

Text Analysis

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



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

Lectures : http://casimpkinsjr.radiantdolphinpress.com/pages/cogs108_ss1_23/

Plan for today

- Announcements
- Review of last time
- Projects
 - Prev. Project review feedback in process
 - Proposal feedback - in process, will appear as issues on github
 - Next checkpoint Friday (“Data”)
- Github invites?
- Upcoming deadlines...

Plan for today II

- Upcoming deadlines
 - Friday @ 11:59pm
 - Q2 (recommend - by tonight)
 - D3 (recommend - tonight)
 - D4 (recommend - by Wed)
 - A2 (recommend - by Wed)
 - CP1: Data (Friday midnight)
- Inference lecture
- Workshop on Inference (D5)

Announcements

- Any remaining github invite issues?
- If you did not have access and emailed the file, please add to your github so we can release feedback on your github

Text Analysis

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?

co-occurring words = topics

i.e. “female fashion” = [“gown”, “silk”, “dress”, “lace”, and “ribbons”]

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

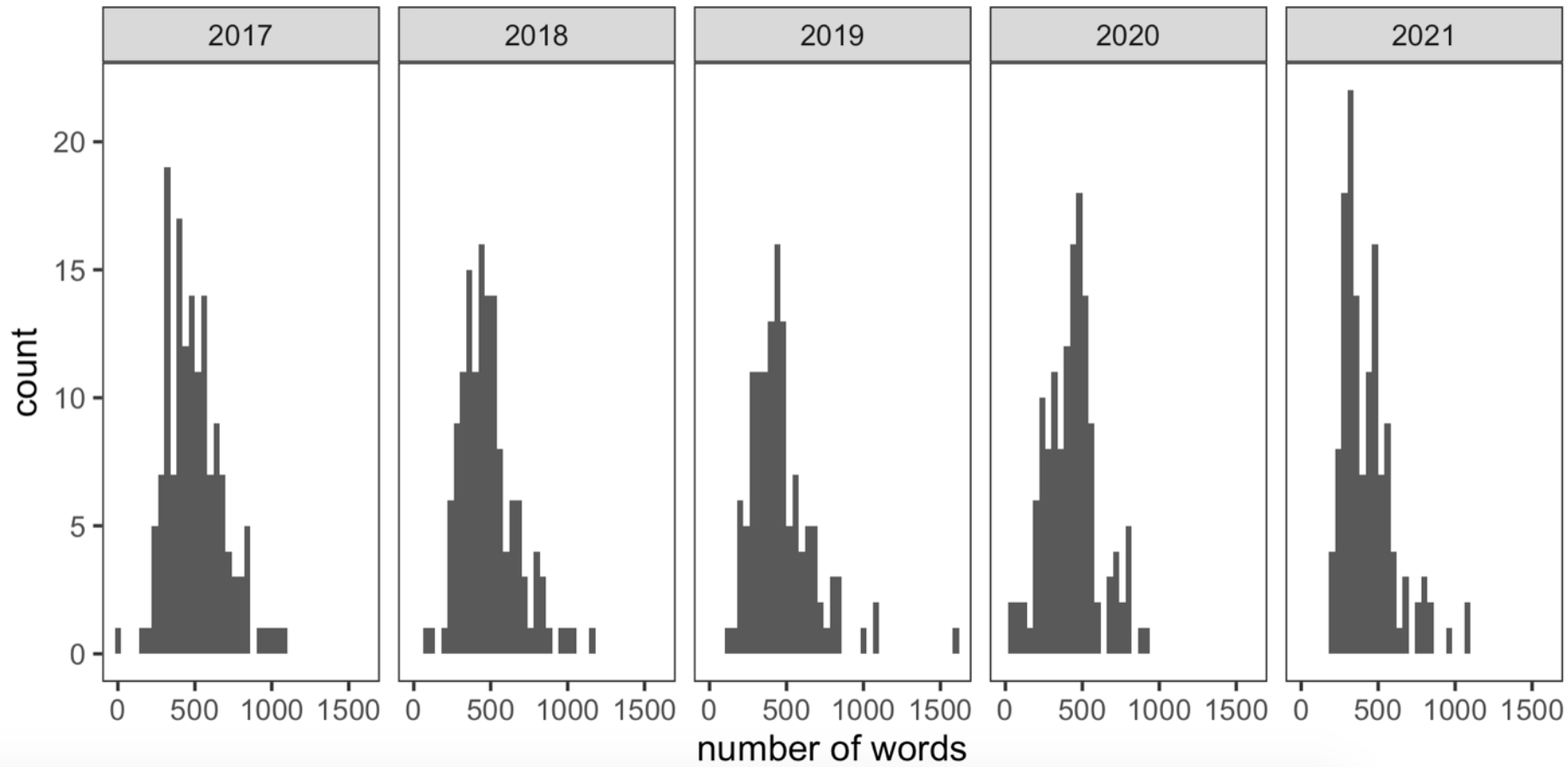
2021: 134 songs

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

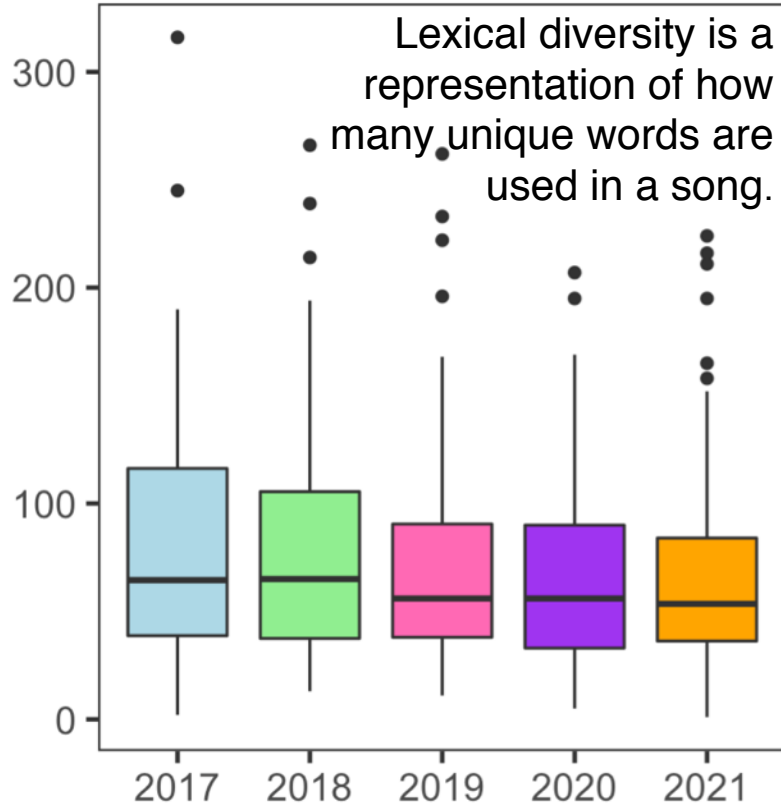


Questions we can ask...

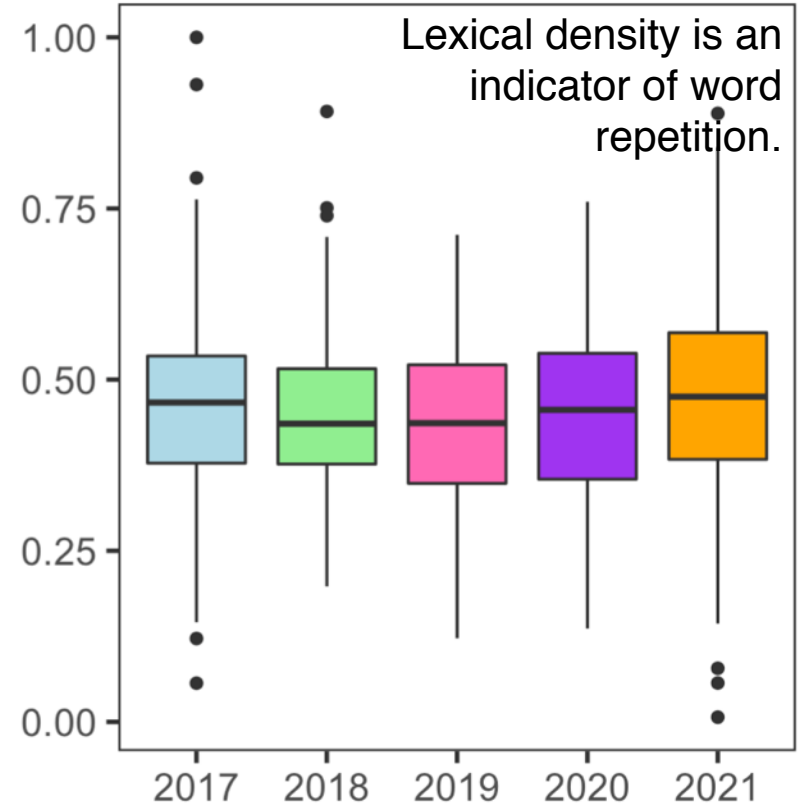
1. Does the total number of words change over time?
2. Does uniqueness change over time?
3. Does the diversity or density change?
4. What words are most common?
5. What words are most unique to each year?
6. What sentiment do songs convey most frequently?
7. Has sentiment changed over time?
8. What are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

- Process of analyzing digital or digitized text in order to determine if the emotional tone is positive, negative or neutral
- Large volumes of text data are now available in forms of
 - Emails
 - Messages
 - Support transcripts
 - Social Media interactions
 - Reviews
 - Digitized phone messages and interaction records (i.e. UCSD)

Sentiment Analysis - goal

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
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	

NRC Sentiment Lexicon

- “The NRC Emotion Lexicon is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). The annotations were manually done by crowdsourcing.”
- <https://saifmohammad.com/WebDocs/Lexicons/NRC-Emotion-Lexicon.zip>
- <https://onlinelibrary.wiley.com/doi/10.1111/j.1467-8640.2012.00460.x>

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???

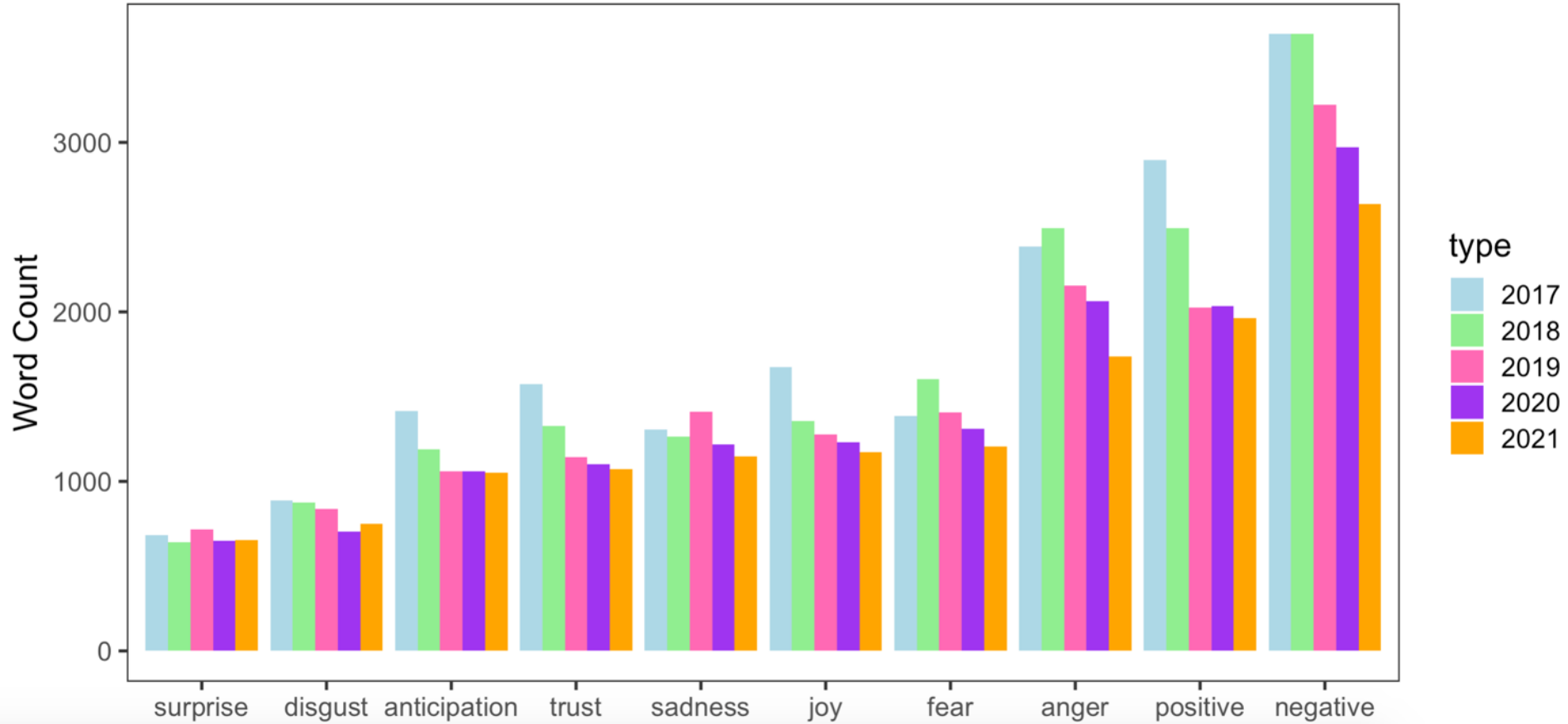
Issues to handle

- Words that are ambiguous
- Jumper can be clothing, a person who jumps on a trampoline, etc.
- Want to figure out how being handled in lexicon and how being handled in your text analysis

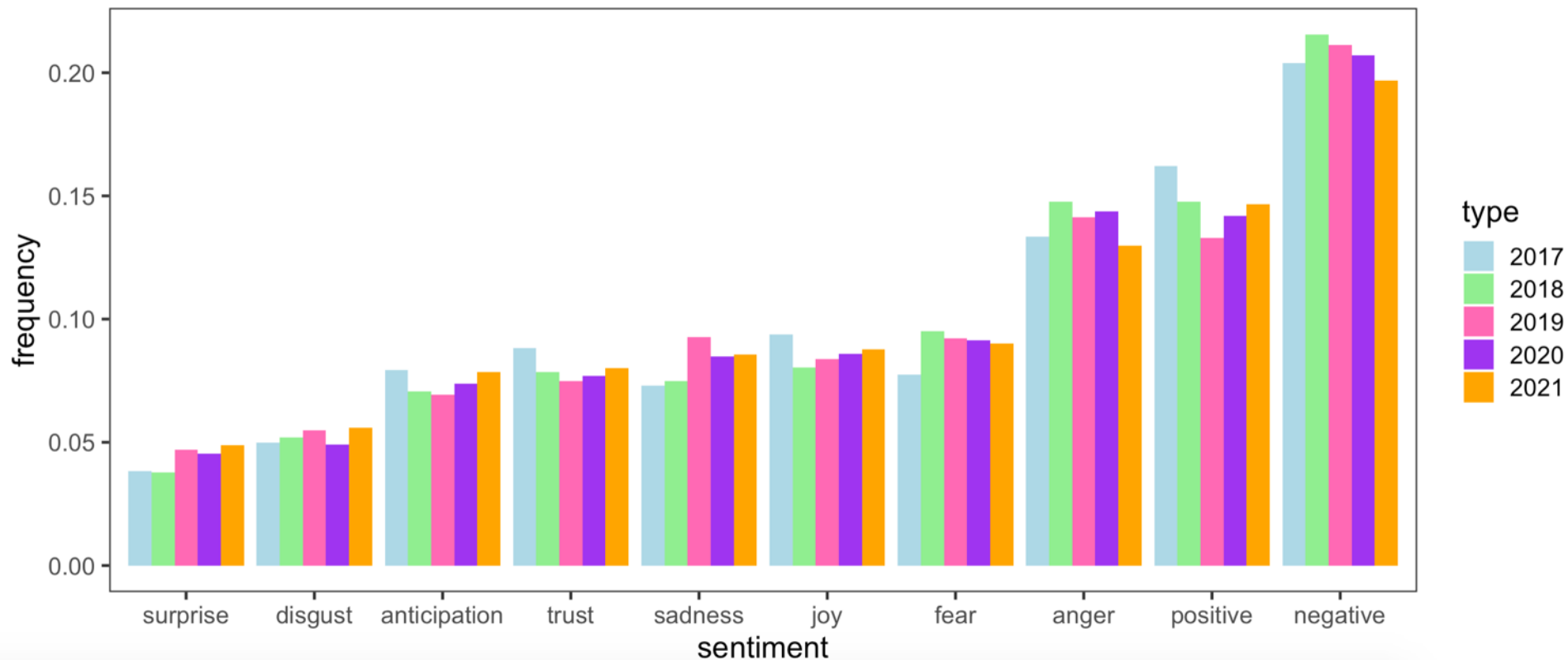
In text analysis, your choices matter:

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

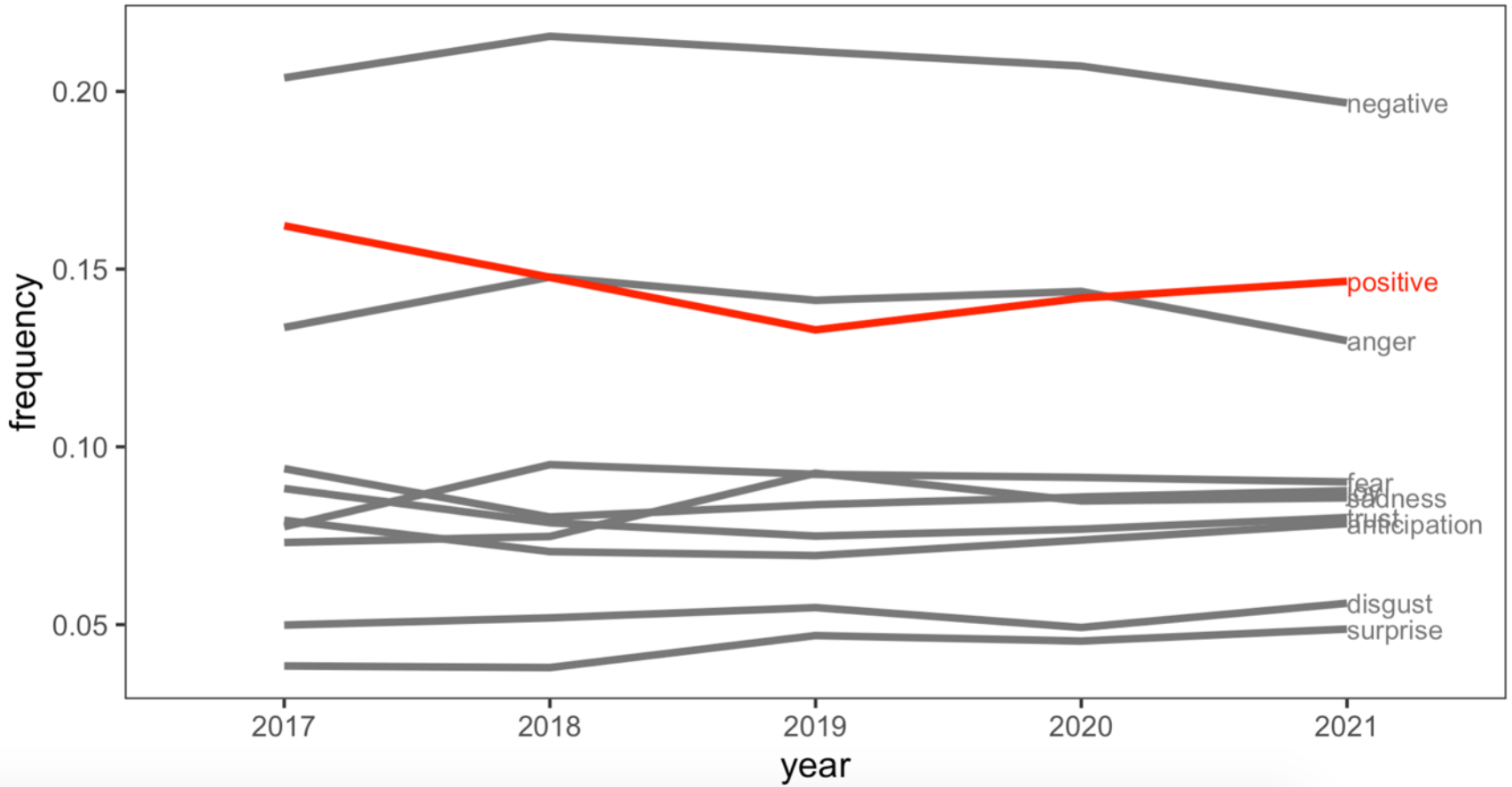
Top Songs Sentiment



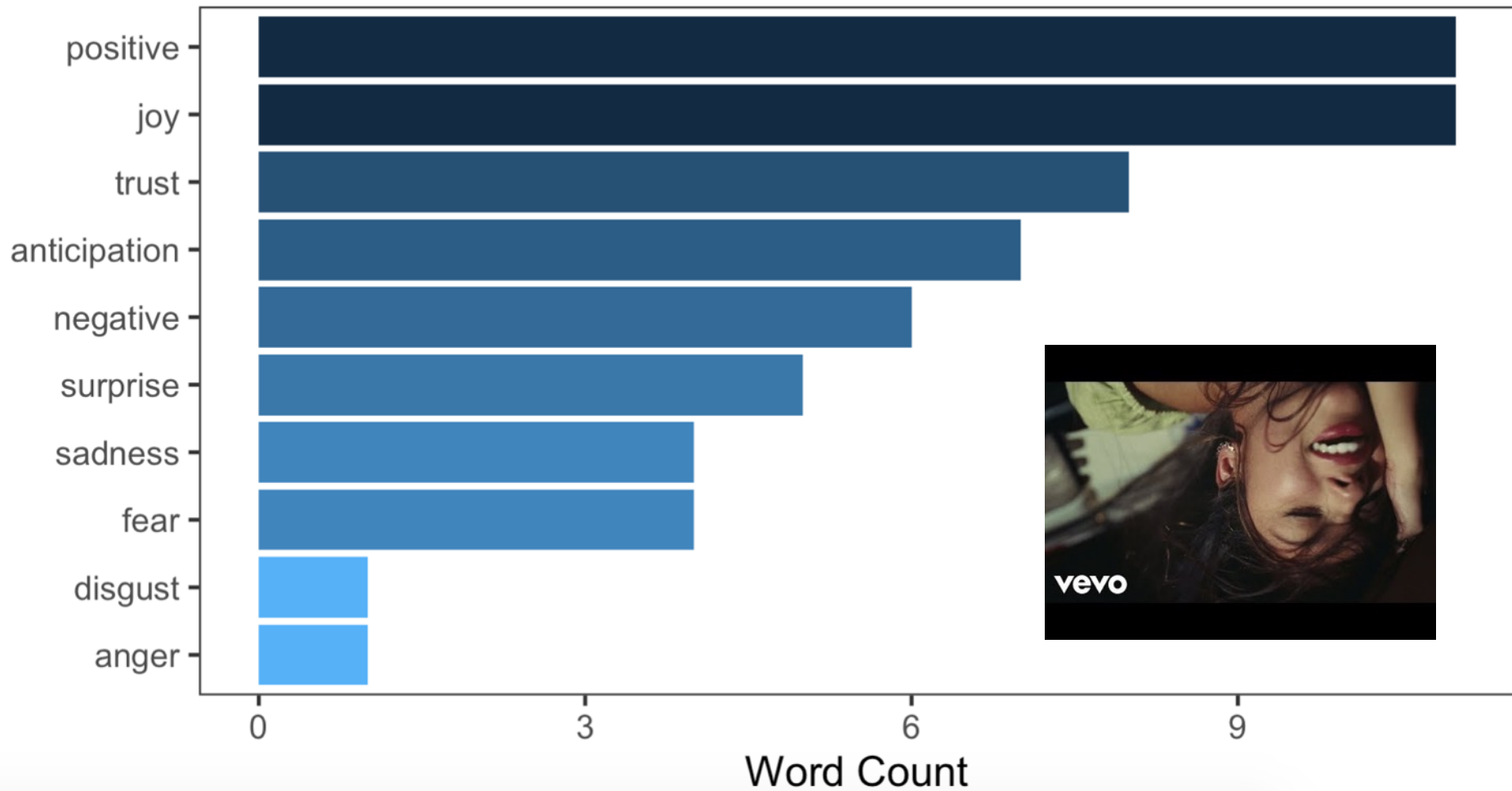
Sentiment by Year



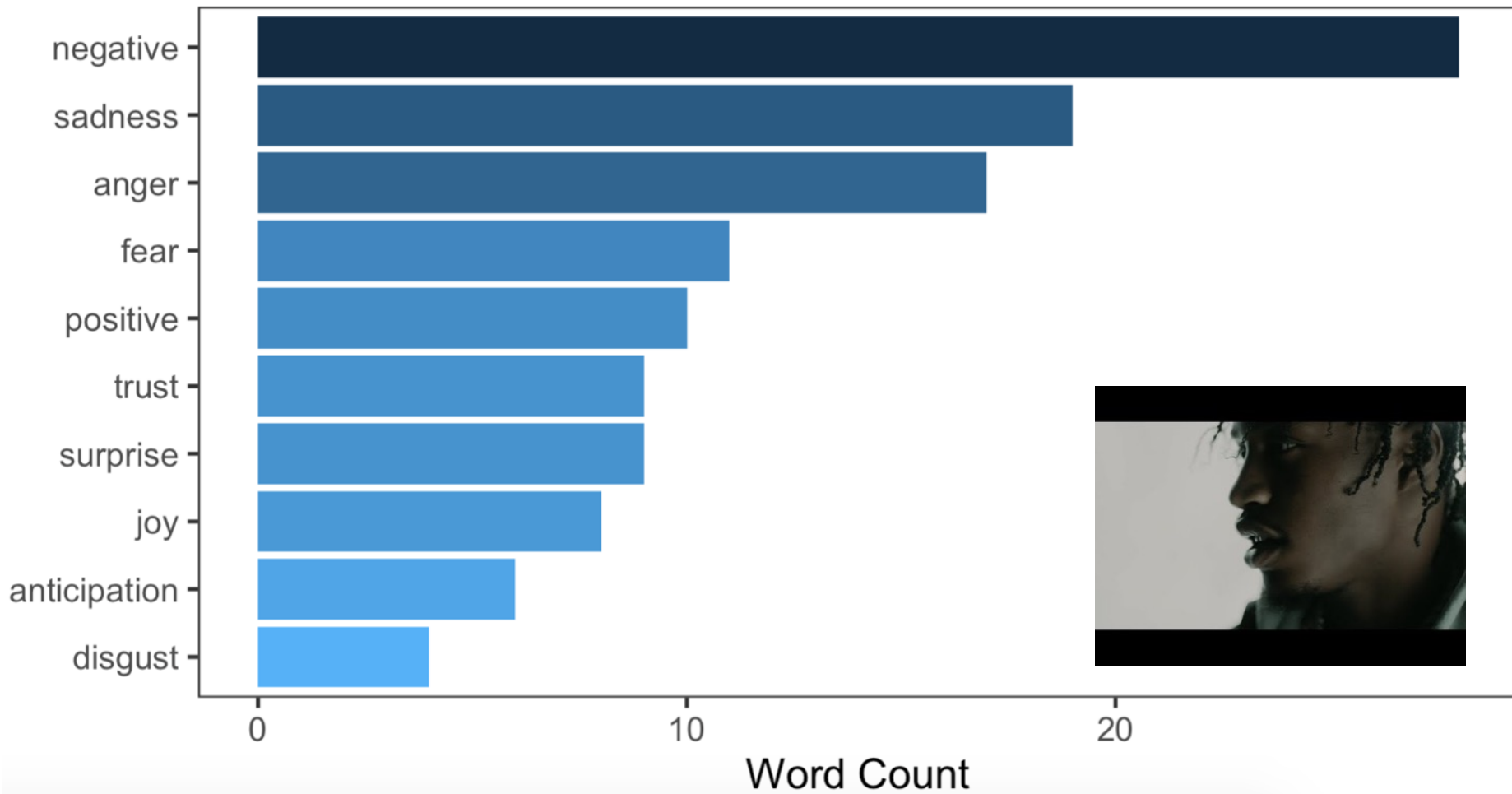
Change in Sentiment over Time



Sentiment: Driver's License



Sentiment: Calling My Name

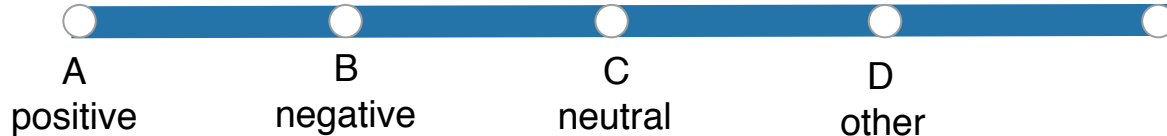




Sentiment Limitations

How would you classify the sentiment of the following sentence?


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





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

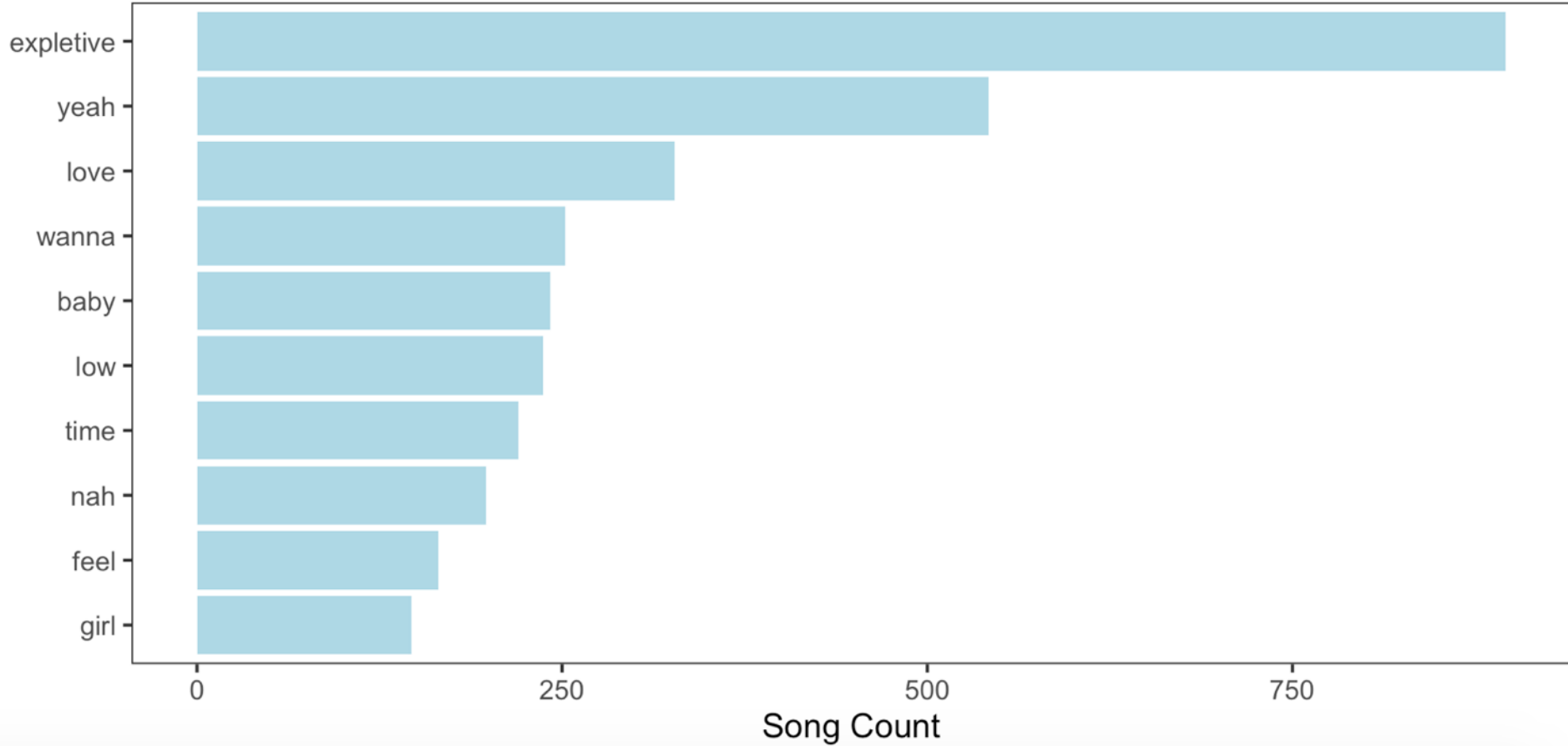
Term Frequency - Inverse Document Frequency

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

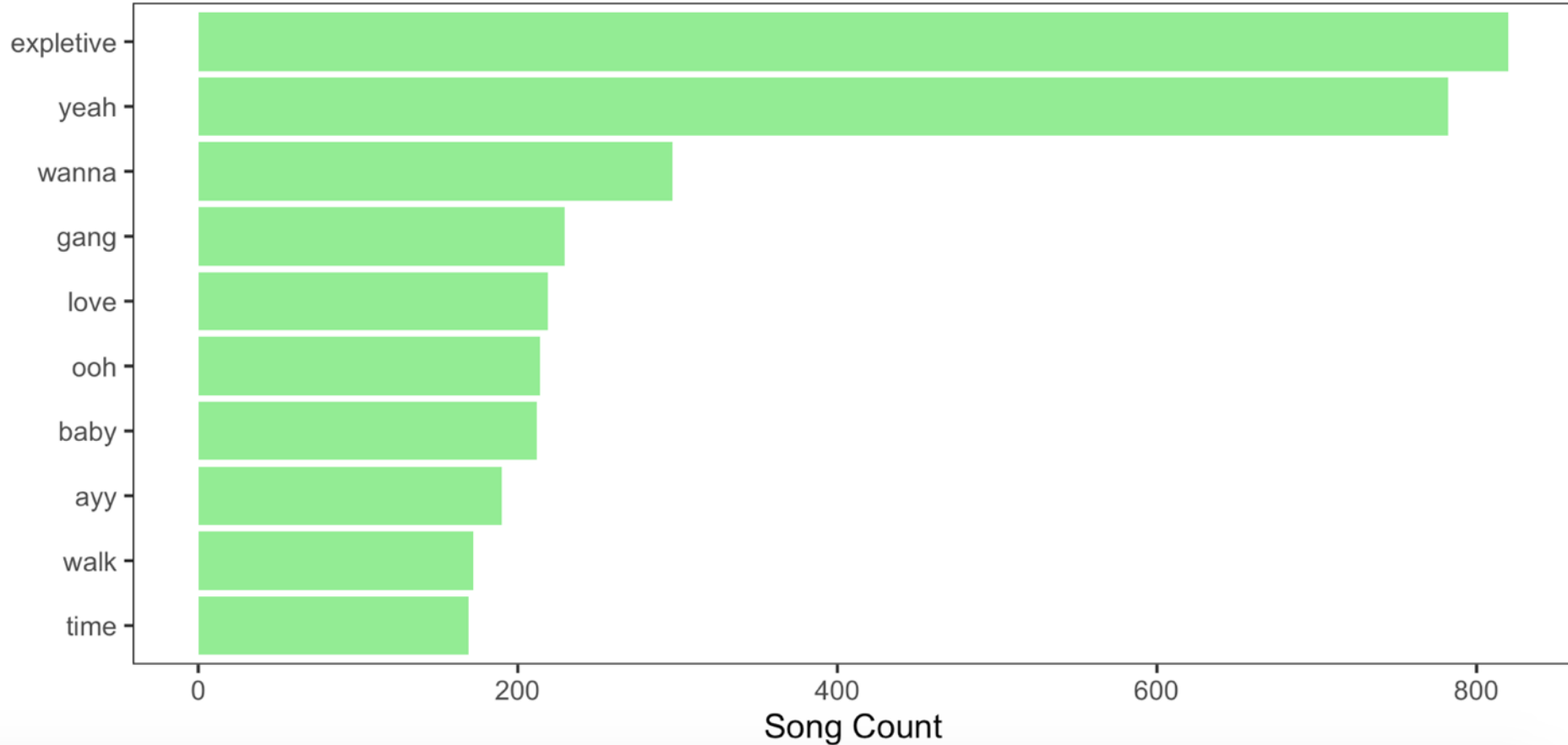
Goal: to use TF-IDF to *find the important words* for the content of each document by decreasing the weight for commonly used words and increasing the weight for words that are not used very much in a collection or corpus of documents

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

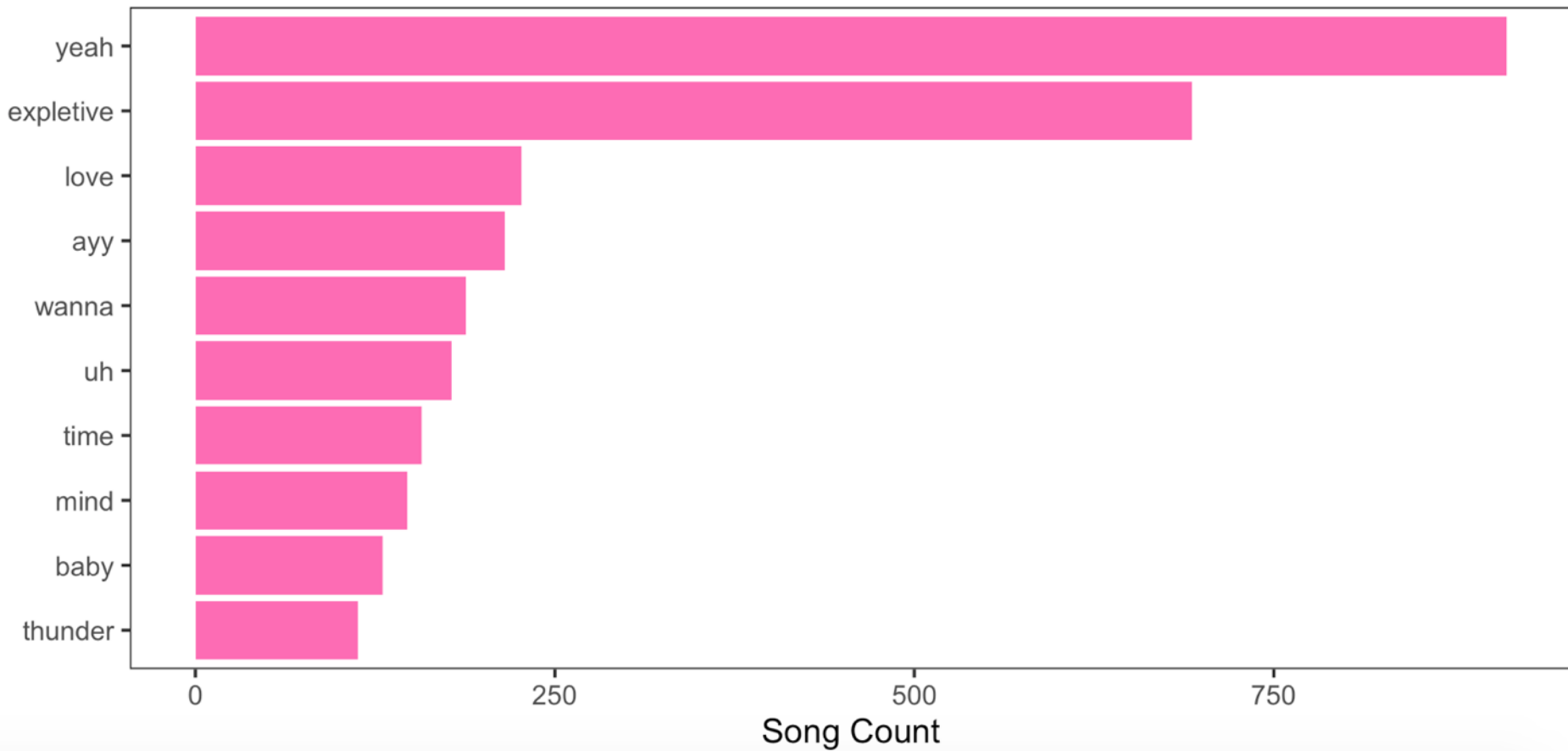
Most Frequently Used Words in top 200 songs (2017)



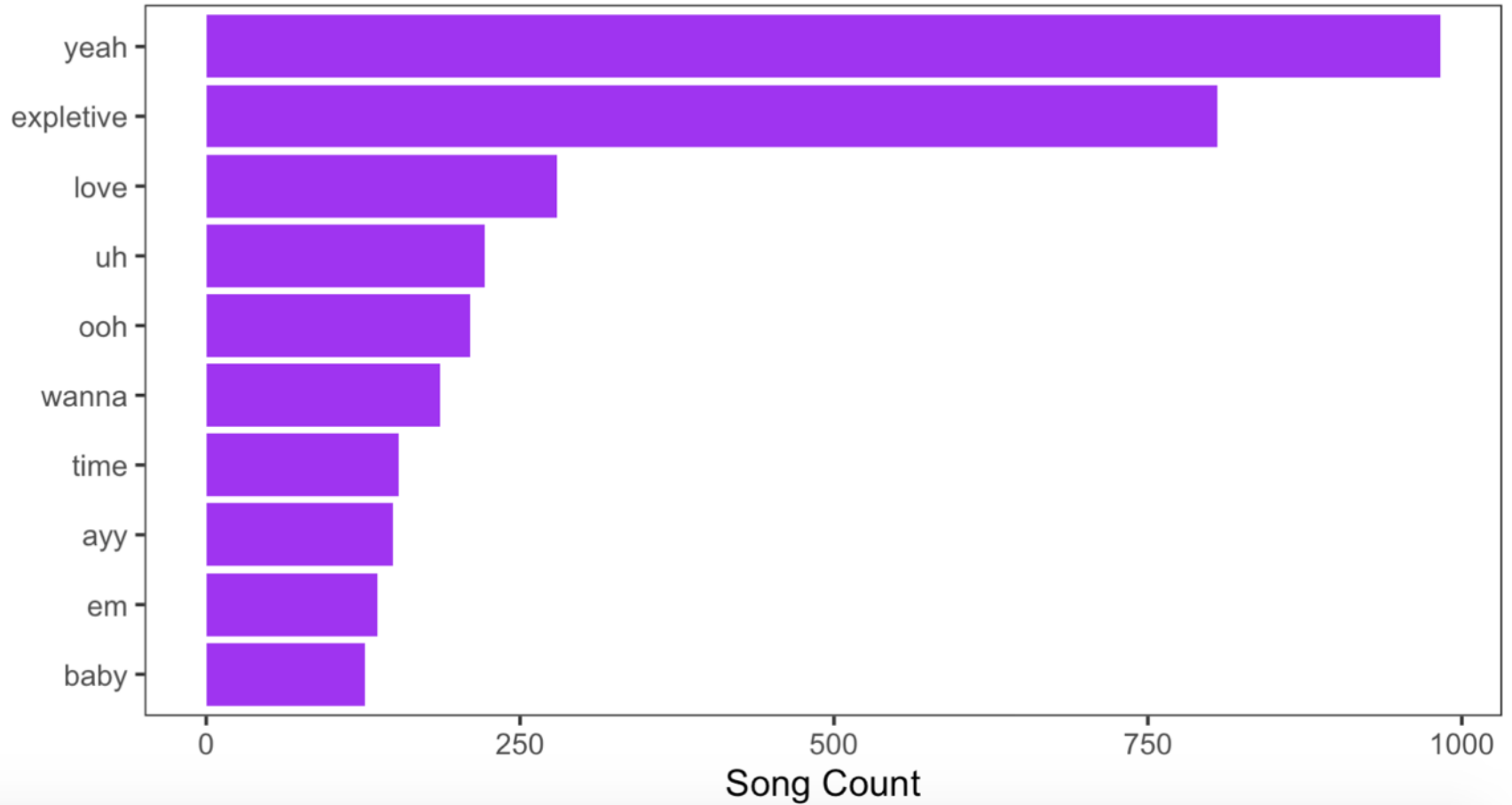
Most Frequently Used Words in top 200 songs (2018)



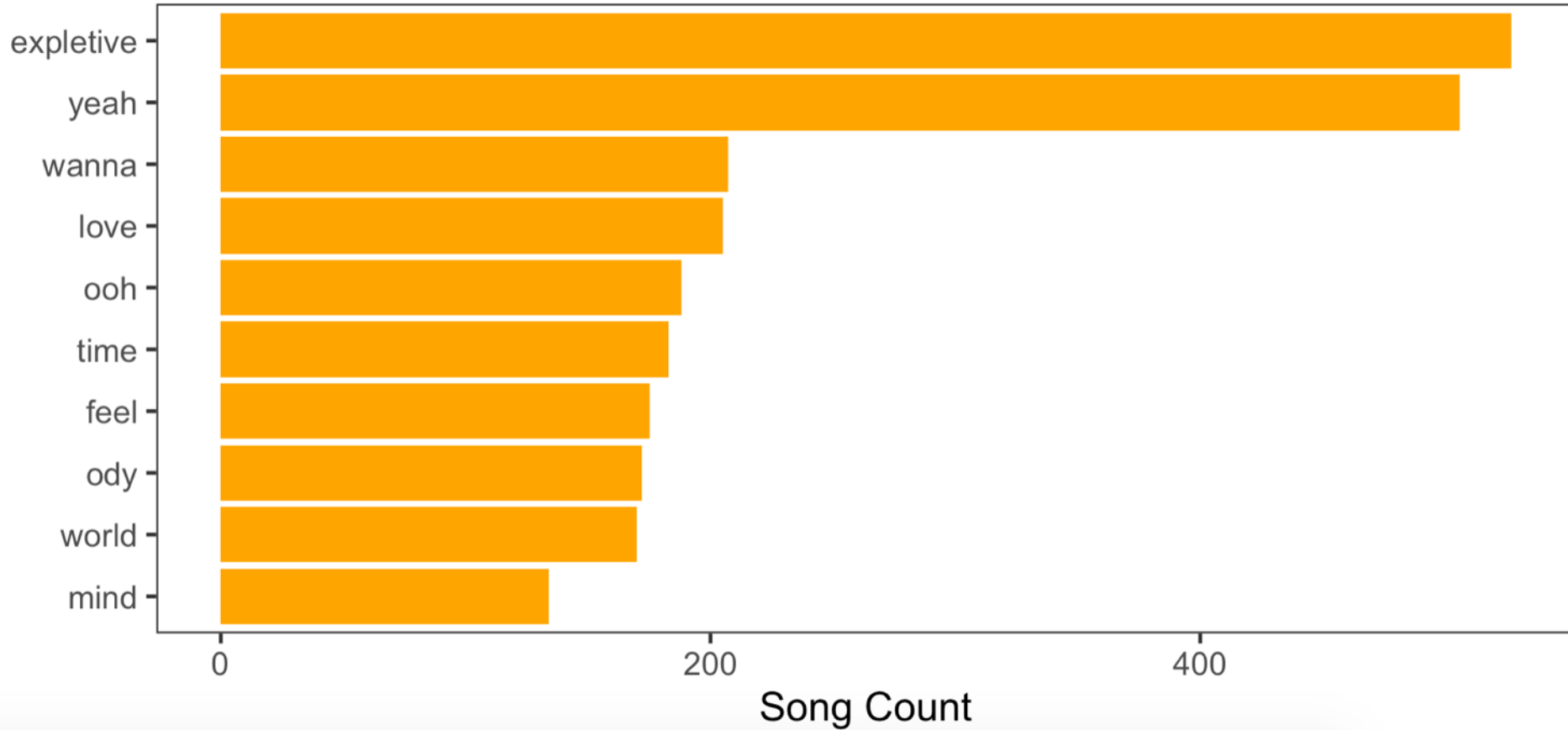
Most Frequently Used Words in top 200 songs (2019)



Most Frequently Used Words in top 200 songs (2020)



Most Frequently Used Words in top 200 songs (2021)



TF-IDF:

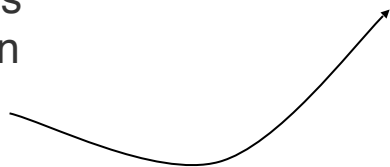
Term Frequency - Inverse Document Frequency

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

$$w_{x,y} = \text{tf}_{x,y} \times \log \left(\frac{N}{\text{df}_x} \right)$$

TF-IDF

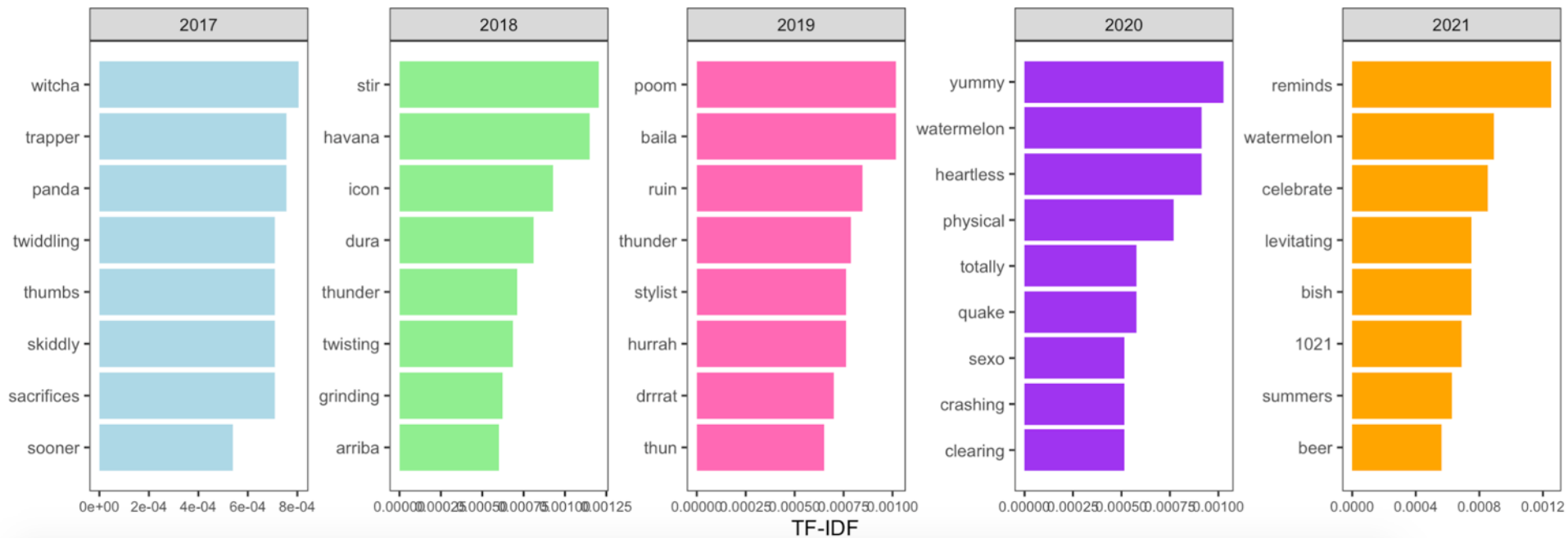
Term x within document y

$\text{tf}_{x,y}$ = frequency of x in y

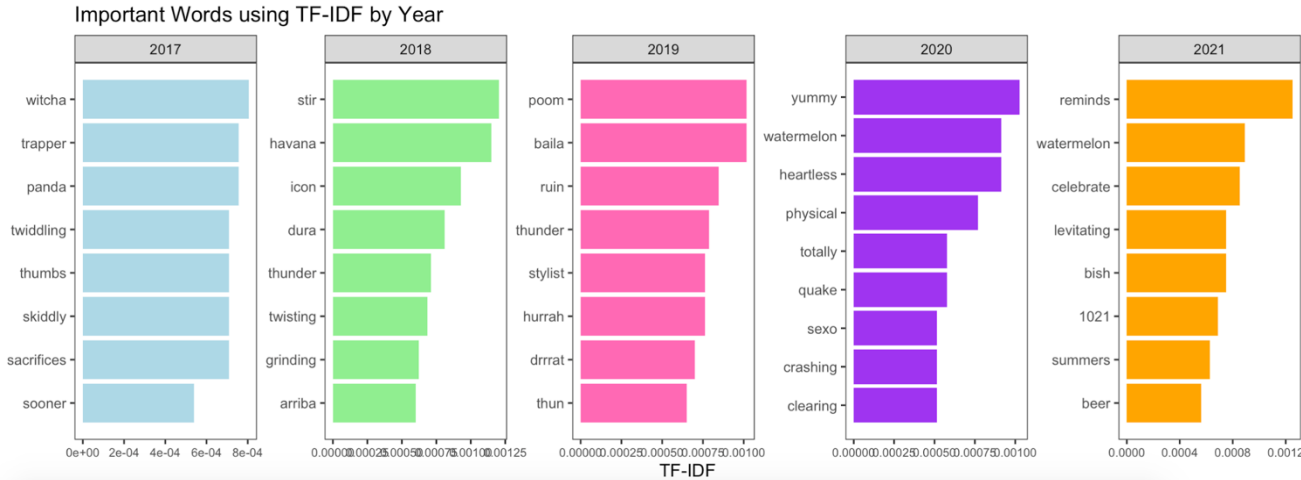
df_x = number of documents containing x

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

nltk

- Natural Language Toolkit
- Easy to use interfaces to various text resources
- free open source
- <https://www.nltk.org>

nlTK - a bit about it

- Interfaces to over 50 corpora and lexical resources such as WordNet
- Libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning,
- Wrappers for industrial-strength NLP libraries
- Active discussion forum

nlTK - a bit about it

- Simple installation: <https://www.nltk.org/install.html>
- Demos: <http://text-processing.com/demo/>
 - Sentiment example: <https://www.nltk.org/howto/sentiment.html>
- We are going to be doing some sentiment analysis coming up!

LISC project

- Open source python module “Literature Scanner”
 - <https://github.com/lisc-tools/lisc>
- Donoghue, Thomas. (2019). LISC: A Python Package for Scientific Literature Collection and Analysis. Journal of Open Source Software. 4. 1674. 10.21105/joss.01674.
- https://www.researchgate.net/publication/336082537_LISC_A_Python_Package_for_Scientific_Literature_Collection_and_Analysis
- LISC is based on BRAIN-SCANR by Voytek (2012)

LISC- Automated methods for digesting vast information

0

- Scientific literature is vast, expanding and beyond a single researcher's ability to digest completely
- By the time an article is read, more are published
- >30M published articles as of 2019 in biomedical sciences alone!
- Automated methods for curation and digestion of literature has been explored to enhance a researcher's abilities to absorb information
- “Knowledge discovery, literature-based discovery, hypothesis generation”

LISC- Automated methods for digesting vast information

0

- Easily accessible
- Connects to several external resources through APIs
- e.g. PubMed, OpenCitations database
- Supports utilities to analyze collected data

LISC- types of data collection

0

- **Counts:** tools to collect and analyze data on the co-occurrence of specified search terms
- **Words:** tools to collect and analyze text and meta-data from scientific articles
- **Citations:** tools to collect and analyze citation and reference data

LISC- includes for supporting use cases

0

- URL management and requesting for interacting with integrated APIs
- Custom data objects for managing collected data
- A database structure, as well as save and load utilities for storing collected data
- Functions and utilities to analyze collected data
- Data visualization for plotting collected data and analysis outputs

LISC vs. Moliere

0

- LISC takes a lightweight, fast and efficient approach to hypothesis generation
- A complement for other tools like Moliere or Meta (www.meta.org)
- More customizable (LISC), tools included for efficient analysis on the results
- Connective interface to Natural Language Processing (NLP) tools such as NLTK
- Moliere/Meta better for more complex analyses

Caveats

- Take care using automated systems since they don't "understand" the literature as a human does
- Programming biases are inevitable
 - Chatbot knowledge biases
 - Programmer biases
- Statistics can be biased
- Use with a grain of salt - it's a tool

• ***“The hammer does not make the building” [Simpkins 2023]***