

Course Reminders

Prev. Project review

Quiz

Look for post on

D3

A2

Proj. Proposal

Grades

Text Analysis

C. Alex Simpkins Jr., Ph.D
UC San Diego, RDPRobotics LLC

...

Department of Cognitive Science
rdrobotics@gmail.com
csimpkinsjr@ucsd.edu

Lectures : <https://github.com/COGS108/Lectures-Wi23>

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. Are interactions via twitter less civil than in person?

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

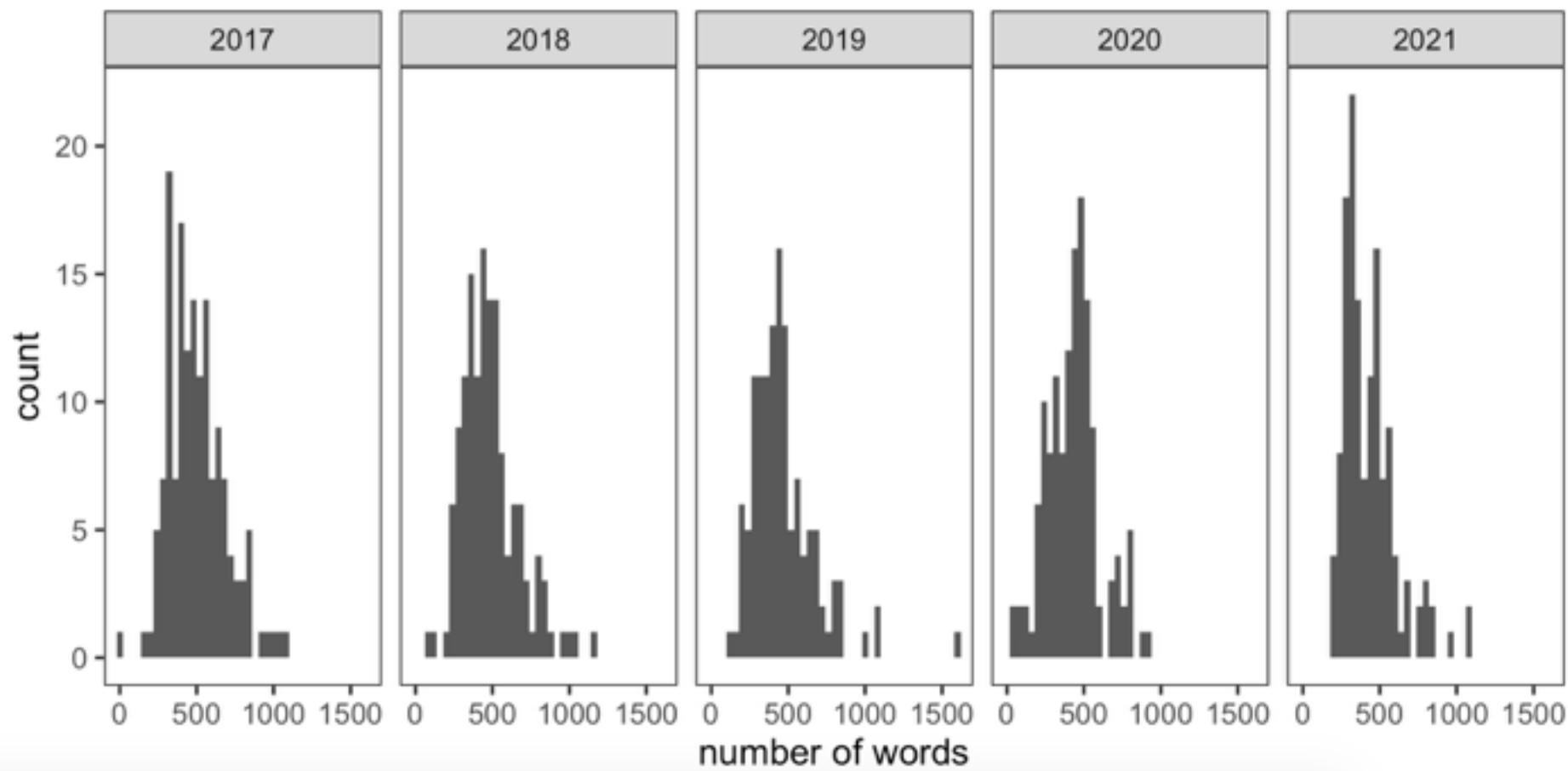
2021: 134 songs

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

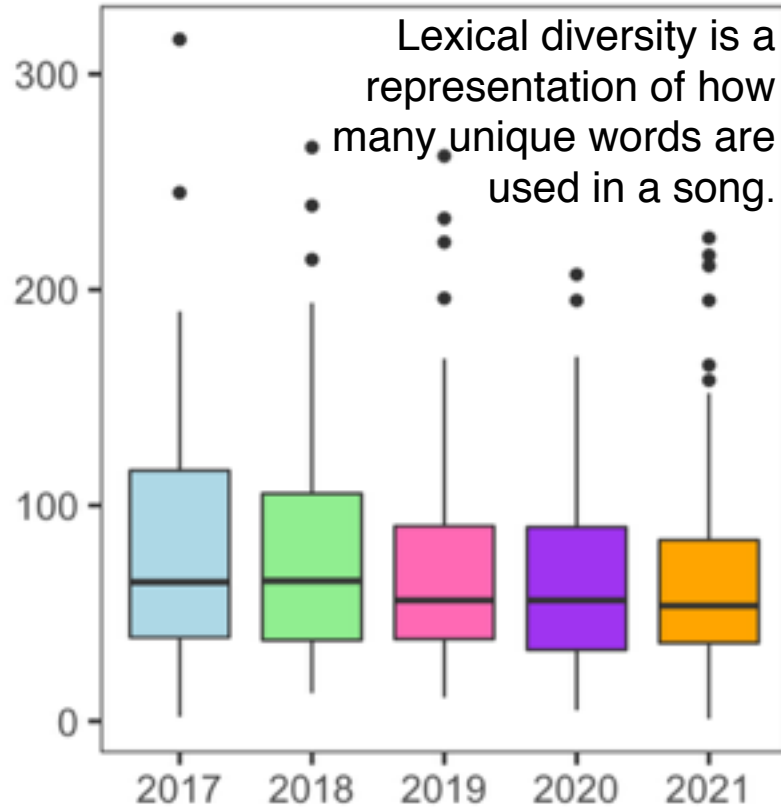


Questions we can ask...

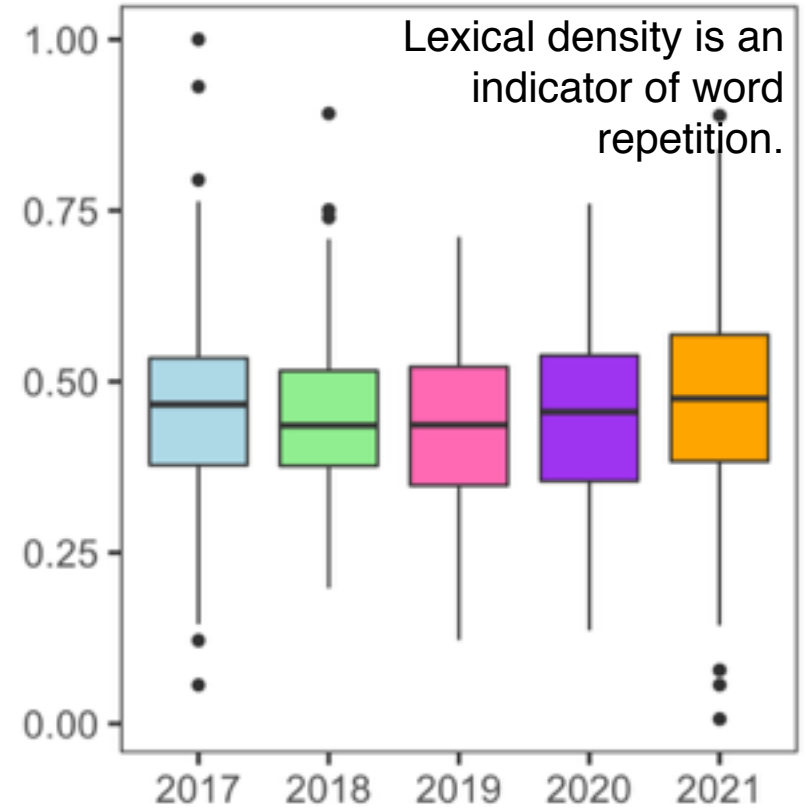
1. Does the total number of words change over time?
2. Does uniqueness change over time?
3. Does the diversity or density change?
4. What words are most common?
5. What words are most unique to each year?
6. What sentiment do songs convey most frequently?
7. Has sentiment changed over time?
8. What are 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

text data text data text data text data
text data text data text data text data
text data text data text data text data
text data text data text data text data
text data text data text data text data
text data text data text data text data
text data text data text data text data
text data text data text data text data
text data text data text data text data
text data text data text data text data




Break down
into an
individual or
combination of
words



compare to a
sentiment lexicon :
dataset containing
words classified by
their sentiment

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...

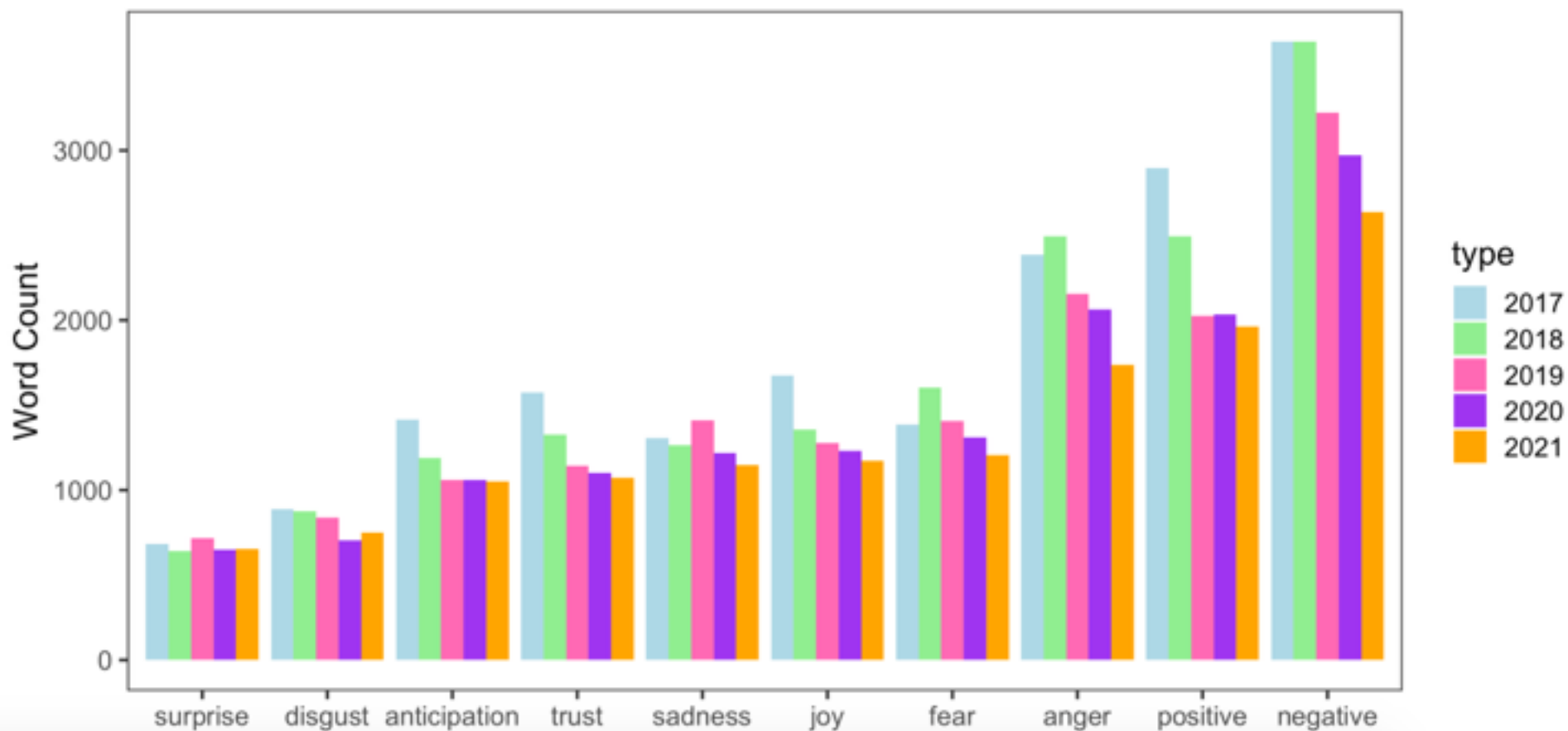
stemming - lexicon normalization

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

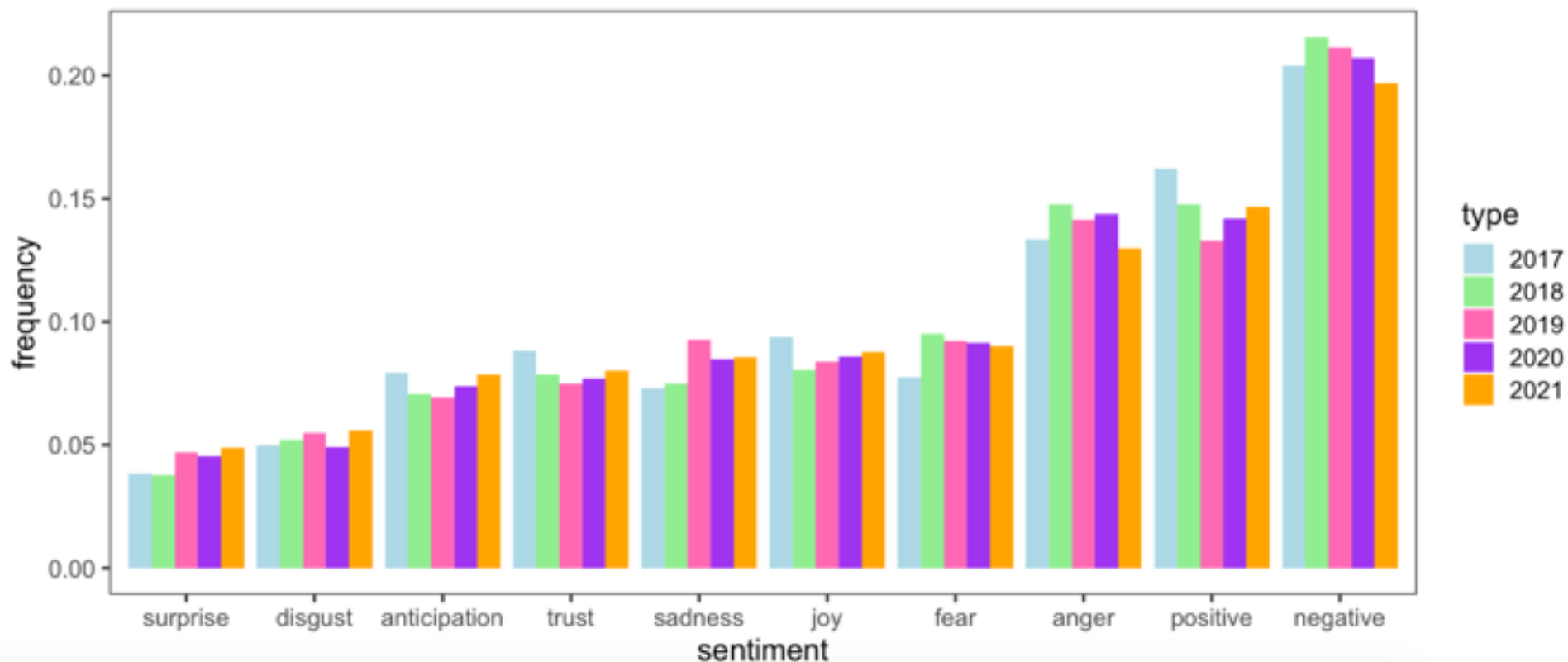
In text analysis, your choices matter:

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

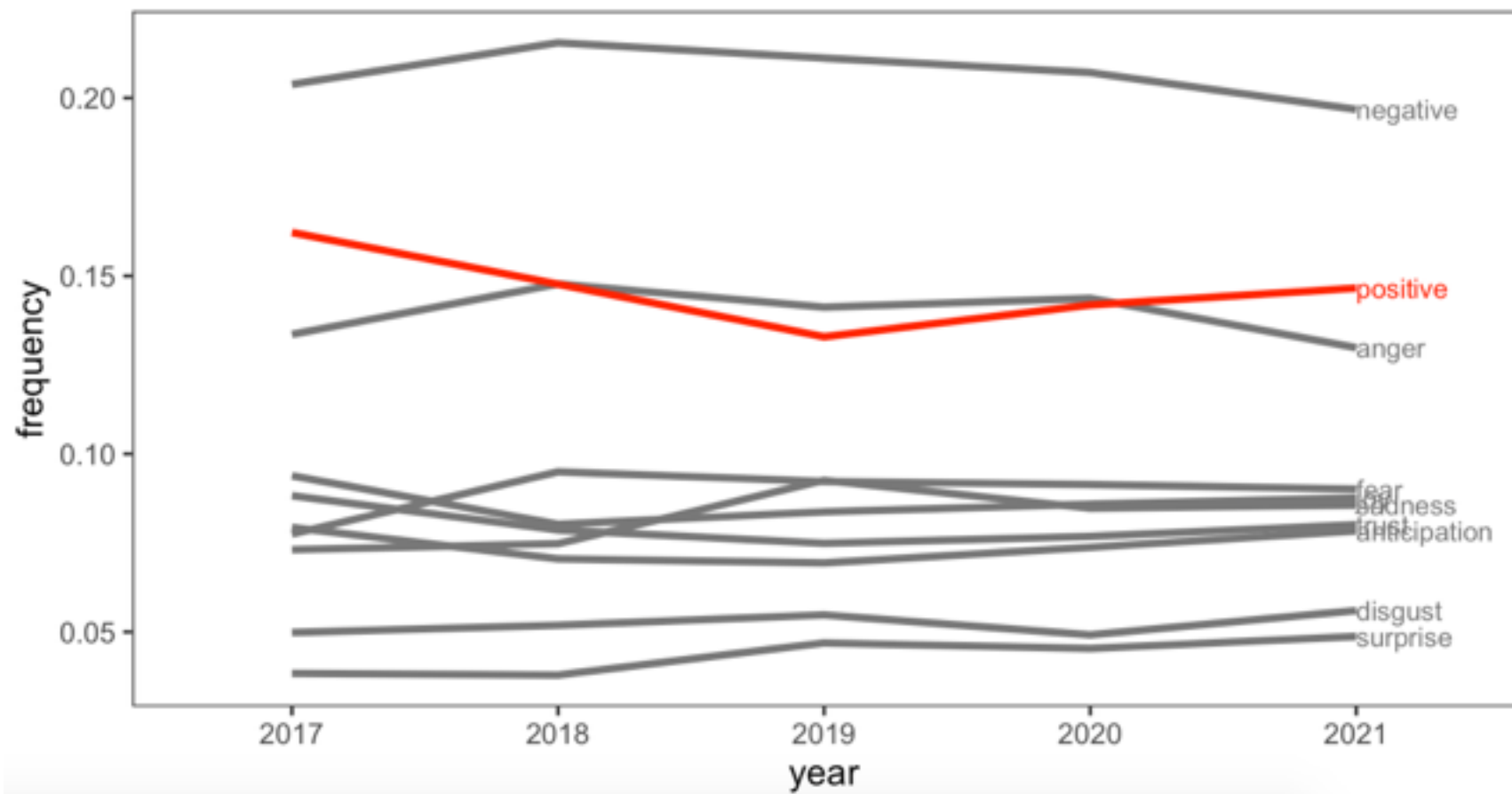
Top Songs Sentiment



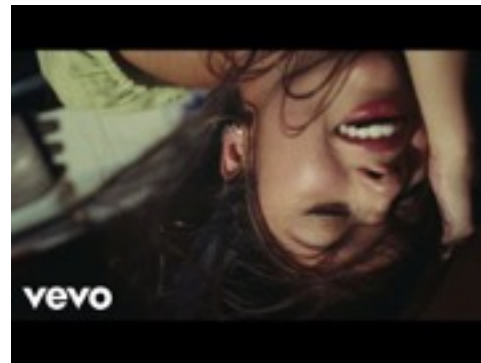
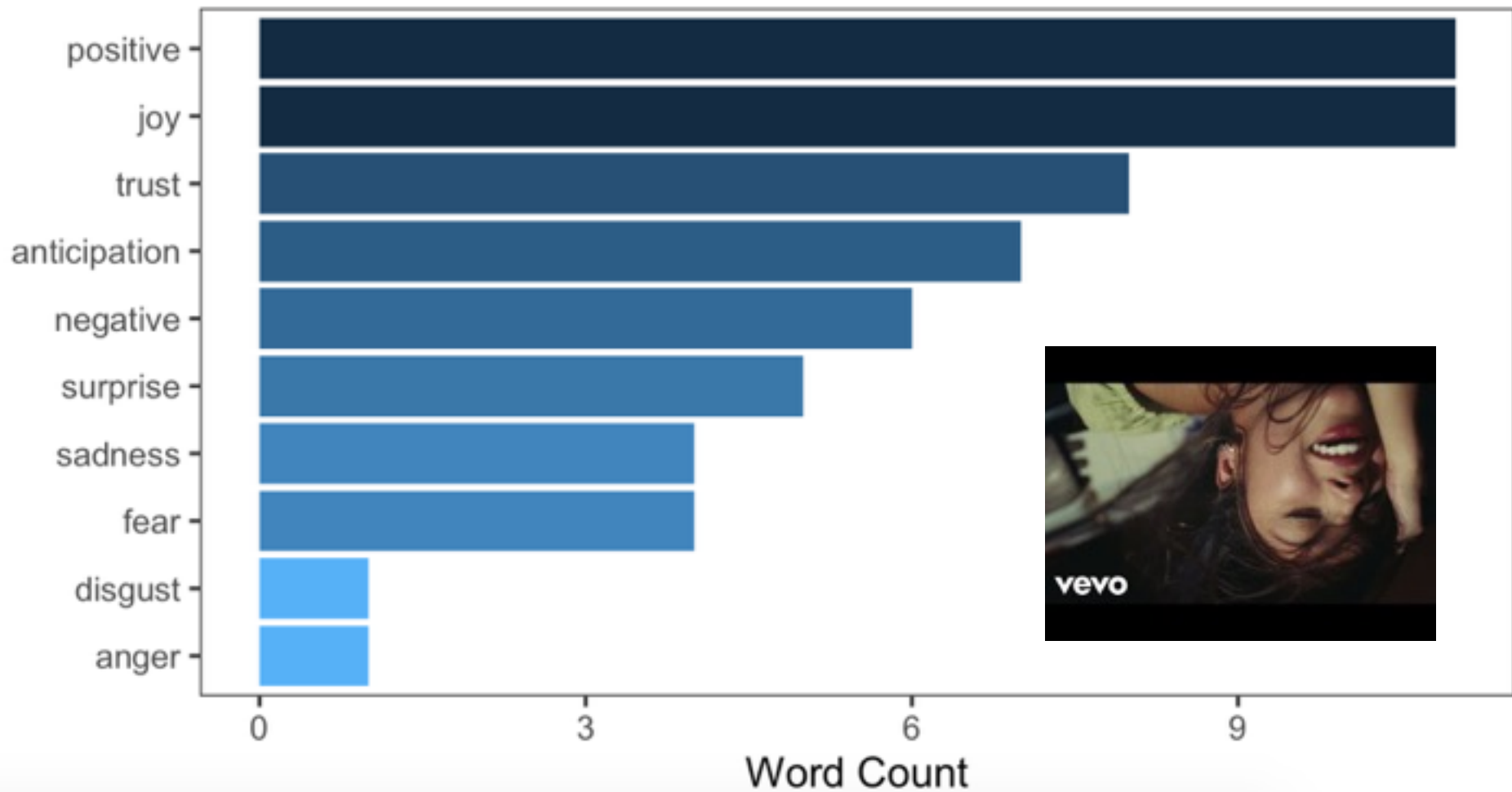
Sentiment by Year



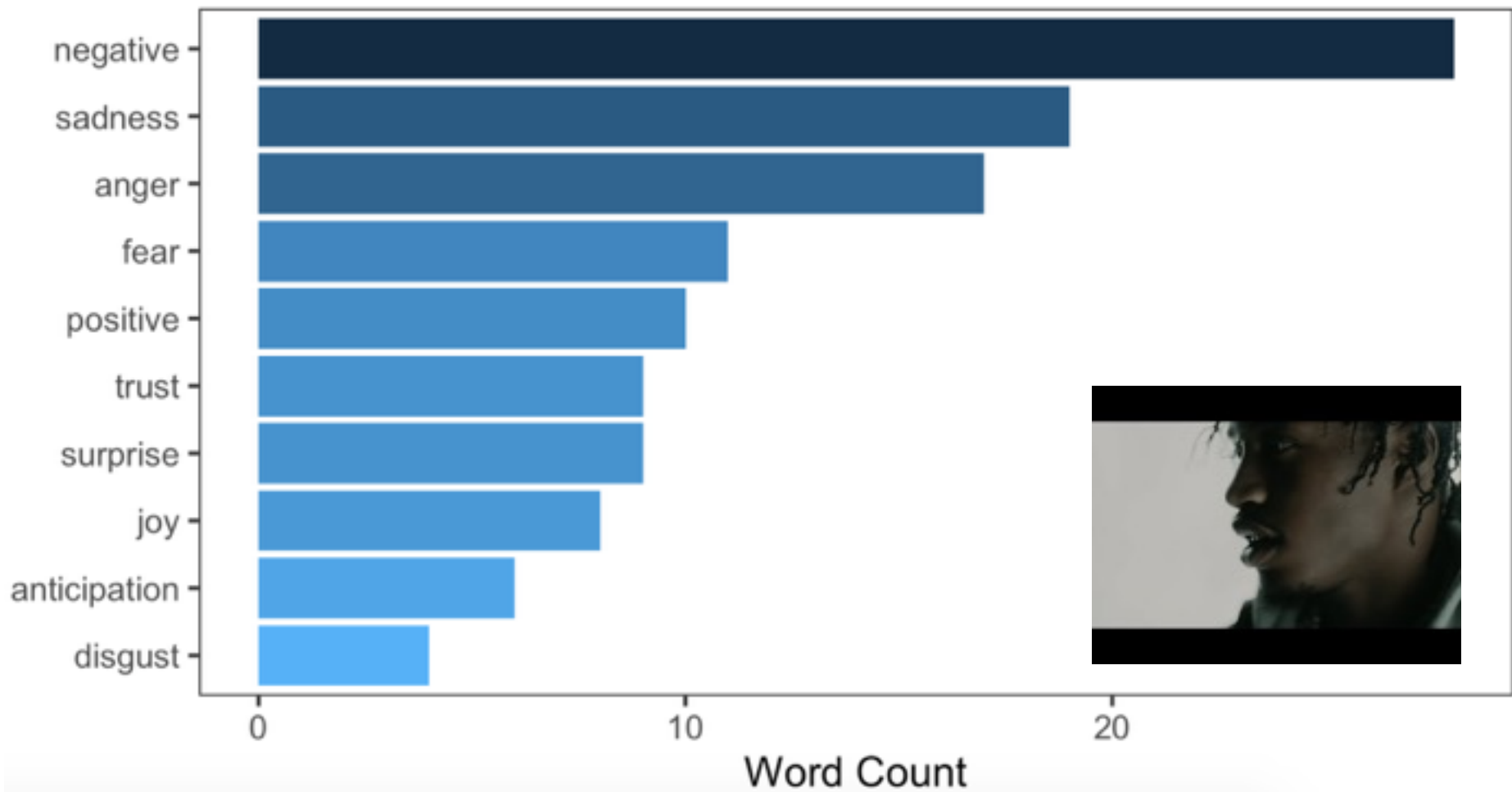
Change in Sentiment over Time



Sentiment: Driver's License



Sentiment: Calling My Name





Sentiment Limitations

How would you classify the sentiment of the following sentence?

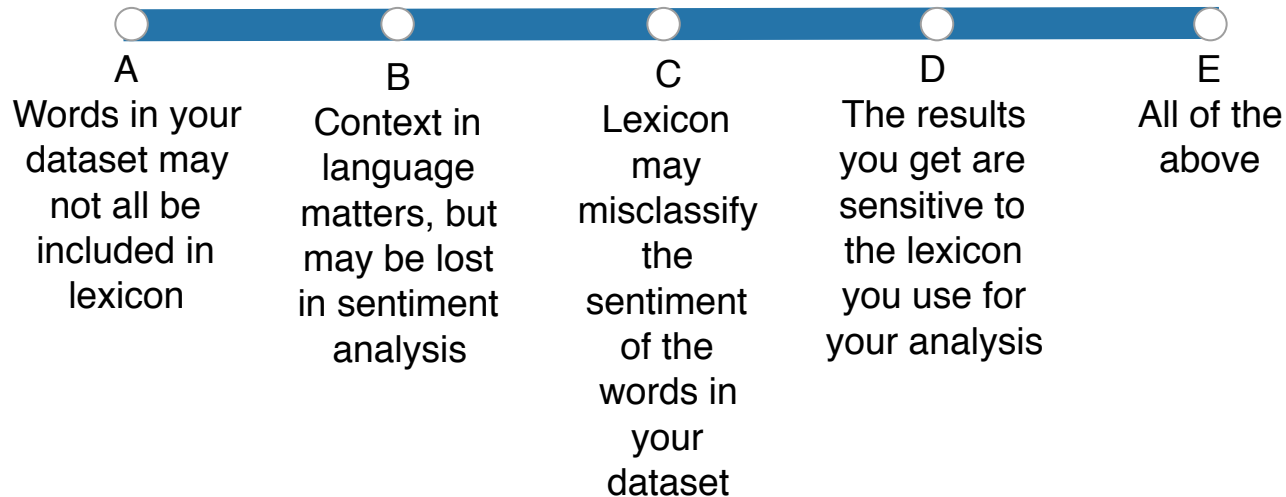
“The idea behind the movie was great, but it could have been better”

A positive B negative C neutral D other



Sentiment Limitations

What is a limitation of sentiment analysis?



TF-IDF

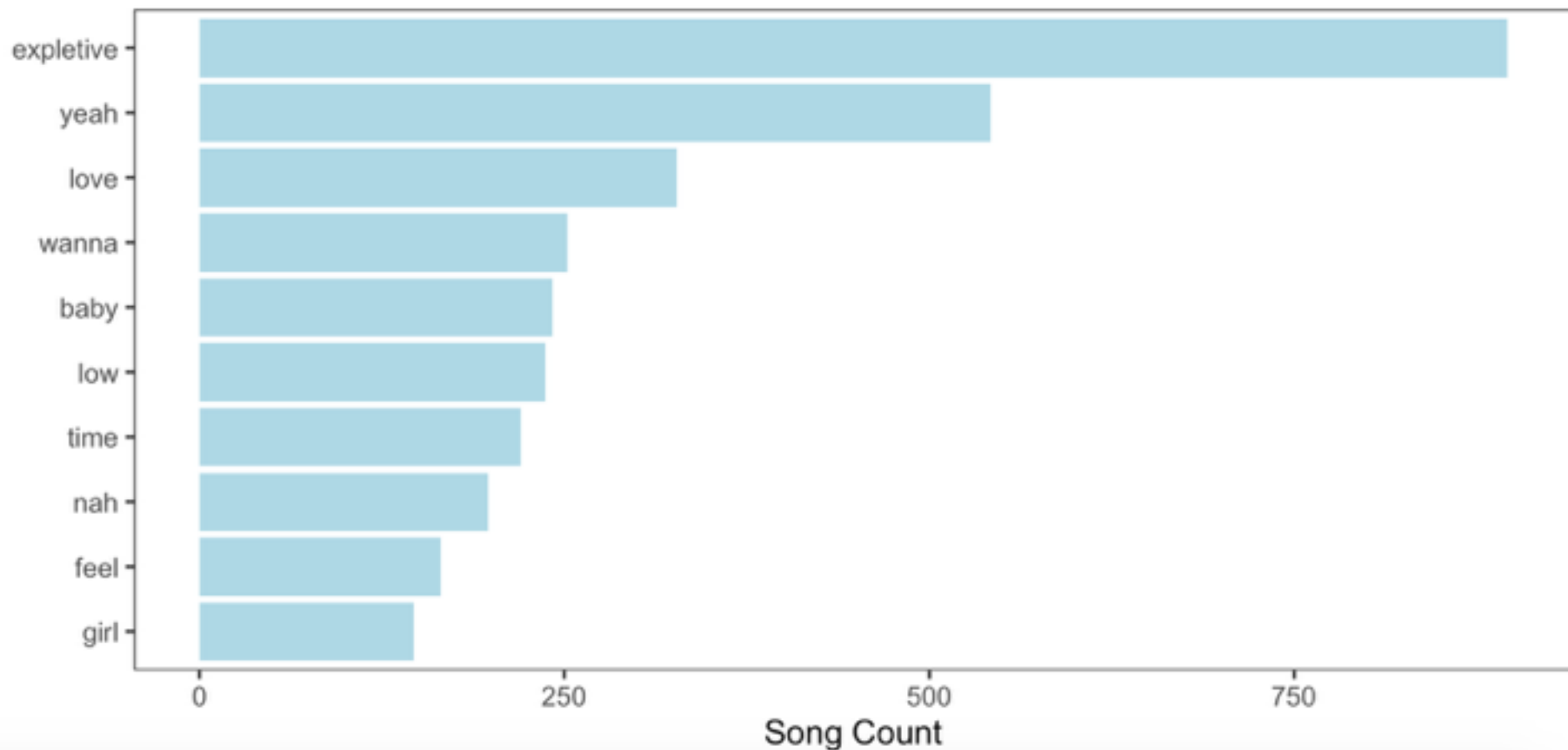
Term Frequency - Inverse Document Frequency

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

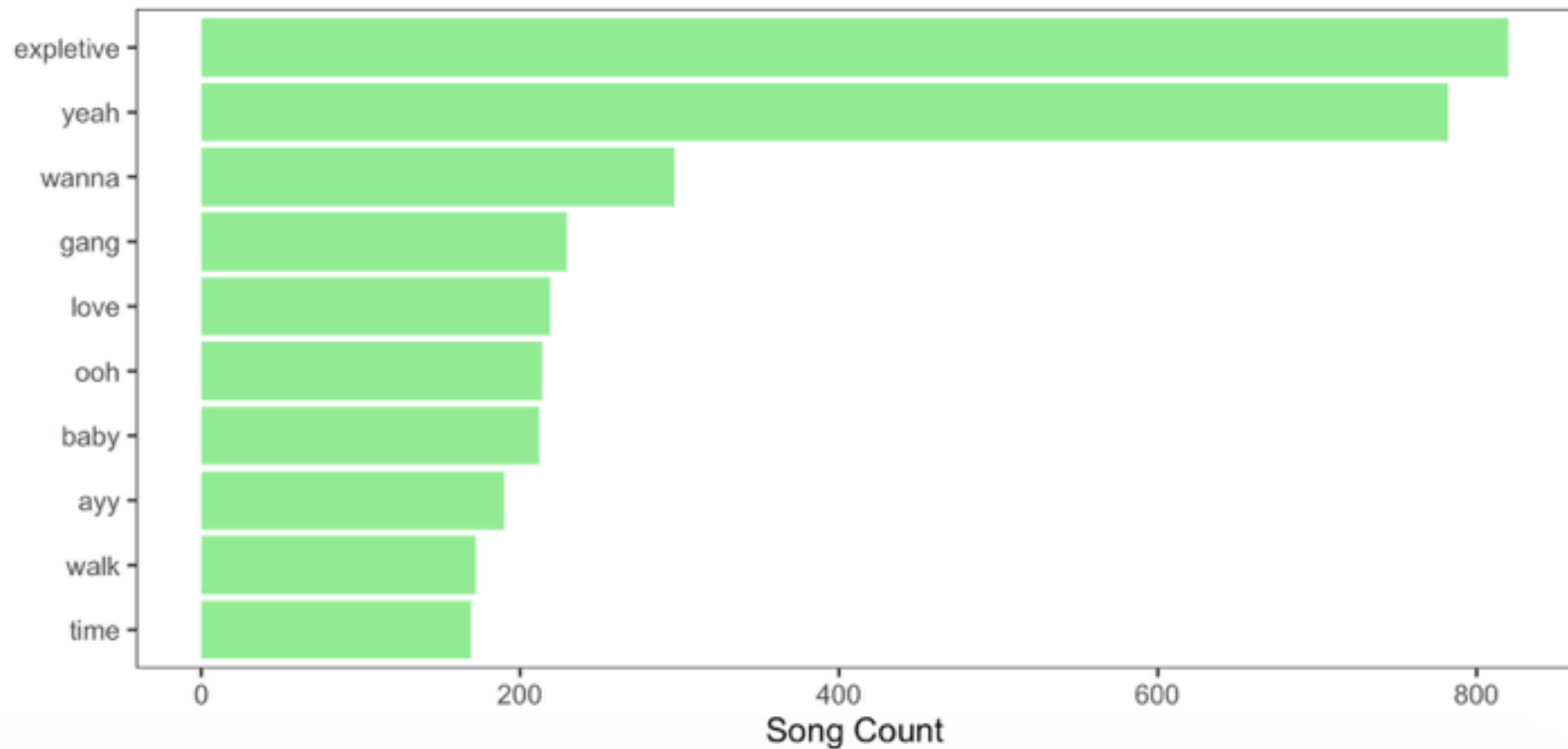
Goal: to use TF-IDF to *find the important words* for the content of each document by decreasing the weight for commonly used words and increasing the weight for words that are not used very much in a collection or corpus of documents

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

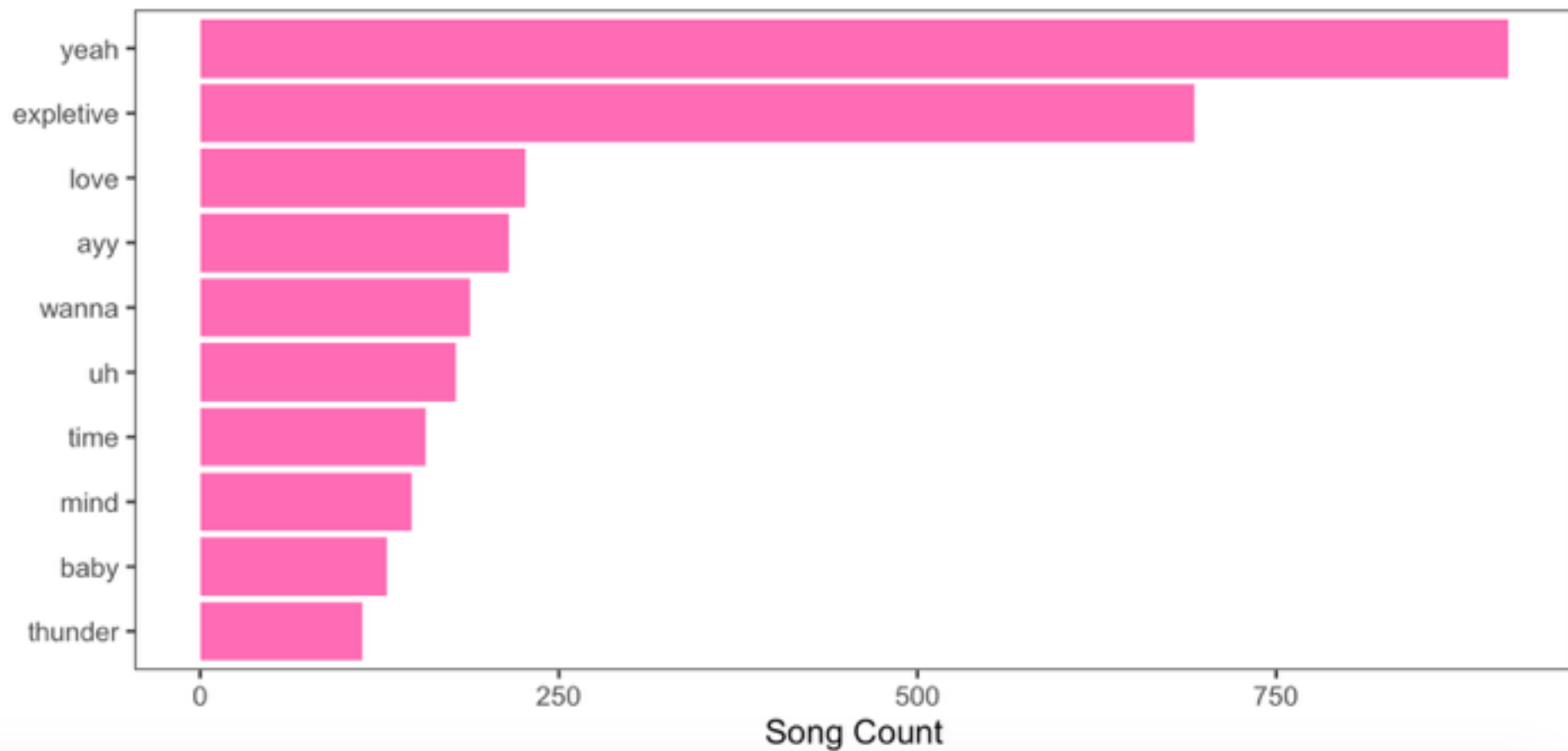
Most Frequently Used Words in top 200 songs (2017)



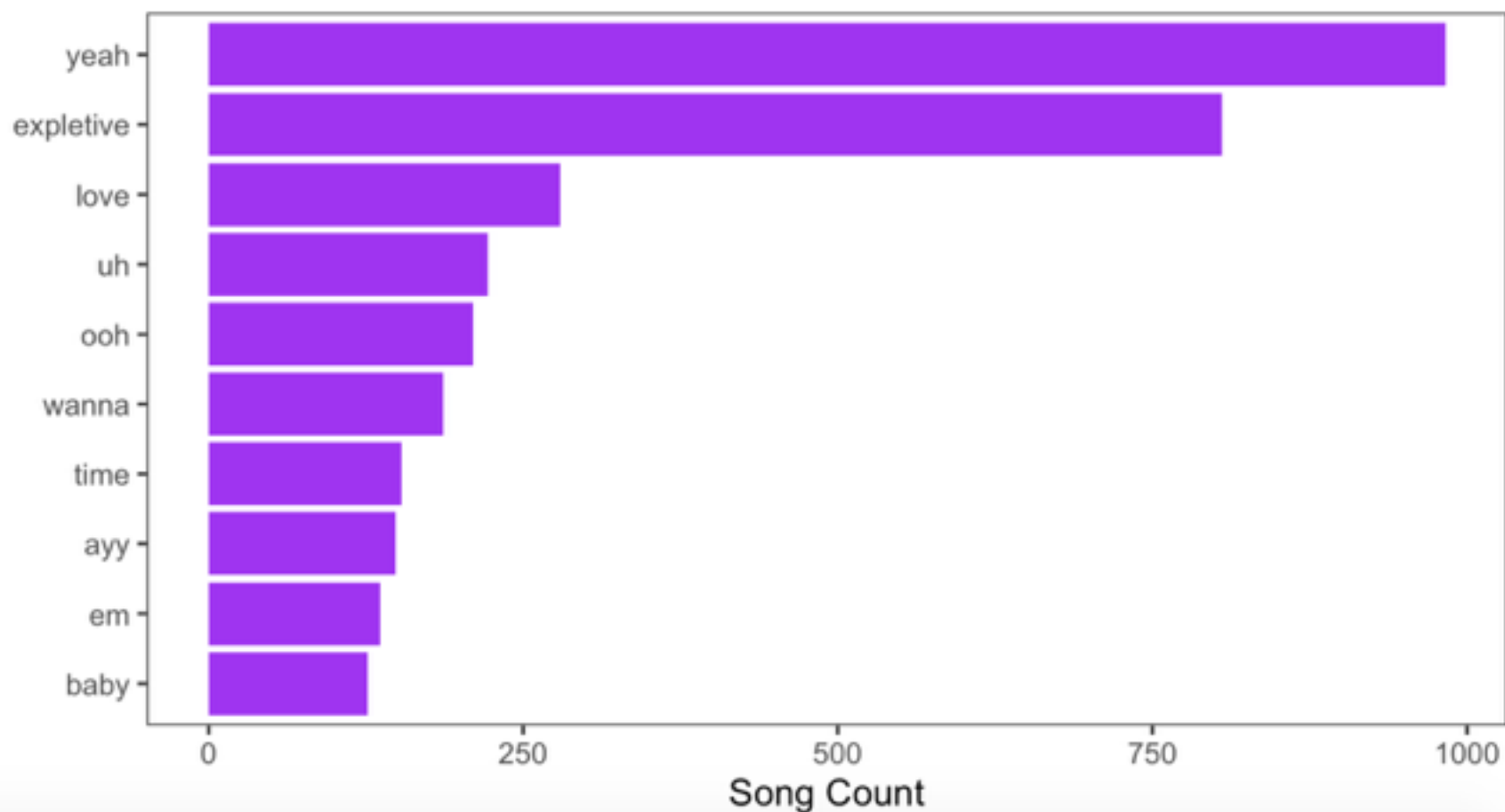
Most Frequently Used Words in top 200 songs (2018)



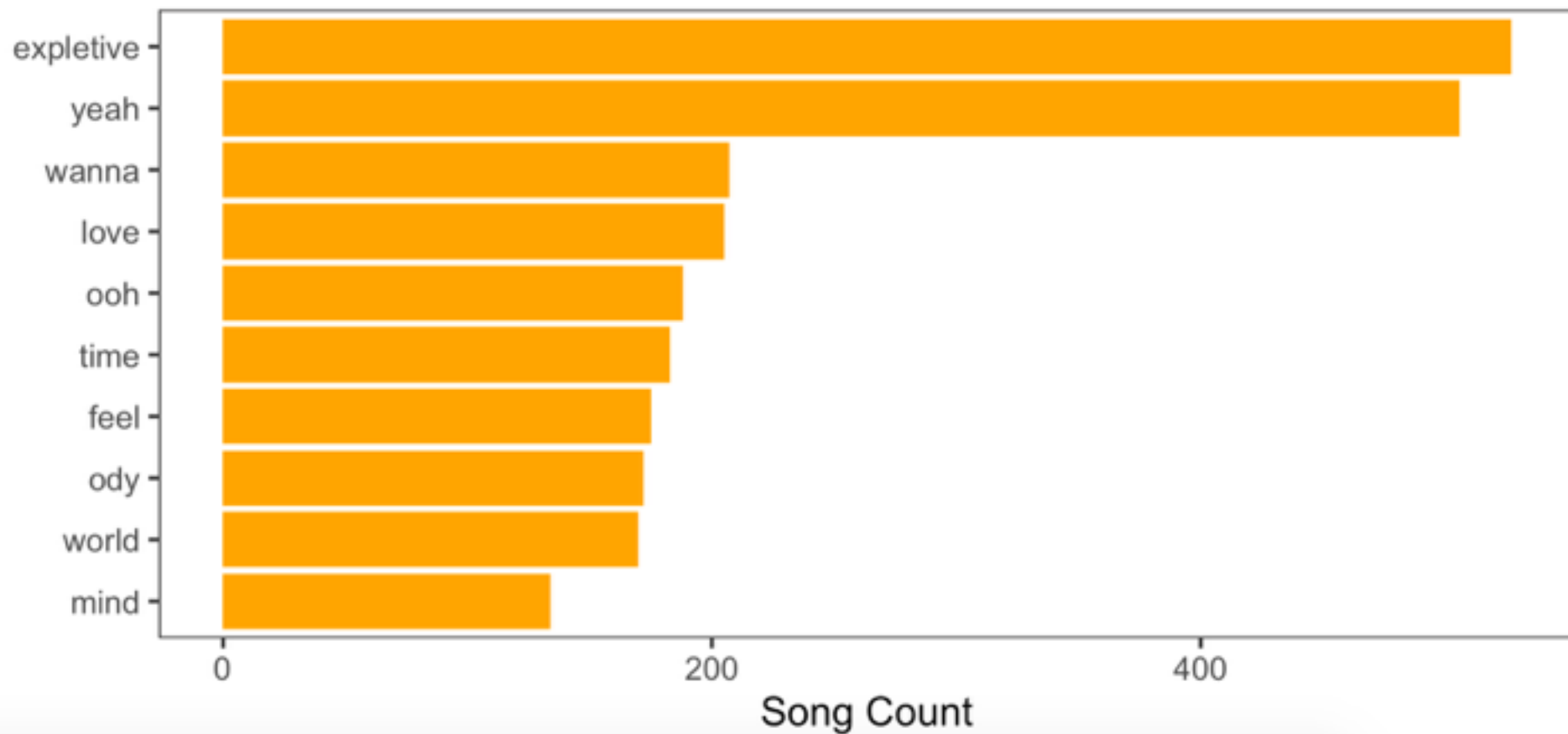
Most Frequently Used Words in top 200 songs (2019)



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

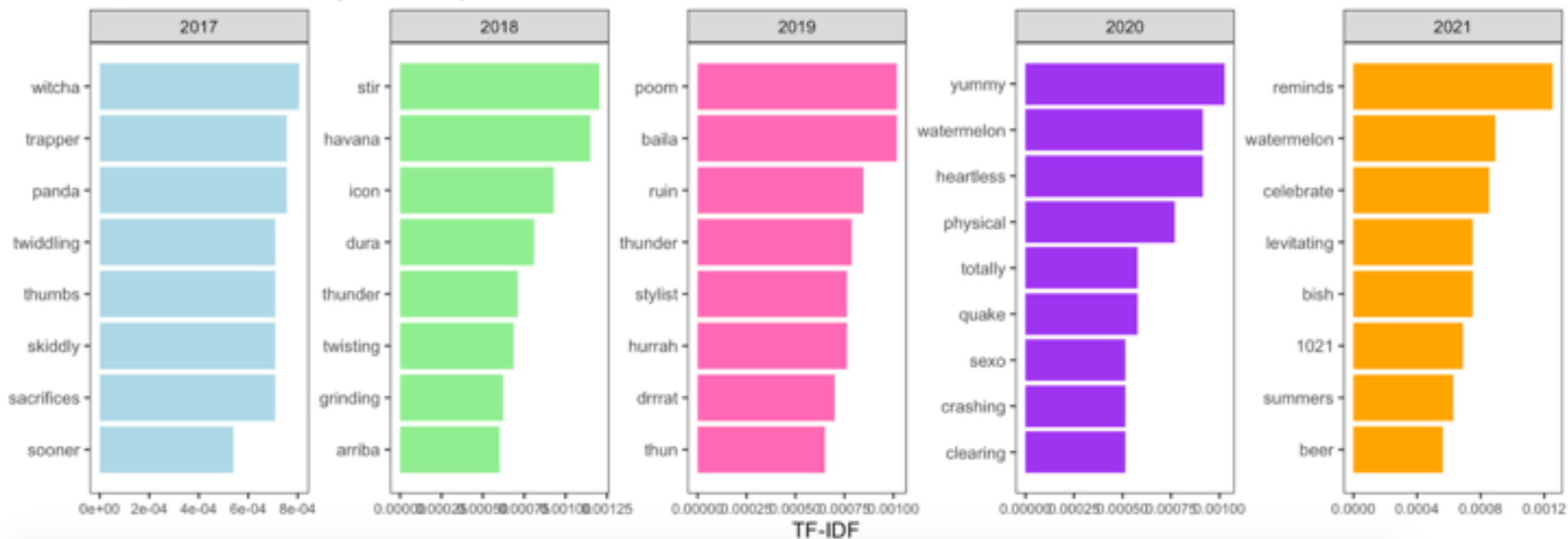
Term x within document y

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

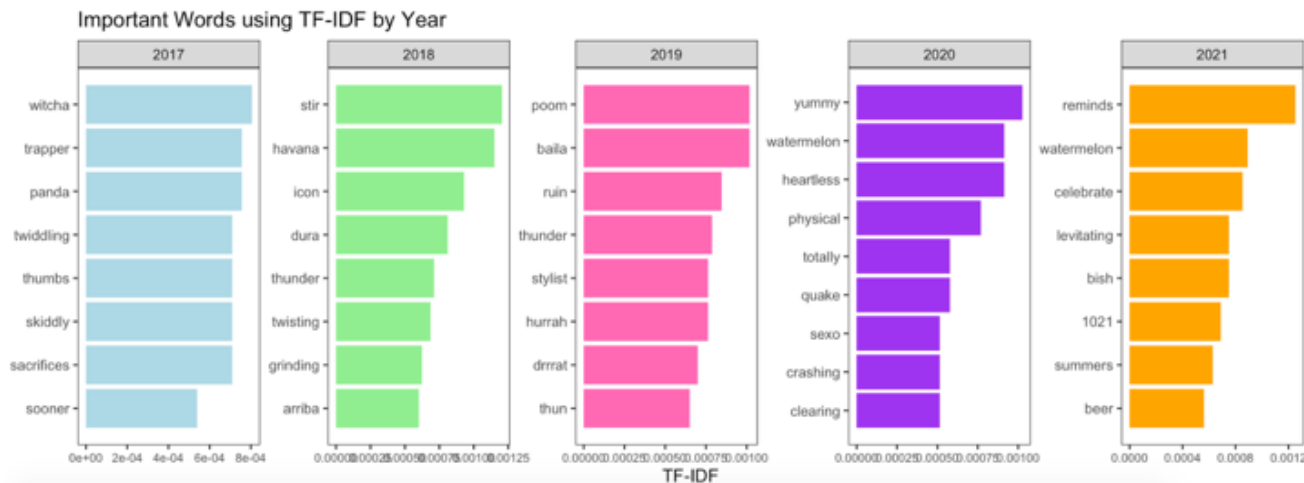
df_x = number of documents containing x

N = total number of documents

Important Words using TF-IDF by Year



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