G B

ology@github

Search this website .

Visualization of Led Zeppelin Lyrics with R

2017-06-01 BY GENE

What can be known about Led Zeppelin lyrics from the standpoint of a computer geek?

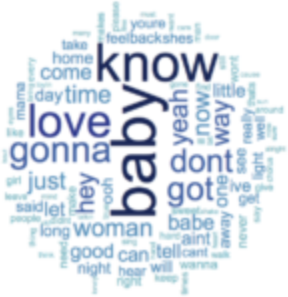
First, I collected and properly named every song with lyrics, with the power of perl and persistence.

Then I found/crafted some R code to process these into a few graphics.

Here is the code for the first graph:

```
library(tm)
library(stringr)
library(wordcloud)

mytext <- Corpus( DirSource('~/Documents/lit/Led-Zeppelin') )

mytext <- Corpus( VectorSource(mytext) )
mytext <- tm_map( mytext, removeNumbers )
mytext <- tm_map( mytext, removePunctuation )
mytext <- tm_map( mytext, content_transformer(tolower) )
mytext <- tm_map( mytext, removeWords, stopwords("english") )
mytext <- tm_map( mytext, stripWhitespace )

pal <- brewer.pal( 9, 'YlGnBu' )
pal <- pal[-(1:4)]

set.seed(123)
wordcloud(
    words = mytext,
    scale = c( 5, 0.1 ),
    max.words = 100,
    random.order = FALSE,
    rot.per = 0.35,
    use.r.layout = FALSE,
    colors=pal
)
```

And here is that “word cloud” graphic showing the most spoken words. And what is the most for Led Zeppelin? It’s “baby” of course!



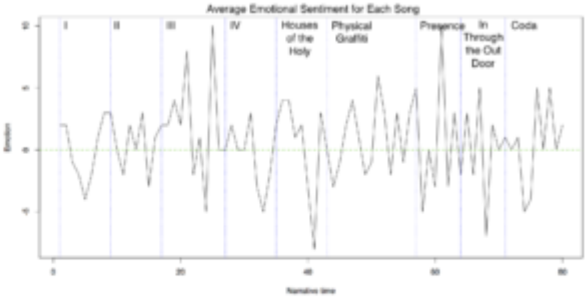
Here are two emotional sentiment charts. And what is “emotional sentiment?” Well, basically it is the assignment of a score to a word based on whether it is positive, neutral or negative. How is this known? Well, a bunch of people (university students, I think) tediously categorized huge word lists as to whether they made them feel good bad or indifferent. And it is these scored word lists that are used in the creation of the following charts.

Here is the code to generate the first graph below:

```
library(syuzhet)

path <- '~/Documents/lit/Led-Zeppelin'
```

```
files <- sort( dir(path) )

df <- data.frame(file = files, stringsAsFactors=FALSE)
df$fullPath <- paste(path, df$file, sep = "/")
df$text <- sapply(df$fullPath, get_text_as_string)

mysentiment <- get_sentiment( df$text, method = 'bing' )
plot( mysentiment, type = 'l', xlab = 'Narrative time', ylab = 'Emotion' )
abline(h = 0, col = 3, lty = 2)

refpoint <- function (string) {
    posn = which( grepl( string, df$text ) );
    abline( v = posn, col = 4, lty = 3 )
}

refpoint('Led Zeppelin I')
refpoint('Houses of the Holy')
refpoint('Physical Graffiti')
refpoint('Presence')
refpoint('In Through the Out Door')
refpoint('Coda')
```

The first shows the averaged positive, neutral and negative sentiment for each song, in order of their release and album play position.

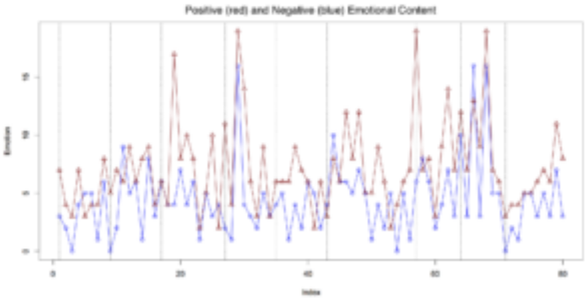


We can see that every album has positive and negative moods.

This bit of code, given the previous program, shows that most positive song is “Bron-Y-Aur Stomp” on III. (The other identical peak on Presence turned out to be an accidental duplicate! Oops.) The most negative song is “No Quarter” from Houses of the Holy.

```
df$file[ which( mysentiment == max(mysentiment) ) ]
[1] "1970-Led-Zeppelin-III-09-Bron_Y_Aur_Stomp.txt"

df$file[ which( mysentiment == min(mysentiment) ) ]
[1] "1973-Houses-of-the-Holy-07-No_Quarter.txt"
```

The second is a line chart showing the amount of positive (red) and negative (blue) emotional sentiment for each song.



We can see that Robert Plant sings slightly more positive words throughout his time with the band.

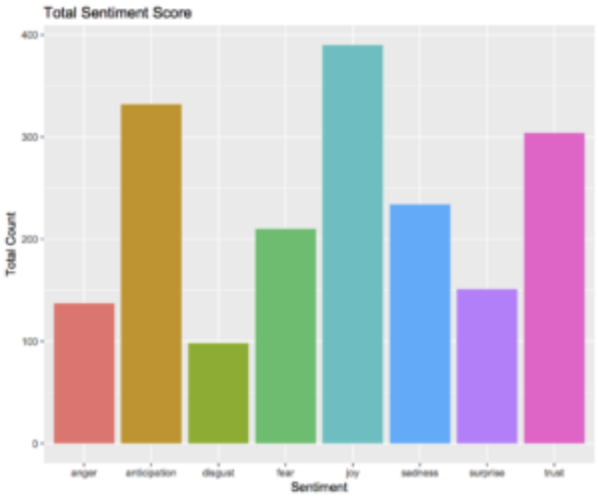
Here is the code for that second graph:

```
library(syuzhet)

path <- '~/Documents/lit/Led-Zeppelin'
files <- sort( dir(path) )

df <- data.frame(file = files, stringsAsFactors=FALSE)
df$fullPath <- paste(path, df$file, sep = "/")
df$text <- sapply(df$fullPath, get_text_as_string)

df <- cbind(df, get_nrc_sentiment(df$text))

max_y <- max(df$positive)
plot(df$negative, type="o", col="blue", ylim=c(0,max_y), ylab="Emotion")
lines(df$positive, type="o", pch=2, col="darkred")

refpoint <- function (string) {
    posn = which( grepl( string, df$text ) );
    abline( v = posn, col = 1, lty = 3 )
}
```

```
refpoint('Led Zeppelin I')
refpoint('Houses of the Holy')
refpoint('Physical Graffiti')
refpoint('Presence')
refpoint('In Through the Out Door')
refpoint('Coda')
```

We can inspect the relative amounts of different emotional categories too! These are anger, anticipation, disgust, fear, joy, sadness, surprise and trust.

```
# With df from above:

library(ggplot2)

sentimentTotals <- data.frame(colSums(df[,4:11]))

names(sentimentTotals) <- "count"
sentimentTotals <- cbind("sentiment" = rownames(sentimentTotals), sentiment
rownames(sentimentTotals) <- NULL

ggplot(data = sentimentTotals, aes(x = sentiment, y = count)) +
    geom_bar(aes(fill = sentiment), stat = "identity") +
    theme(legend.position = "none") +
    xlab("Sentiment") +
    ylab("Total Count") +
    ggtitle("Total Sentiment Score")
```



Anticipation, joy and trust are outstanding feelings in Led Zeppelin songs, according to this chart.

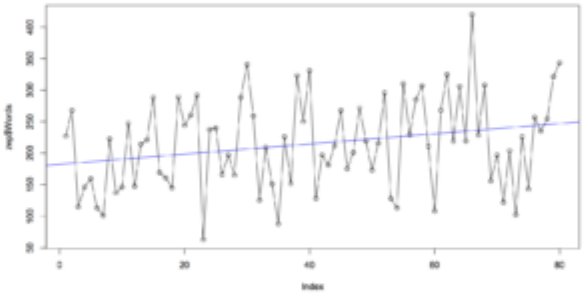
Finally, I wanted to see the basic stats of each song, from my fathom program. In tabular format this shows that lyrics have very high and skewed readability scores. This might be because songs are generally not punctuated, so the sentence complexity is not useful. Anyway, here is an example of this output:

```
$ perl fathom ~/Documents/lit/Led-Zeppelin > ~/sandbox/dev/Lex/Led-Zeppelin
```

```
Chars   Words   Complex Senten  SylW    WpS     FOG     Flesch  Kincaid Nam
1434    285     2.11    4       1.21    71.25   29.34   32.11   26.48   197
1490    307     1.63    10      1.20    30.70   12.93   74.54   10.49   197
1076    211     4.74    6       1.22    35.17   15.96   68.10   12.50   197
545     108     5.56    1       1.31    108.00  45.42   -13.23  41.94   197
1304    268     1.12    14      1.16    19.14   8.10    89.23   5.57    197
1598    325     0.92    5       1.18    65.00   26.37   41.16   23.67   197
1049    219     0.91    7       1.18    31.29   12.88   75.41   10.51   197
```

But I am curious about the words per song over time. So I whipped up a bit more R:

```
zep <- read.table('~/sandbox/dev/Lex/Led-Zeppelin-stats.txt', header=T)

plot(zep$Words, type="o")

index <- 1:nrow(zep)
fit <- lm(zep$Words ~ index)
abline(fit, col="blue")
```



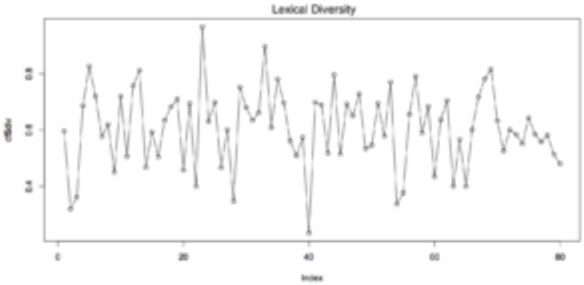
We can see that Robert wrote slightly more words per song over his career with Zeppelin. Notice that the last album, Coda is full of songs from earlier days with shorter songs. This pulls down the trend line on the right end.

Since we are here, what is the song with the most words?

```
zep[ which( zep$Words == max(zep$Words) ), ]

... 1979-In-Through-the-Out-Door-03-Fool_In_The_Rain
```

What about the richness of the language (“lexical diversity”) used? Does that change over time?

```
library(syuzhet)

library(tm)
library(NLP)
library(openNLP)

path <- '~/Documents/lit/Led-Zeppelin'
files <- sort( dir(path) )

df <- data.frame(file = files, stringsAsFactors=FALSE)
df$fullPath <- paste(path, df$file, sep = "/")
df$text <- sapply(df$fullPath, get_text_as_string)

my_get_words <- function (intext) {
    processed <- Corpus( VectorSource(intext) )
    processed <- tm_map( processed, removeNumbers )
    processed <- tm_map( processed, removePunctuation )
    processed <- tm_map( processed, content_transformer(tolower) )
    processed <- tm_map( processed, removeWords, stopwords("english") )
    processed <- as.String( as.character( processed[[1]] ) ) # Necessary fo
    sent_ann <- Maxent_Sent_Token_Annotator()
    word_ann <- Maxent_Word_Token_Annotator()
    myannotation <- annotate( processed, list( sent_ann, word_ann ) )
    processed_doc <- AnnotatedPlainTextDocument( processed, myannotation )
    mywords <- words(processed_doc)
    lex <- length( unique(mywords) ) / length(mywords)
    return(lex)
}

df$div <- sapply(df$text, my_get_words)
plot(df$div, type="o")
```



Nope!

FILED UNDER: MUSIC, SOFTWARE
TAGGED WITH: LED ZEPPELIN, R, VISUALIZATION