Course Reminders

Due today (11:59 PM):

- D7
- Q7
- Project Checkpoint #2: EDA
- Weekly Project Survey

Notes:

- If you cannot see your A3 feedback on datahub, DM me on Campuswire (or post to campuswire) and I'll get it to you
- Regrades: make request for feedback by EOD today; have 3d after you receive your feedback to submit a regrade request

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 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

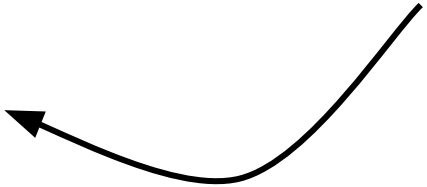
2017: 152 songs

2018: 139 songs

2019: 127 songs

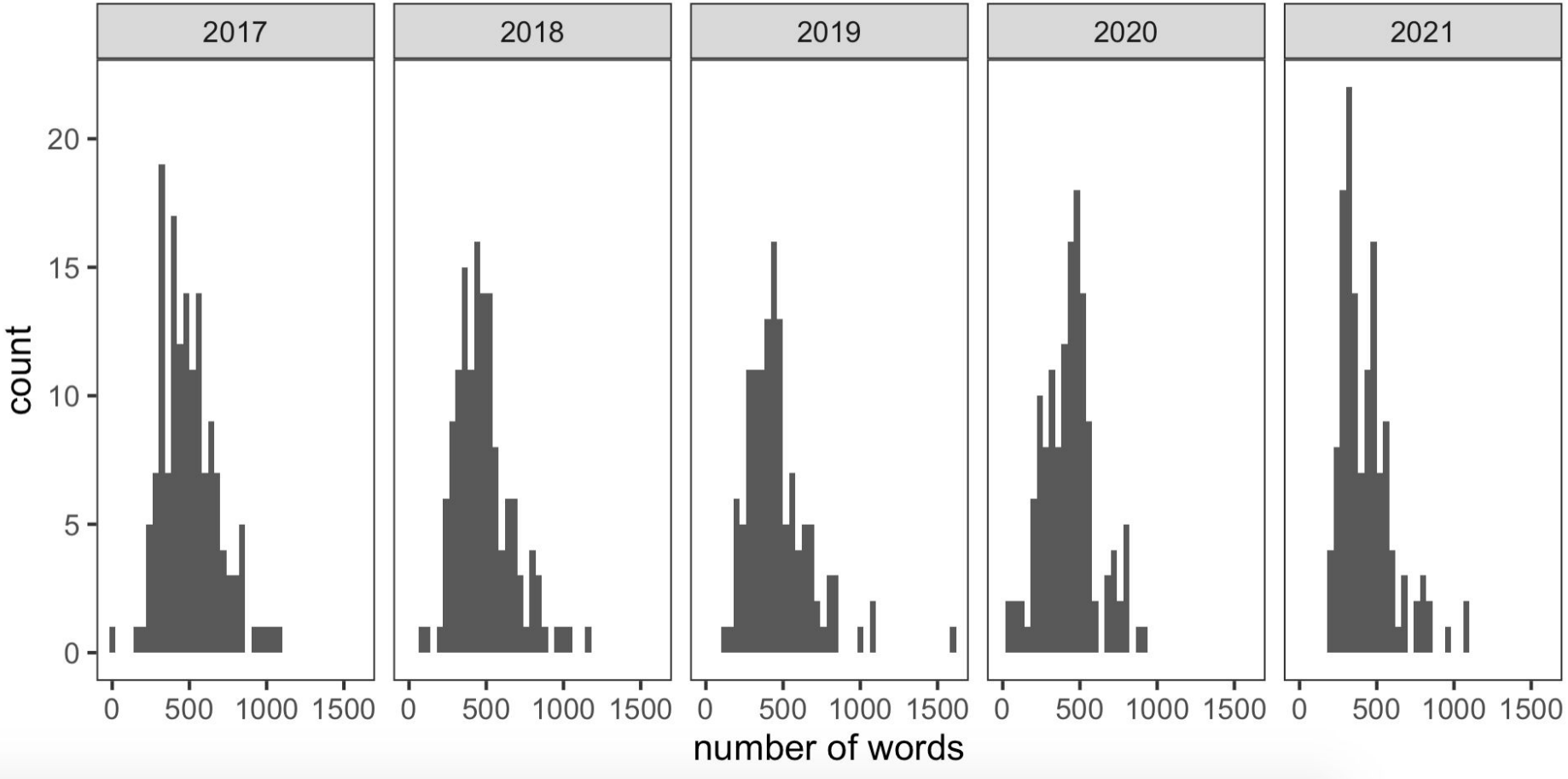
2020: 137 songs

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

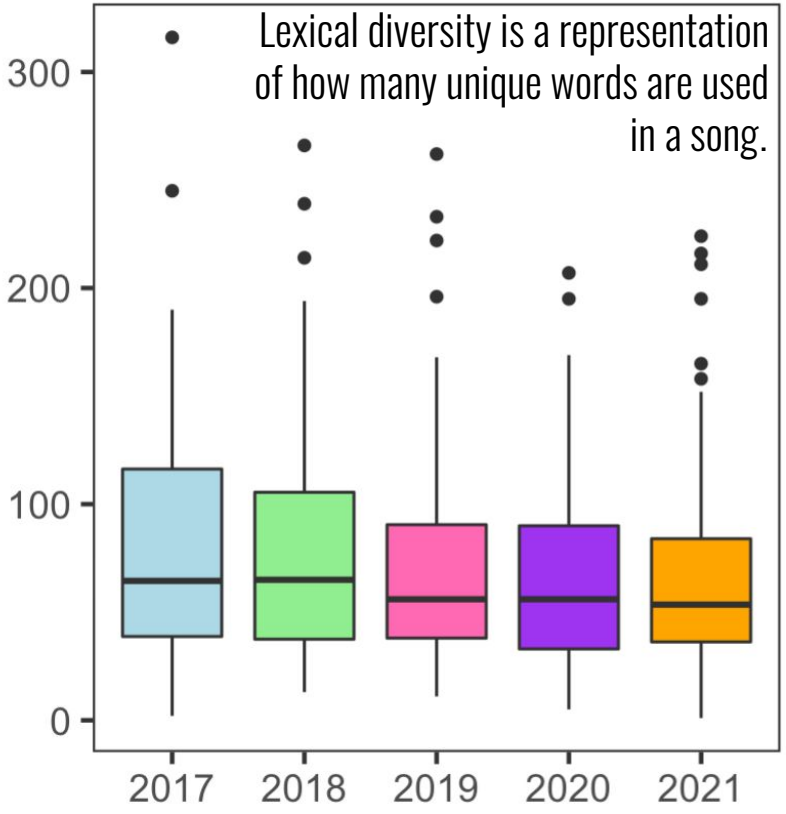


Questions we can ask...

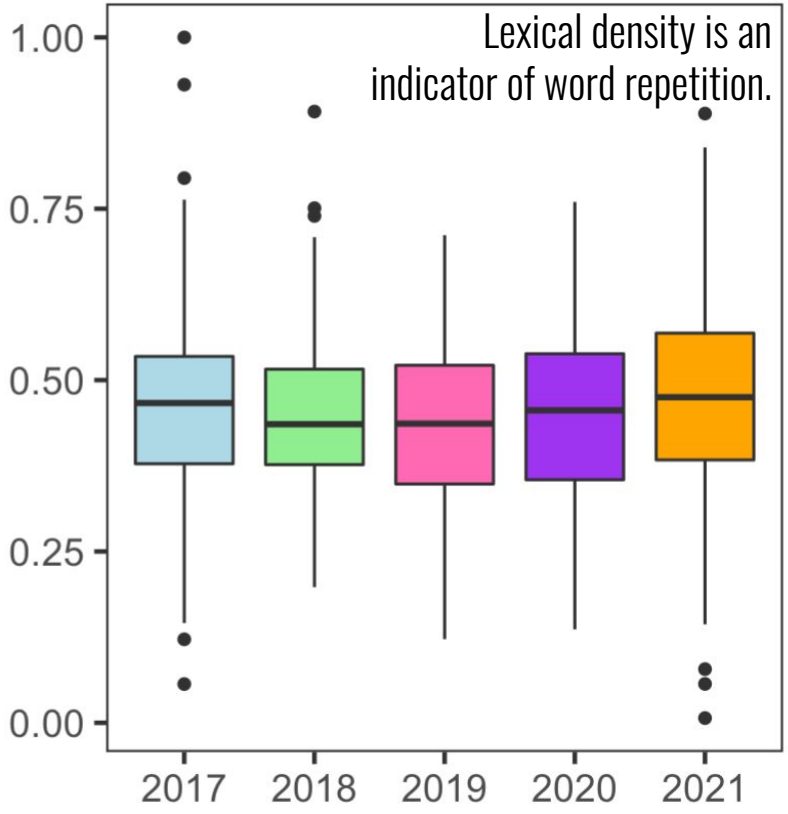
1. Does the total number of words change over time?
2. Does uniqueness change over time?
3. Does the diversity or density change?
4. What words are most common?
5. What words are most unique to each year?
6. What sentiment do songs convey most frequently?
7. Has sentiment changed over time?
8. What are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


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

Lexical Diversity



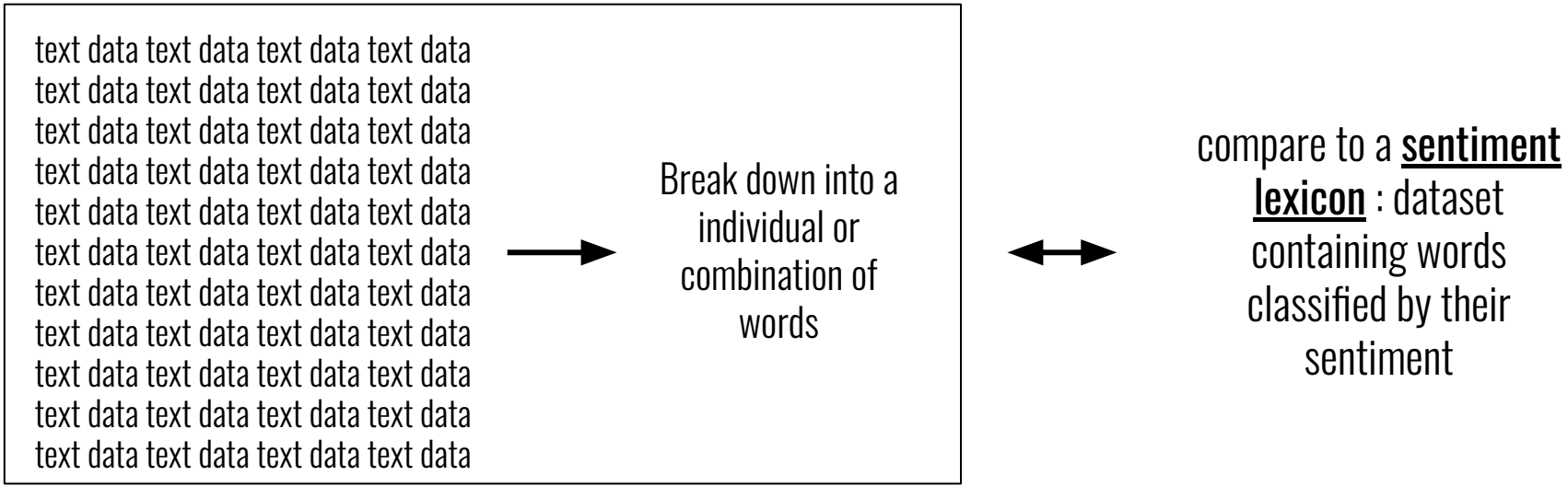
Lexical Density



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

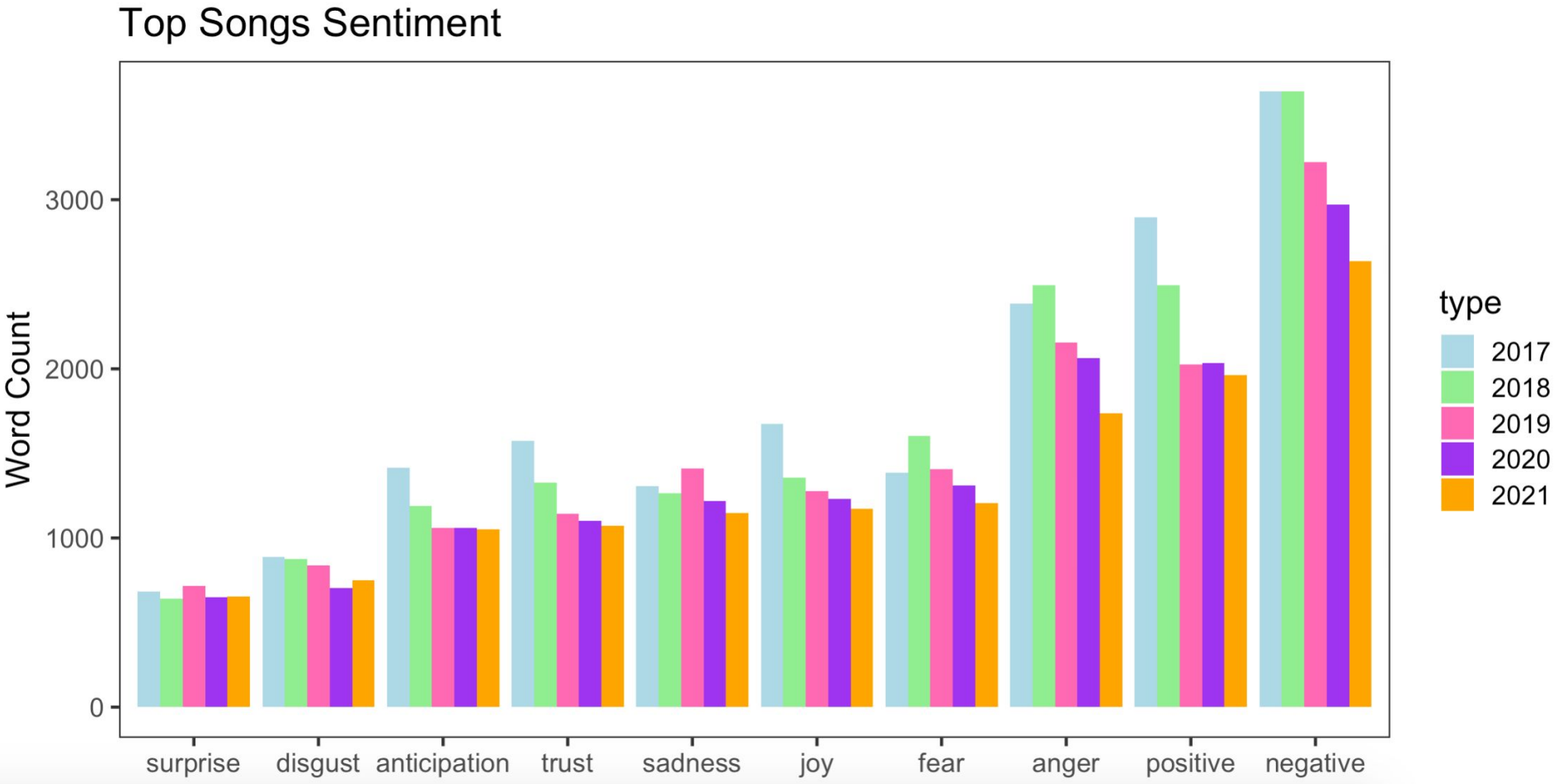
When doing sentiment analysis...

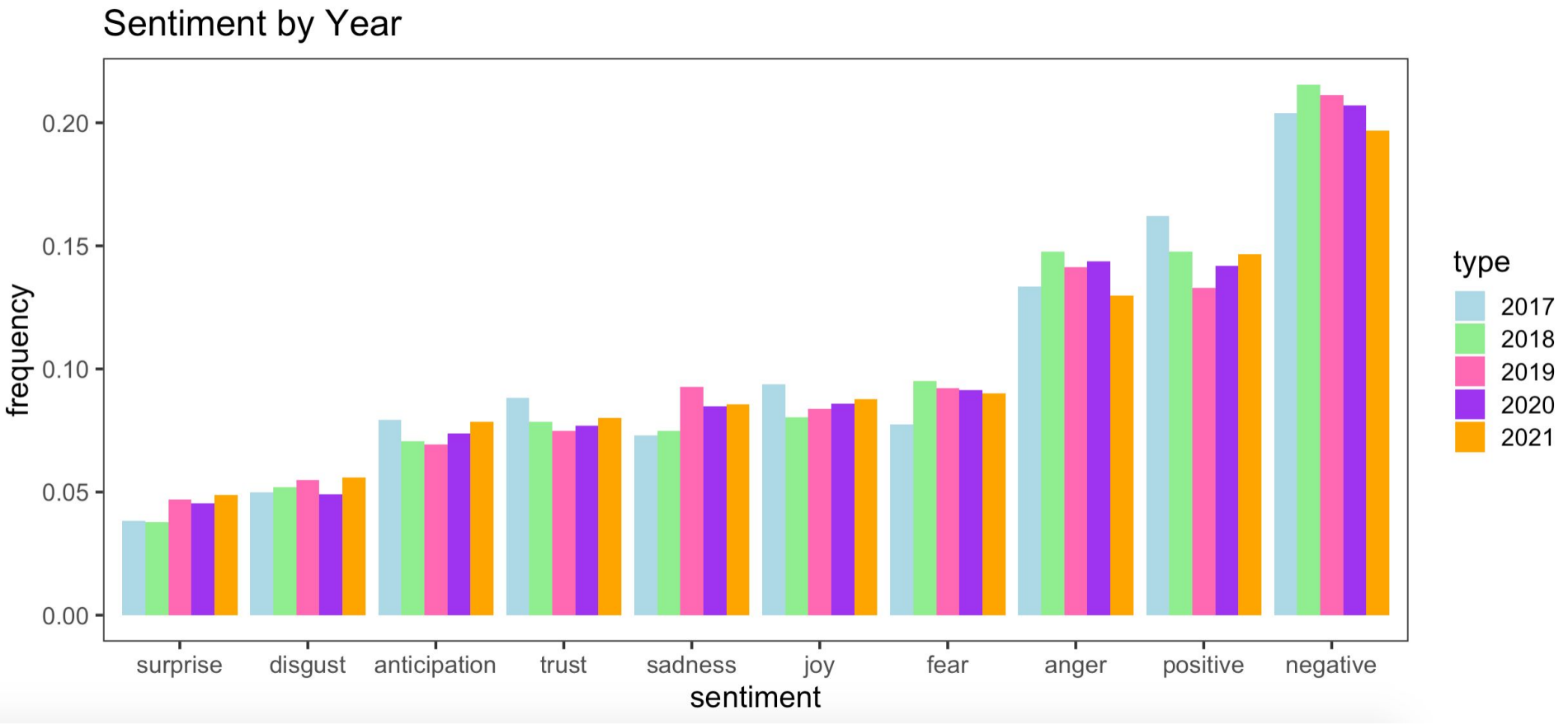
stemming - lexicon normalization

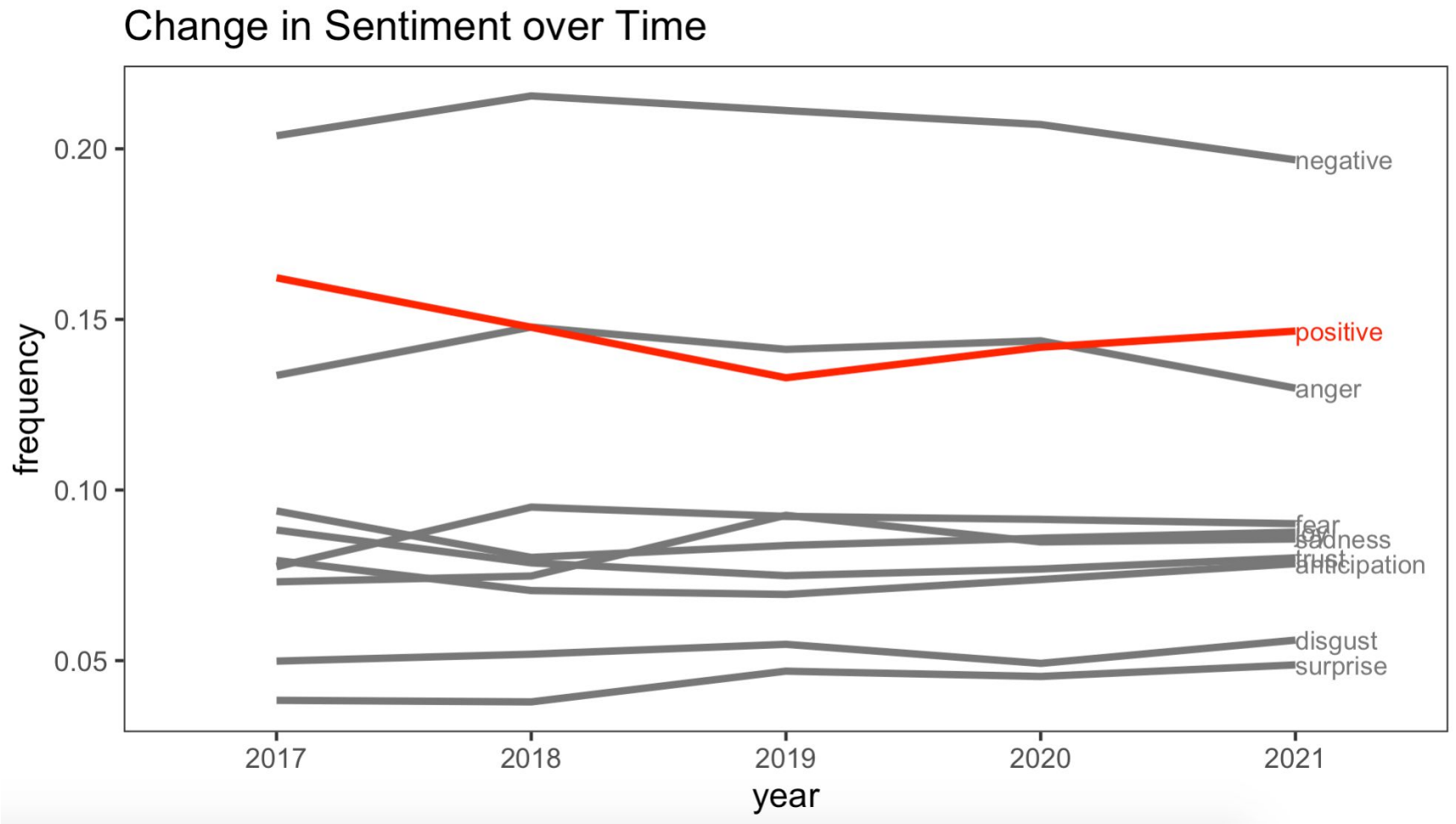
- Identifying the root for each token
- Jumping, jumped, jumps, jump all have the same root ‘jump’
- Where things get tricky: jumper???

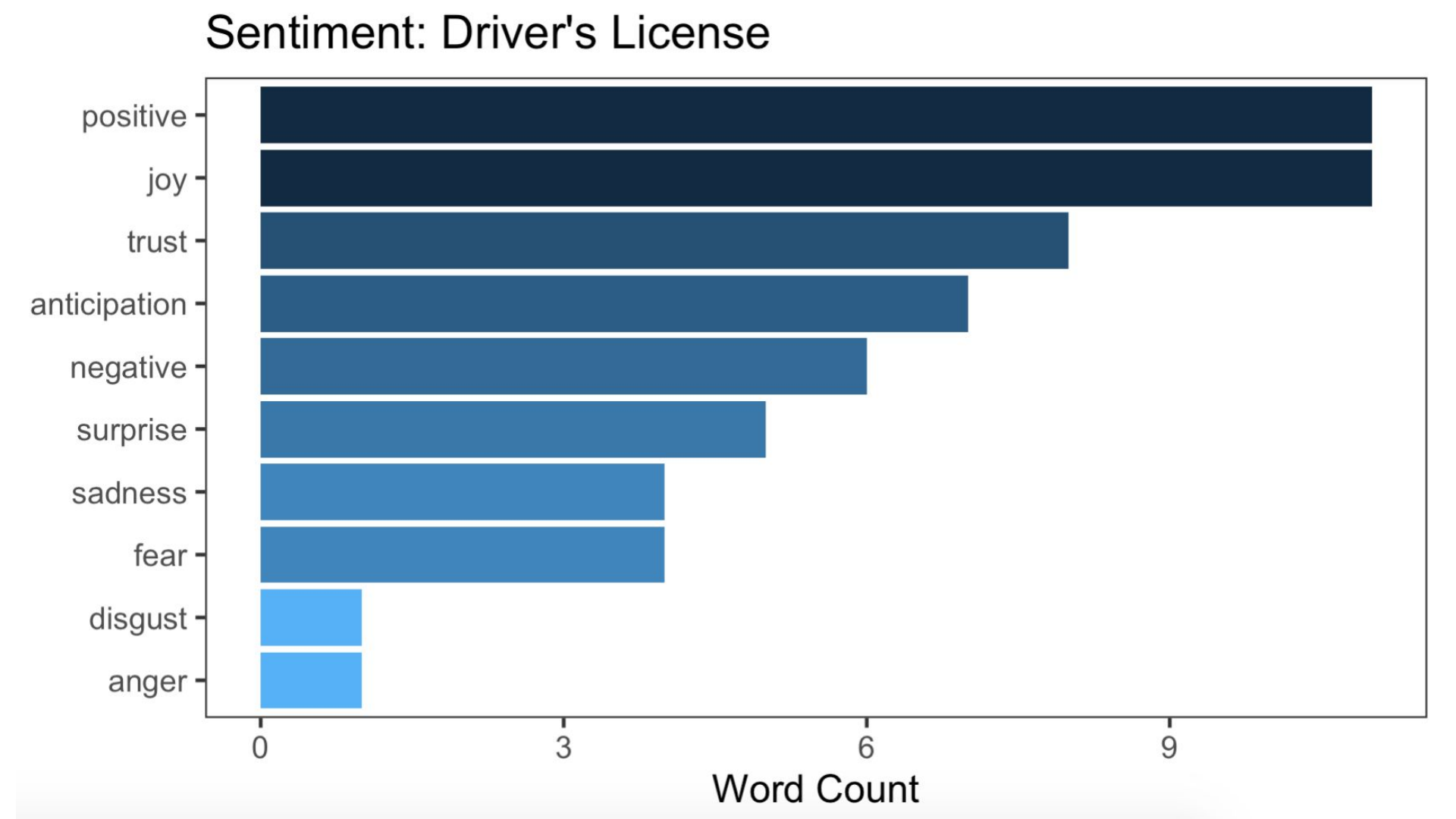
In text analysis, your choices matter:

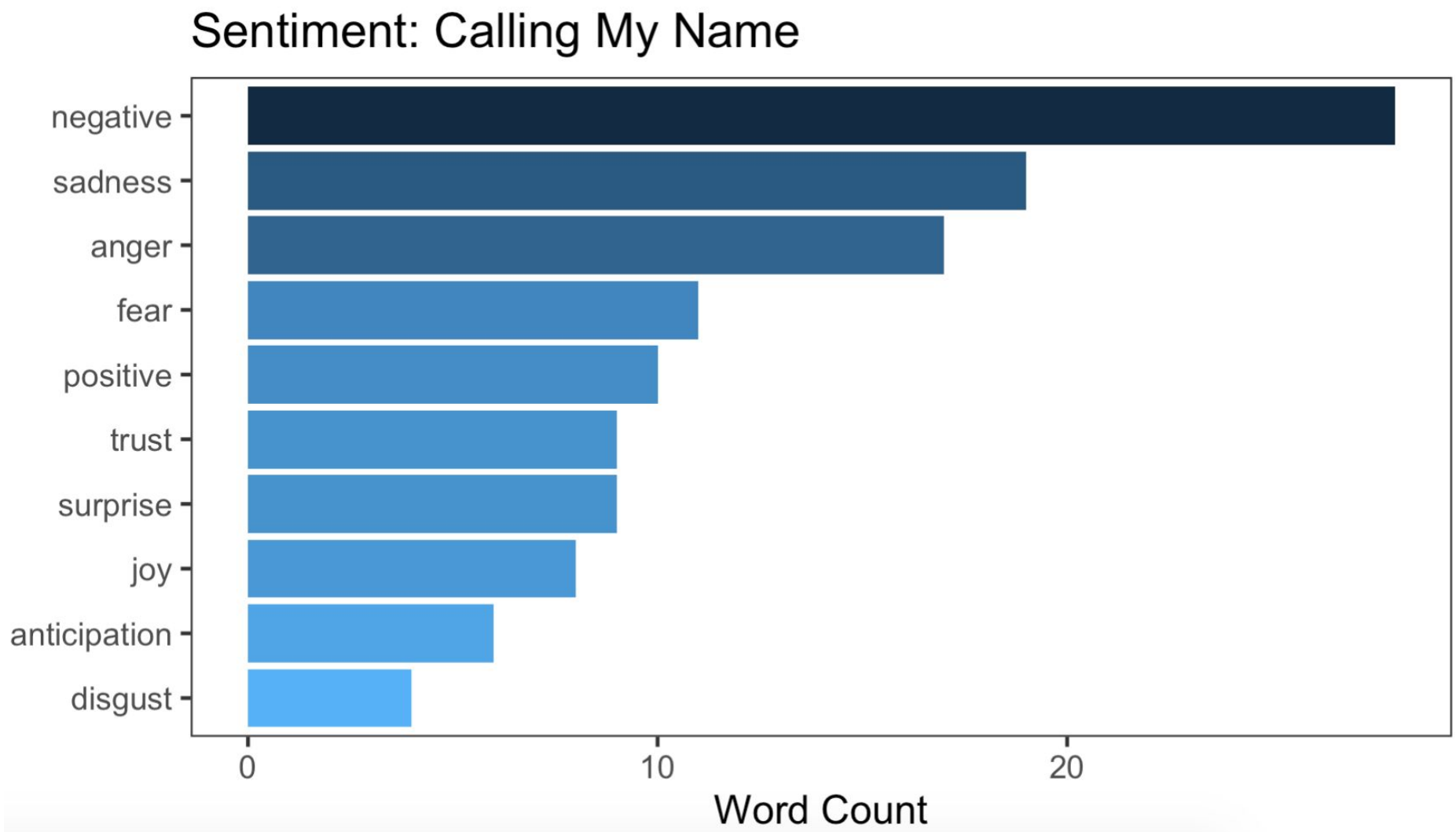
1. How to tokenize?
2. What lexicon to use?
3. Remove stop words? Remove common words?
4. Use stemming?


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

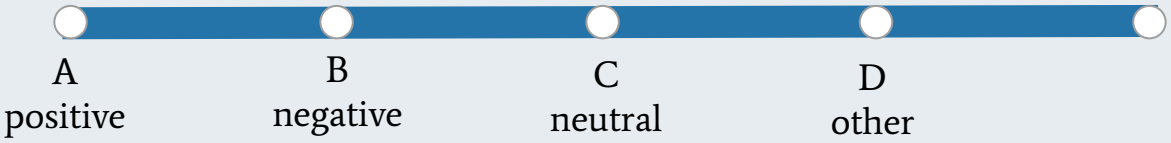

Sentiment: Driver's License
positive
joy
trust
anticipation
negative
surprise
sadness
fear
disgust
anger
0
3
6
9
Word Count


Sentiment: Calling My Name
negative
sadness
anger
fear
positive
trust
surprise
joy
anticipation
disgust
0
10
20
Word Count

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

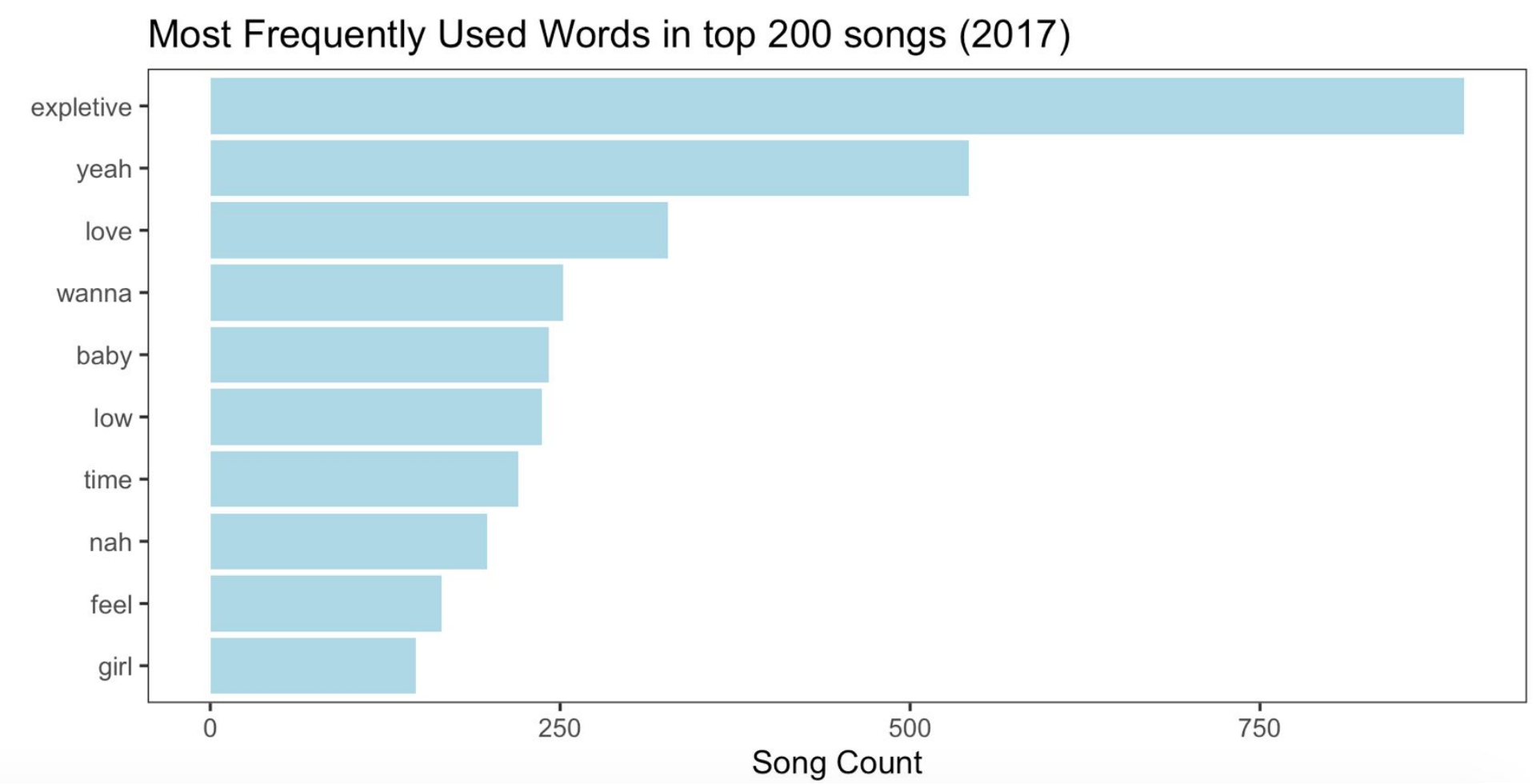
TF-IDF

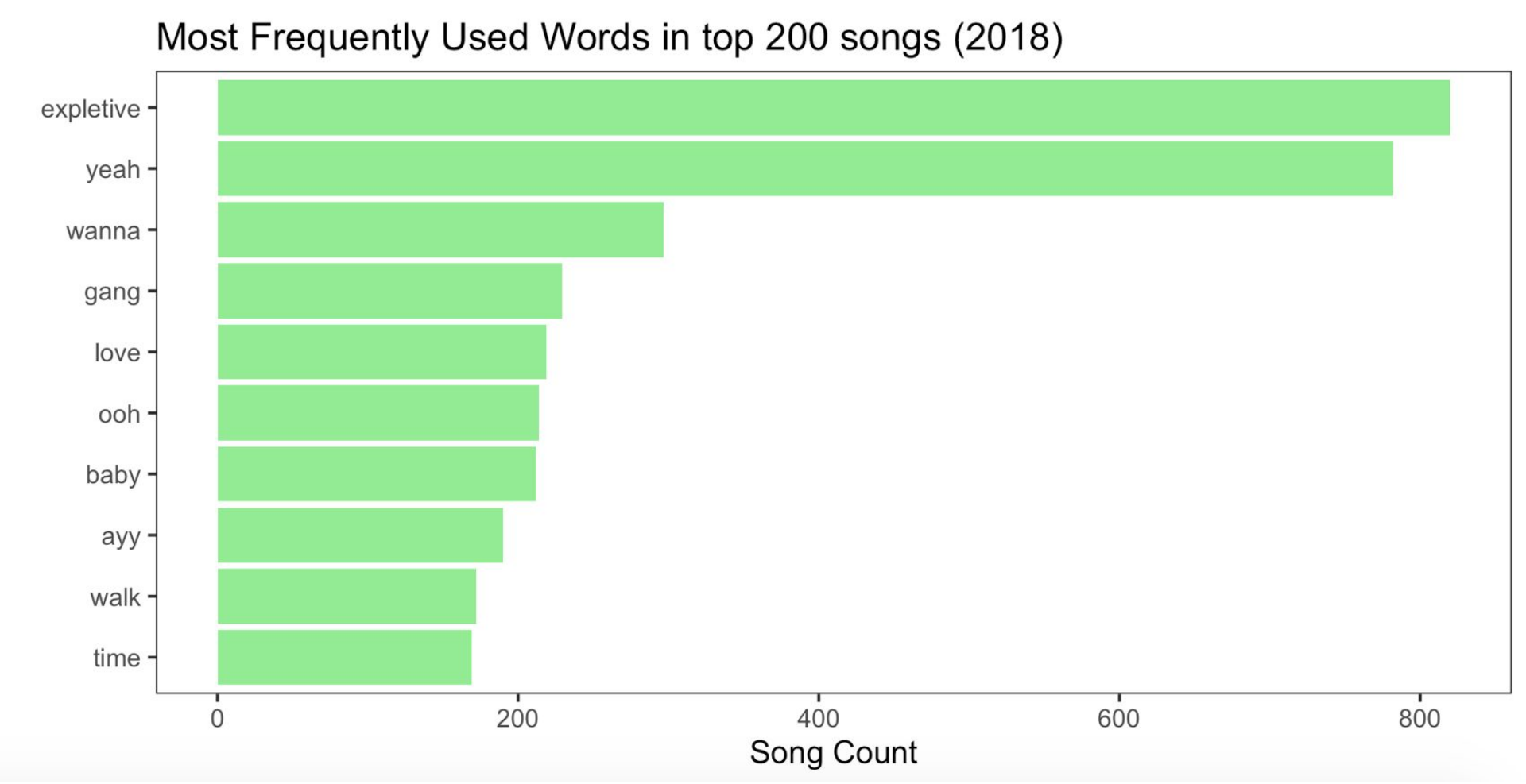
Term Frequency - Inverse Document Frequency

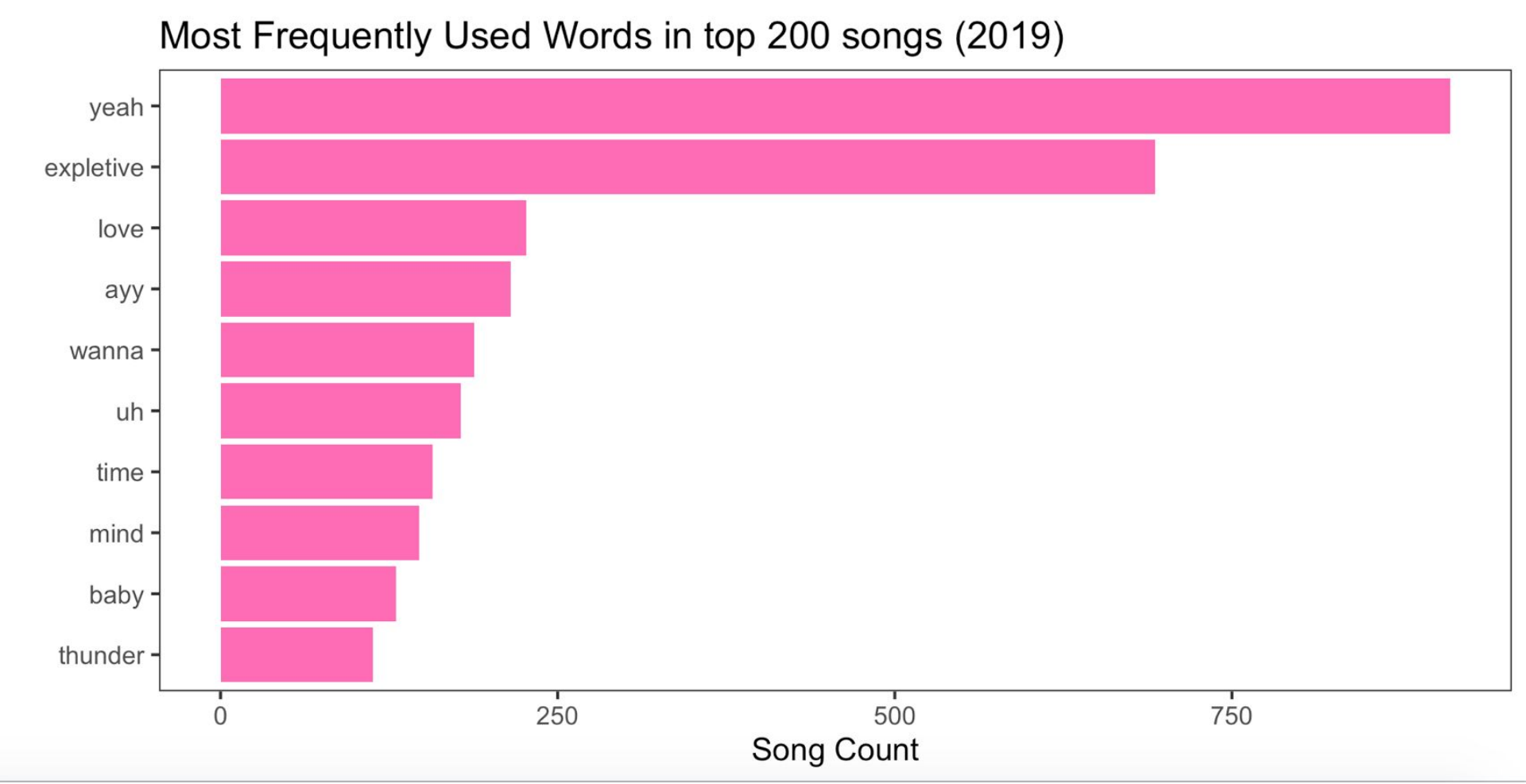
What words are the most unique to the lyrics of each year's top hits?

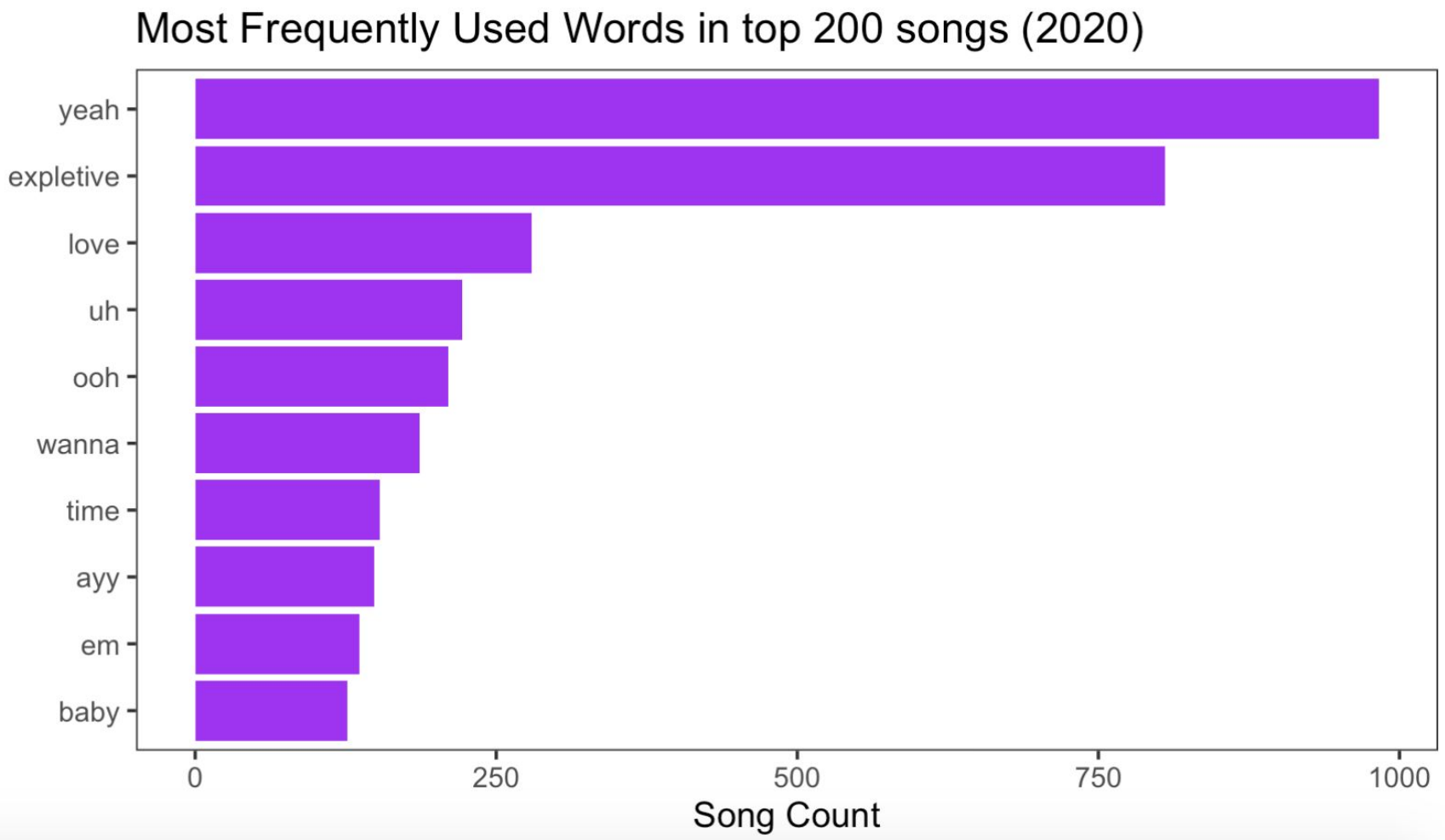
Goal: to use TF-IDF to *find the important words* for the content of each document by decreasing the weight for commonly used words and increasing the weight for words that are not used very much in a collection or corpus of documents

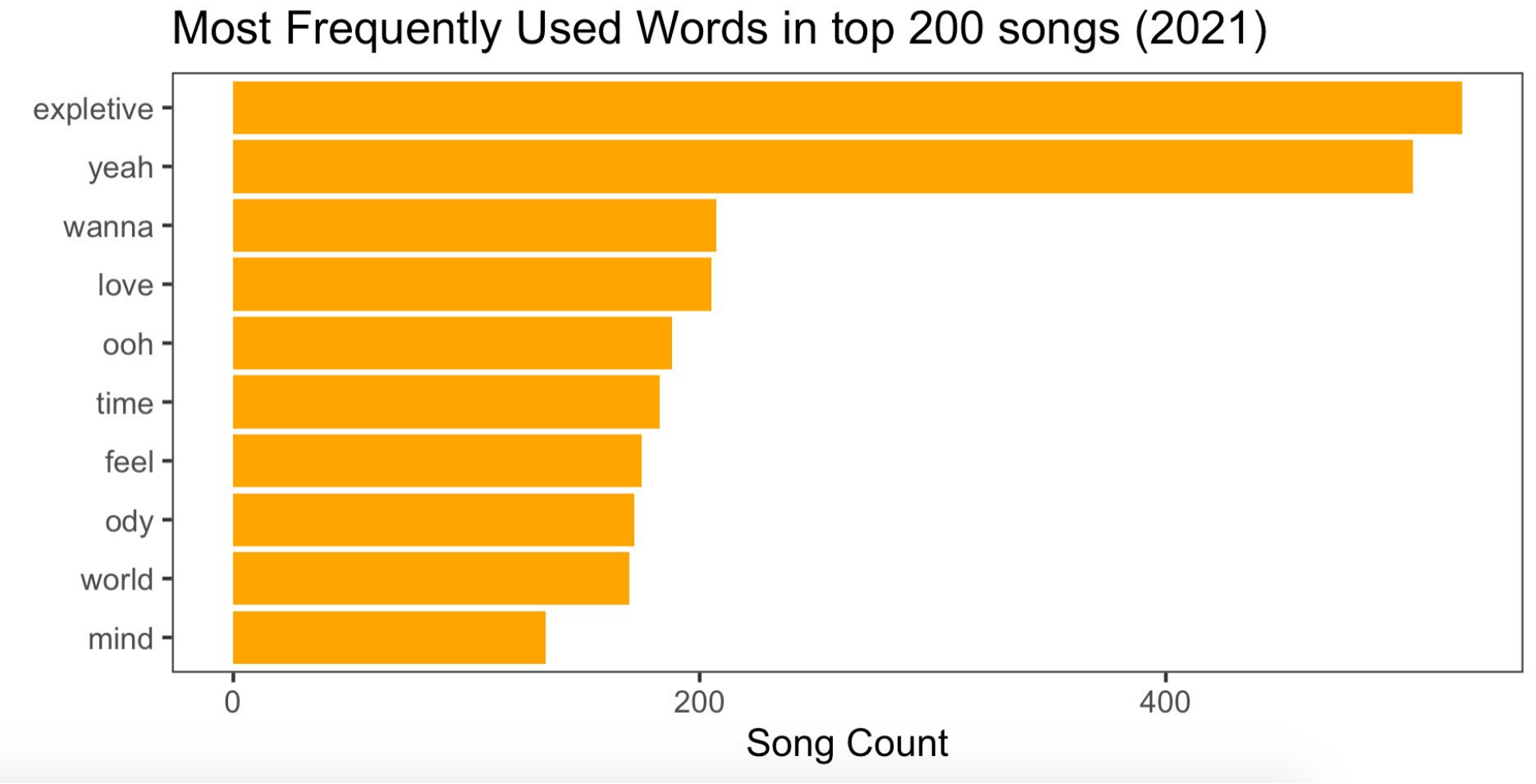
Calculating TF-IDF attempts to find the words that are important (i.e., common) in a text, but not *too* common


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

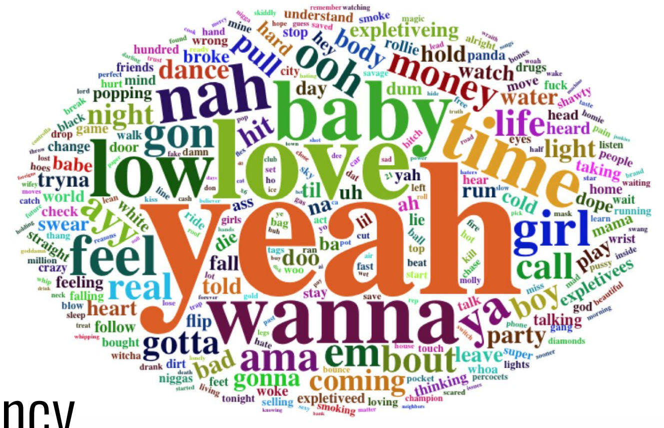

Most Frequently Used Words in top 200 songs (2018)
expletive
yeah
wanna
gang
love
ooh
baby
ayy
walk
time
0
200
400
600
800
Song Count


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....

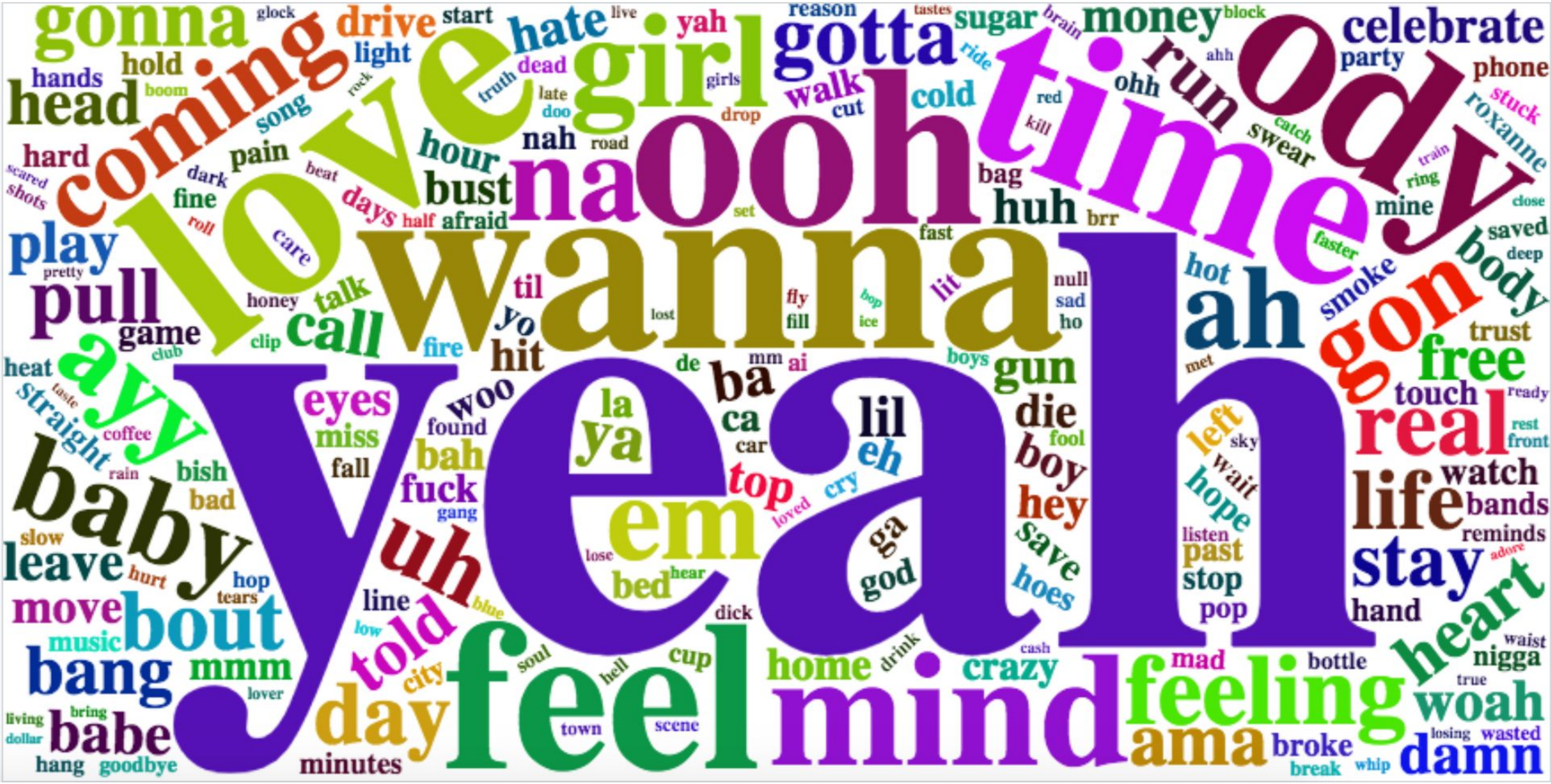


2017

2018

2019

2020


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

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

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

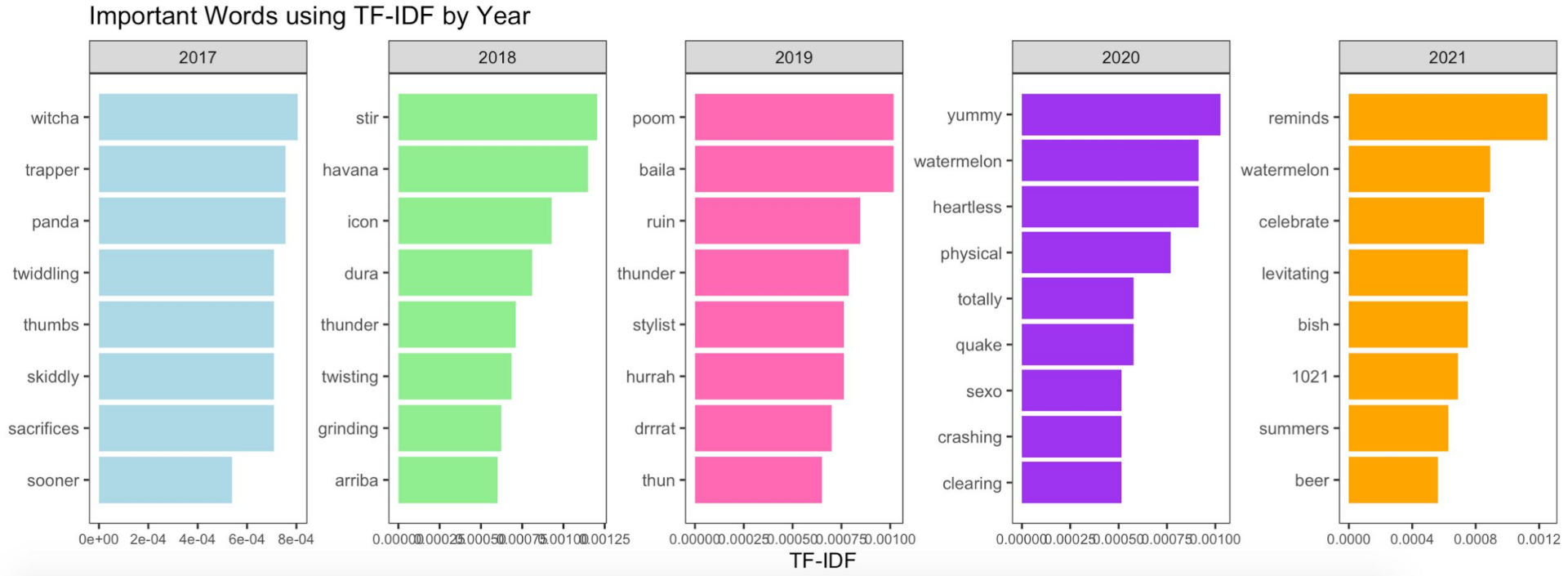
TF-IDF

Term x within document y

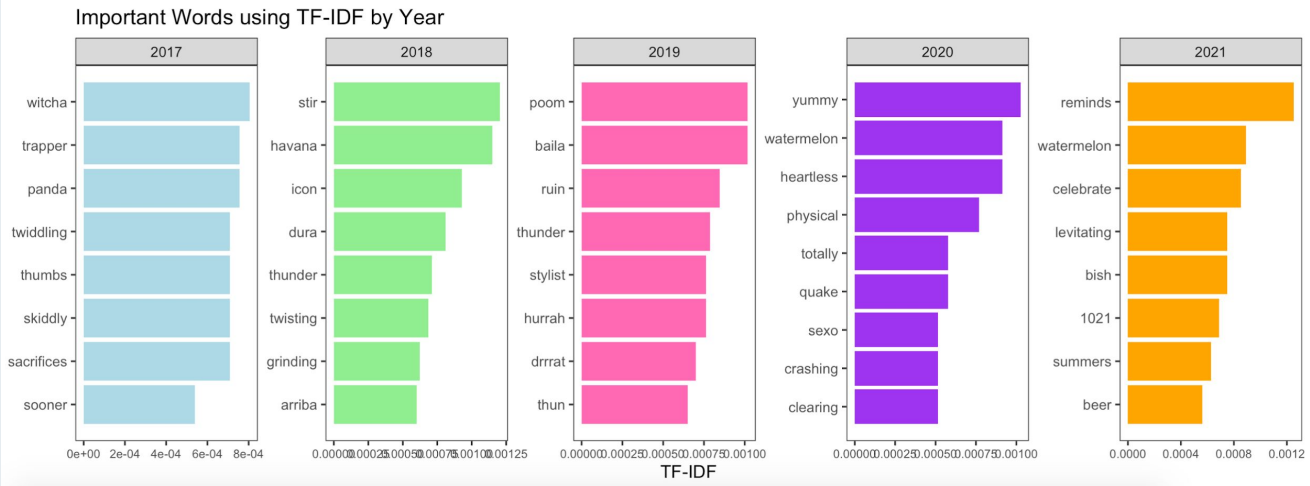
$tf_{x,y}$ = frequency of x in y

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 2020 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