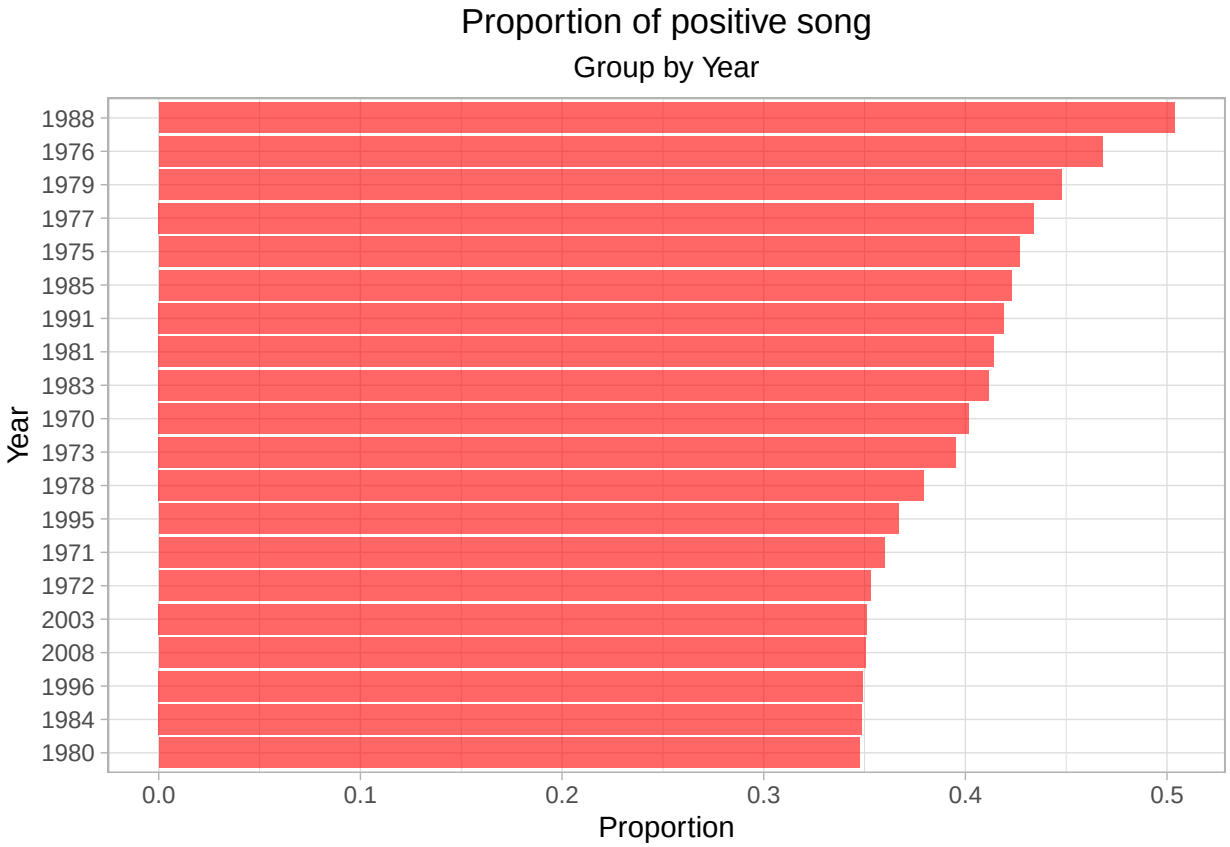

Proportion of negative song
Group by Genre
Genre
Metal
Hip−Hop
Rock
Indie
Other
Folk
Electronic
Not Available
Country
Pop
R&B
Jazz
0.00
0.25
0.50
0.75
Proportion


Proportion of positive song
Group by Genre
Genre
Jazz
R&B
Pop
Country
Not Available
Folk
Other
Electronic
Indie
Rock
Hip−Hop
Metal
0.0
0.2
0.4
Proportion


Proportion of negative song
Group by Region
Region
Delaware
Germany
Turkey
British Columbia
Barbados
Maryland
Wirral
Finland
South Yorkshire
West Virginia
Minnesota
Netherlands
Victoria
Arizona
Missouri
Japan
Quebec
Vermont
Oregon
Alberta
0.00
0.25
0.50
0.75
Proportion


Proportion of positive song
Group by Region
Region
Columbia
Colombia
Philippines
Russia
Iowa
Lebanon
Trinidad
Egypt
Puerto Rico
Oxfordshire
South Carolina
Holland
Alabama
India
Lancashire
Indiana
Tanzania
Nova Scotia
Mississippi
Cuba
0.00
0.25
0.50
0.75
1.00
Proportion

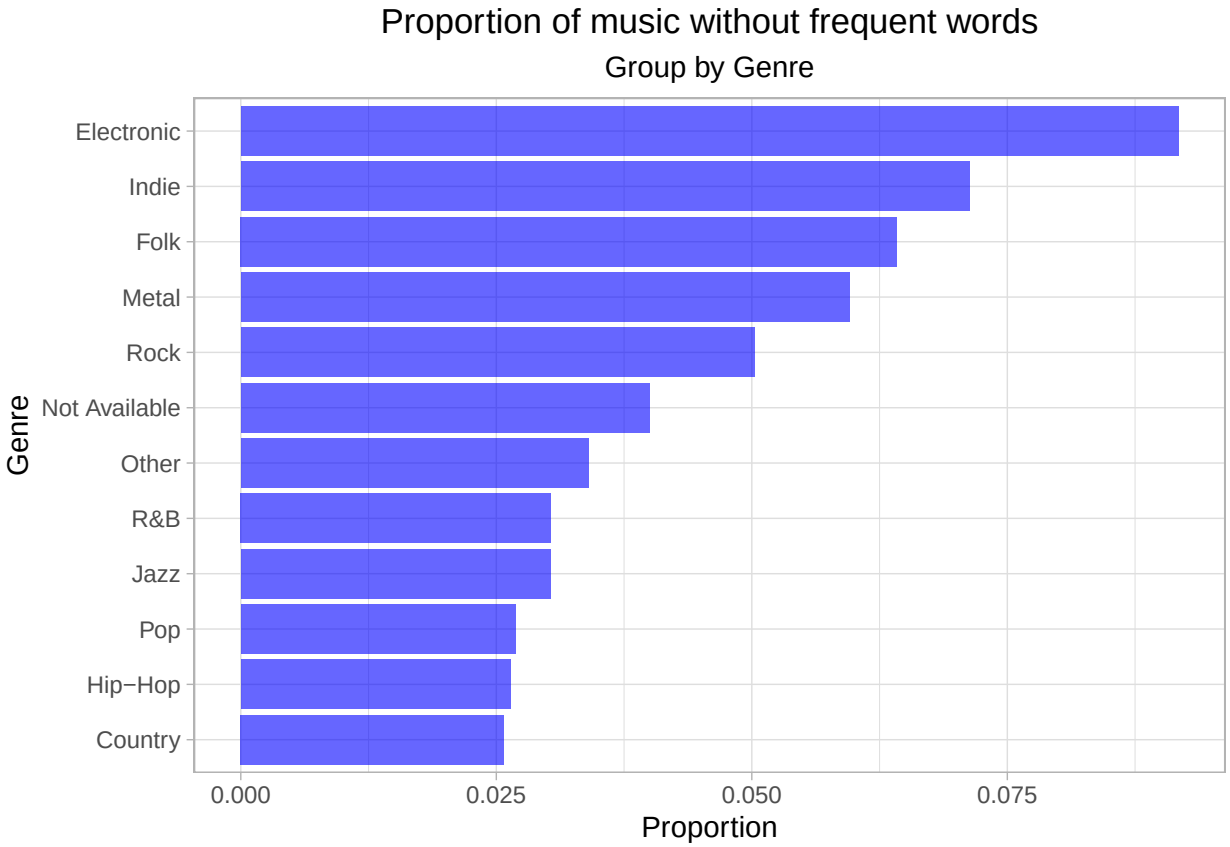

Proportion of negative song
Group by Year
Year
1968
1986
1982
1998
1974
2007
2005
2010
2011
2009
1987
2000
2006
2016
2015
2013
2004
2002
1994
2012
0.00
0.25
0.50
0.75
1.00
Proportion

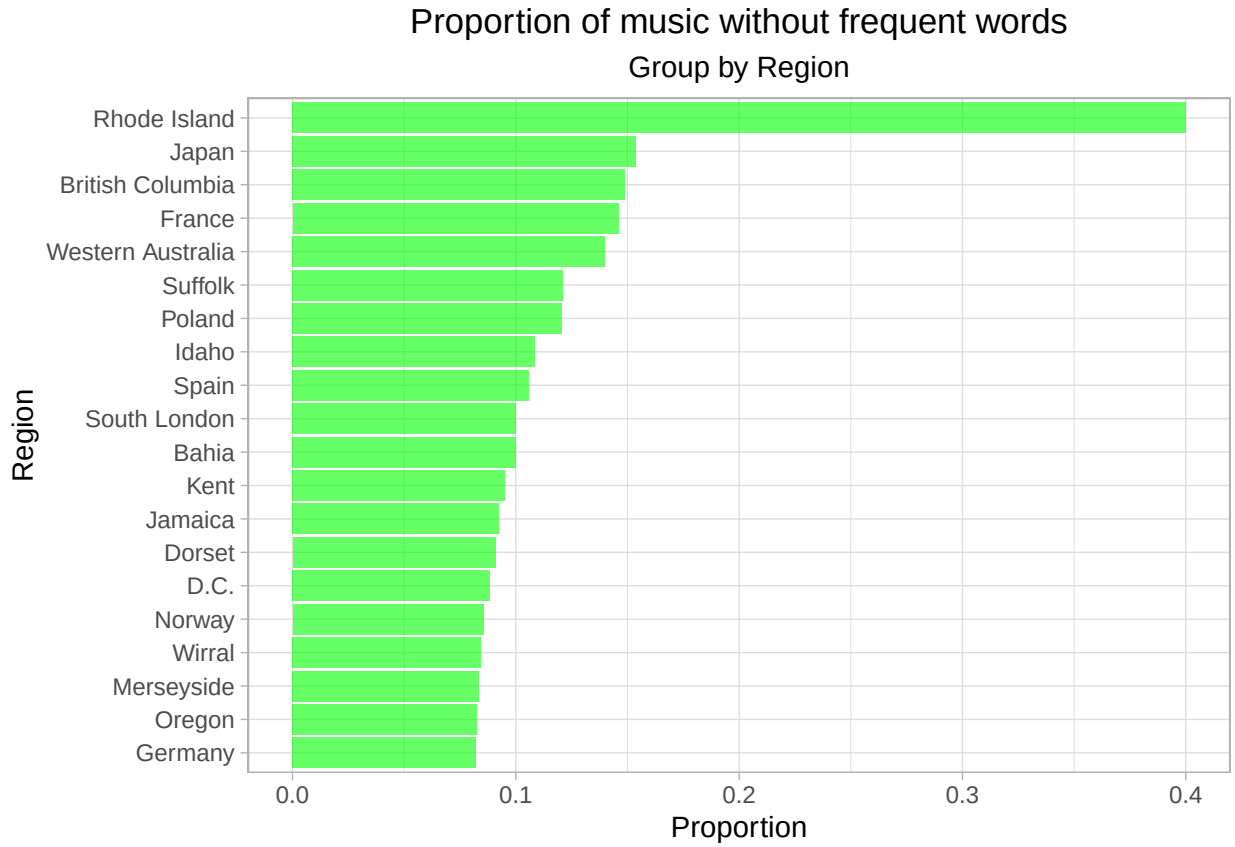


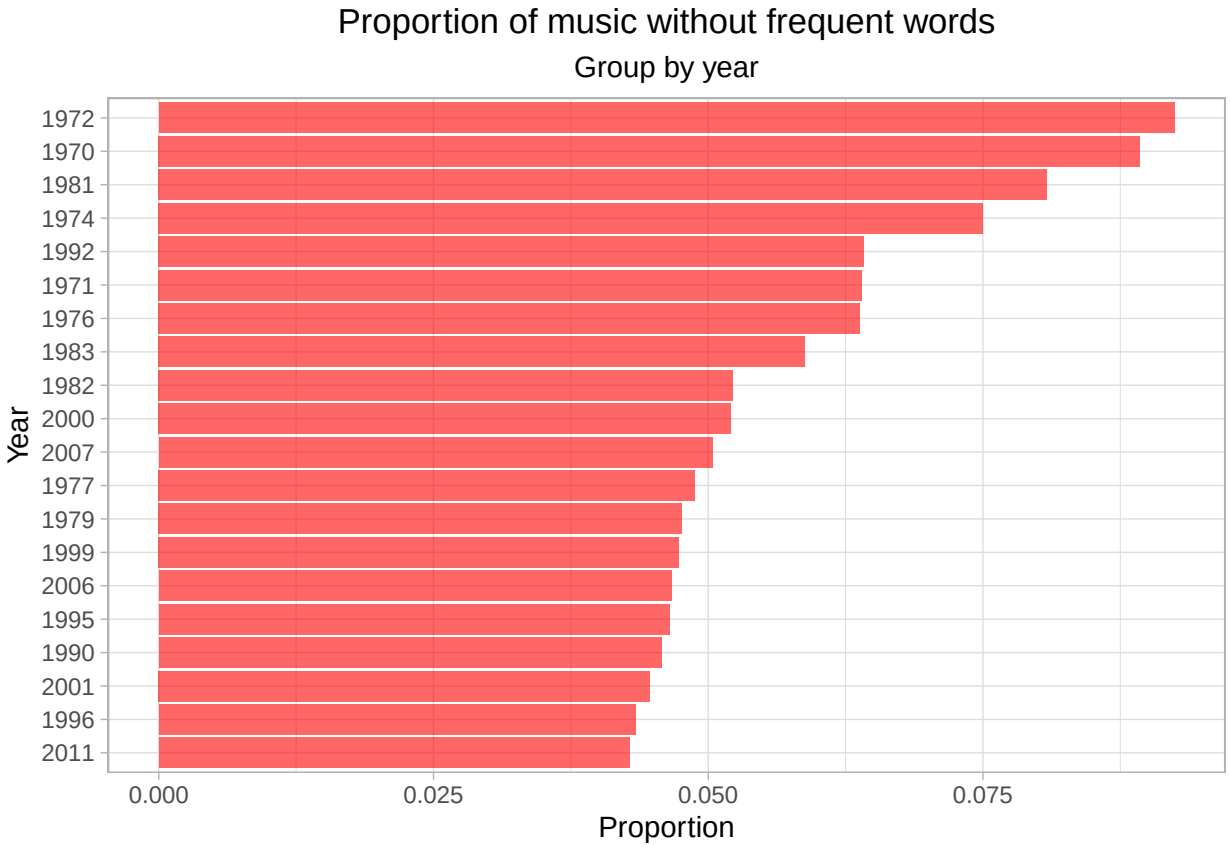
Metal and Hip-hop are the music genres with the most negative words and Jazz is with the least. Columbia, Phillippines and Russia are the regional using positive words while most of others use negative words more. Since there was only one song in 1968, the proportion reached 100%. In addition, the negative 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 15 words, we can get music statistics in different categories as follows.


Proportion of music without frequent words
Group by Genre
Genre
Electronic
Indie
Folk
Metal
Rock
Not Available
Other
R&B
Jazz
Pop
Hip−Hop
Country
0.000
0.025
0.050
0.075
Proportion


Proportion of music without frequent words
Group by Region
Region
Rhode Island
Japan
British Columbia
France
Western Australia
Suffolk
Poland
Idaho
Spain
South London
Bahia
Kent
Jamaica
Dorset
D.C.
Norway
Wirral
Merseyside
Oregon
Germany
0.0
0.1
0.2
0.3
0.4
Proportion



These statisical graphs show that electronic music uses the least high-frequency words when classifying music genres. Rhode Island, Japan, British Columbia, France and Western Australia use the least frequent words when classifying by country and people used the least frequent words in 1972, 1970, 1981 and 1974.Then, what are the musical keywords for these genres, places and times?

If we don't remove high-frequency words, taking Electronic genre as an example, then even the most distinctive genre, its main words are still 'love' and 'time'. We can't see the essence behind lyrics just by categorizing them like this, because the characteristics of these words determine that they are suitable to every music. So in the following analysis, I removed the 15 most frequent words.


remember
arms
forever
coming
game
breath
mine
wrong
till
chorus
sing
change
people
blue
makes
found
hope
cry
play
lie
dark
fire
fear
black
beat
gotta
stay
leave
body
tears
cold
tonight
head
sun
run
soul
door
call
start
bring
care
feeling
stand
burn
walk
light
friend
takes
line
real
stop
pain
god
hide
id
hear
dream
dance
sky
waiting
inside
fight
girl
hold
fall
lost
hard
left
x
move
true
sleep
fly
die
kiss
words
youll
watch
home
listen
break
free
youve
close
shine
round
whats
touch
stars
talk
chance
deep
sound
boy
lose
told
smile
set
understand
sweet


friend
understand
sing
begin
reason
silence
cold
set
coming
lose
youve
stars
change
tears
human
tonight
fight
run
kill
break
learn
lay
move
stay
call
stand
sun
air
wrong
blood
dead
dark
fear
fill
dream
words
pay
sleep
fire
left
soul
fall
chorus
close
illusion
grow
head
light
die
watch
kiss
waiting
burn
inside
pain
hold
alive
youll
god
forget
hope
feeling
lie
lost
finally
force
shes
free
hear
moment
deep
leave
sky
start
stop
home
walk
earth
hate
body
ground
wall
strong
sight
cry
power
shadows
bring
takes
play
hide
door
sound
rise
late
breath
control
forever
hell
skin

The hue of the graphs above are consistent with the previous graphs. We can see that 'feeling', 'light', 'dream' are more used by Electronic music. These words also match the entertaining nature of this kind of music. 'burn', 'fall', '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.