Plan:

1. Work through a text analysis of song lyrics

Text Analysis: Example

Shannon E. Ellis, Ph.D
UC San Diego

Department of Cognitive Science
sellis@ucsd.edu

Today's example question: How has pop music changed in the last four years?

Goal: Understand the basics of sentiment analysis and TF-IDF

What data would we need to answer this question?

How has pop music changed in the last four years?

Data: Lyrics to the most popular songs from each year

The data : Top songs from Feb music charts 2017-2020

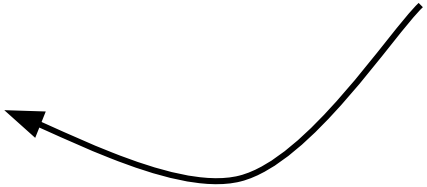
2017: 152 songs

2018: 139 songs

2019: 127 songs

2020: 137 songs

Song data from **Spotify**.
Lyrics from **genius.com**



Questions we can ask...

1. Does the total number of words change over time?
2. Does uniqueness change over time?
3. Does the diversity or density change?
4. What words are most common?
5. What words are most unique to each year?
6. What sentiment do songs convey most frequently?
7. Has sentiment changed over time?
8. What are the sentiment of the #1 songs?
9. What words contribute to the sentiment of these #1 songs?
10. ...what about bigrams? N-grams?

Questions we can ask...

1. Does the total number of words change over time?
2. Does uniqueness change over time?
3. Does the diversity or density change?

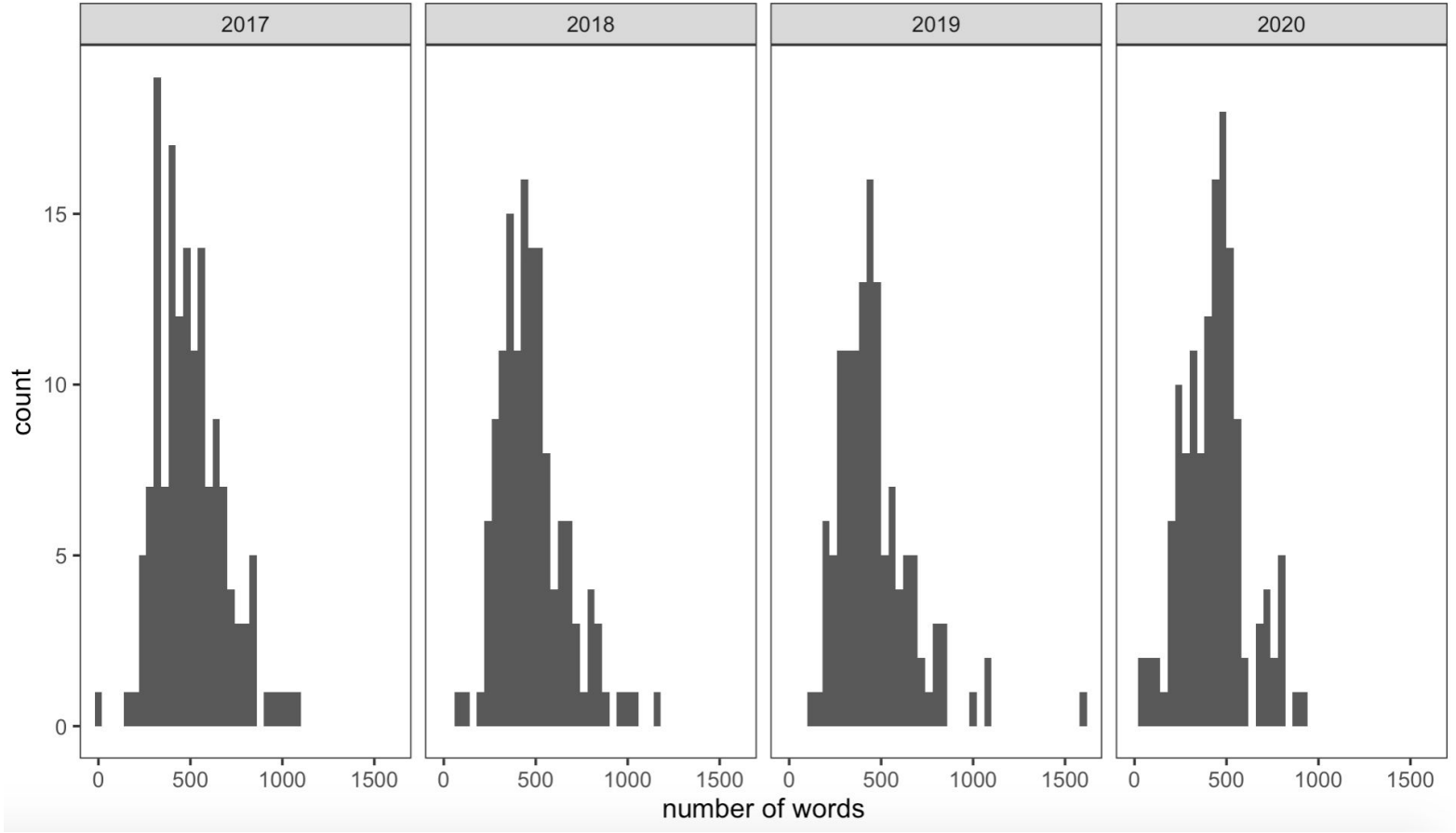
EDA

4. What words are most common?
5. What words are most unique to each year?

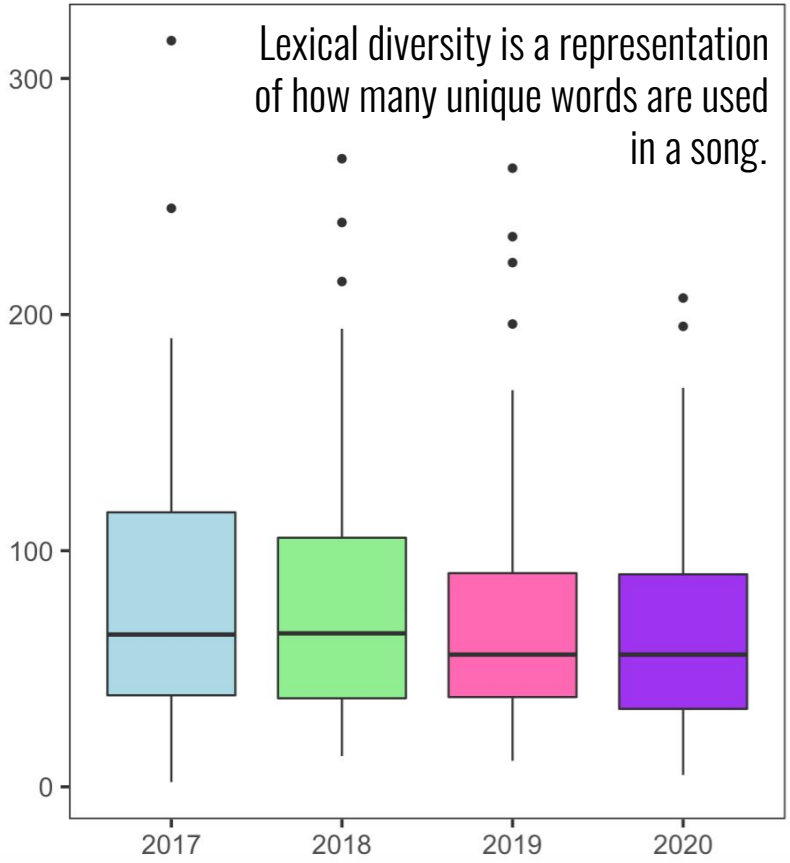
TF-IDF

6. What sentiment do songs convey most frequently?
7. Has sentiment changed over time?
8. What are the sentiment of the #1 songs?
9. What words contribute to the sentiment of these #1 songs?
10. ...what about bigrams? N-grams?

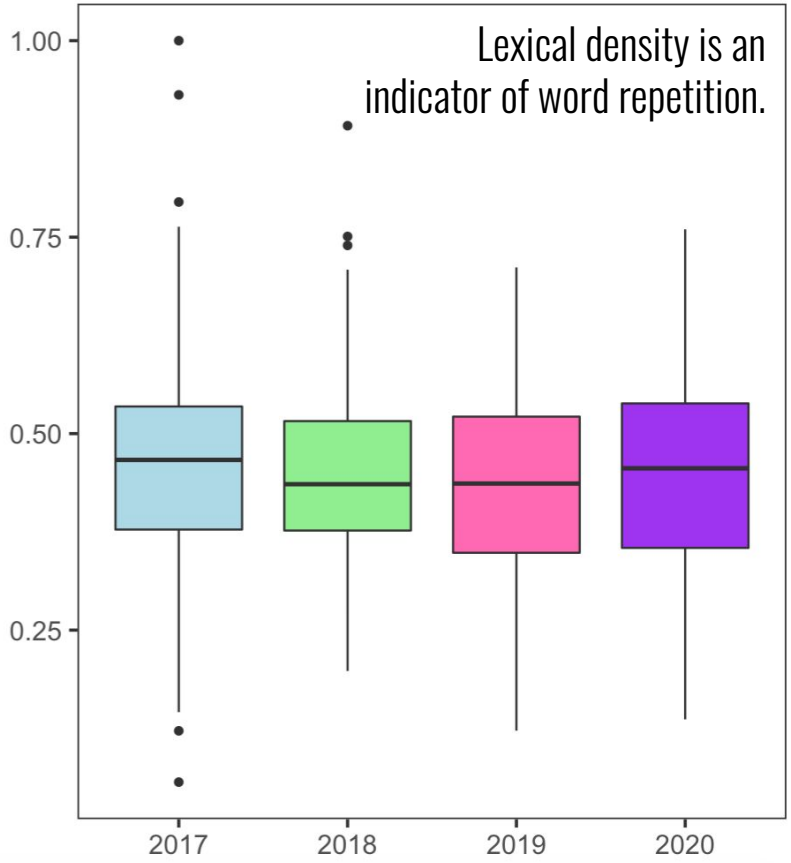
Sentiment Analysis


2017
2018
2019
2020
count
0
5
10
15
0
500
1000
1500
number of words

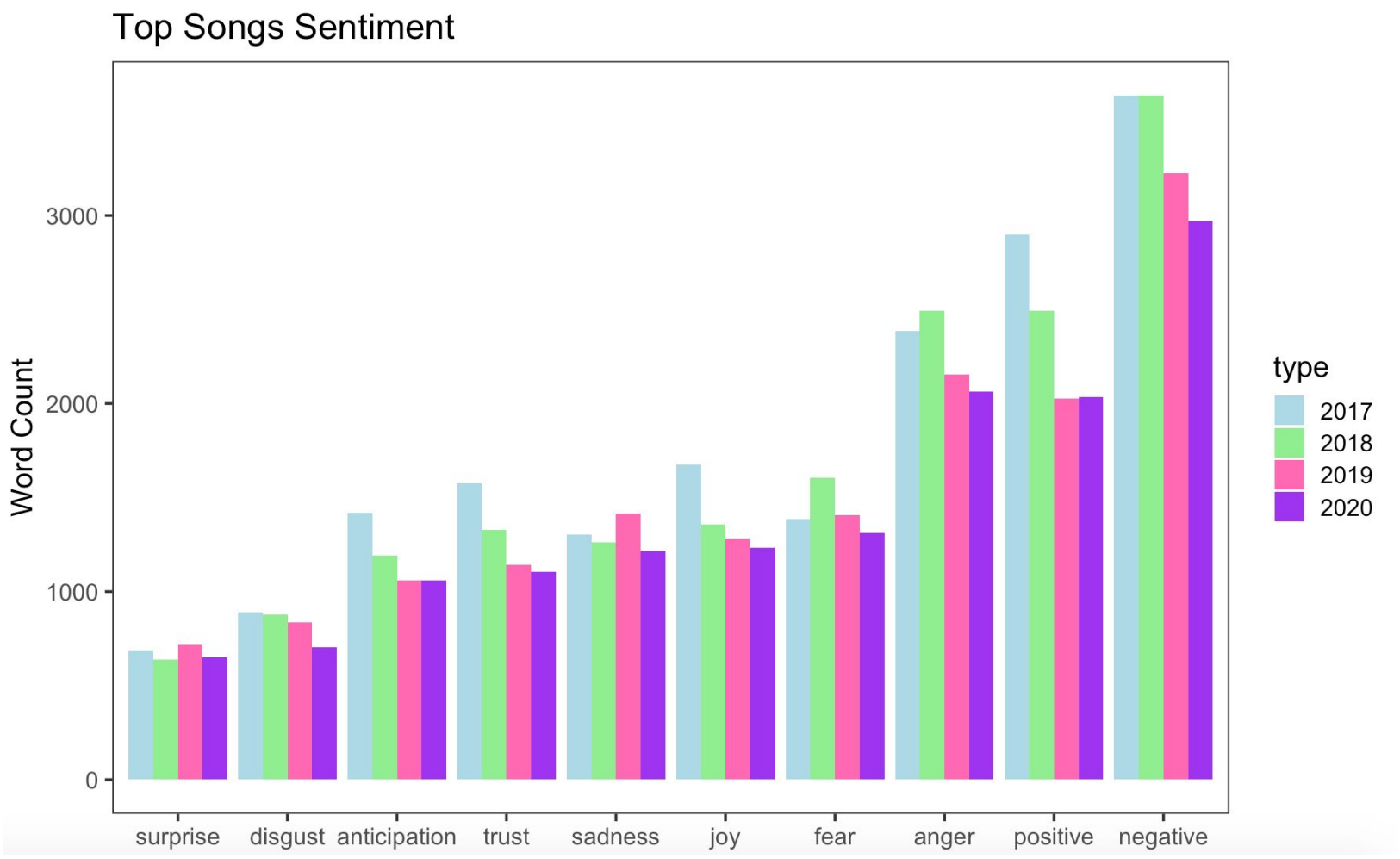
Lexical Diversity

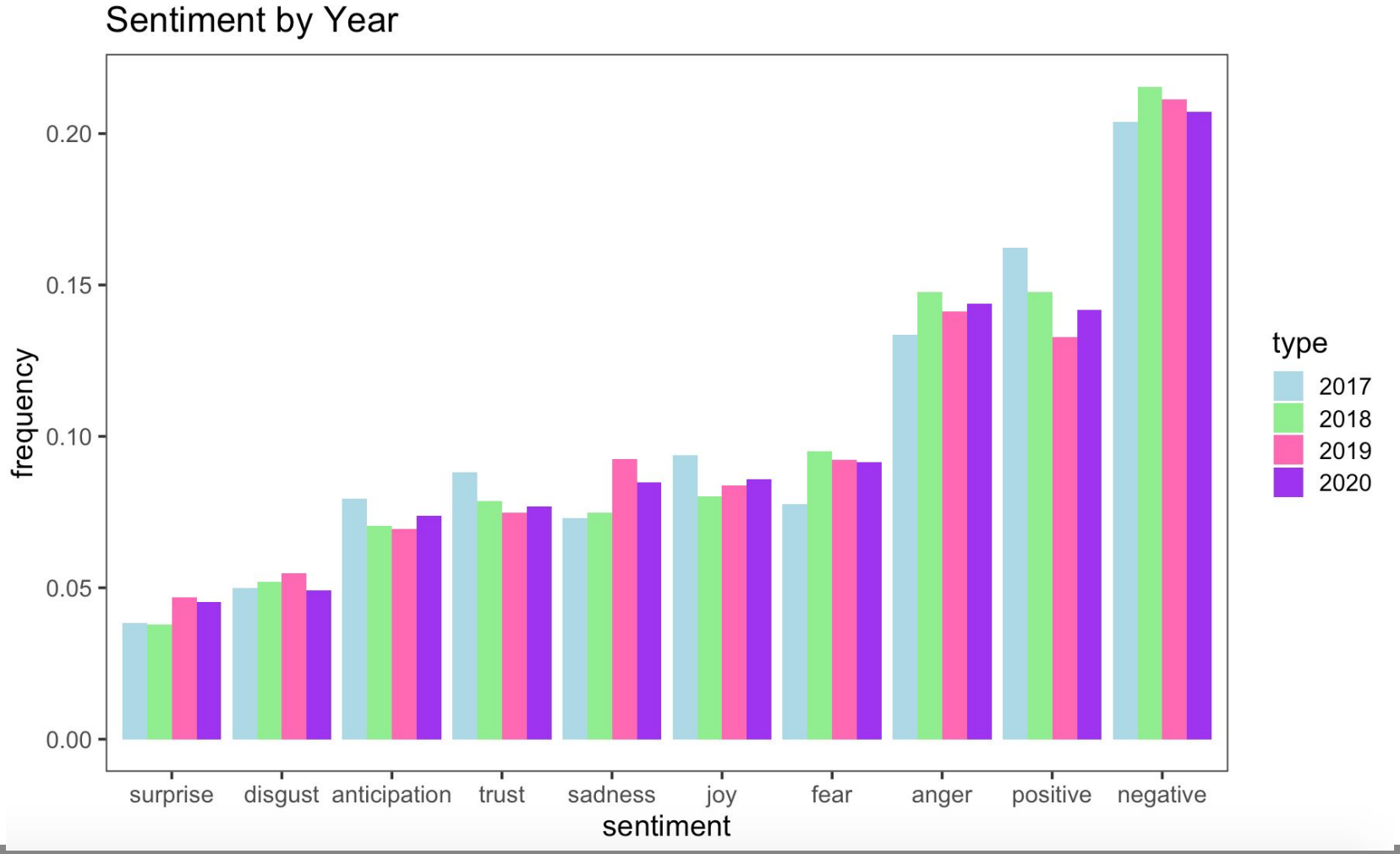


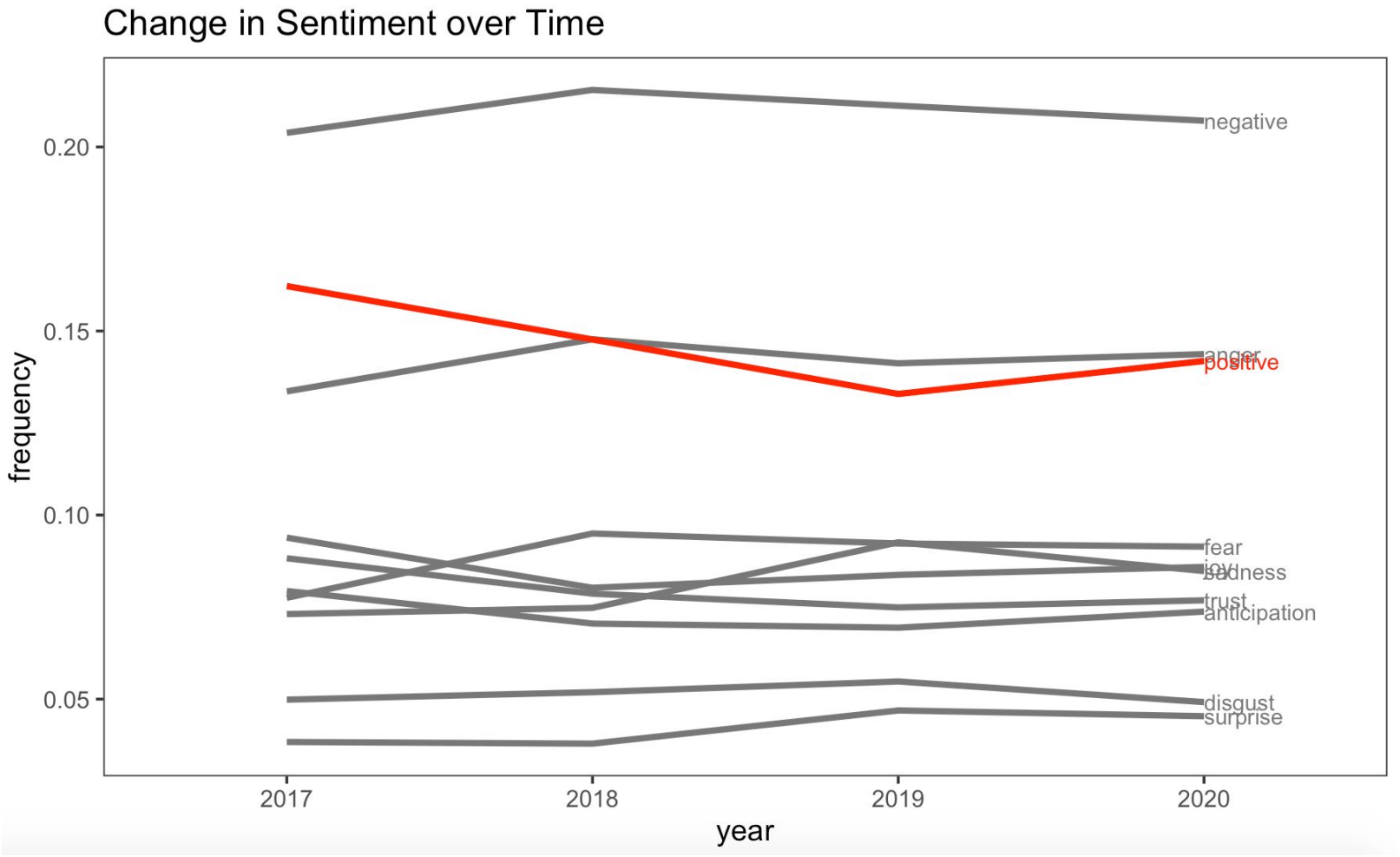
Lexical Density

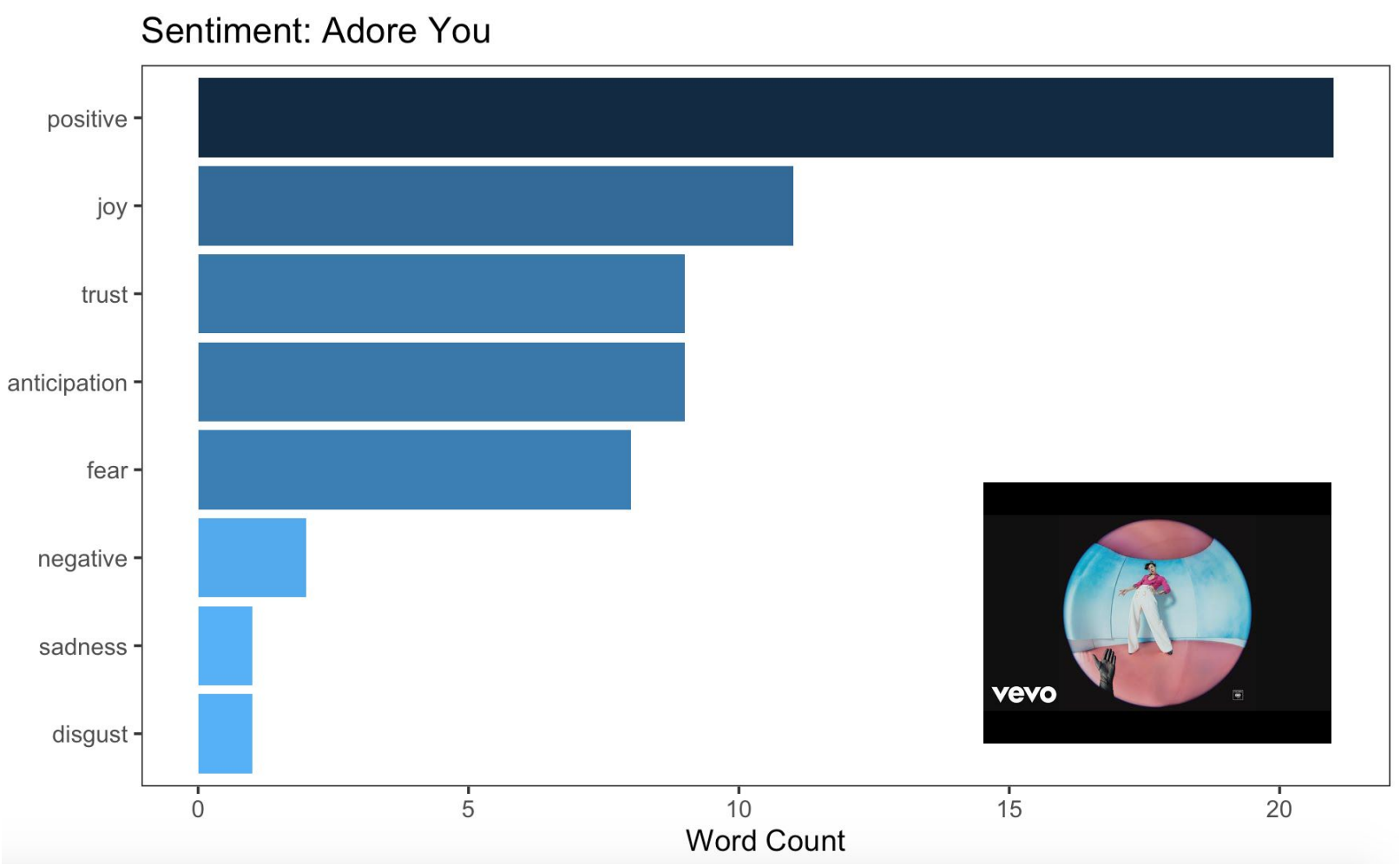


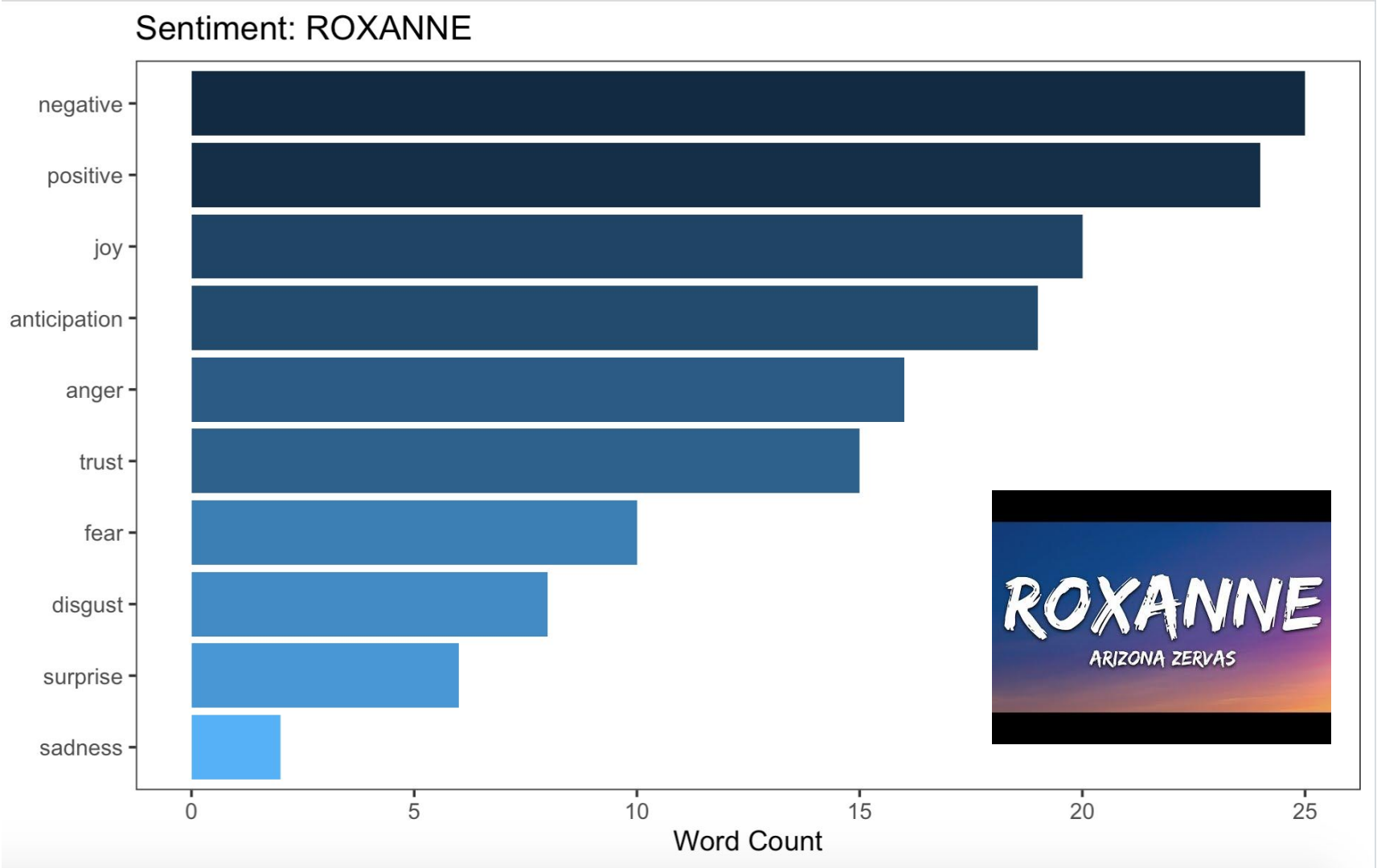
Sentiment Analysis


Top Songs Sentiment
Word Count
0
1000
2000
3000
surprise
disgust
anticipation
trust
sadness
joy
fear
anger
positive
negative
type
2017
2018
2019
2020


Sentiment by Year
frequency
0.20
0.15
0.10
0.05
0.00
surprise
disgust
anticipation
trust
sadness
joy
fear
anger
positive
negative
sentiment
type
2017
2018
2019
2020


Change in Sentiment over Time
0.20
0.15
0.10
0.05
frequency
2017
2018
2019
2020
year
negative
anger
positive
fear
joy
sadness
trust
anticipation
disgust
surprise


Sentiment: Adore You
positive
joy
trust
anticipation
fear
negative
sadness
disgust
0
5
10
15
20
Word Count
vevo


Sentiment: ROXANNE
negative
positive
joy
anticipation
anger
trust
fear
disgust
surprise
sadness
0
5
10
15
20
25
Word Count
ROXANNE
ARIZONA ZERVAS

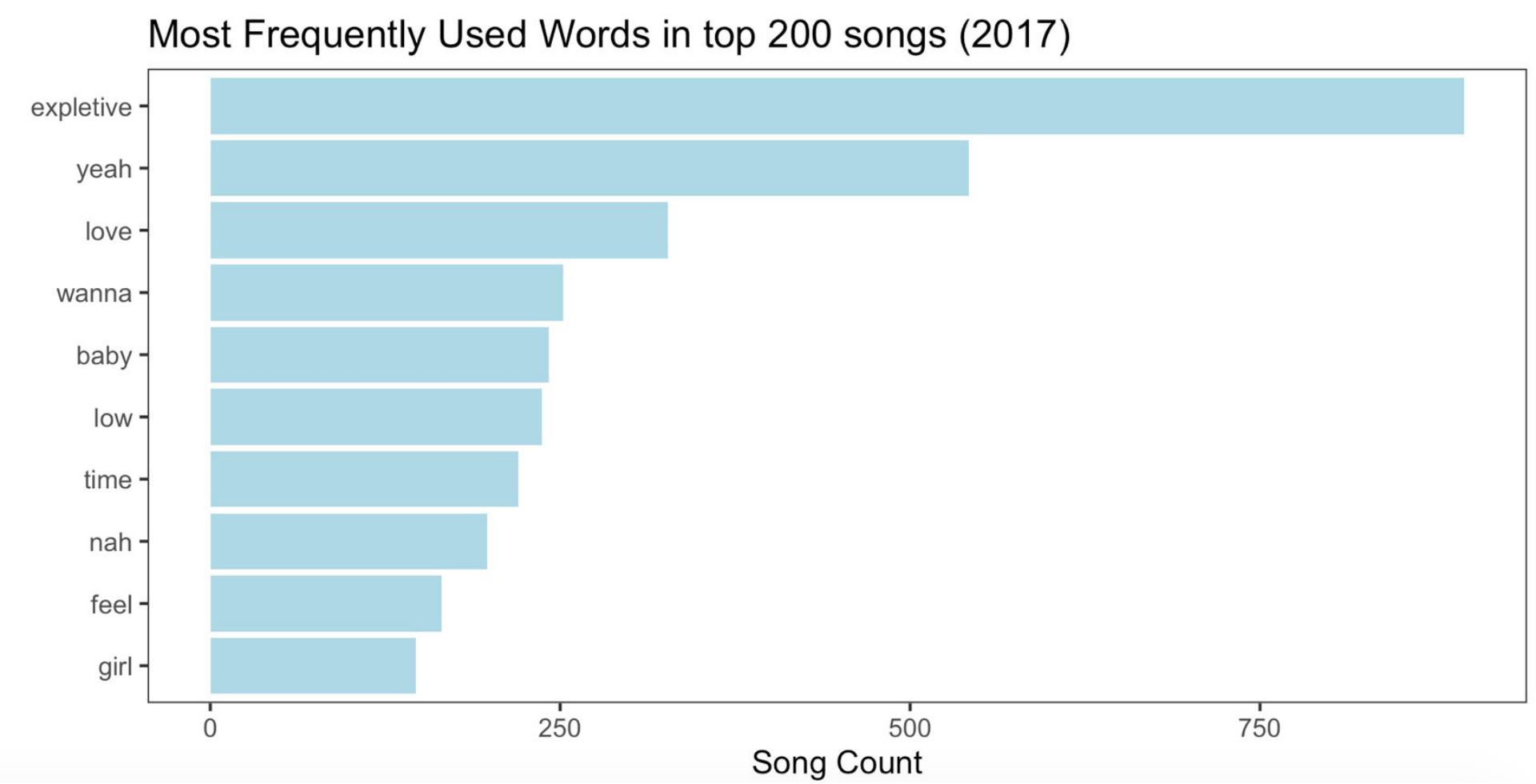
TF-IDF

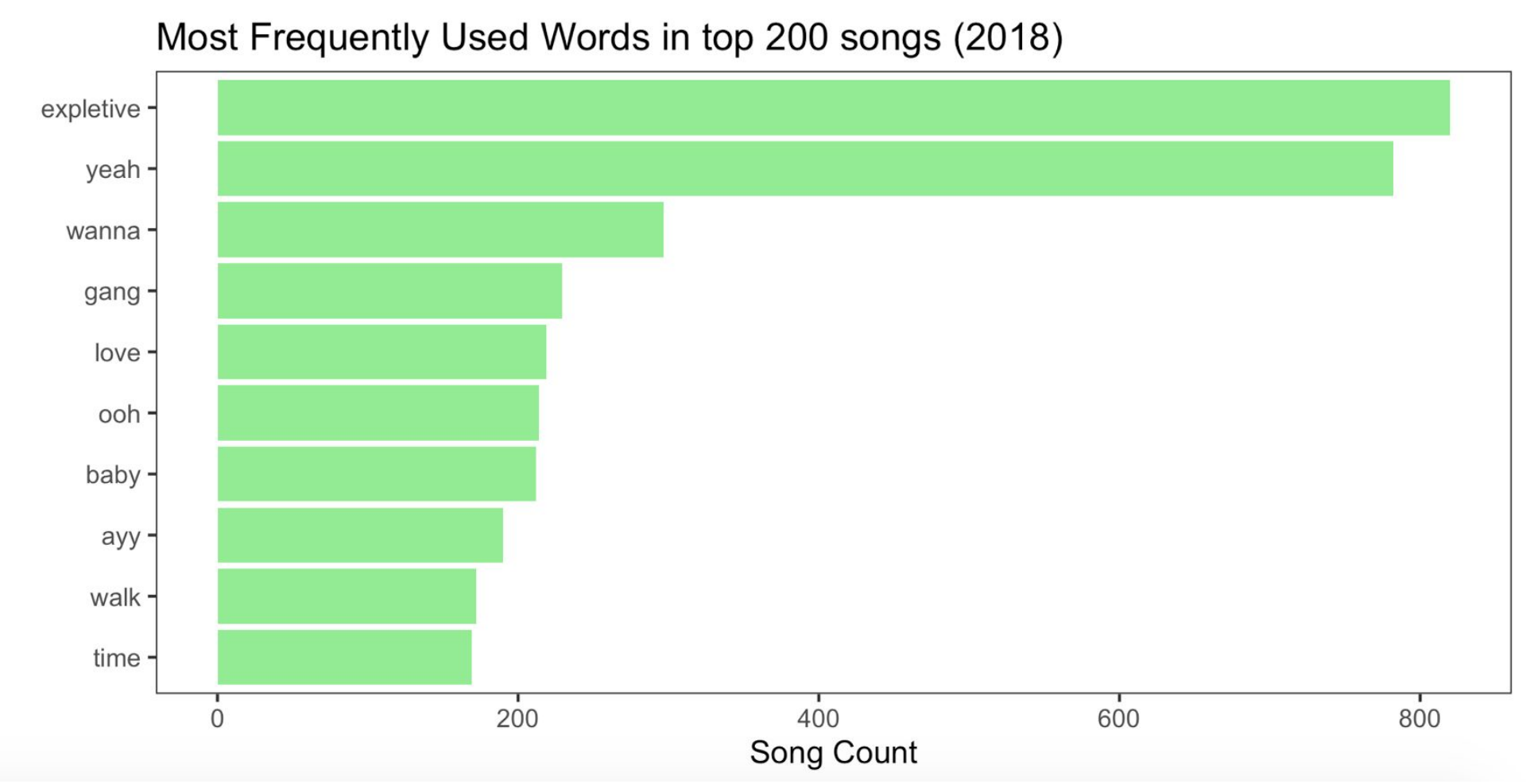
Term Frequency - Inverse Document Frequency

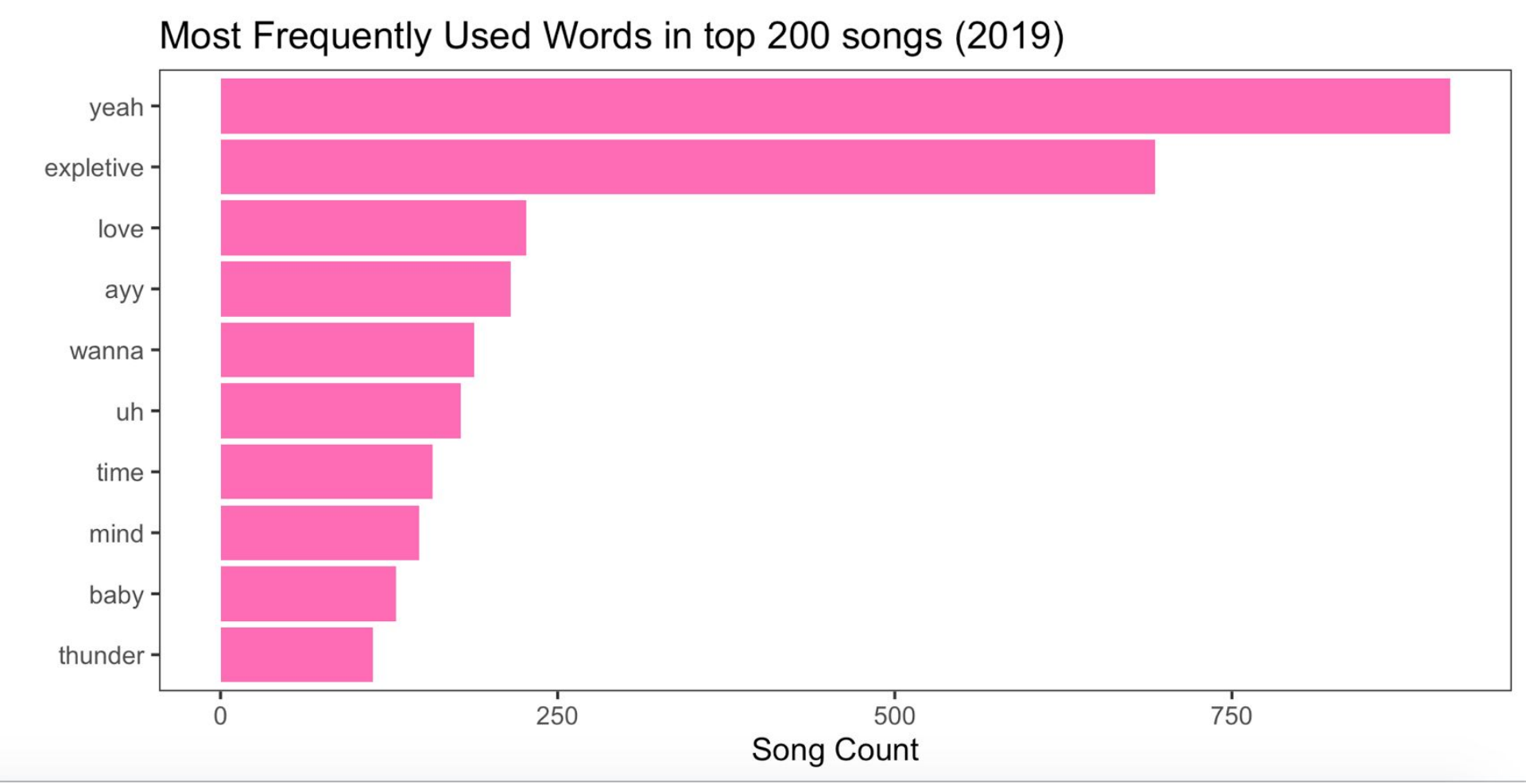
What words are the most unique to the lyrics of each year's top hits?

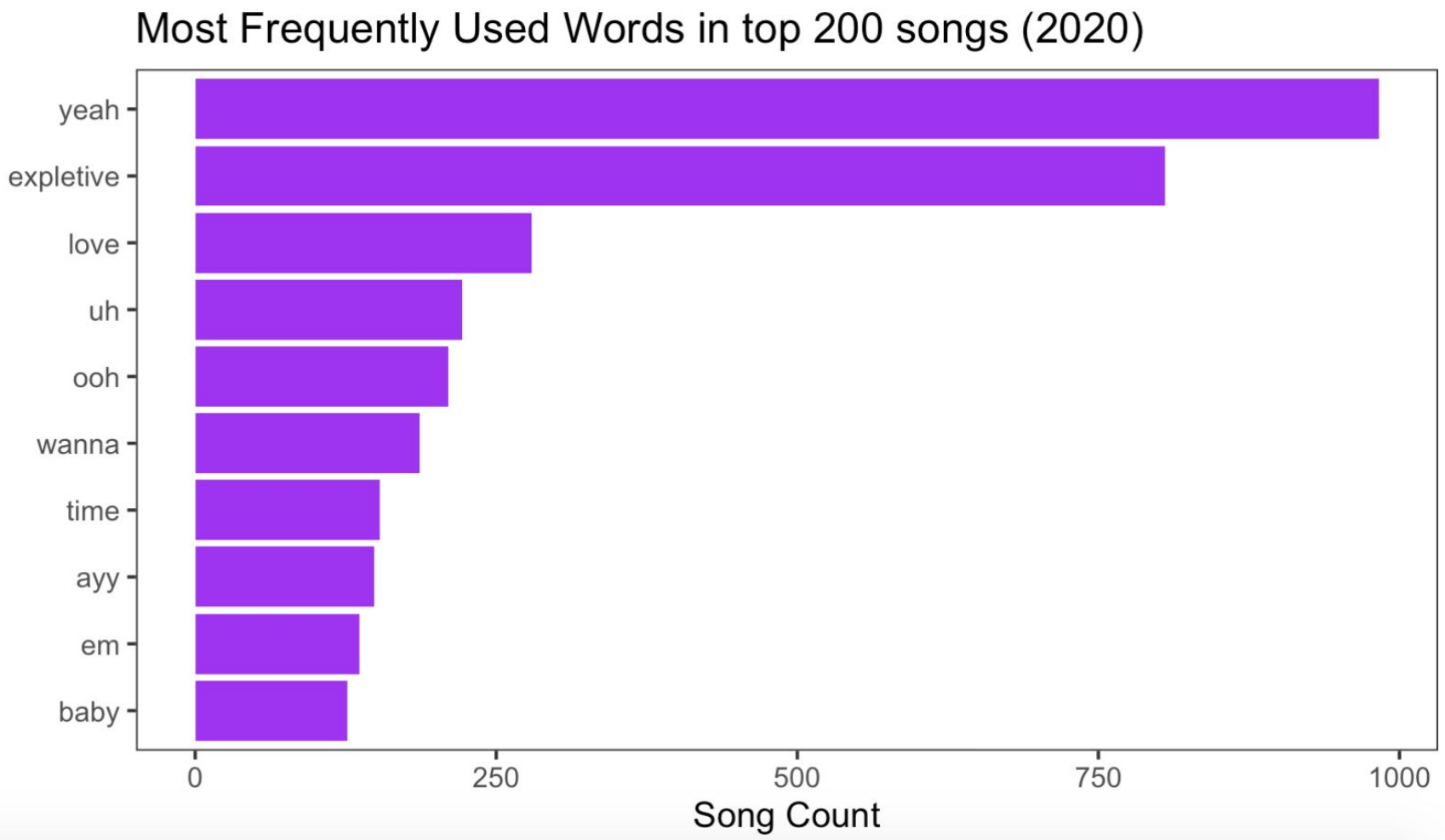
Goal: to use TF-IDF to *find the important words* for the content of each document by decreasing the weight for commonly used words and increasing the weight for words that are not used very much in a collection or corpus of documents

Calculating TF-IDF attempts to find the words that are important (i.e., common) in a text, but not *too* common

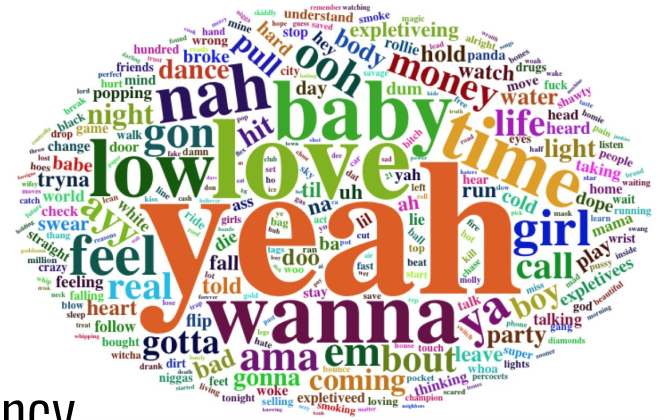

Most Frequently Used Words in top 200 songs (2017)
expletive
yeah
love
wanna
baby
low
time
nah
feel
girl
0
250
500
750
Song Count


Most Frequently Used Words in top 200 songs (2018)
expletive
yeah
wanna
gang
love
ooh
baby
ayy
walk
time
0
200
400
600
800
Song Count


Most Frequently Used Words in top 200 songs (2019)
yeah
expletive
love
ayy
wanna
uh
time
mind
baby
thunder
0
250
500
750
Song Count

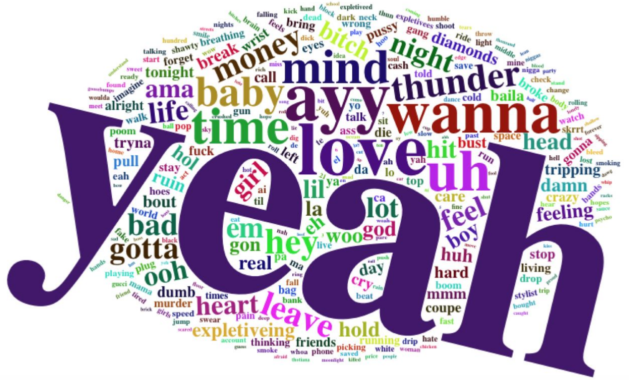

Most Frequently Used Words in top 200 songs (2020)
yeah
expletive
love
uh
ooh
wanna
time
ayy
em
baby
0
250
500
750
1000
Song Count

Term Frequency can only tell us so much....



2017

2018



2019

2020

TF-IDF:

Term Frequency - Inverse Document Frequency

the frequency of a term adjusted for how rarely it is used

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

Important Words using TF-IDF by Year