

Text analysis

Jason G. Fleischer, Ph.D.

Asst. Teaching Professor

Department of Cognitive Science, UC San Diego

jfleischer@ucsd.edu



@jasongfleischer

<https://jgfleischer.com>

Examples of questions that require text analysis

1. Did J.K. Rowling write The Cuckoo's Calling under the pen name Robert Galbraith?
2. What themes are common in 19th century literature?
3. Can we tell the difference between tweets that come from Trump himself or a staffer?
4. Is Hillary the most poisoned name in US History?

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

Data: Lyrics to the most popular songs
from each year

The data : Top songs from Feb music charts 2017-2021

2017: 152 songs

2018: 139 songs

2019: 127 songs

2020: 137 songs

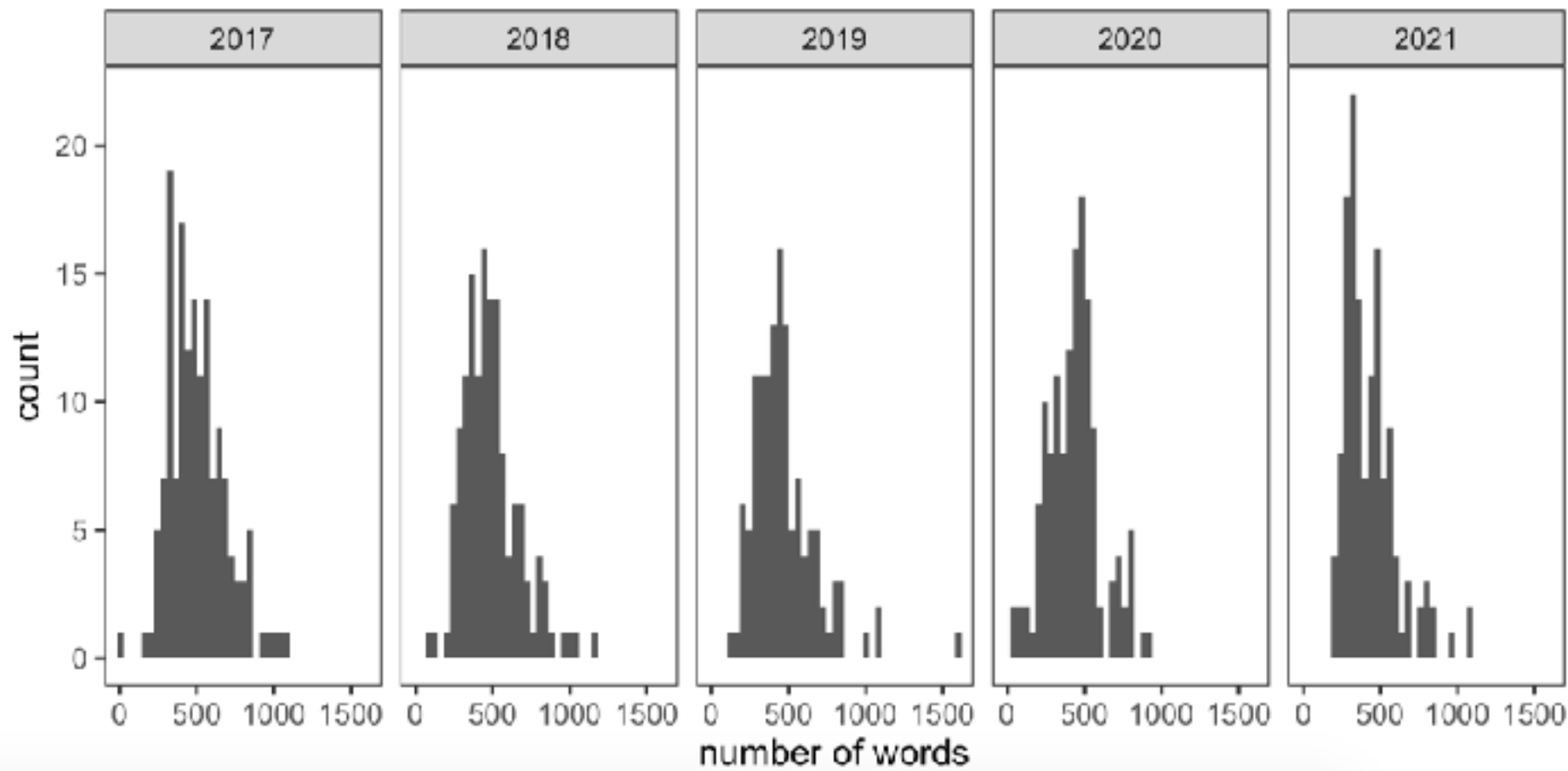
2021: 134 songs

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

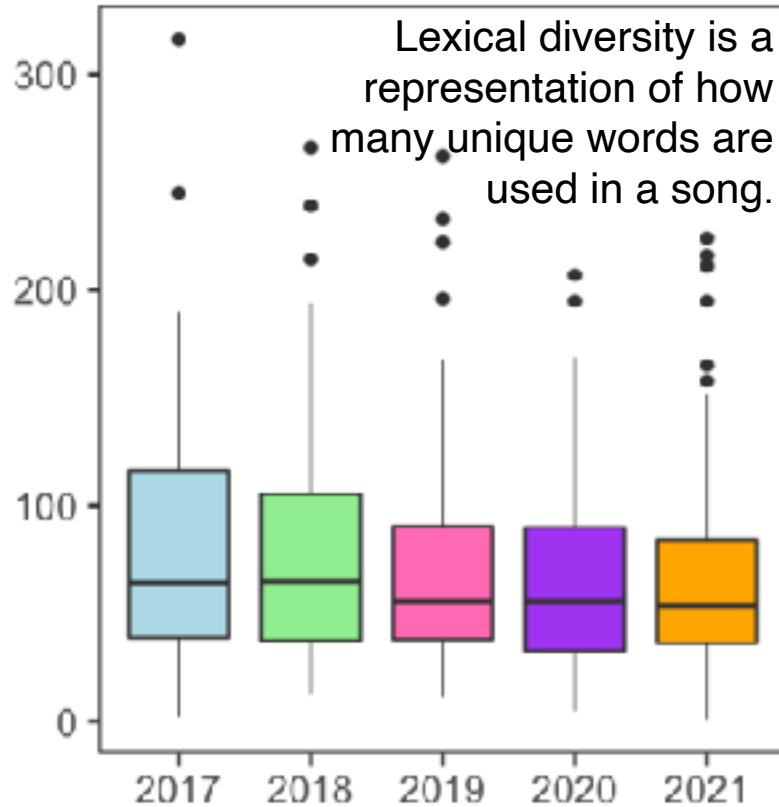


Questions we can ask...

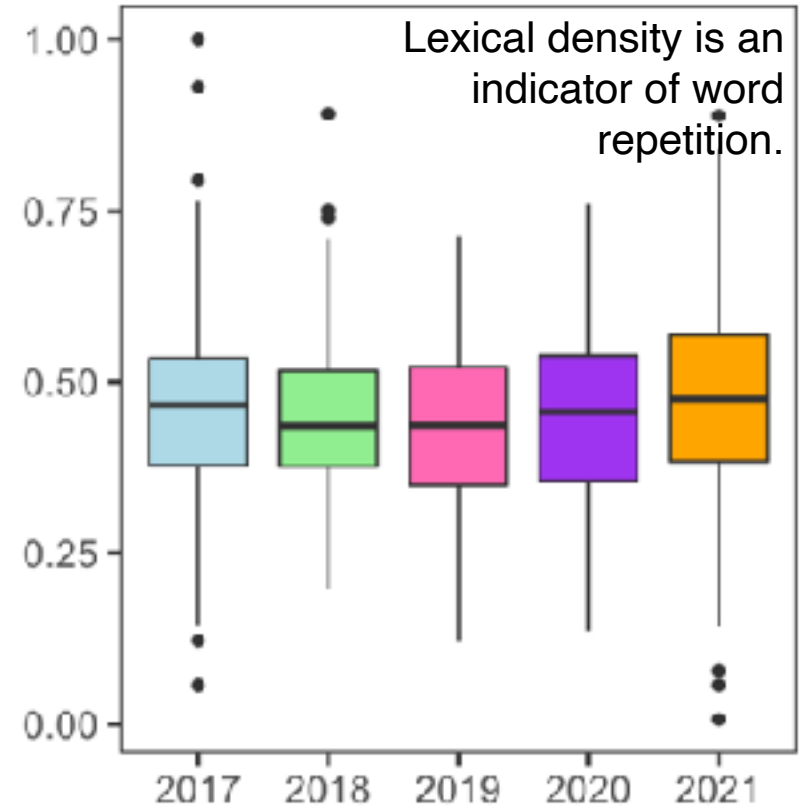
1. Does the total number of words change over time?
2. Does uniqueness change over time?
3. Does the diversity or density change?
4. What words are most common?
5. What words are most unique to each year?
6. What sentiment do songs convey most frequently?
7. Has sentiment changed over time?
8. What are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



Lexical Diversity



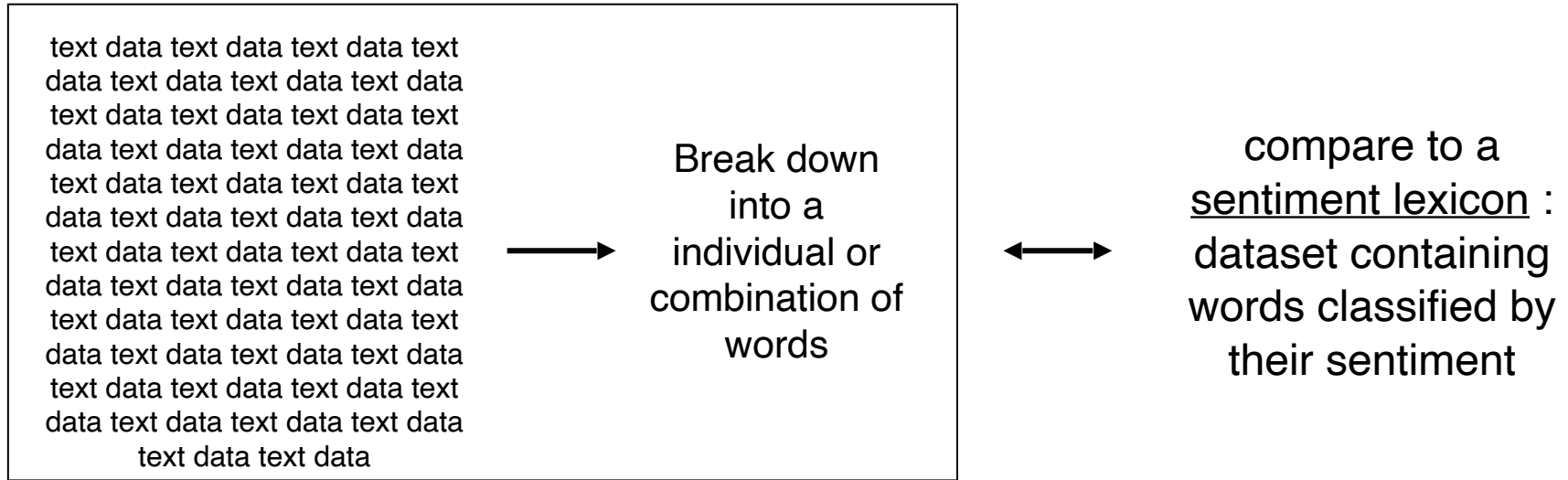
Lexical Density




Sentiment Analysis

Sentiment Analysis

Programmatically infer emotional content of text



Part of the
“NRC”
sentiment
lexicon:



word	sentiment	lexicon
<chr>	<chr>	<chr>
abacus	trust	nrc
abandon	fear	nrc
abandon	negative	nrc
abandon	sadness	nrc
abandoned	anger	nrc
abandoned	fear	nrc
abandoned	negative	nrc
abandoned	sadness	nrc
abandonment	anger	nrc
abandonment	fear	nrc

... with 27,304 more rows

When doing sentiment analysis...

token - a meaningful unit of text

- what you use for analysis
- *tokenization* takes corpus of text and splits it into tokens (words, bigrams, etc.)

stop words - words not helpful for analysis

- extremely common words such as “the”, “of”, “to”
- are typically removed from analysis

When doing sentiment analysis...

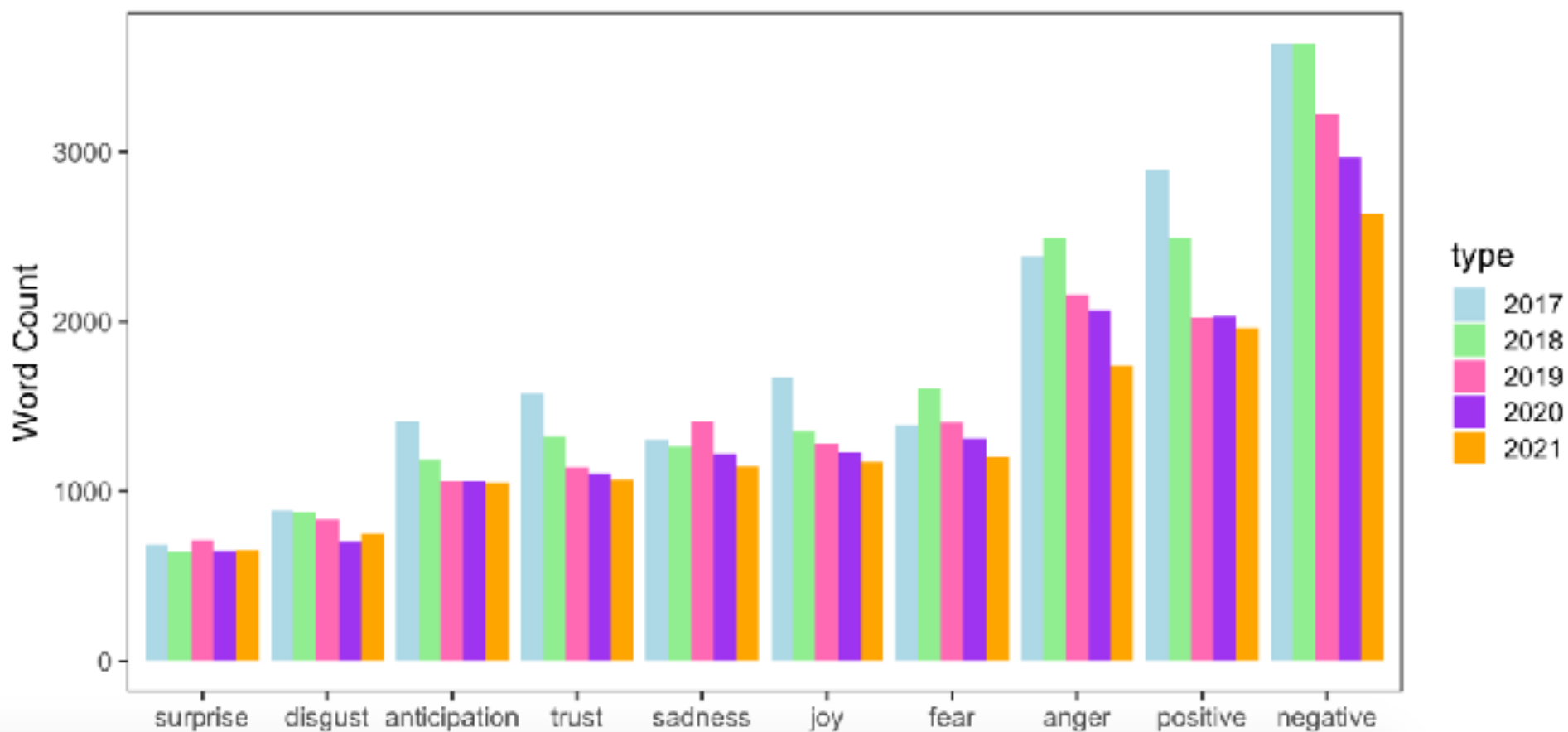
stemming - lexicon normalization

- Identifying the root for each token
- Jumping, jumped, jumps, jump all have the same root 'jump'
- Where things get tricky: jumper???

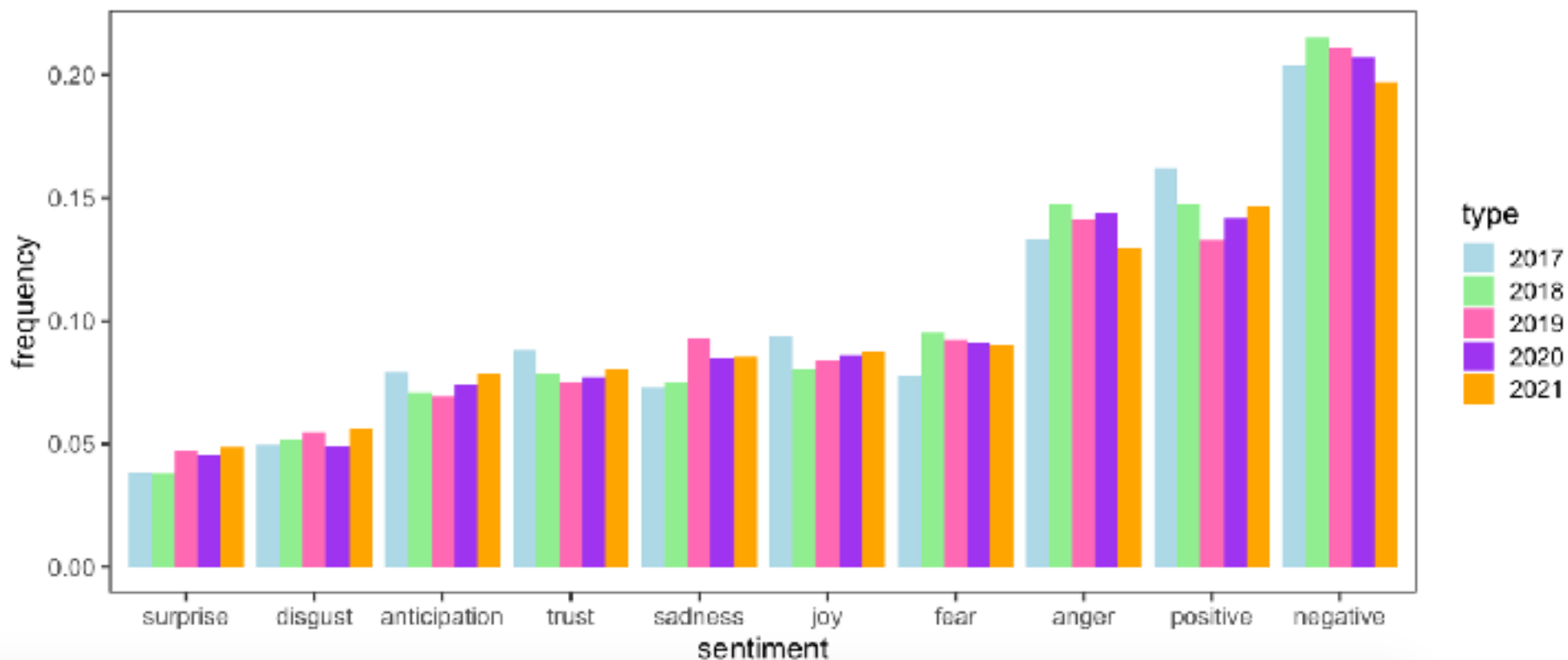
In text analysis, your choices matter:

1. How to tokenize?
2. What lexicon to use?
3. Remove stop words? Remove common words?
4. Use stemming?

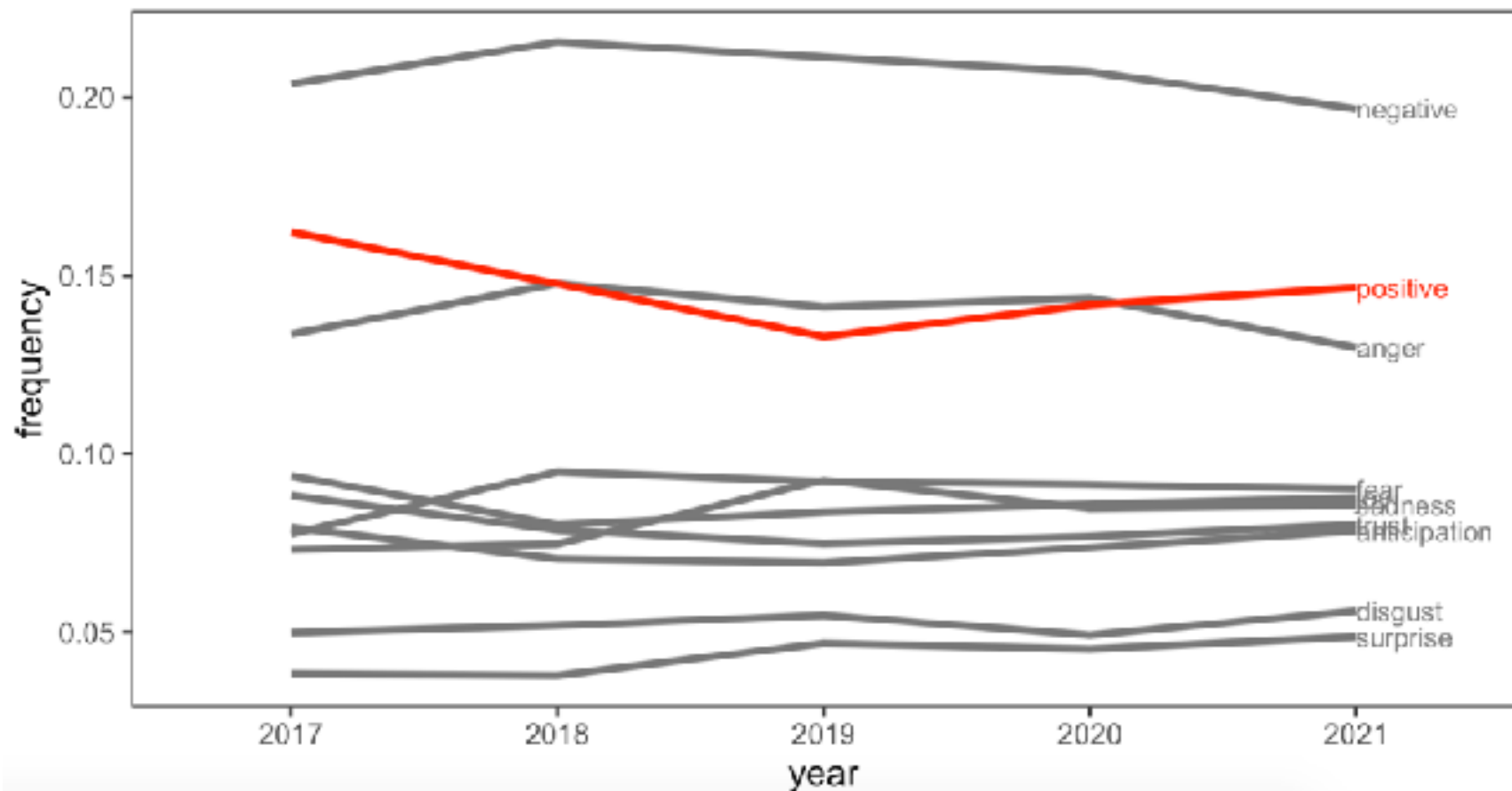
Top Songs Sentiment



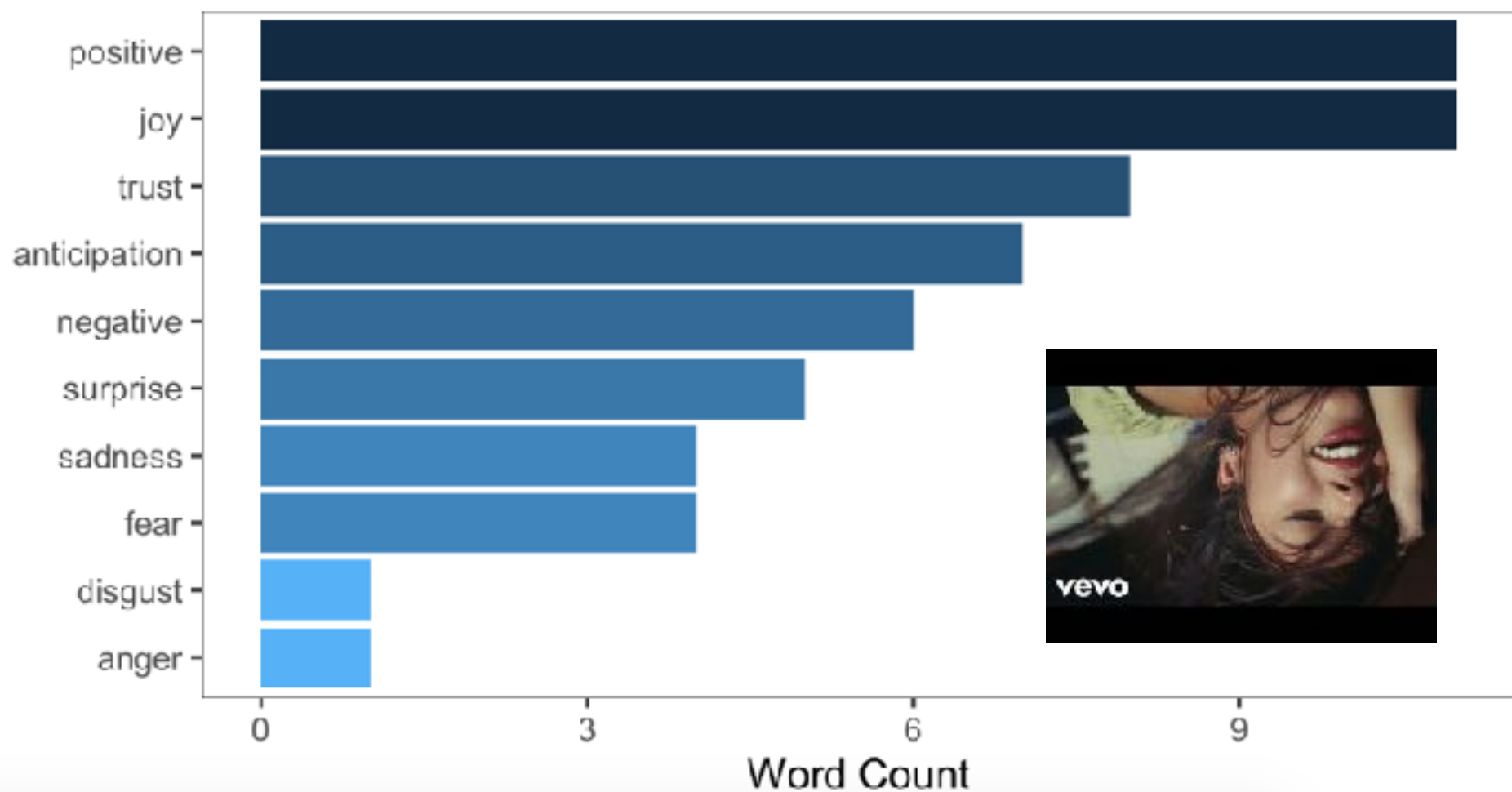
Sentiment by Year



Change in Sentiment over Time

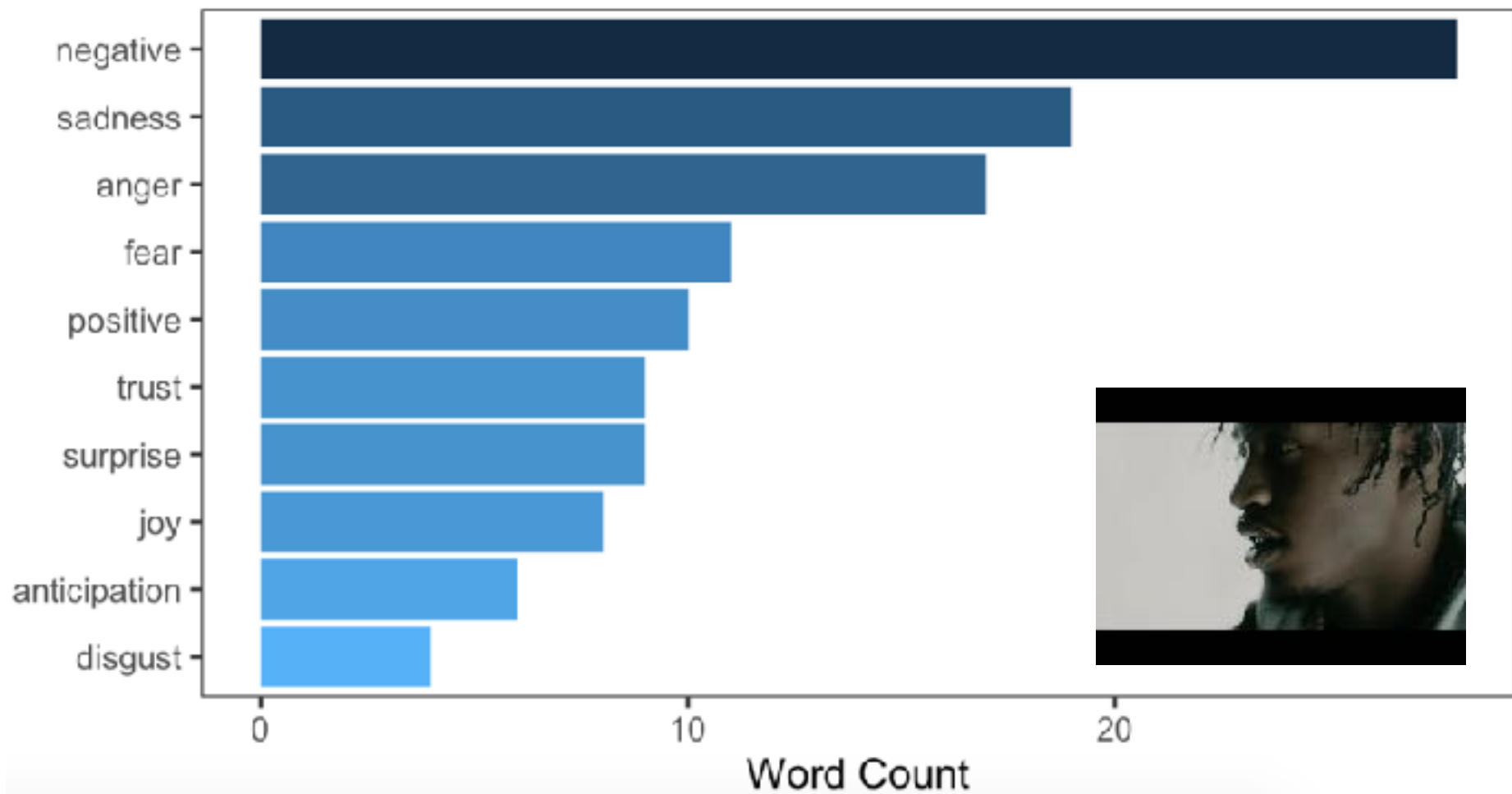


Sentiment: Driver's License



I got my driver's license last week
Just like we always talked about
'Cause you were so excited for me
To finally drive up to your house
But today I drove through the suburbs
Crying 'cause you weren't around
And you're probably with that blonde girl
Who always made me doubt
She's so much older than me
She's everything I'm insecure about
Yeah, today I drove through the suburbs
'Cause how could I ever love someone else?
And I know we weren't perfect but I've never felt this way for no one
And I just can't imagine how you could be so okay now that I'm gone
Guess you didn't mean what you wrote in that song about me
'Cause you said forever, now I drive alone past your street
And all my friends are tired
Of hearing how much I miss you, but
I kinda feel sorry for them
'Cause they'll never know you the way that I do, yeah
Today I drove through the suburbs
And pictured I was driving home to you

Sentiment: Calling My Name



I ain't tryna play these game no more
I don't wanna be textin' your name no more
I ain't tryna feel this pain no more
And I'm sorry but my feelings ain't the same no more (no)
Used to be my homie, you ain't gang no more (no)
I am not a nigga you could claim you no more (no)
Traumatized, hoping it don't rain no more
You done put me through some things that done changed my aura
Now all around the world, I explore, no Dora
New bitch, I might drip out and Dior her
Ass fat, shawty straight heat, no Florida
Bad and she know it, for herself, I applaud her
No needs, yeah, I'm talkin' my boo
So please, leave 'lone, I'm through
And it's all 'cause of what you started
I been told you I won't lose (mmh)
Steady callin' my phone (brrt)
I done told you before that it's over, leave me 'lone
Know it's hurtin' you to see me gone
Dark clouds, you gon' see me storm
I won't go back (go back)
But trust me, you're gon' hold that
Hold that (mmh, mmh)



Sentiment Limitations

How would you classify the sentiment of the following sentence?

“The idea behind the movie was great, but it could have been better”

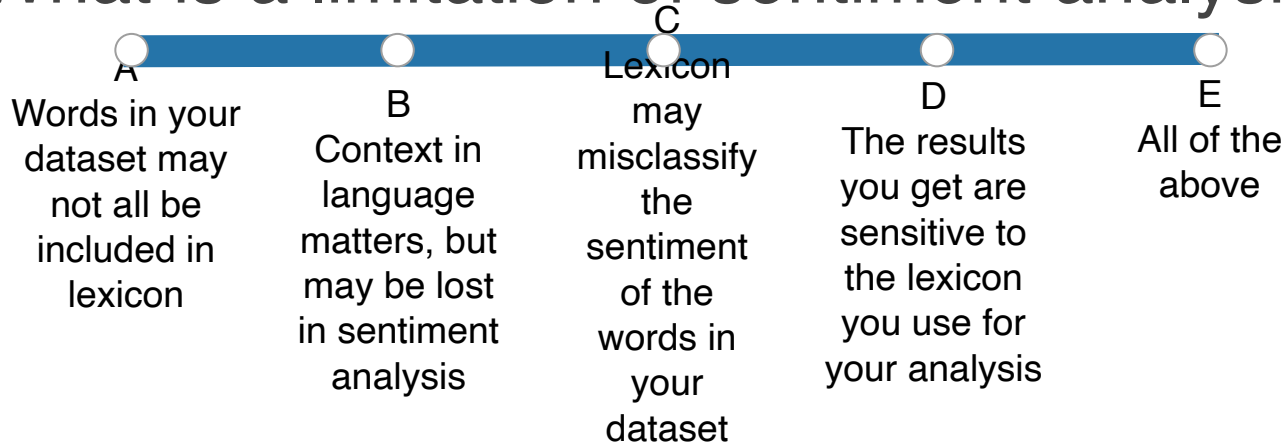
A B C D

positive negative neutral other



Sentiment Limitations

What is a limitation of sentiment analysis?



TF-IDF

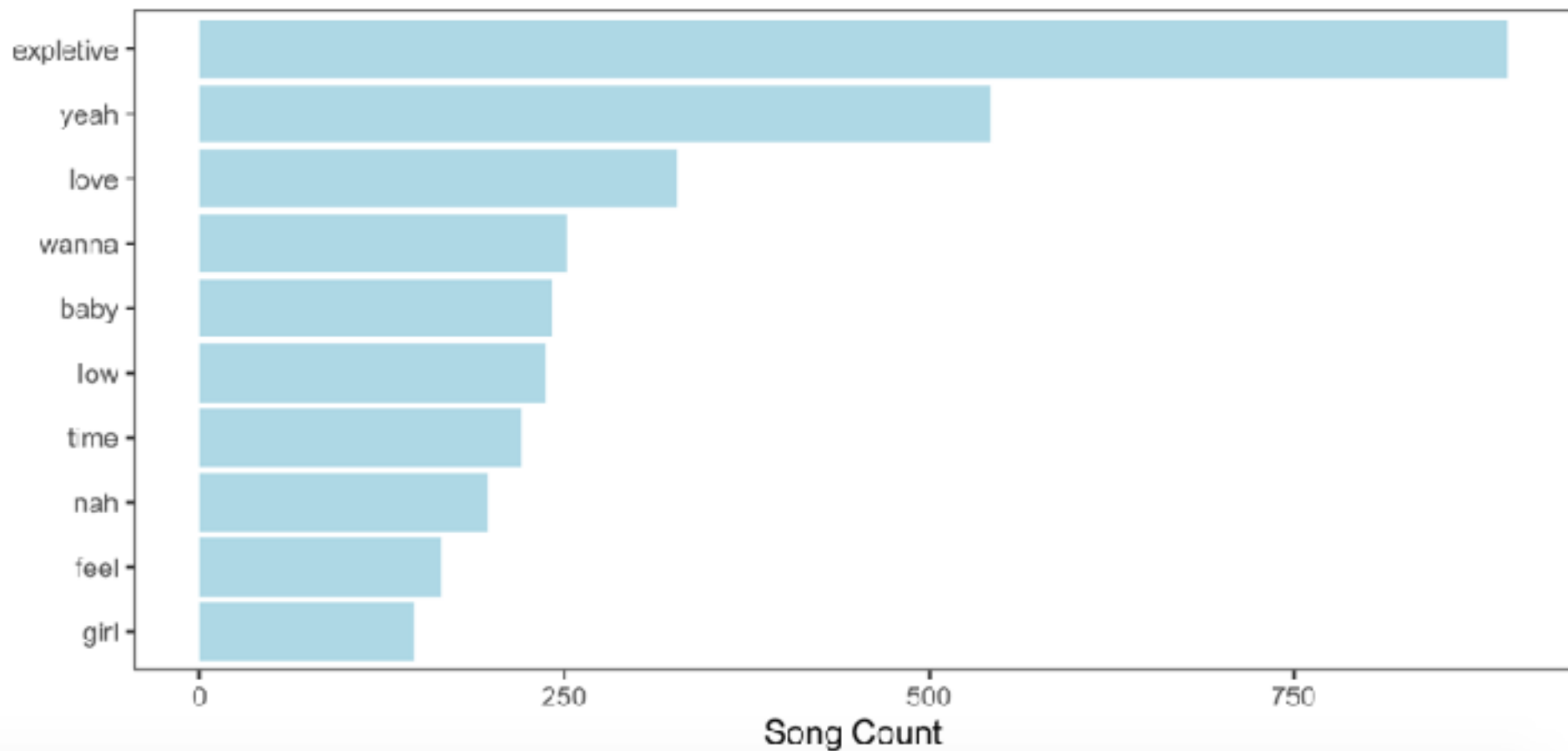
Term Frequency - Inverse Document Frequency

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

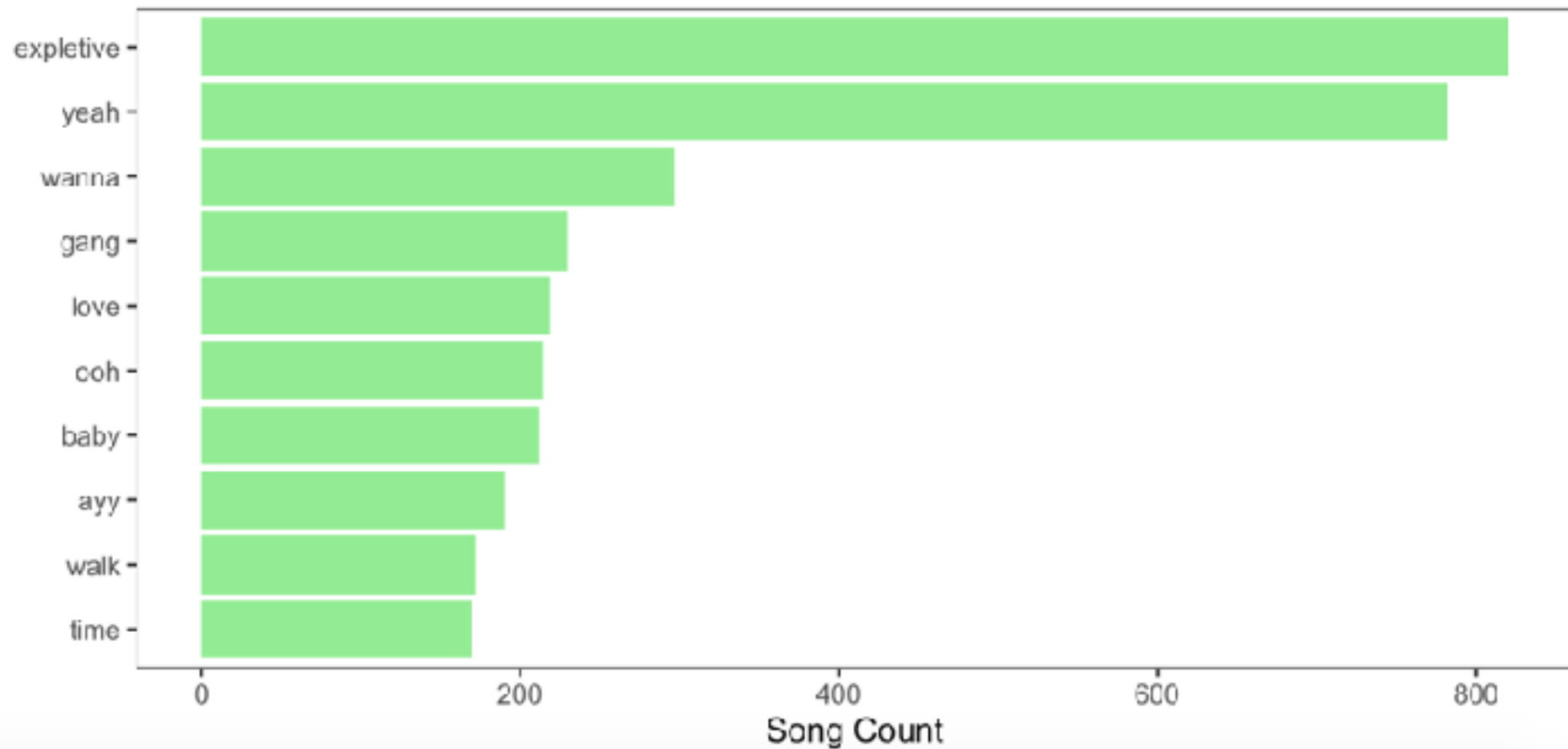
Goal: to use TF-IDF to *find the important words* for the content of each document by decreasing the weight for commonly used words and increasing the weight for words that are not used very much in a collection or corpus of documents

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

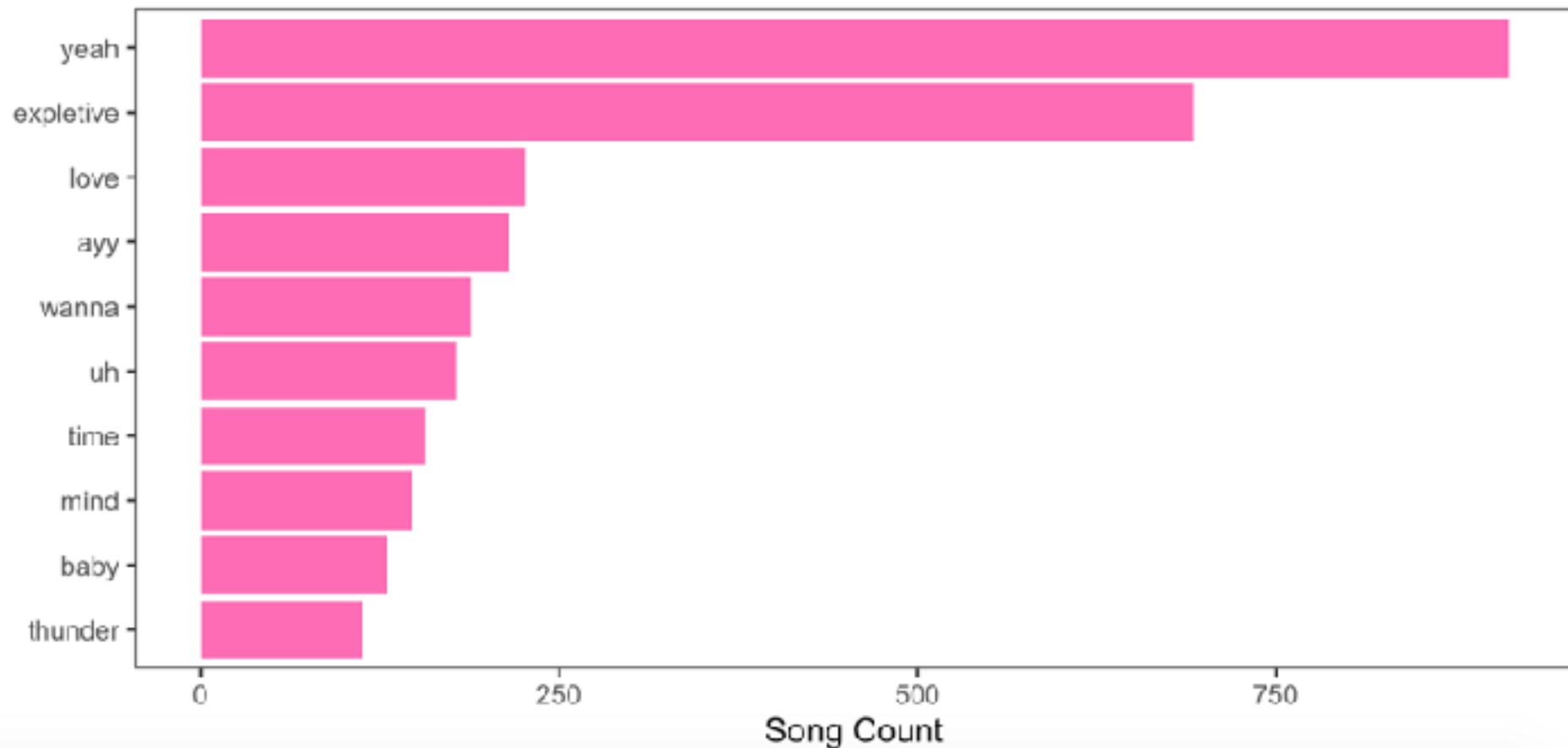
Most Frequently Used Words in top 200 songs (2017)



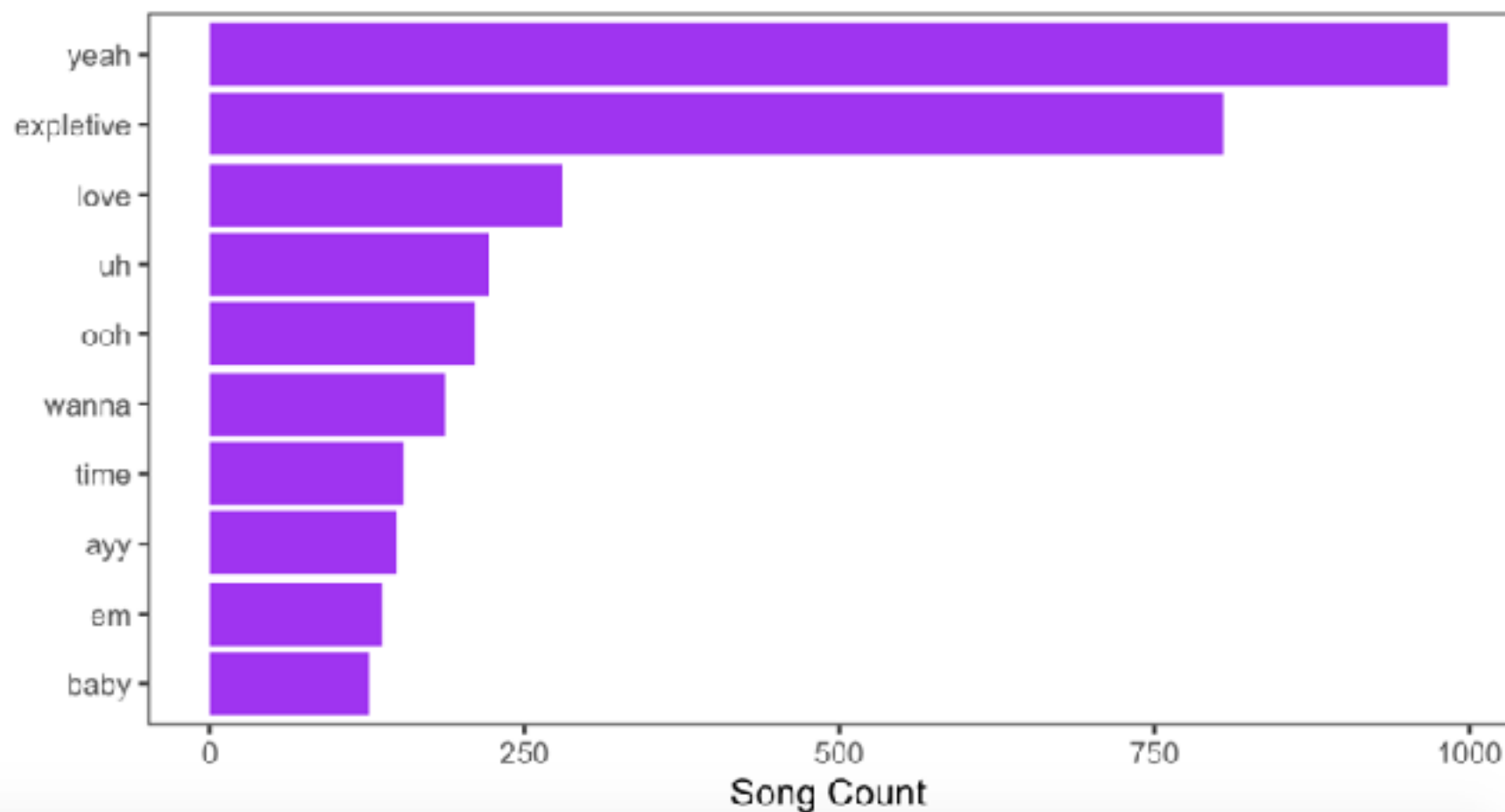
Most Frequently Used Words in top 200 songs (2018)



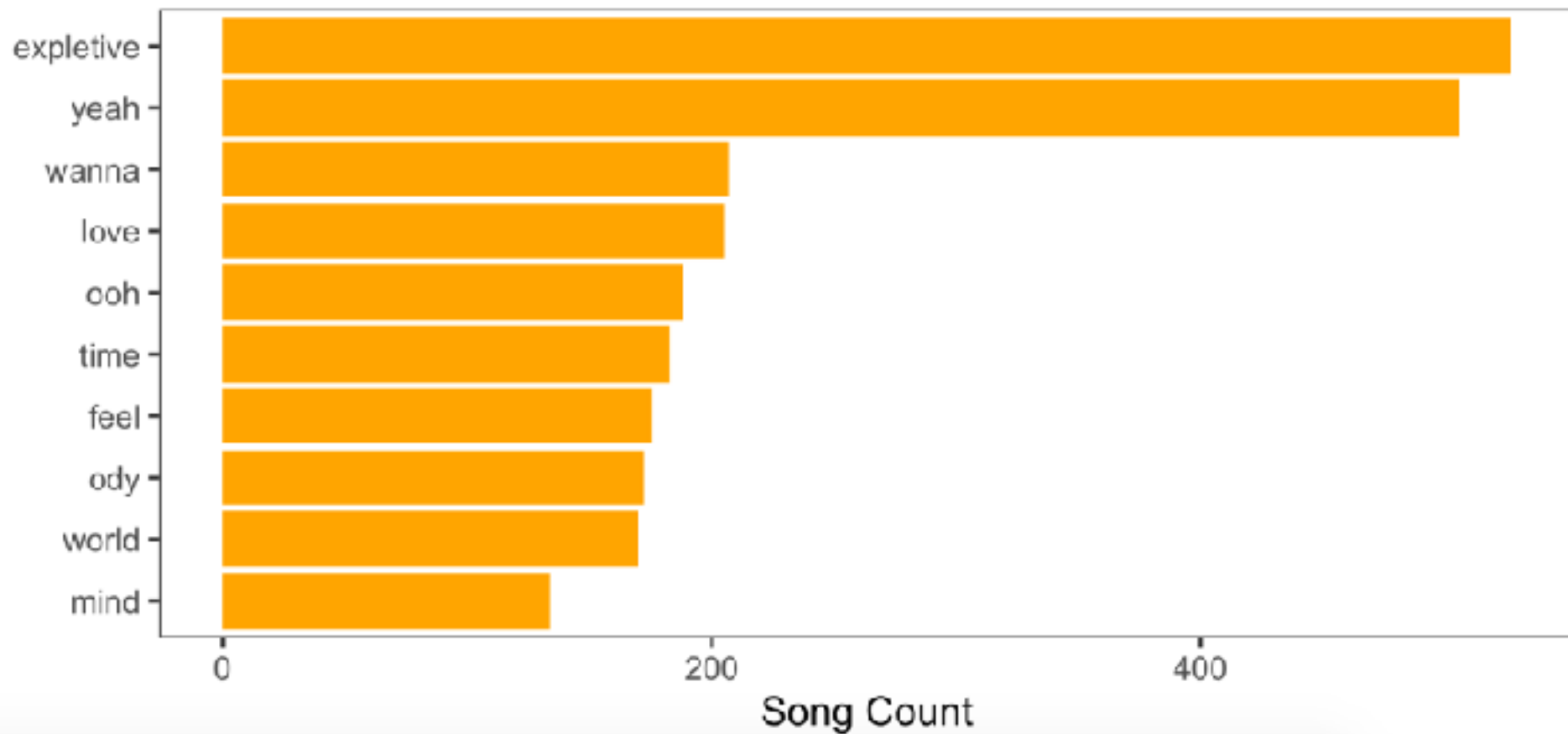
Most Frequently Used Words in top 200 songs (2019)



Most Frequently Used Words in top 200 songs (2020)



Most Frequently Used Words in top 200 songs (2021)





2017



2018



2019



2020

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

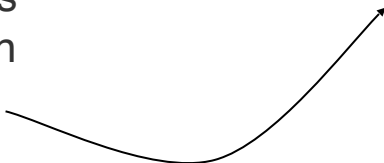


TF-IDF:
Term Frequency - Inverse Document Frequency

Term Frequency (TF) : how frequently a word occurs in a document

Inverse document frequency (IDF) : intended to measure how important a word is to a document

decreases the weight for
commonly used words and
increases the weight for words
that are not used very much in
a collection of documents

$$idf(\text{term}) = \ln \left(\frac{n_{\text{documents}}}{n_{\text{documents containing term}}} \right)$$


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

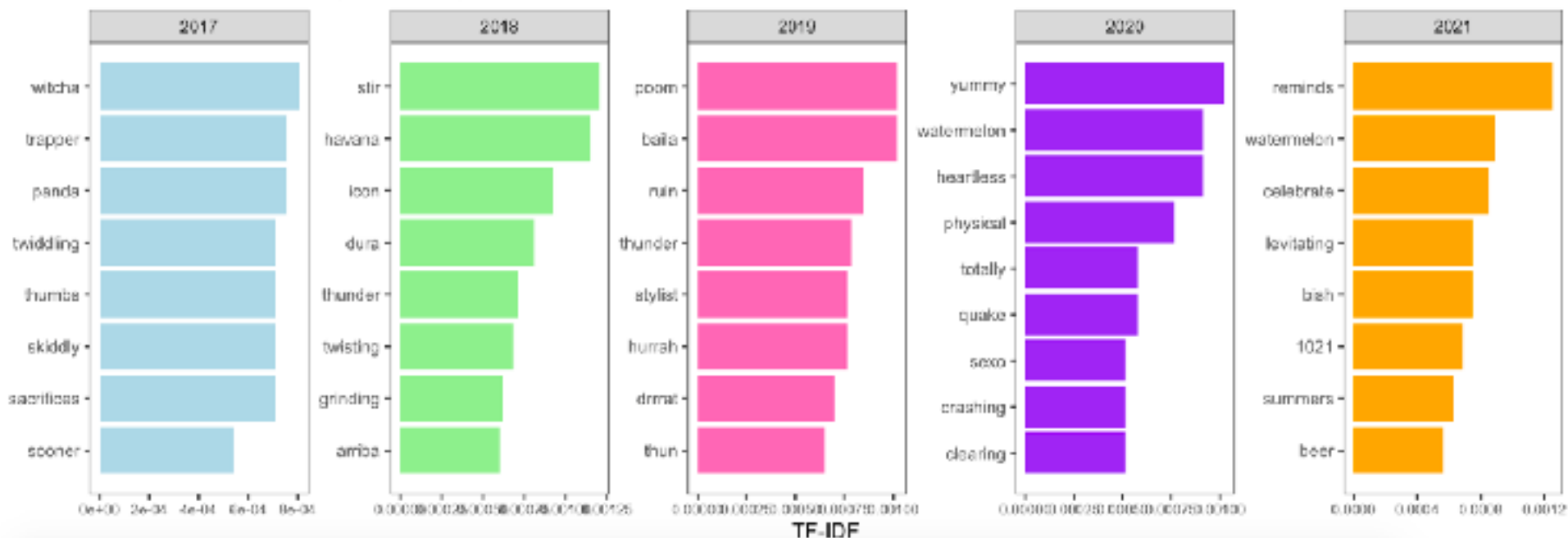
Term x within document y

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

df_x = number of documents containing x

N = total number of documents

Important Words using TF-IDF by Year



Father and Son
Flight of the Conchords



Live In London
2019

Just you and I
Driving round in the car
I even let you drive
Eating dinner from a can
dad's **yummy** can delight

I'm Me
Choi



Punk
2019

ビーフ アンド チャオズ フィッシュ オレンジ
Everything **yummy** foods!

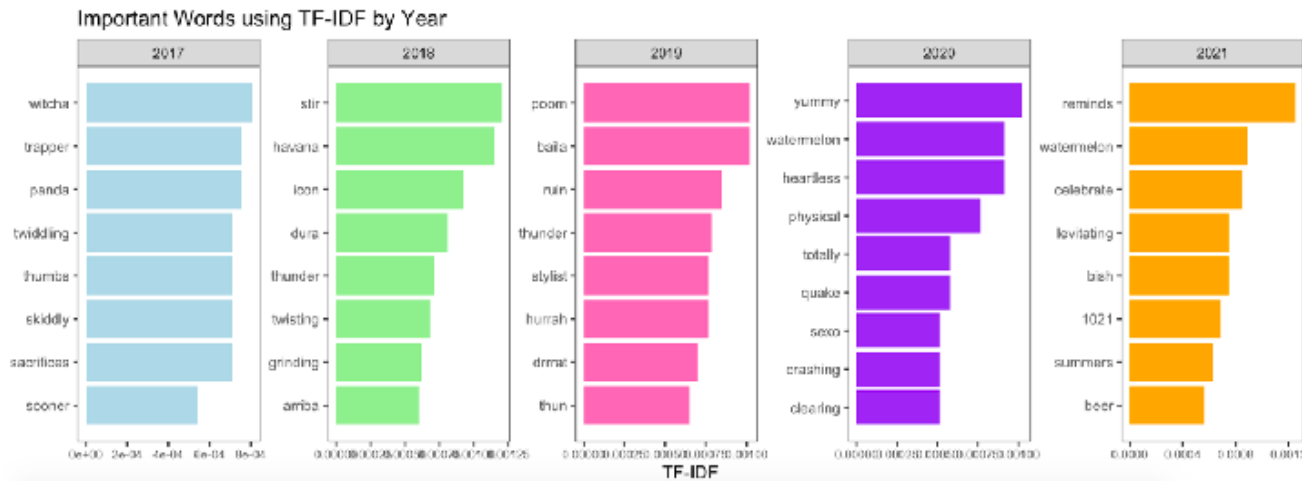
Sex Talk
Megan Thee Stallion



Sex Talk
2019

He eat from the back to the front
Bitch I got green like I'm Buttercup (ah)
He say I got that lil **yummy** yum (ah ah)

What can you conclude from this TF-IDF plot?



A No words overlap across the years in these data

B 'reminds' and 'watermelon' are the most unique words to the 2021 data

C 'watermelon' is the most common word in this dataset

D A-C (all of the above)

E None of the above

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?

EDA

TF-IDF

Sentiment
Analysis