Study Abroad
UC San Diego

WHY STUDY ABROAD?

- Across all academic disciplines, students who study abroad graduate on-time at a higher rate than students who do not study abroad
- 38.5% of UCSD Study Abroad alumni say their experience abroad led to their first job
- 99% of UCSD Study Abroad alumni say they learned valuable job skills while abroad
- Improve your language skills (programs offered in English too!)
- Intern or do research abroad
- Meet your degree requirements faster by doing a summer program
- Learn about another culture by spending a semester or year abroad

studyabroad.ucsd.edu

First Steps

Mondays 2:00pm-3:00pm*
Tuesdays 1:00pm-2:00pm*
Wednesdays 11:00am-12:00pm*
Thursdays 3:00pm-4:00pm*
Fridays 12:00pm-1:00pm*
Link: ucsd.zoom.us/my/studyabroaducsd

Virtual Advising

Scheduled Appointment ~30 minutes
Schedule through Virtual Front Desk
Link: ucsd.zoom.us/my/studyabroaducsd
Monday & Thursday 9:00am-7:00pm*
Tuesday, Wednesday & Friday 9:00am-4:00pm*
Drop In Advising ~15 minutes
Mondays 1:00pm-2:00pm*
Thursdays 2:30pm-3:30pm*
Link: ucsd.zoom.us/my/studyabroaducsd

*Hours may change

Course Reminders

Due Friday (11:59 PM):

- D5
- Q5
- Project Checkpoint #1: Data
- Mid-course Survey (*optional*)
- Weekly Project Survey (*optional*)

Note: Project Proposal Feedback Now Available

- Scores and Feedback on GitHub repo (as an issue)
- Score should match score on Canvas

Text Analysis

Shannon E. Ellis, Ph.D
UC San Diego

Department of Cognitive Science
sellis@ucsd.edu

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. Do the angriest tweets come from Trump himself?
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

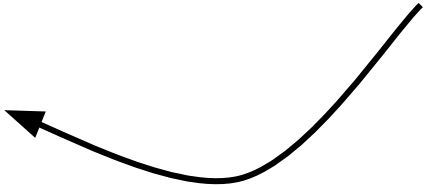
2018: 139 songs

2019: 127 songs

2020: 137 songs

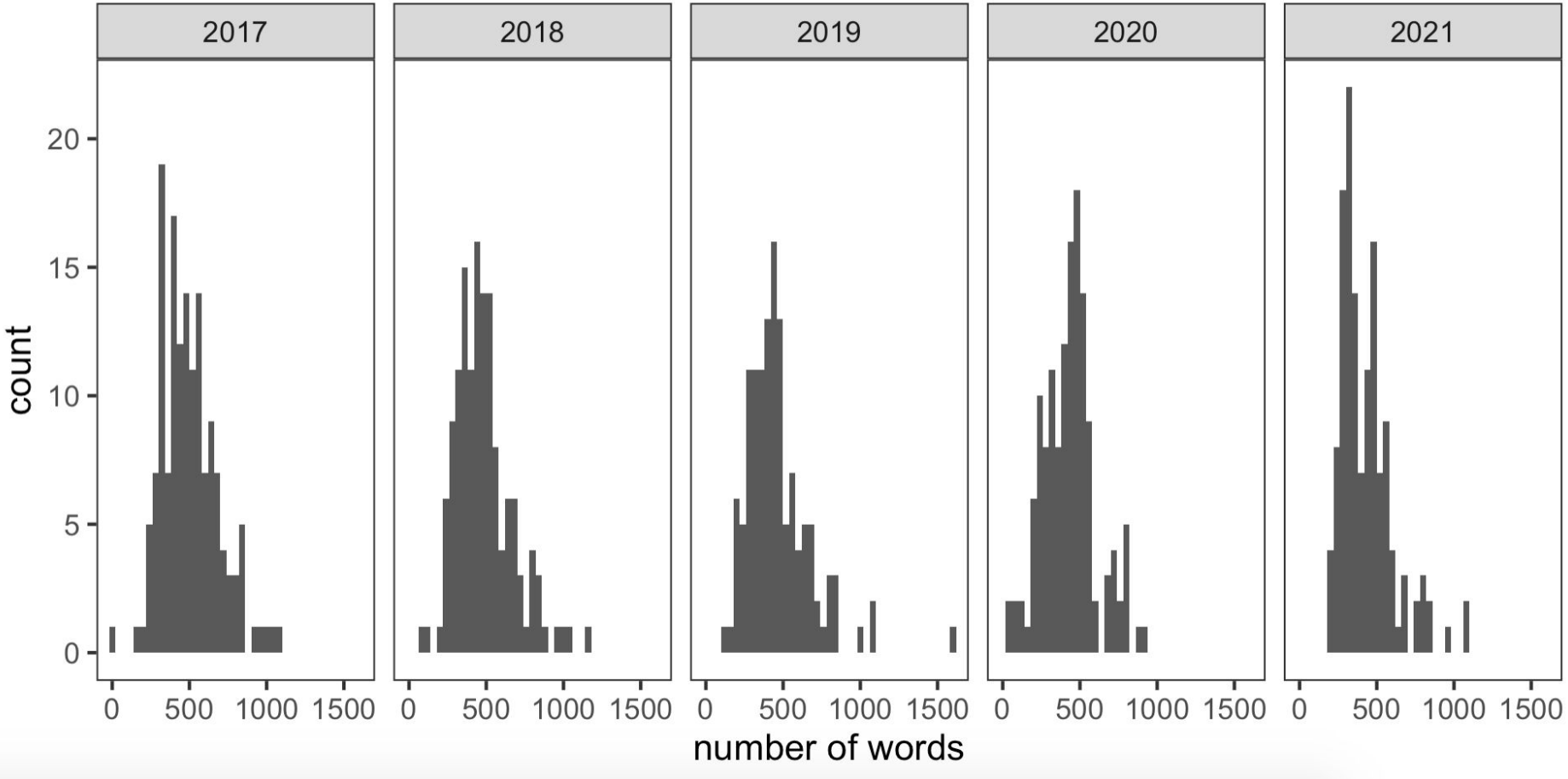
2021: 134 songs

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

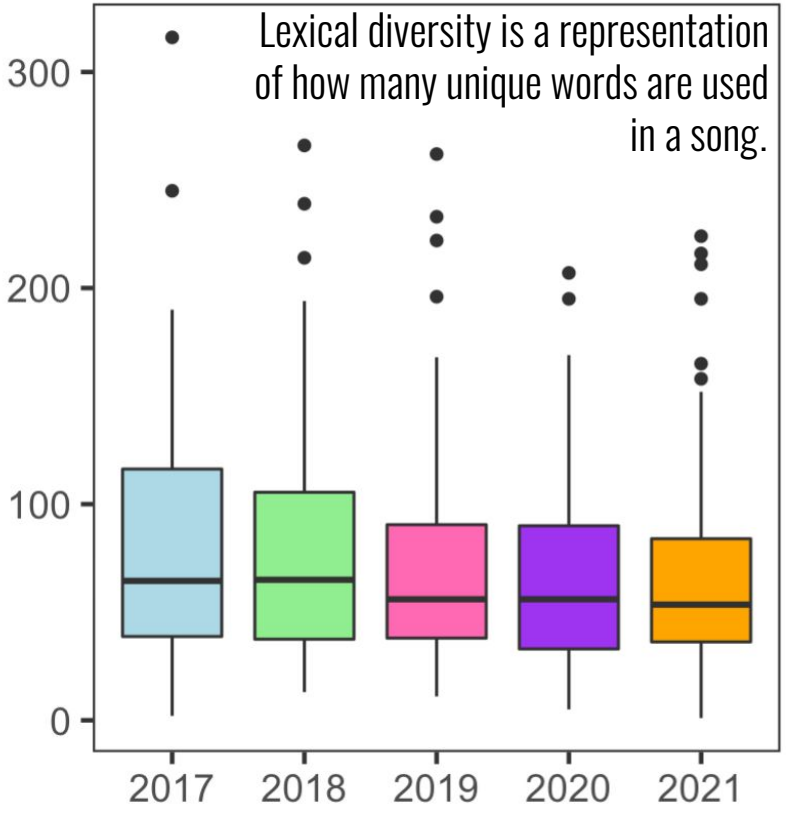


Questions we can ask...

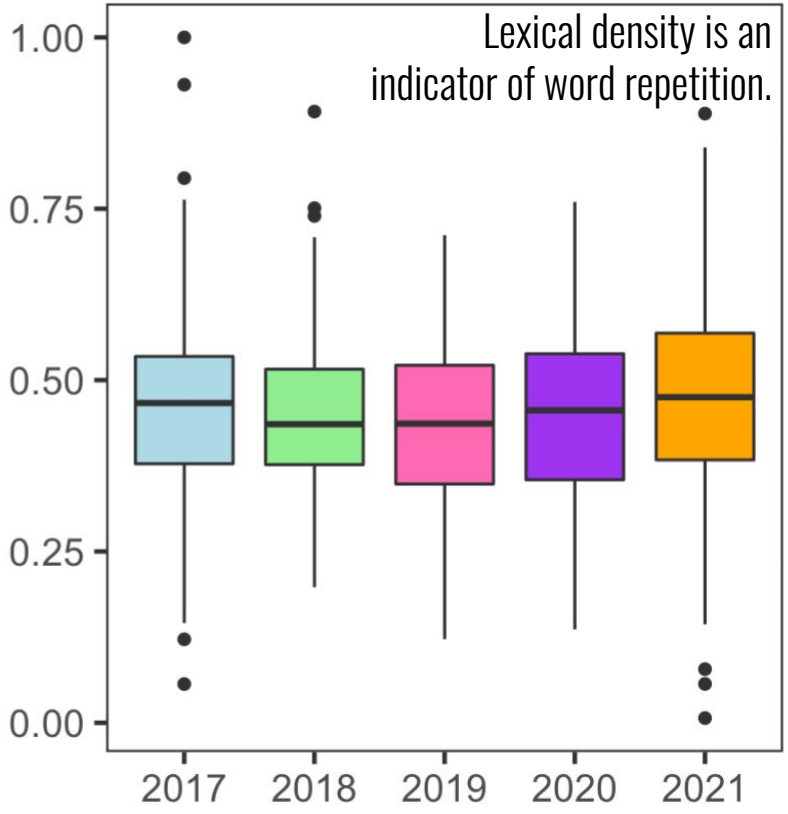
1. Does the total number of words change over time?
2. Does uniqueness change over time?
3. Does the diversity or density change?
4. What words are most common?
5. What words are most unique to each year?
6. What sentiment do songs convey most frequently?
7. Has sentiment changed over time?
8. What are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


2017
2018
2019
2020
2021
count
0
5
10
15
20
0
500
1000
1500
number of words

Lexical Diversity



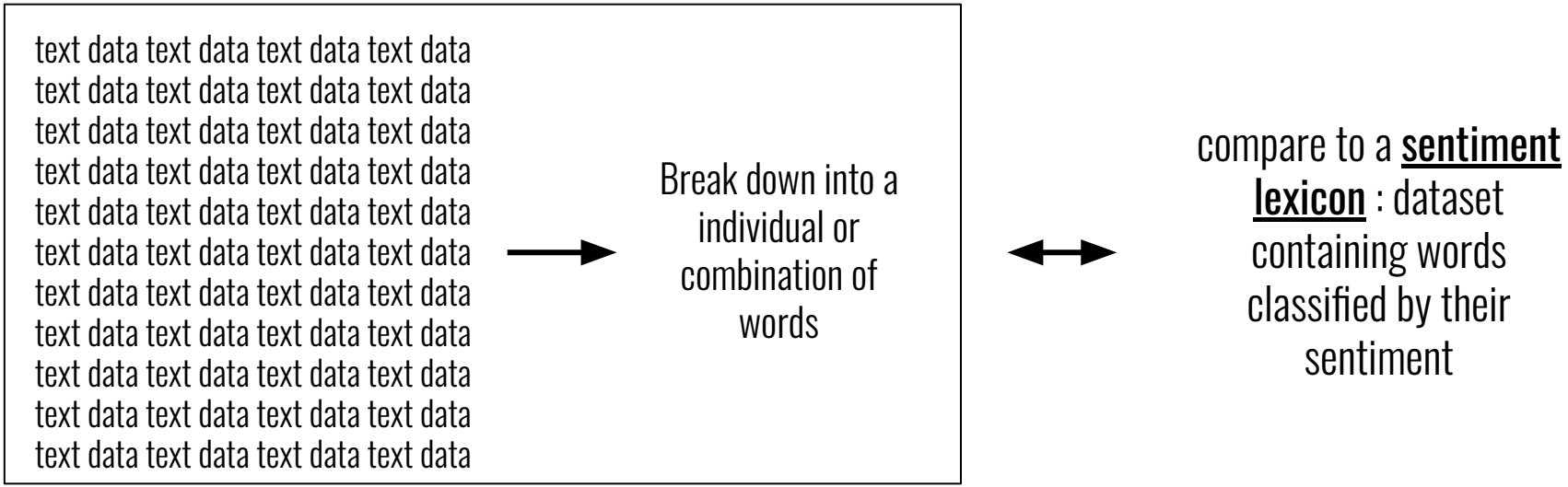
Lexical Density



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

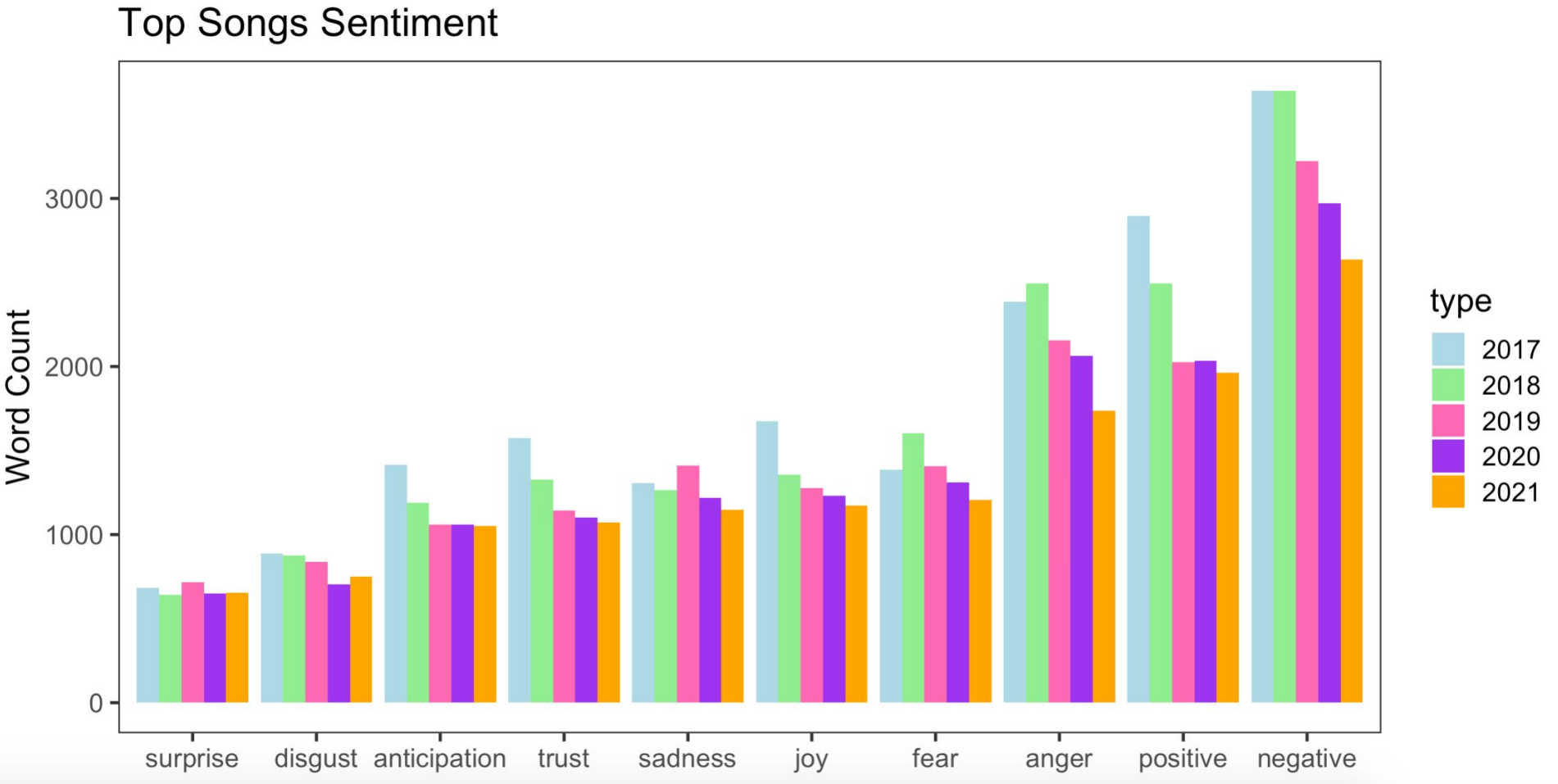
When doing sentiment analysis...

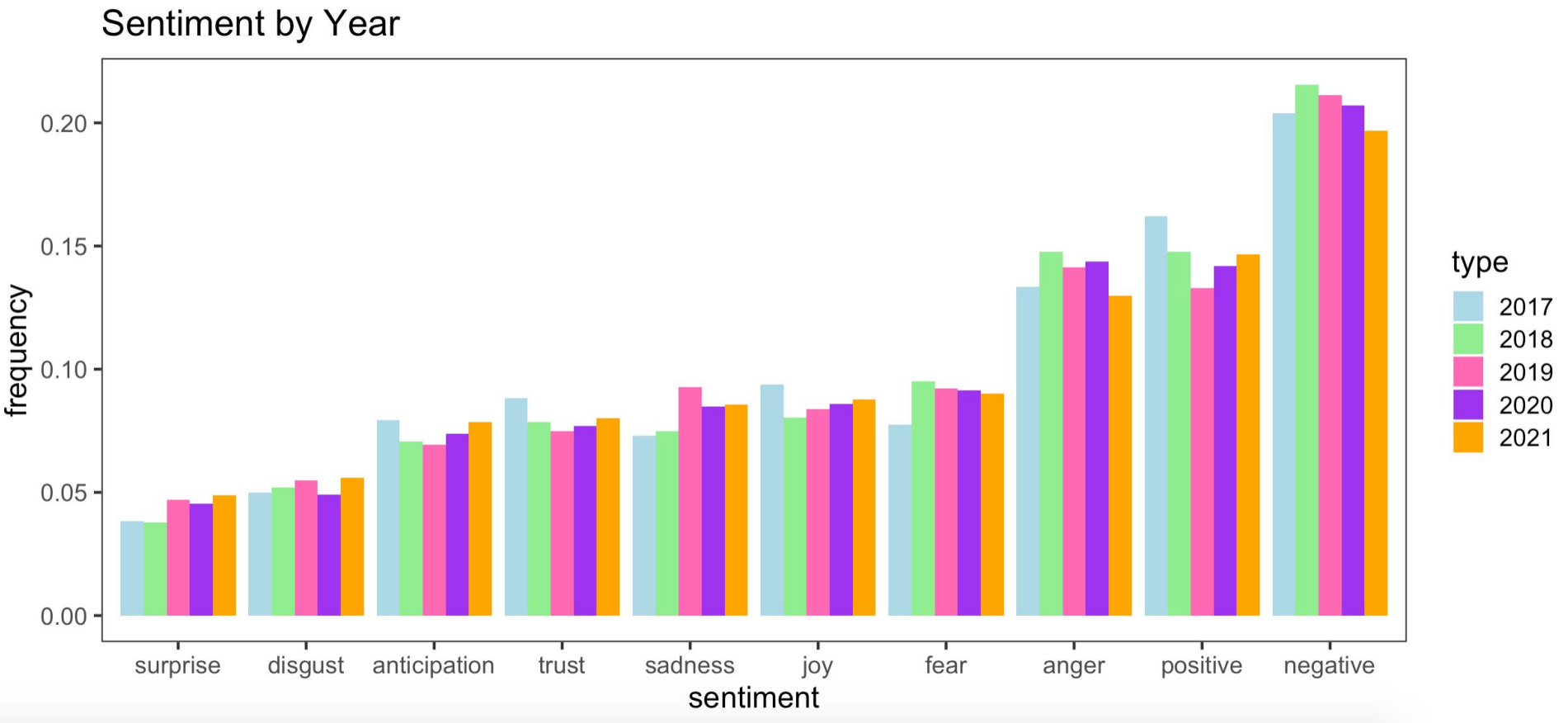
stemming - lexicon normalization

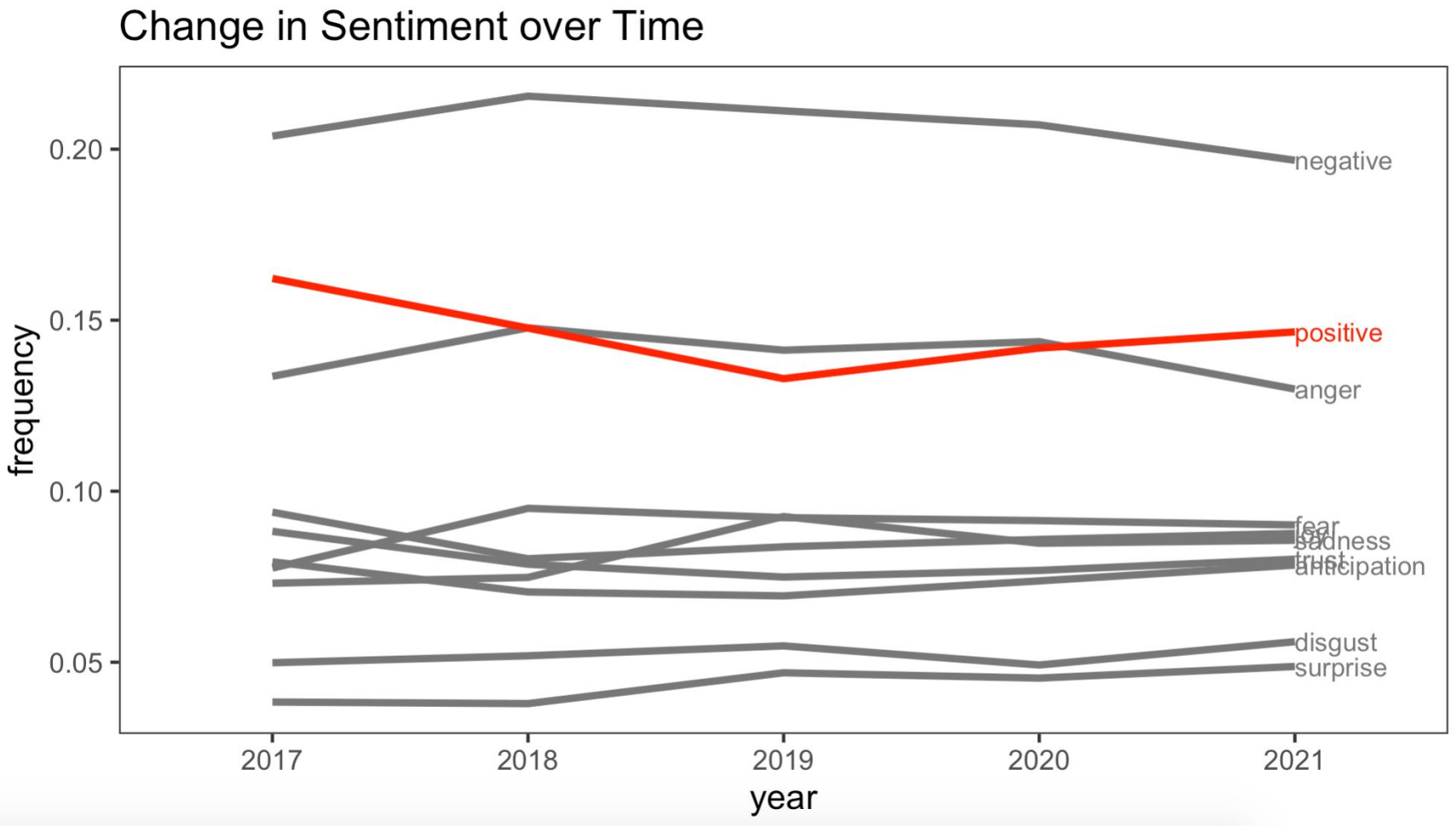
- Identifying the root for each token
- Jumping, jumped, jumps, jump all have the same root ‘jump’
- Where things get tricky: jumper???

In text analysis, your choices matter:

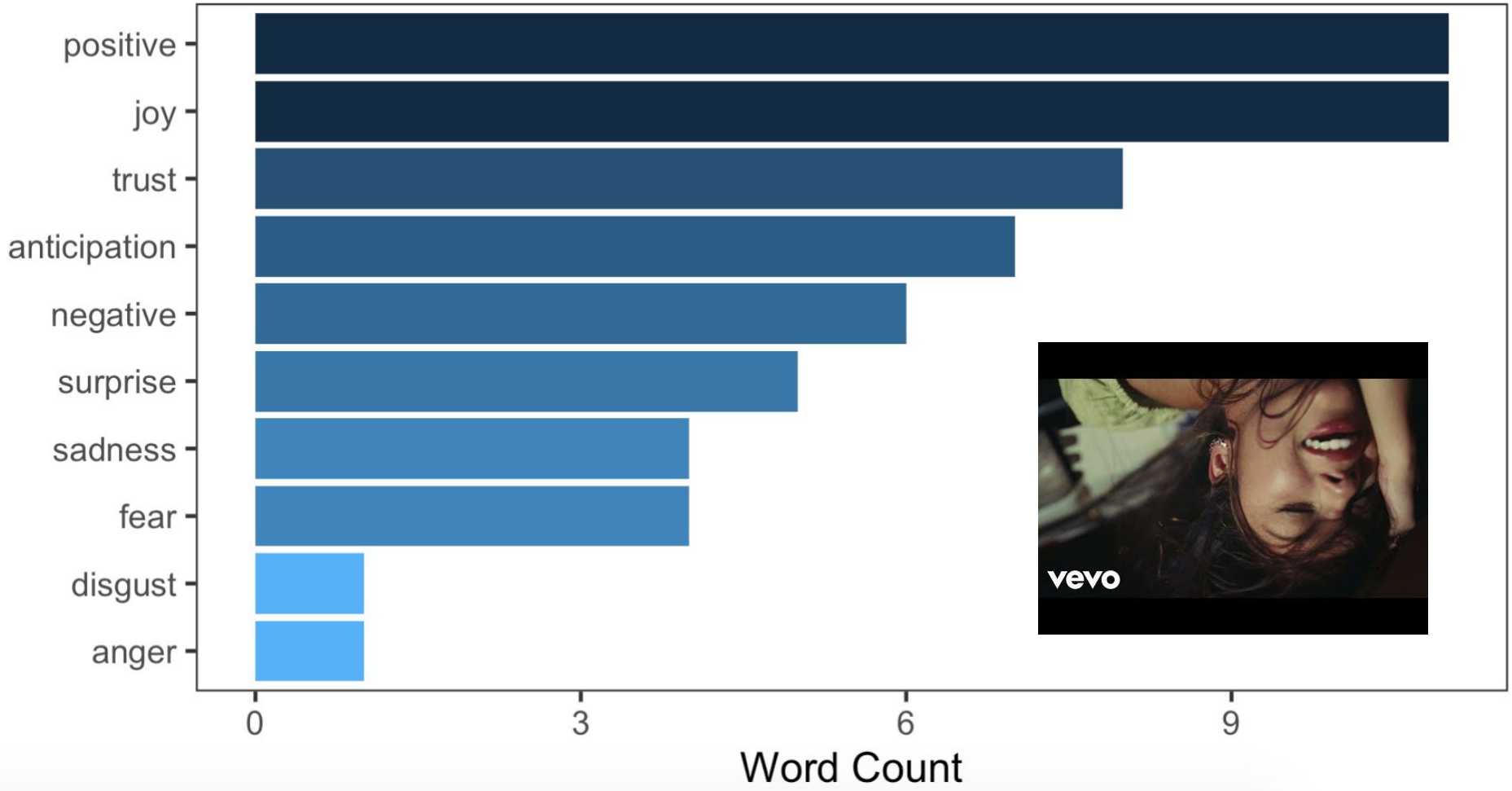
1. How to tokenize?
2. What lexicon to use?
3. Remove stop words? Remove common words?
4. Use stemming?


Top Songs Sentiment
Word Count
3000
2000
1000
0
surprise
disgust
anticipation
trust
sadness
joy
fear
anger
positive
negative
type
2017
2018
2019
2020
2021

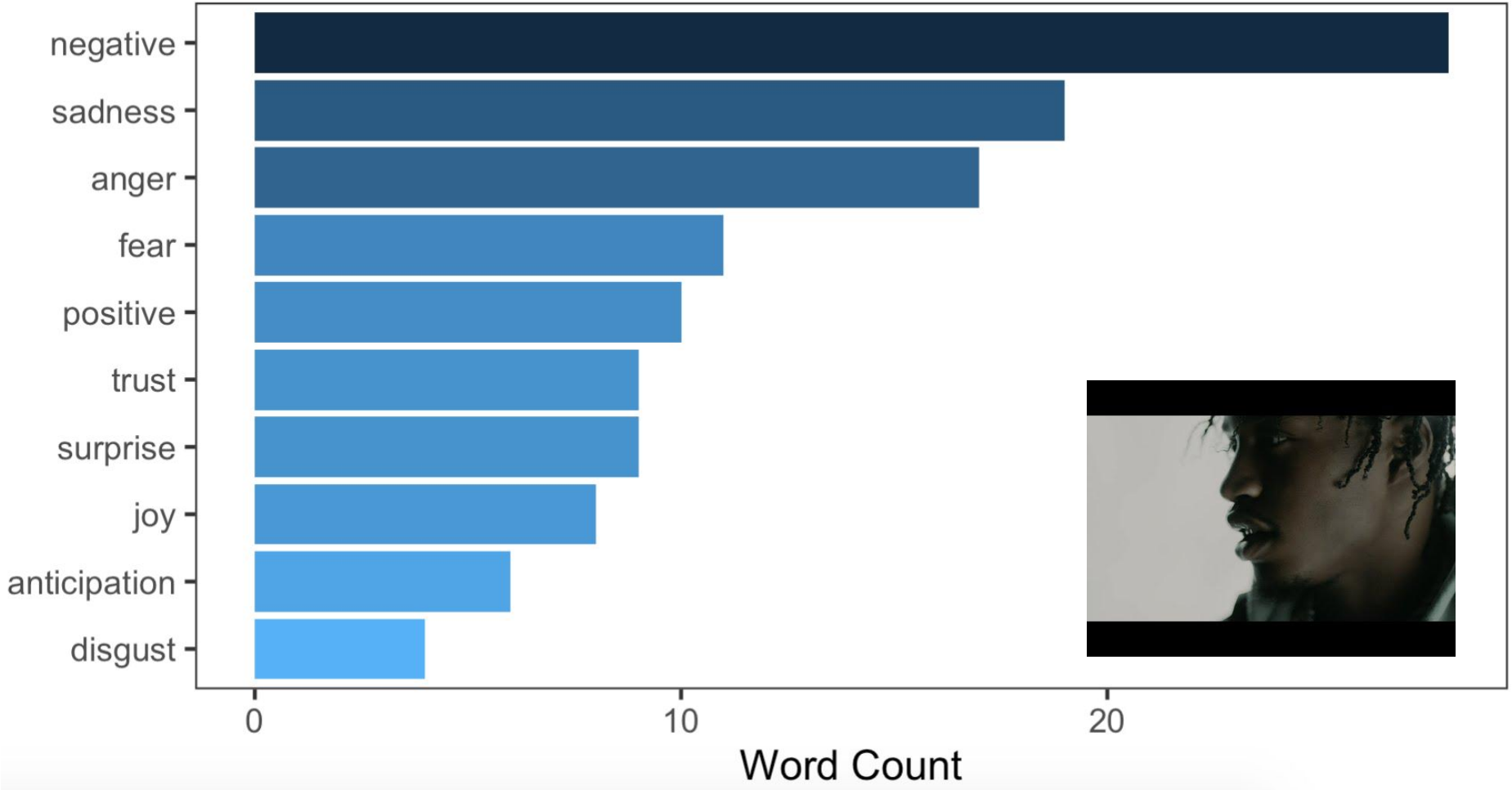

Sentiment by Year
frequency
0.20
0.15
0.10
0.05
0.00
surprise
disgust
anticipation
trust
sadness
joy
fear
anger
positive
negative
sentiment
type
2017
2018
2019
2020
2021


Change in Sentiment over Time
0.20
0.15
0.10
0.05
frequency
negative
positive
anger
fear
joy
sadness
trust
anticipation
disgust
surprise
2017
2018
2019
2020
2021
year

Sentiment: Driver's License



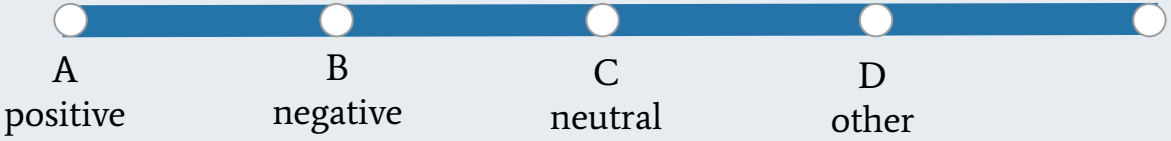
Sentiment: Calling My Name



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"



Sentiment Limitations

What is a limitation of sentiment analysis?

A
Words in your dataset may not all be included in lexicon

B
Context in language matters, but may be lost in sentiment analysis

C
Lexicon may misclassify the sentiment of the words in your dataset

D
The results you get are sensitive to the lexicon you use for your analysis

E
All of the above

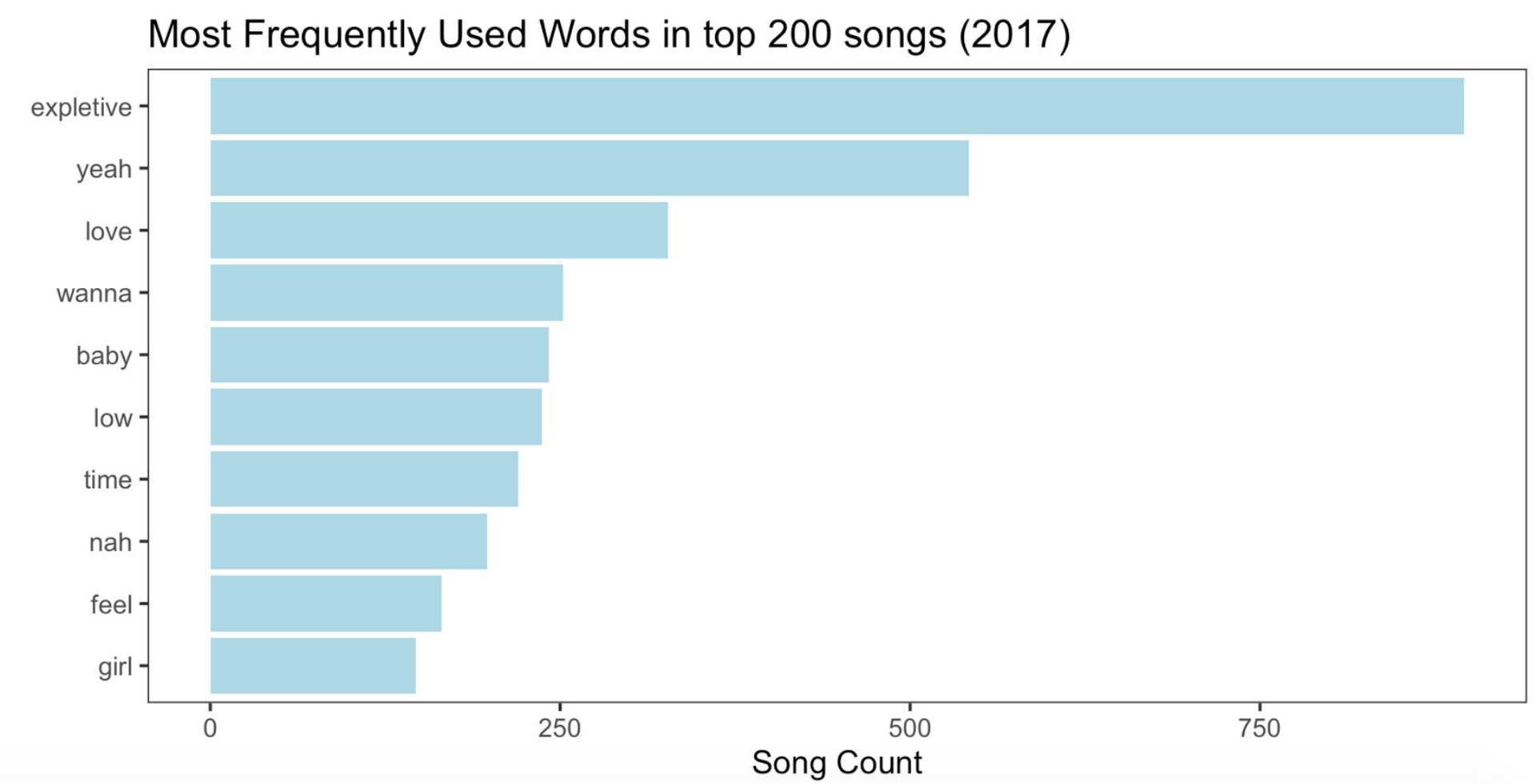
TF-IDF

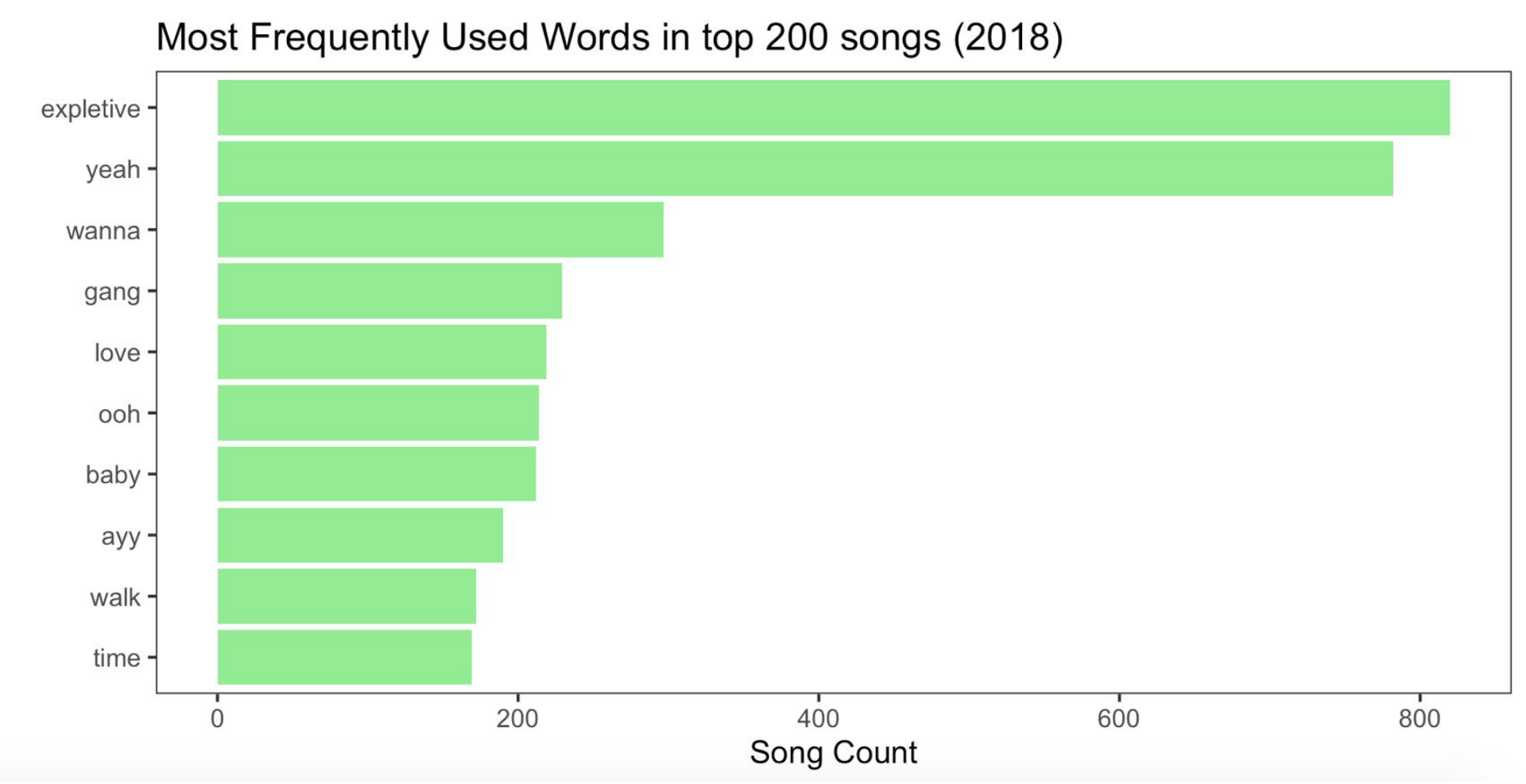
Term Frequency - Inverse Document Frequency

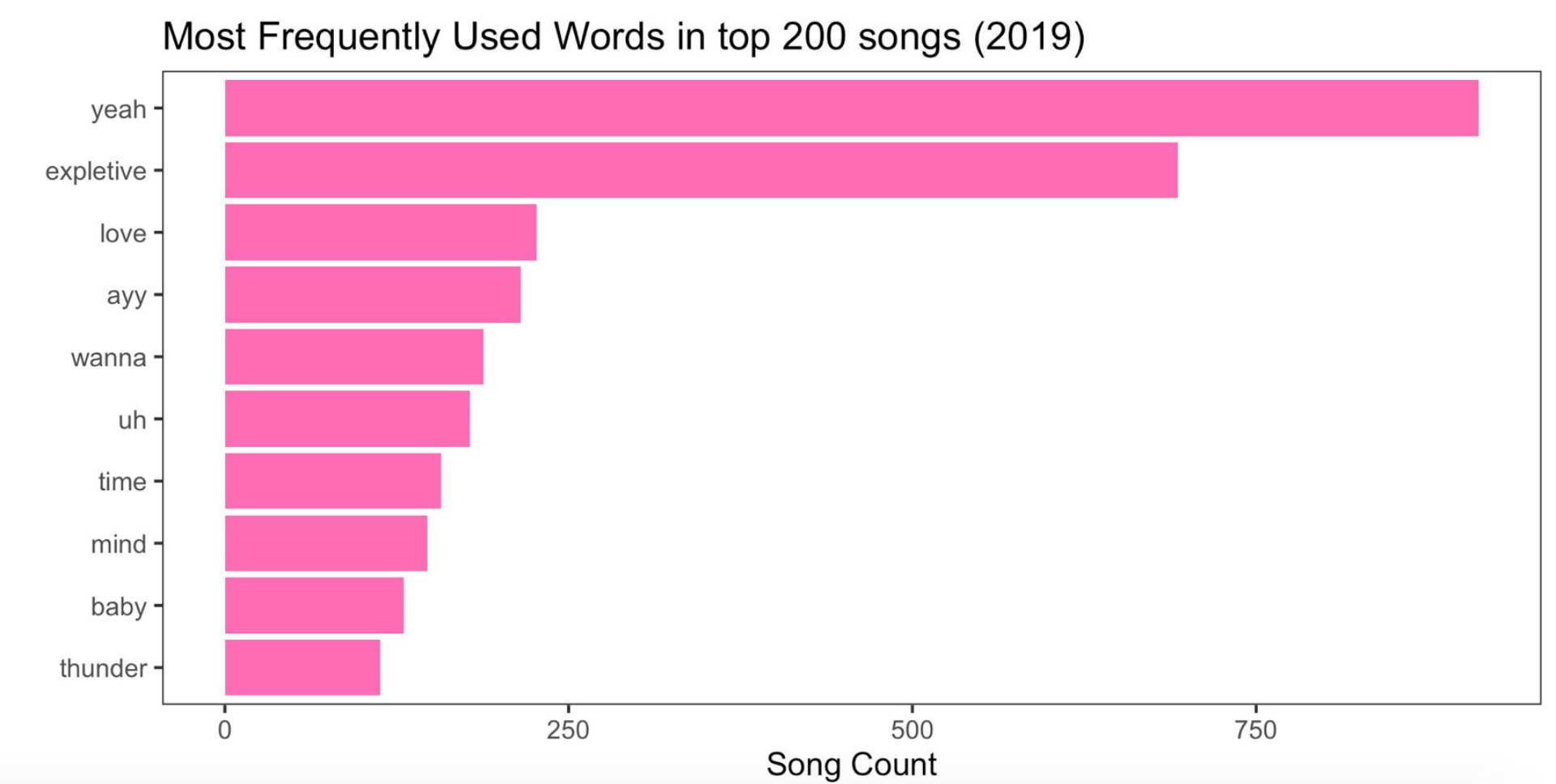
What words are the most unique to the lyrics of each year's top hits?

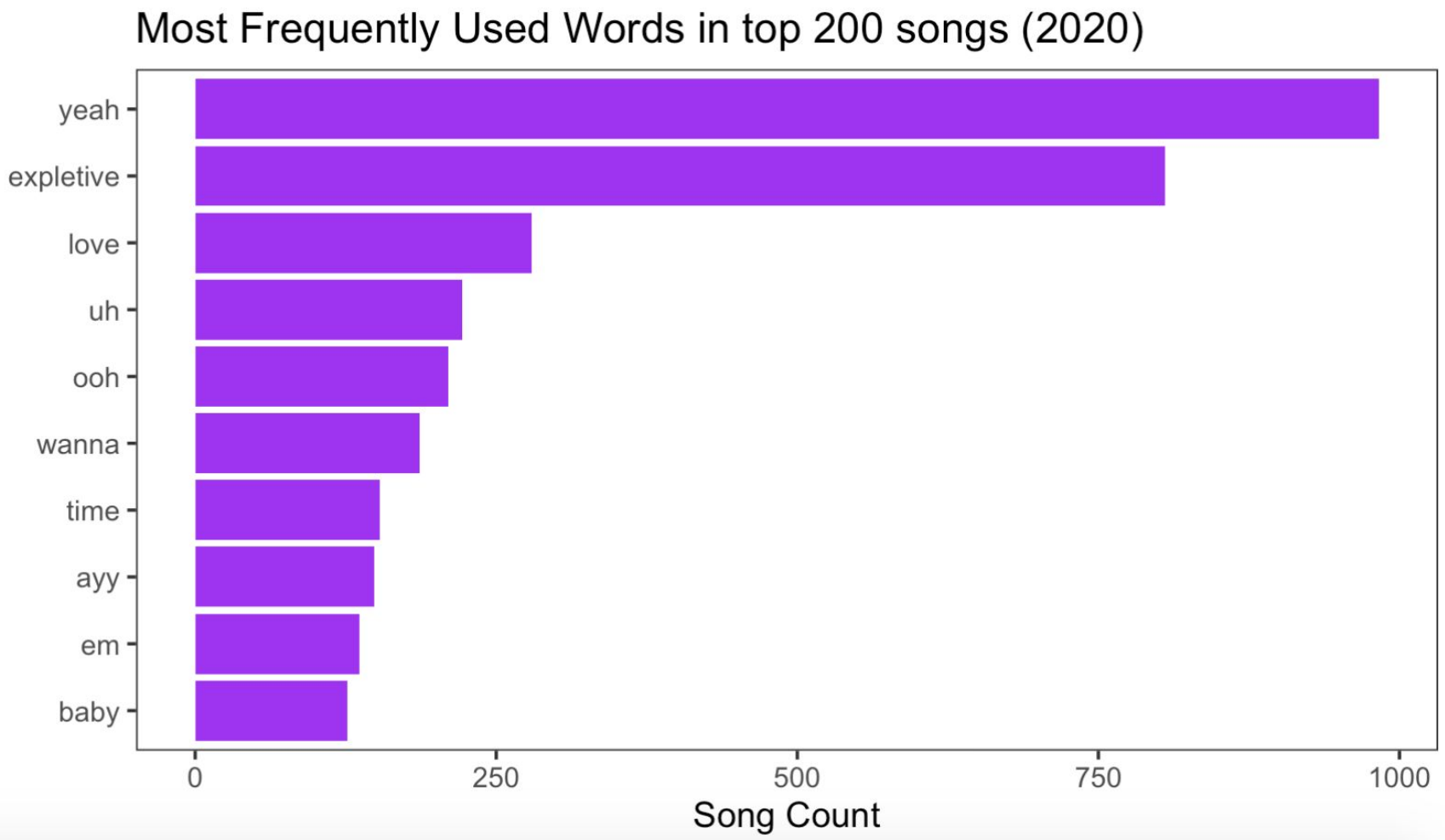
Goal: to use TF-IDF to *find the important words* for the content of each document by decreasing the weight for commonly used words and increasing the weight for words that are not used very much in a collection or corpus of documents

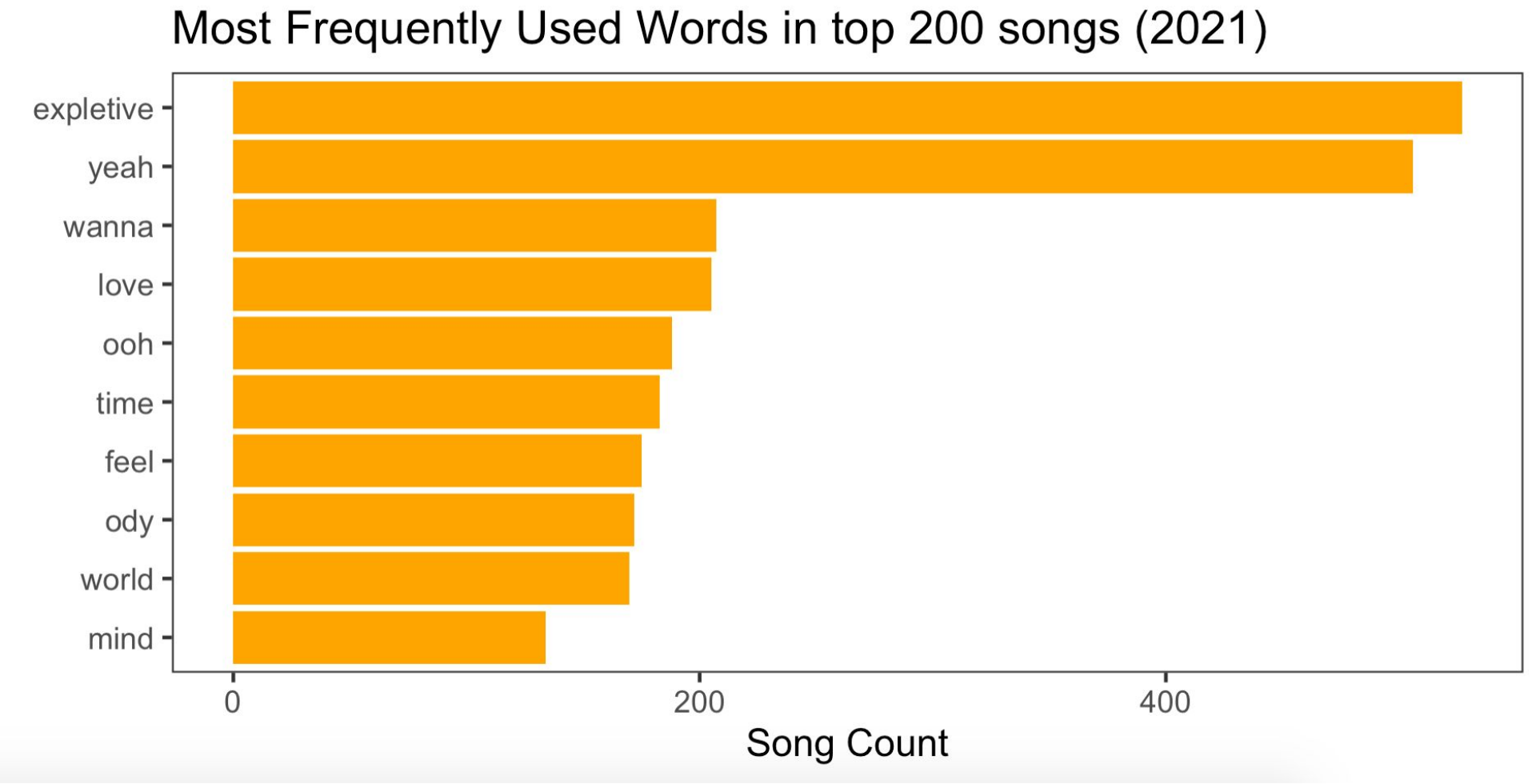
Calculating TF-IDF attempts to find the words that are important (i.e., common) in a text, but not *too* common


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

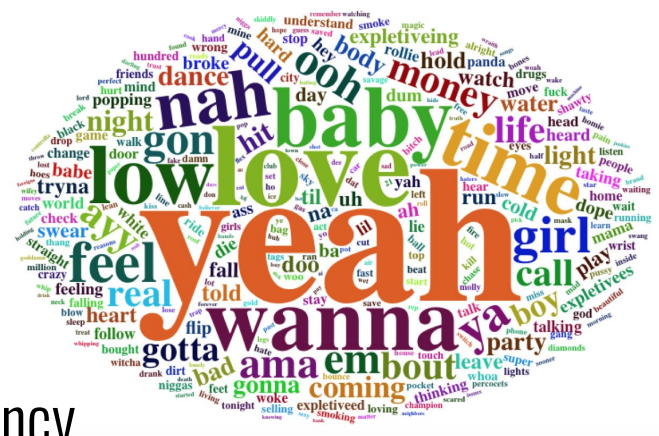

Most Frequently Used Words in top 200 songs (2018)
expletive
yeah
wanna
gang
love
ooh
baby
ayy
walk
time
0
200
400
600
800
Song Count


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


Most Frequently Used Words in top 200 songs (2021)
expletive
yeah
wanna
love
ooh
time
feel
ody
world
mind
0
200
400
Song Count

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

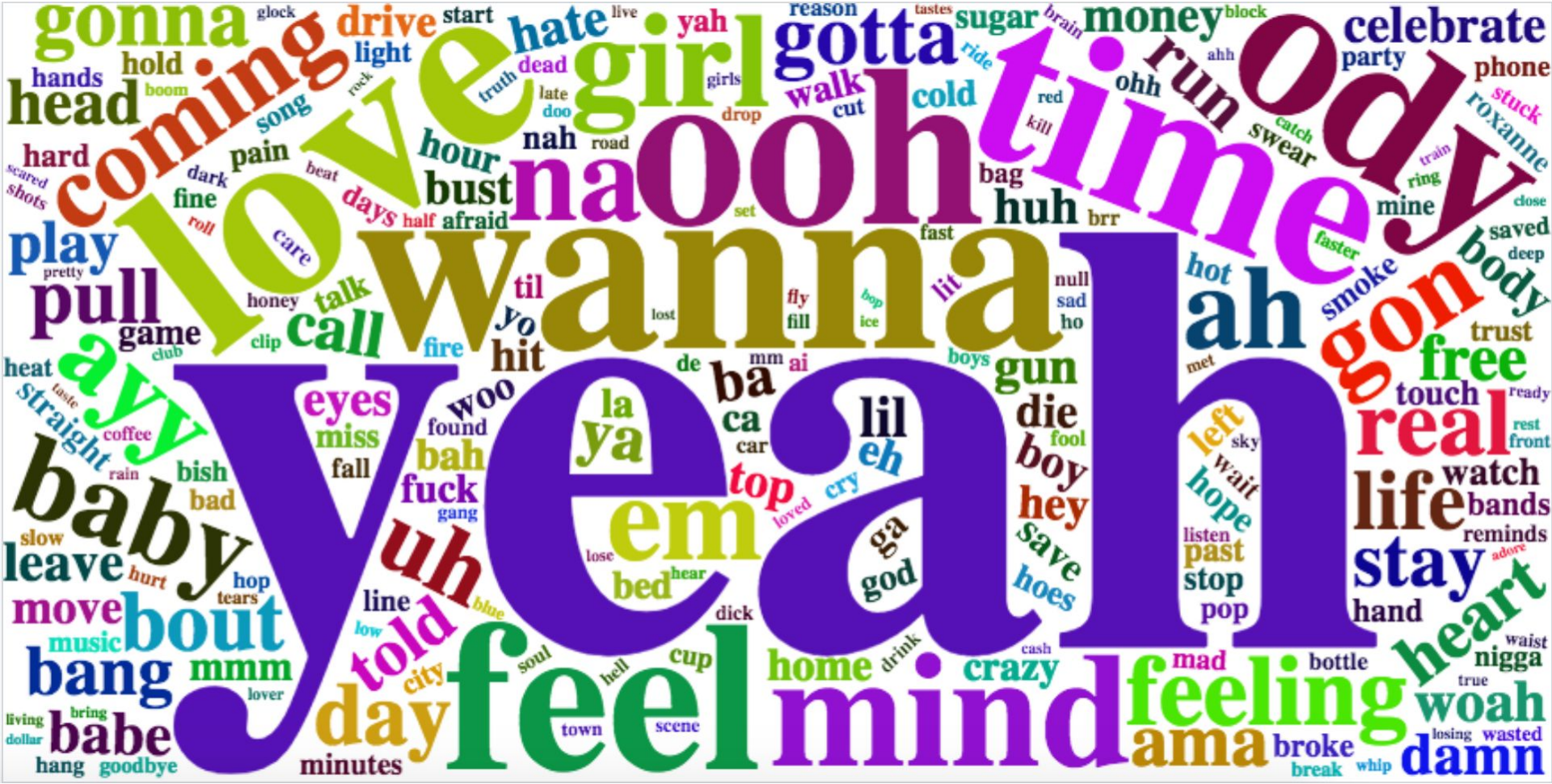


2017

2018

2019

2020


yeah
wanna
love
ooh
time
ody
coming
girl
gotta
na
gonna
head
hold
hands
hard
play
pull
ayy
baby
leave
move
bout
bang
babe
day
feel
mind
feeling
ama
damn
woah
heart
stay
life
real
gon
ah
huh
bust
hour
hate
drive
money
celebrate
run
walk
cold
hit
call
talk
eyes
miss
fuck
bah
uh
em
ya
la
lil
eh
top
hey
boy
die
gun
ba
hope
left
wait
watch
bands
reminds
touch
free
trust
smoke
body
hot
mine
swear
roxanne
stuck
phone
party
ohh
sugar
yah
reason
tastes
brain
block
ahh
ride
red
kill
catch
train
ring
close
saved
deep
faster
brr
fast
bag
nah
road
girls
drop
cut
dead
truth
late
doo
light
start
glock
live
rock
song
pain
dark
beat
days
half
afraid
fine
roll
care
scared
shots
pretty
game
club
clip
honey
fire
yo
til
lost
fly
fill
bop
ice
lit
null
sad
ho
boys
met
ai
mm
de
ca
car
found
fool
sky
wait
hope
listen
past
stop
pop
save
hoes
ga
god
cry
loved
lose
hear
bed
dick
cash
drink
crazy
mad
bottle
waist
nigga
true
losing
wasted
broke
break
whip
home
cup
soul
hell
town
scene
minutes
line
blue
low
told
city
lover
bring
living
dollar
hang
goodbye
hop
tears
hurt
bad
slow
bish
rain
coffee
straight
taste
heat
ready
rest
front
hand
adore
mmm
music
bring

2021

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

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

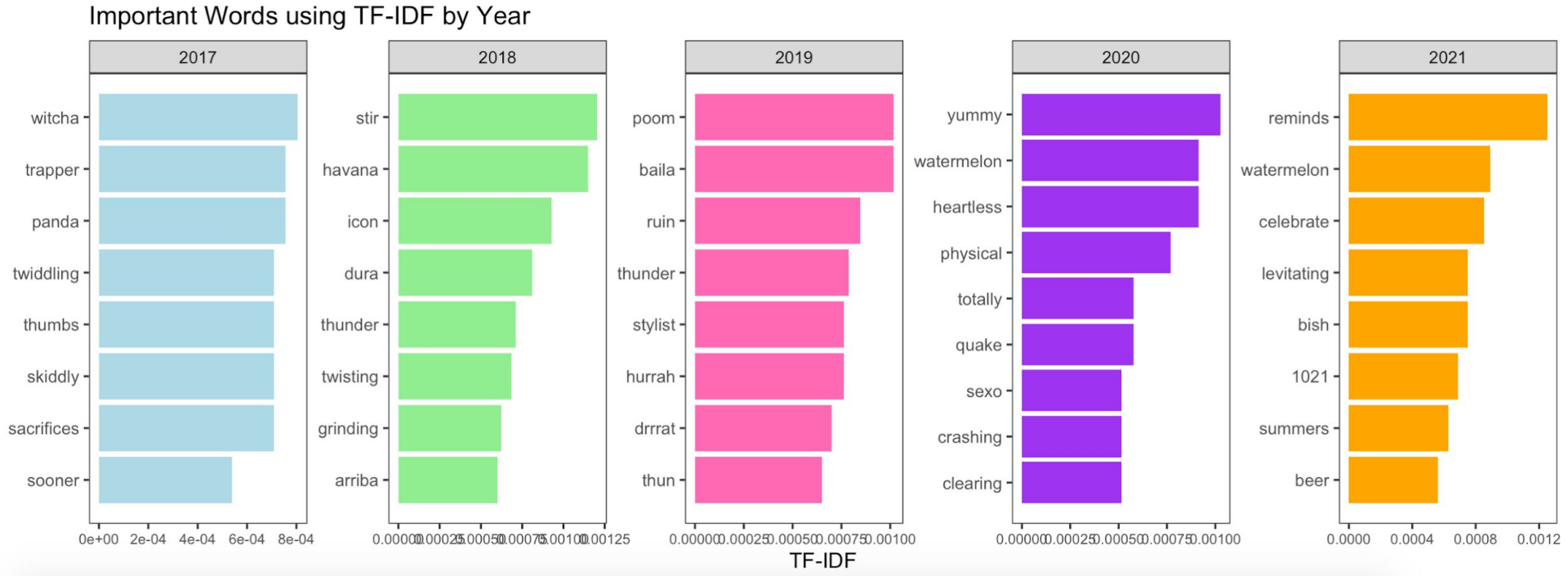
TF-IDF

Term x within document y

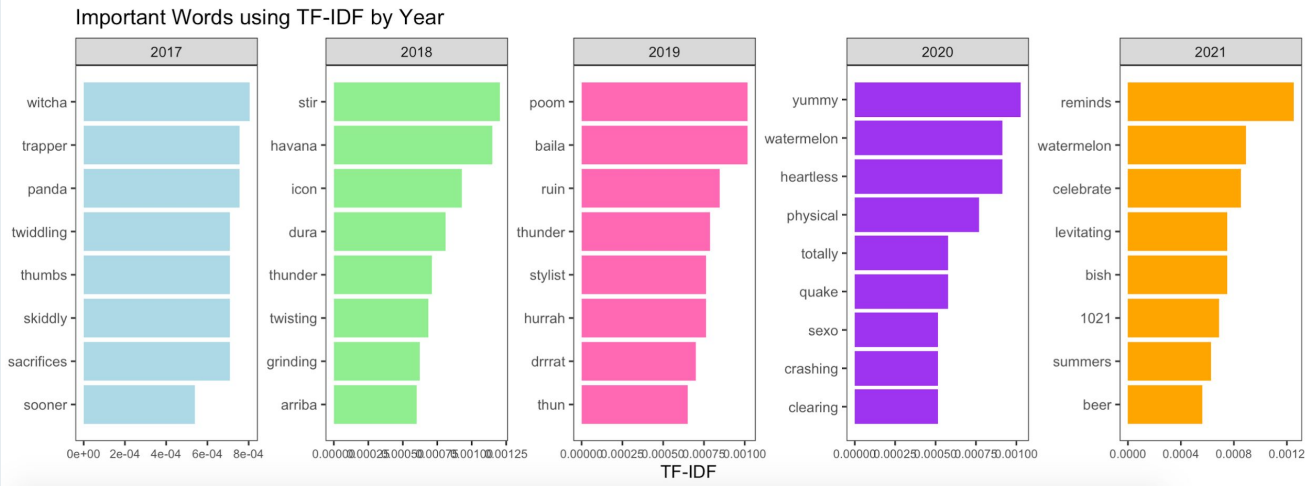
$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents


Important Words using TF-IDF by Year
2017
witcha
trapper
panda
twiddling
thumbs
skiddly
sacrifices
sooner
0e+00
2e-04
4e-04
6e-04
8e-04
2018
stir
havana
icon
dura
thunder
twisting
grinding
arriba
0.00000
0.00025
0.00050
0.00075
0.00100
0.00125
2019
poom
baila
ruin
thunder
stylist
hurrah
drrrat
thun
0.00000
0.00025
0.00050
0.00075
0.00100
2020
yummy
watermelon
heartless
physical
totally
quake
sexo
crashing
clearing
0.00000
0.00025
0.00050
0.00075
0.00100
2021
reminds
watermelon
celebrate
levitating
bish
1021
summers
beer
0.0000
0.0004
0.0008
0.0012
TF-IDF

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?

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