Course Reminders

Important Dates:

- A3 due Sunday (11:59 PM)

Notes:

- A2 grades released & feedback provided
  - Answer to 2d regraded after initial grade release
- Mid-course survey EC added on TritonEd
  - We'll use the responses to this in class on Friday

A2: Question 2d

Question 2d: Suppose that missing incomes did not occur at random, and that individuals with incomes below $10000 a year are less likely to report their incomes. If so, one of the statements is true. Record your choice in the variable q2d_answer.

1. df['income'].mean() will likely output a value that is smaller than the population's average income.
2. **df['income'].mean() will likely output a value that is larger than the population's average income.**
3. df['income'].mean() will likely output a value that is the same as the population's average income
4. df['income'].mean() will raise an error.

This was initially marked incorrectly (totally my fault).
Your TAs regraded and 0.25 pt returned to all with correct answer (2).
My apologies!

New rule: Courtesy Rule

For the rest of the quarter, if I don't have to:

1. Stop mid-lecture for people talking to one another and distracting myself and your classmates.
    a. whispers happen. If I'm distracted, your classmates are *certainly* distracted.
2. Ask you to not pack up your things as I'm still speaking.
    a. A few people having to skip out early happens.
    b. *Many people* starting to pack up should not.

I'll add 1% extra credit to everyone's grade.

This has to happen across both lectures.

Credit to Mary Boyle for this idea.

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?

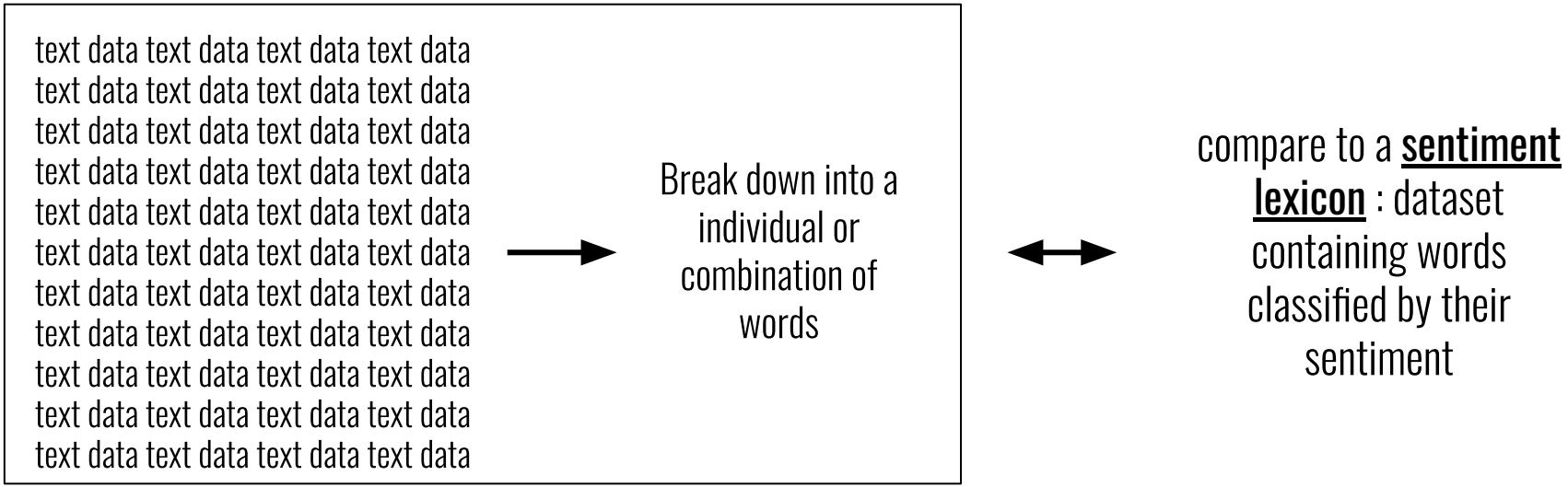
Examples of questions that require text analysis

1. Did J.K. Rowling write *The Cuckoo's Calling* under the pen name Robert Galbraith?
   - distribution of word lengths
   - 100 most common words in the text
   - distribution of character 4-grams
   - word bigrams
2. What themes are common in 19th century literature?
   - co-occurring words = topics
   - i.e. "female fashion" = ["gown", "silk", "dress", "lace", and "ribbons"]

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???

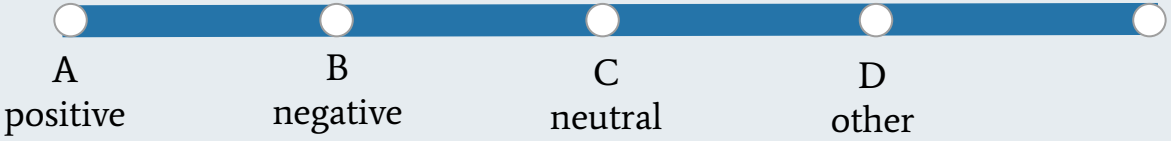
In text analysis, your choices matter:

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

Sentiment Limitations

How would you classify the sentiment of the following sentence?

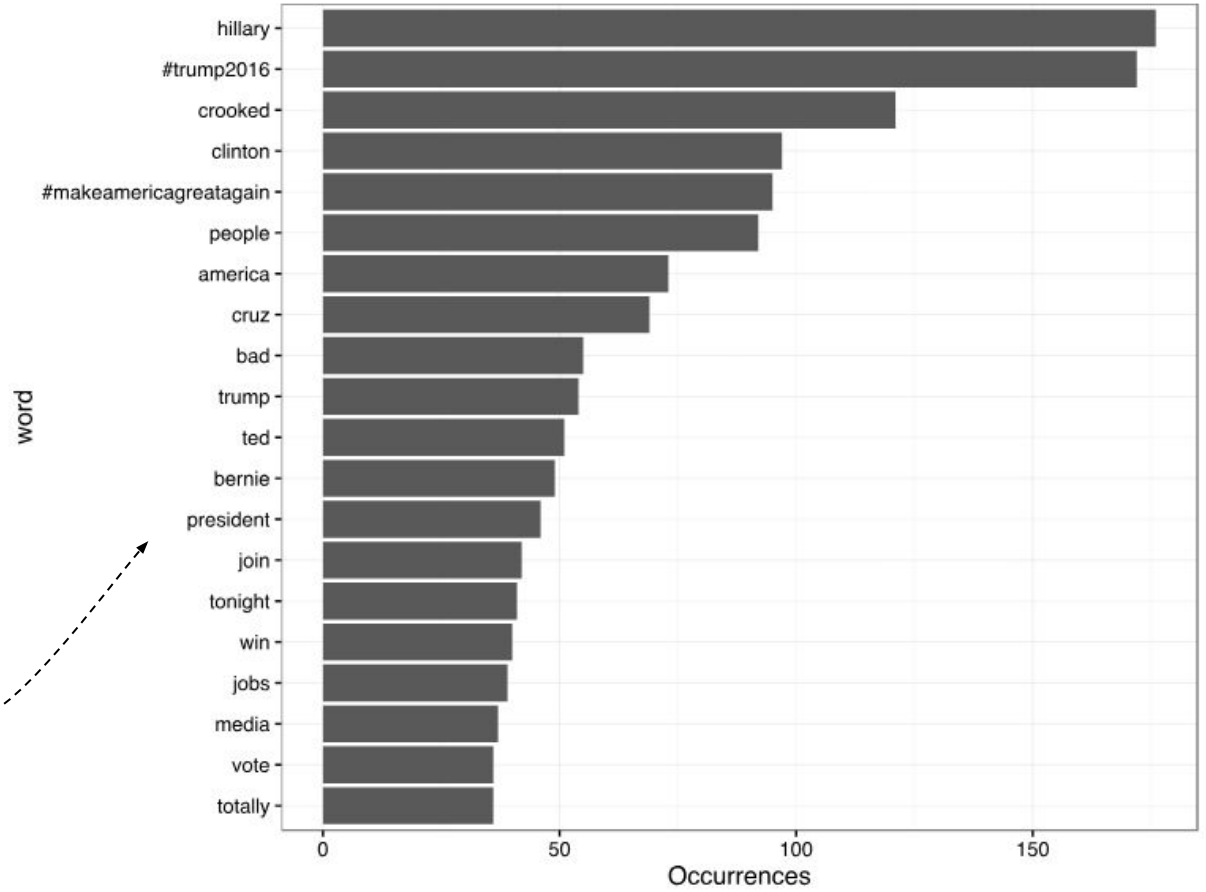
"The idea behind the movie was great, but it could have been better"



Are the angrier and more hyperbolic tweets from Trump himself (rather than his staff)?

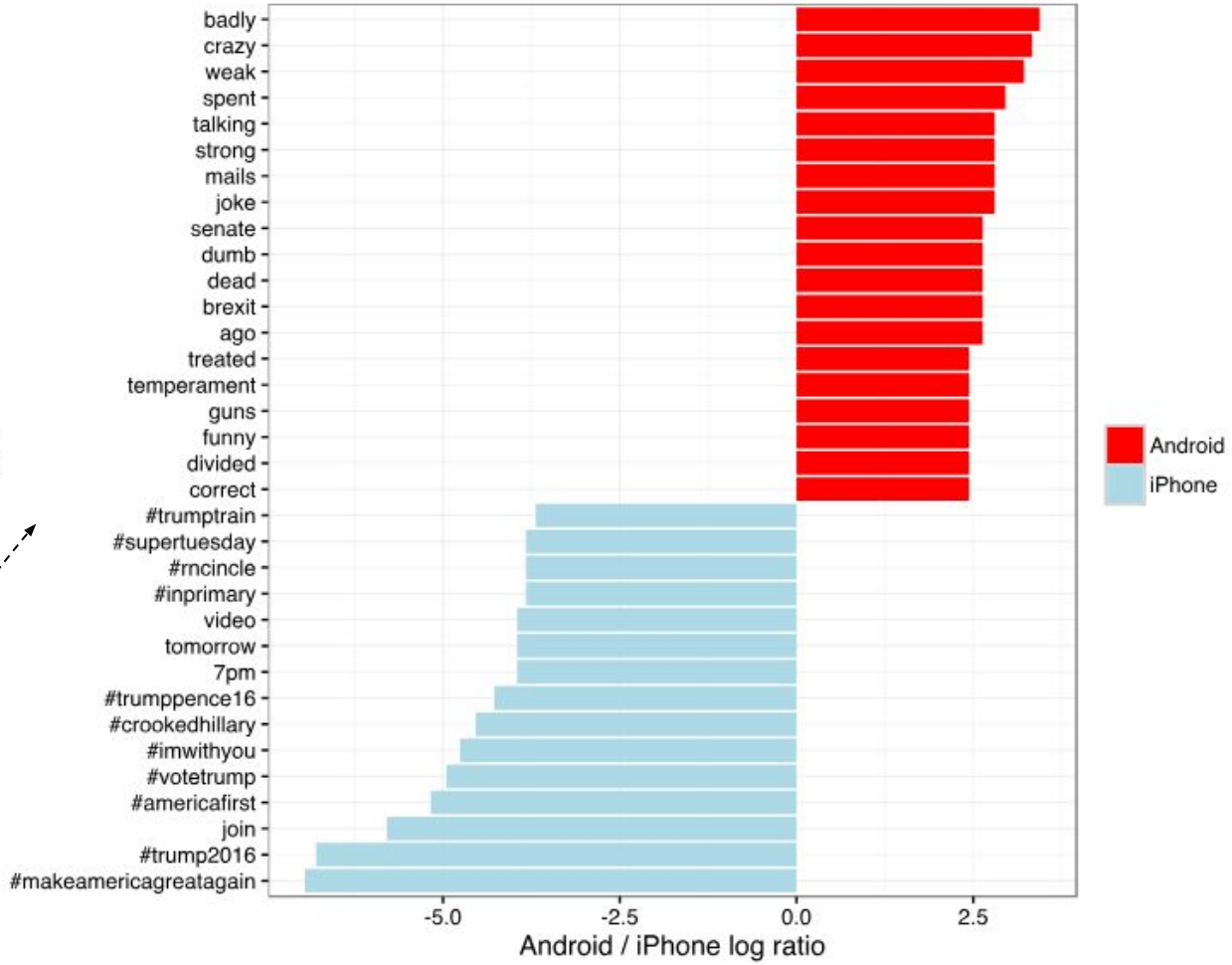
(Note: we knew Trump was using a Samsung Galaxy)

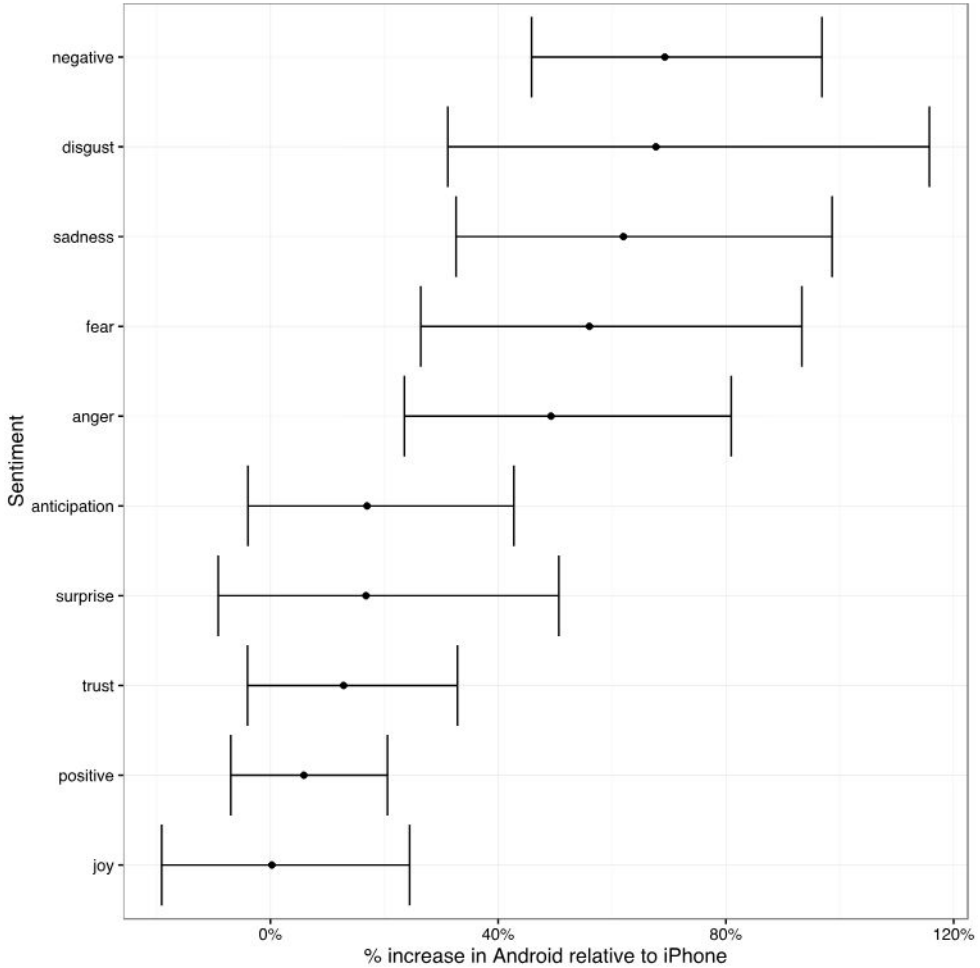
Most common words in Trump's tweets



http://varianceexplained.org/r/trump-tweets/

Frequency broken down by device



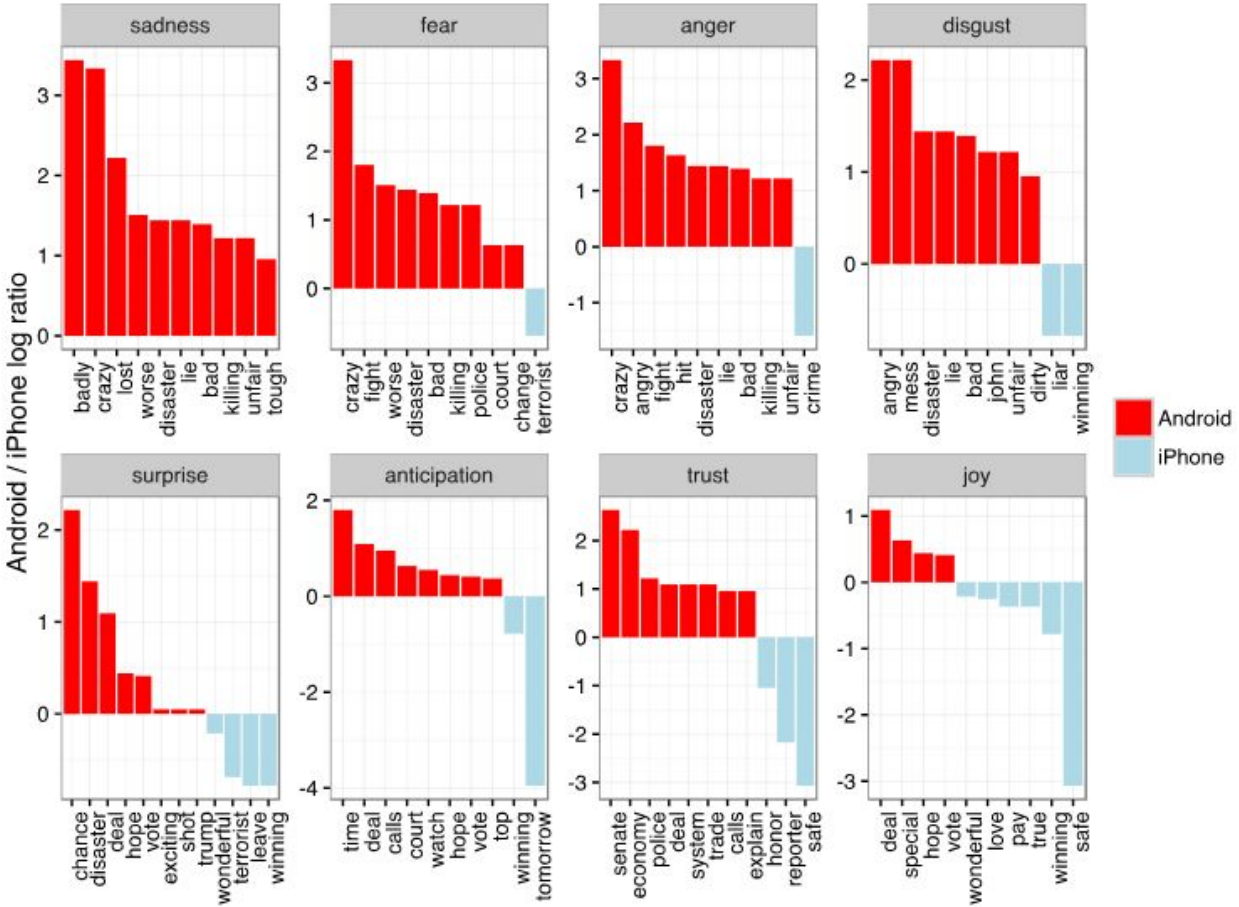


Emotionally charged negative sentiment words more frequently sent from an android

"Trump's Android account uses about 40-80% more words related to disgust, sadness, fear, anger, and other "negative" sentiments than the iPhone account does"

http://varianceexplained.org/r/trump-tweets/

Display of words driving this increase in negative sentiment



http://varianceexplained.org/r/trump-tweets/

Sentiment Limitations

What is a limitation of sentiment analysis?

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

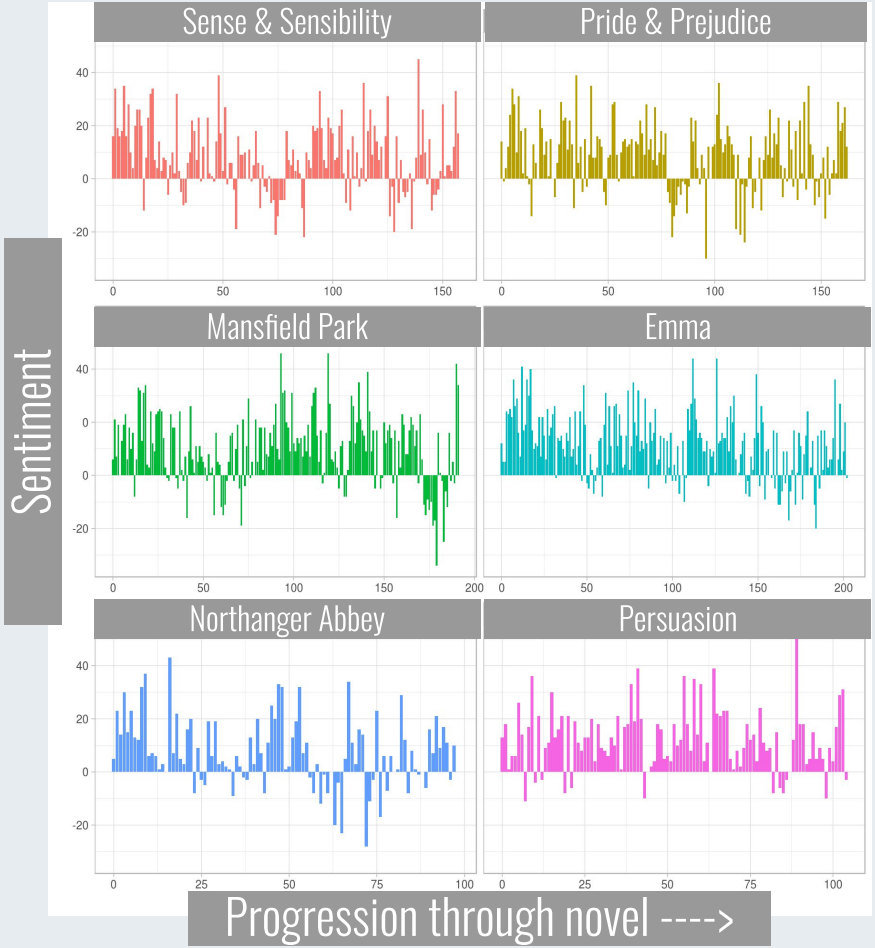
B
Context in language matters, but may be lost in sentiment analysis

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

Which of the following is true?



A Novels are overwhelmingly negative

B Mansfield Park is more negative toward the end of the novel

C Emma takes a dark turn in the middle of the novel

D Sense & Sensibility has a negative tone at its start

E Northanger Abbey is Austen's most positive novel

TF-IDF

Term Frequency - Inverse Document Frequency

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

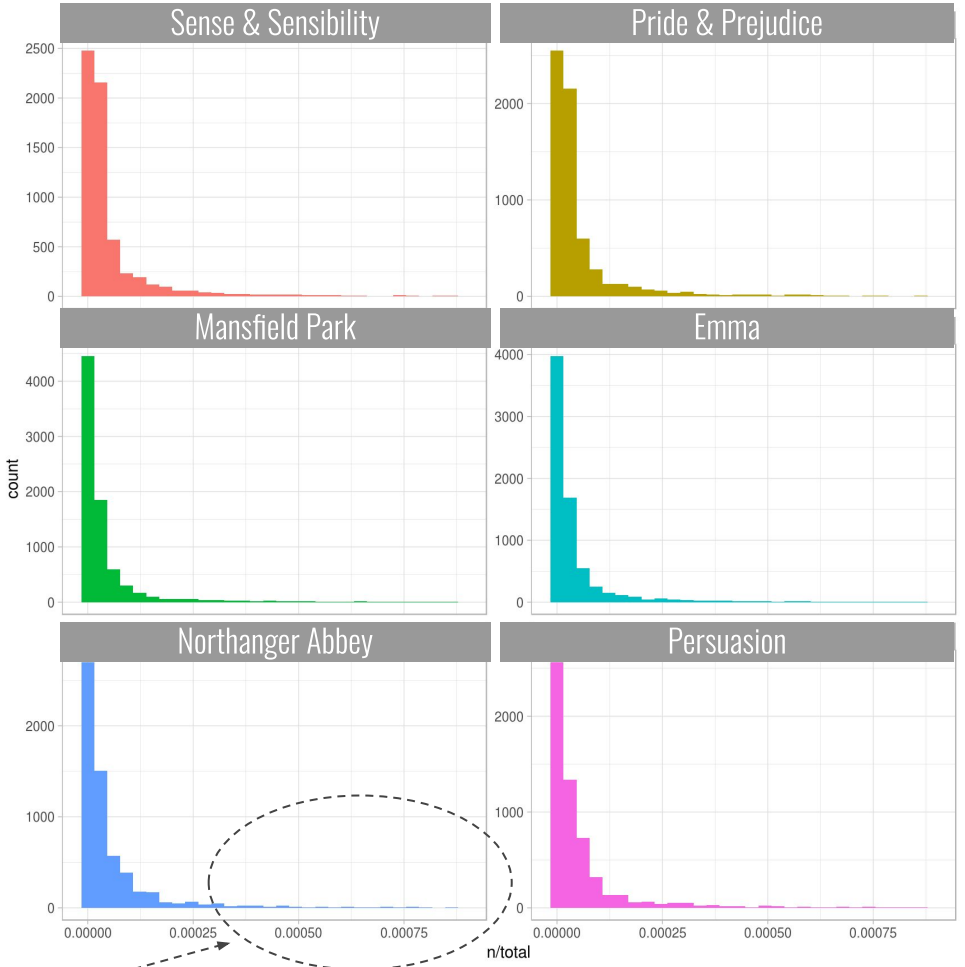
N = total number of documents

What are the most commonly used words in Jane Austen's novels?

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

Frequency Distribution in Jane Austen's Novels



The long tails in each plot are those very frequent words

	book	word	n	total	tf	idf	tf_idf
	<fct>	<chr>	<int>	<int>	<dbl>	<dbl>	<dbl>
1	Mansfield Park	the	6206	160460	0.0387	0	0
2	Mansfield Park	to	5475	160460	0.0341	0	0
3	Mansfield Park	and	5438	160460	0.0339	0	0
4	Emma	to	5239	160996	0.0325	0	0
5	Emma	the	5201	160996	0.0323	0	0
6	Emma	and	4896	160996	0.0304	0	0
7	Mansfield Park	of	4778	160460	0.0298	0	0
8	Pride & Prejudice	the	4331	122204	0.0354	0	0
9	Emma	of	4291	160996	0.0267	0	0
10	Pride & Prejudice	to	4162	122204	0.0341	0	0

... with 40,369 more rows

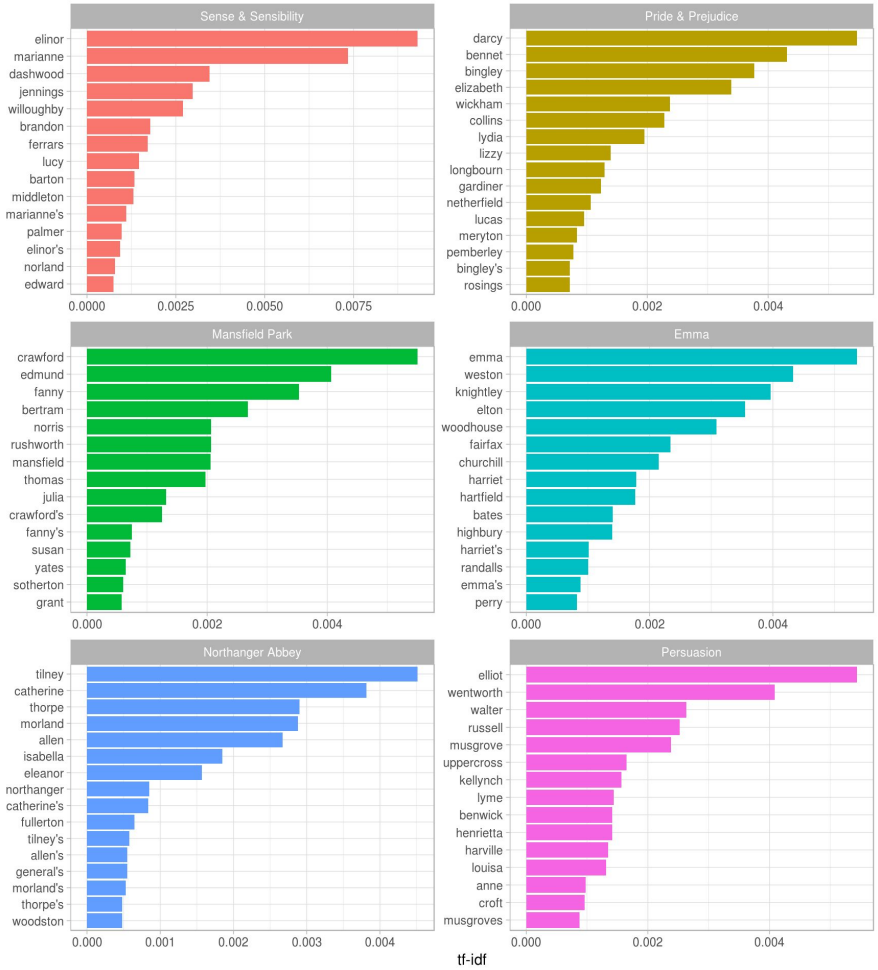
Super common words will have TF-IDF of zero....since they occur frequently across all documents

	book	word	n	tf	idf	tf_idf
	<fct>	<chr>	<int>	<dbl>	<dbl>	<dbl>
1	Sense & Sensibility	elinor	623	0.00519	1.79	0.00931
2	Sense & Sensibility	marianne	492	0.00410	1.79	0.00735
3	Mansfield Park	crawford	493	0.00307	1.79	0.00551
4	Pride & Prejudice	darcy	373	0.00305	1.79	0.00547
5	Persuasion	elliot	254	0.00304	1.79	0.00544
6	Emma	emma	786	0.00488	1.10	0.00536
7	Northanger Abbey	tilney	196	0.00252	1.79	0.00452
8	Emma	weston	389	0.00242	1.79	0.00433
9	Pride & Prejudice	bennet	294	0.00241	1.79	0.00431
10	Persuasion	wentworth	191	0.00228	1.79	0.00409

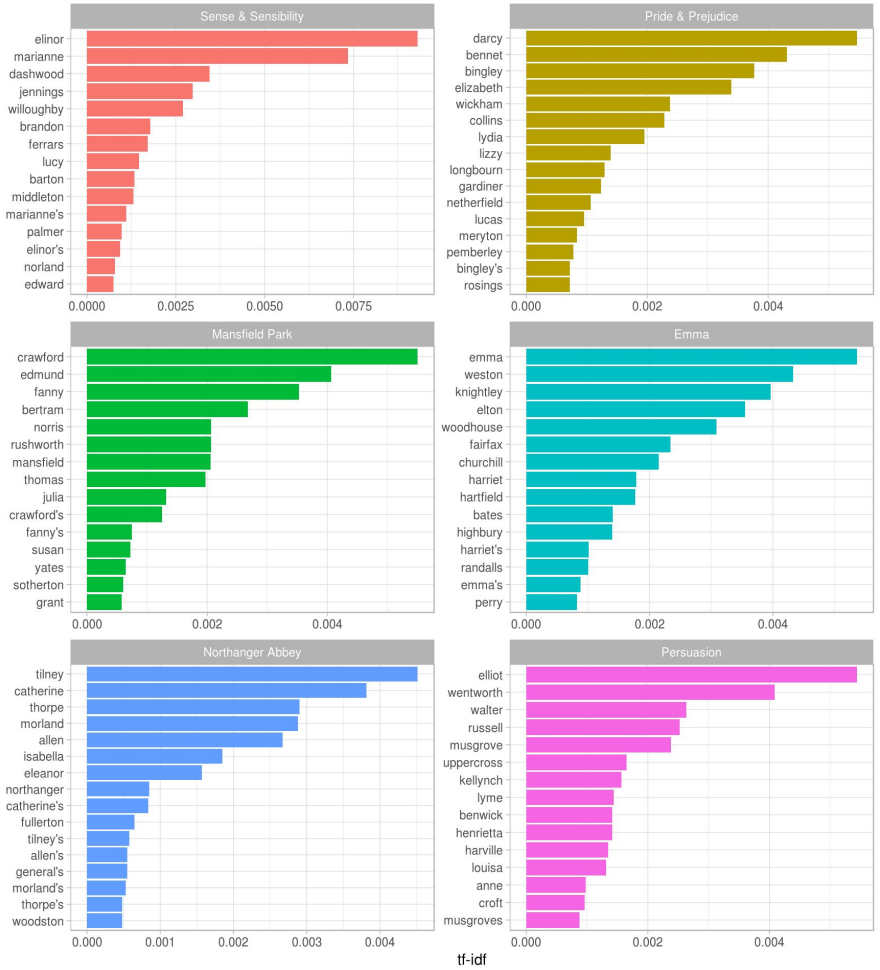
... with 40,369 more rows

Proper nouns, like character names, are important to a specific novel, and have a higher TF-IDF

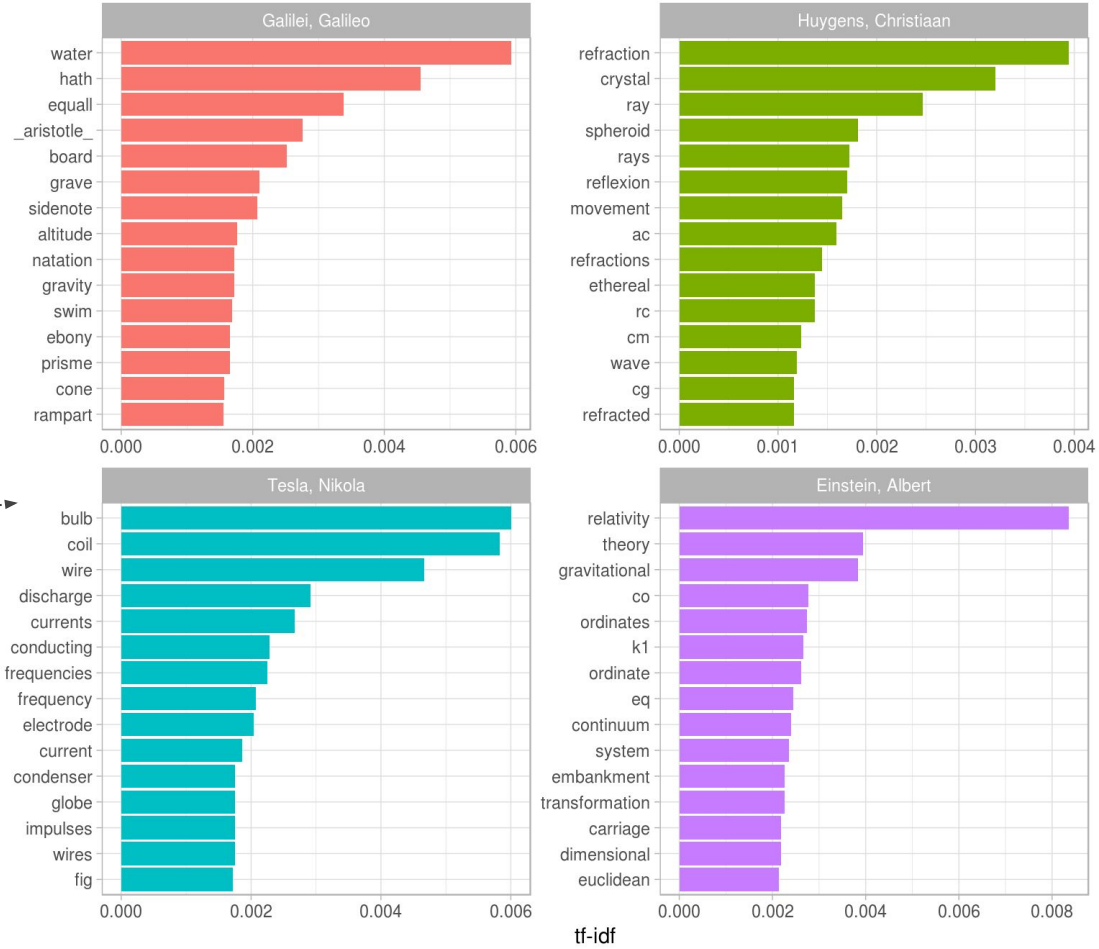
High TF-IDF words broken down by Austen novel



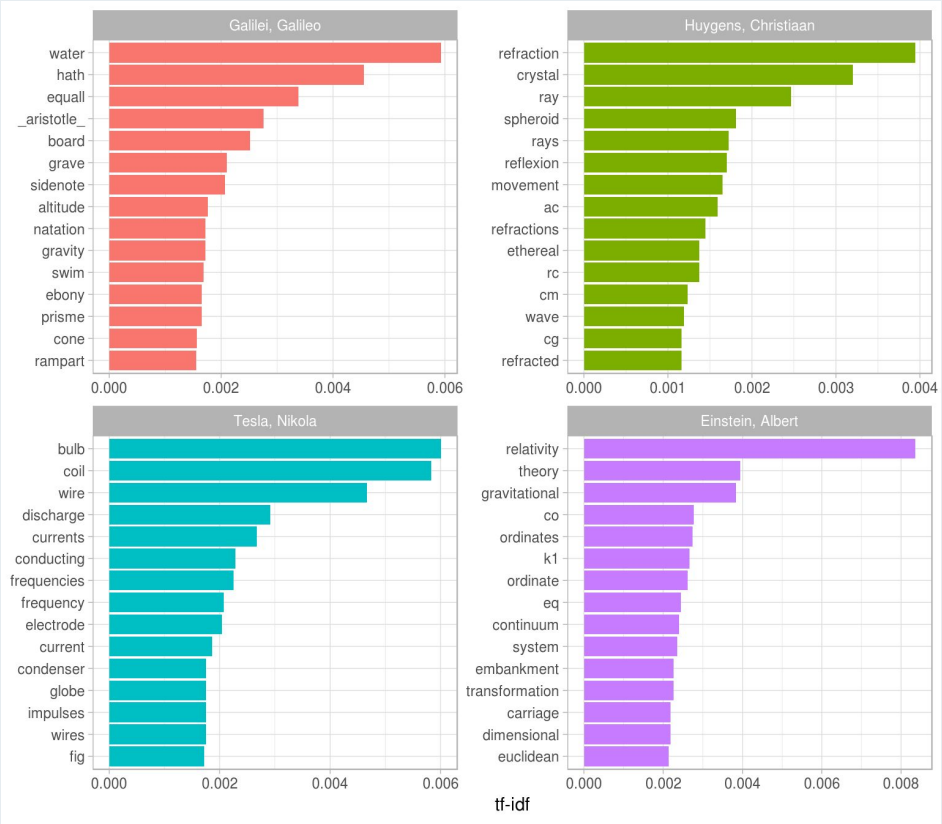
Can conclude that "Jane Austen used similar language across her six novels, and *what distinguishes one novel from the rest within the collection of her works are the proper nouns*, the names of people and places"



A quick look at TF-IDF in another corpus: classic physics texts from Project Gutenberg



Which word is most uniquely “Einstein”?



A relativity

B theory

C refraction

D euclidean

E water

How has pop music changed in the last three years?

What data would we need to answer this question?

How has pop music changed in the last three years?

Data: Lyrics to the (200) most popular songs from 2017-2019

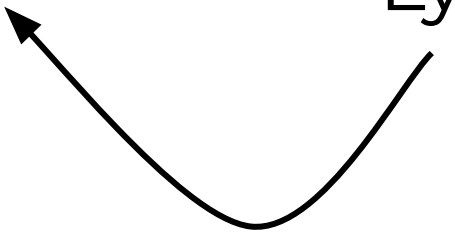
The data : Top songs from Feb music charts 2017-2019

2017: 152 songs

2018: 139 songs

2019: 127 songs

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



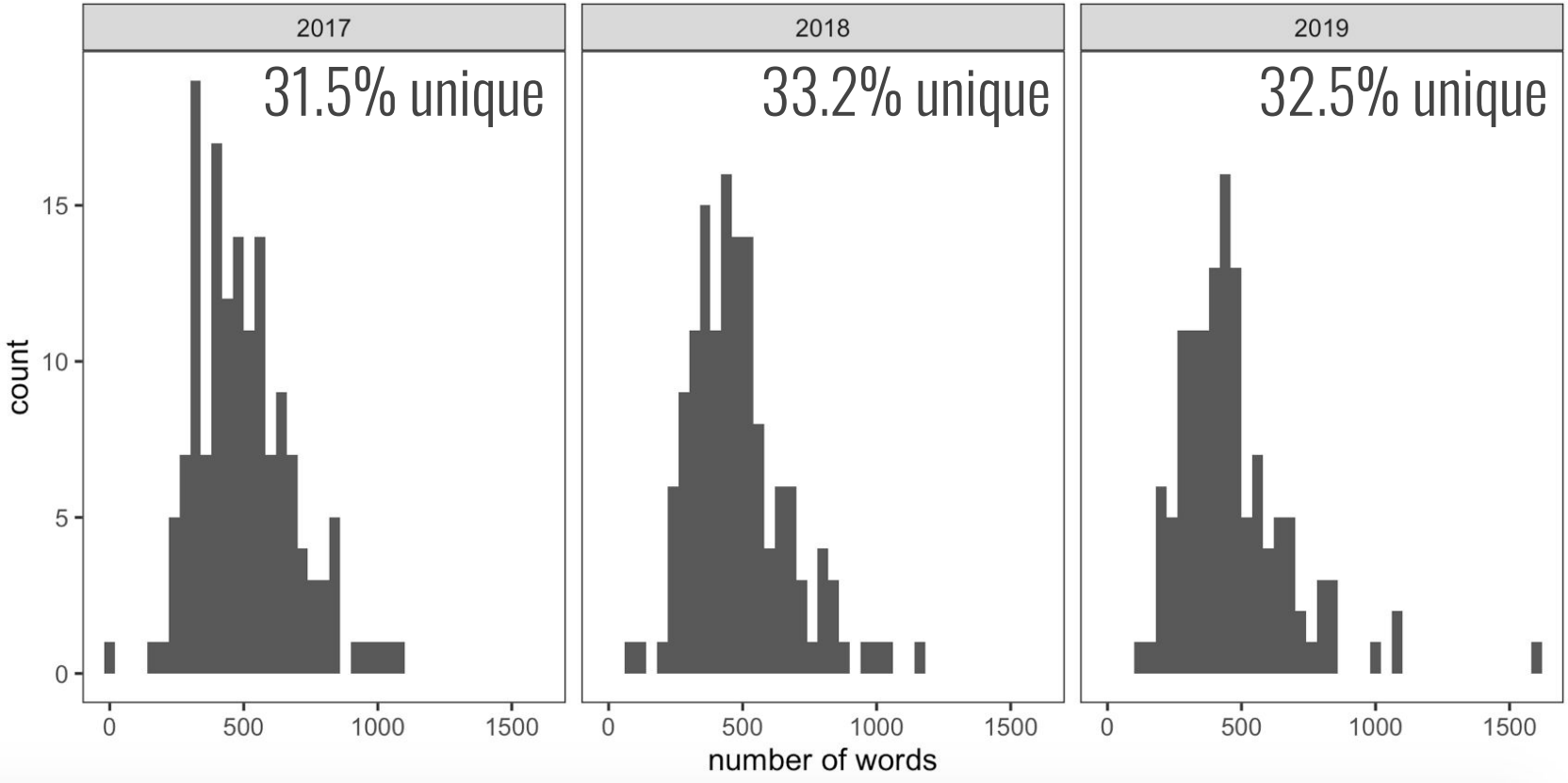
Questions we can ask...

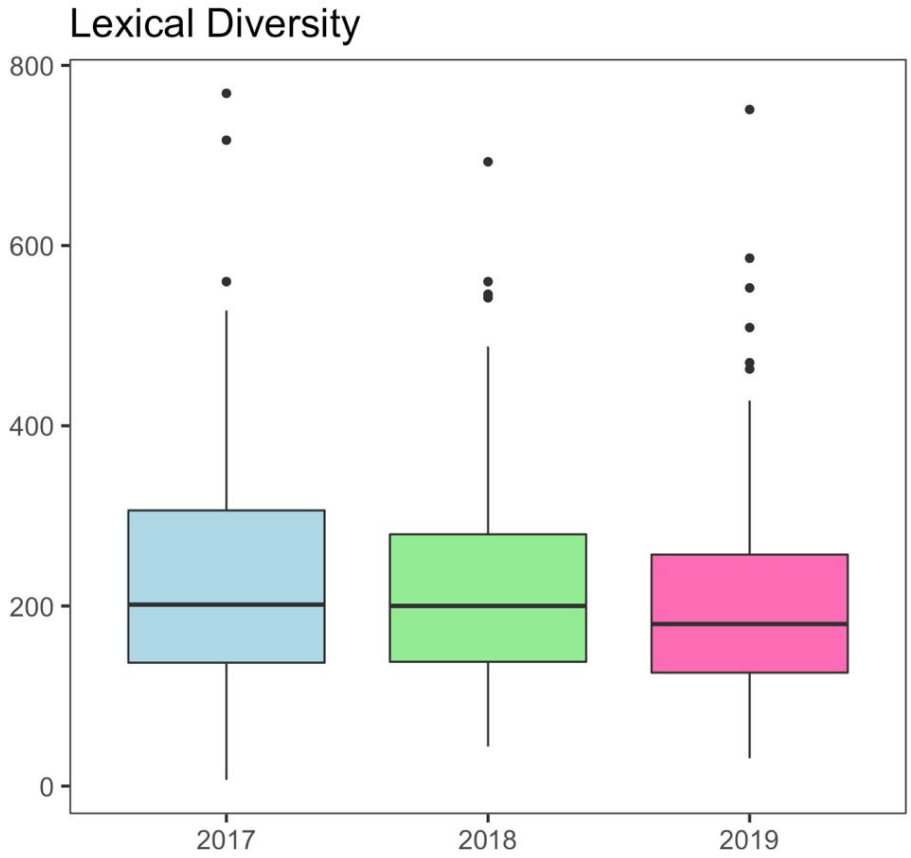
1. Does the total number of words change?
2. Does uniqueness change over time?
3. Does the diversity or density change?
4. What words are most common?
5. What words are most unique to each year?
6. What sentiment do songs convey most frequently?
7. Has sentiment changed over time?
8. What are the sentiment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

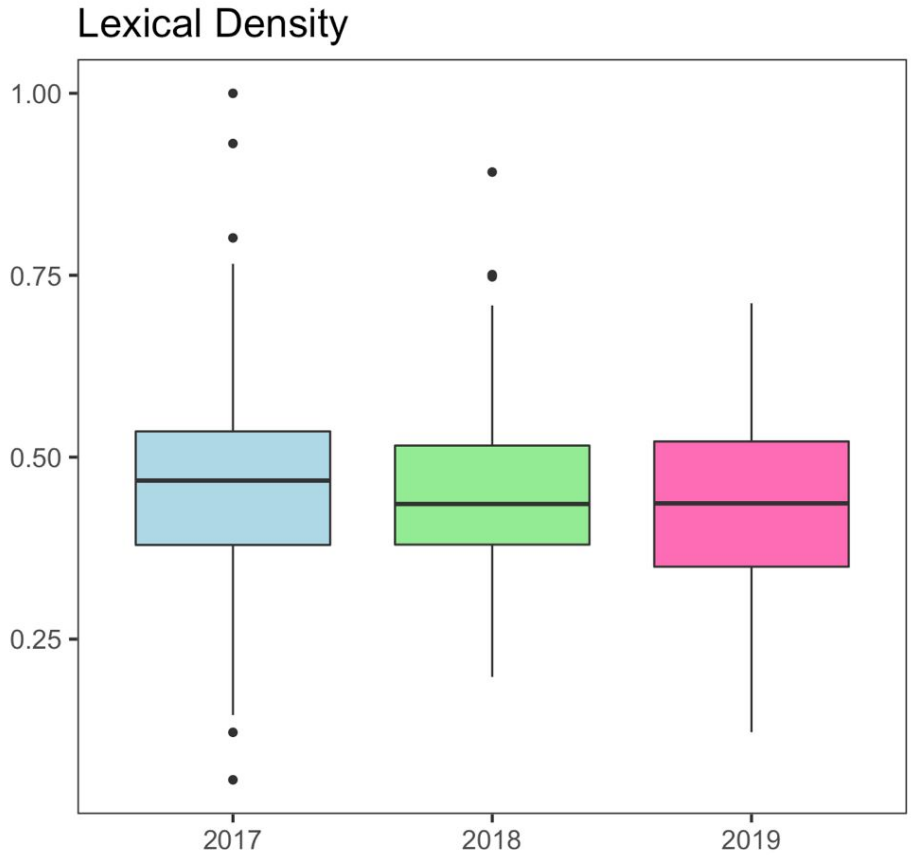
Questions we can ask...

1. Does the total number of words change?
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?

Words per song

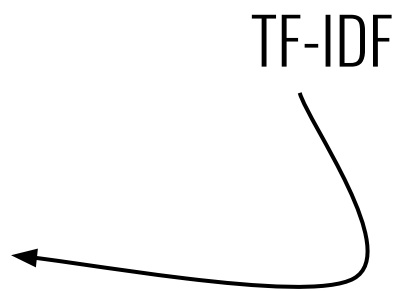


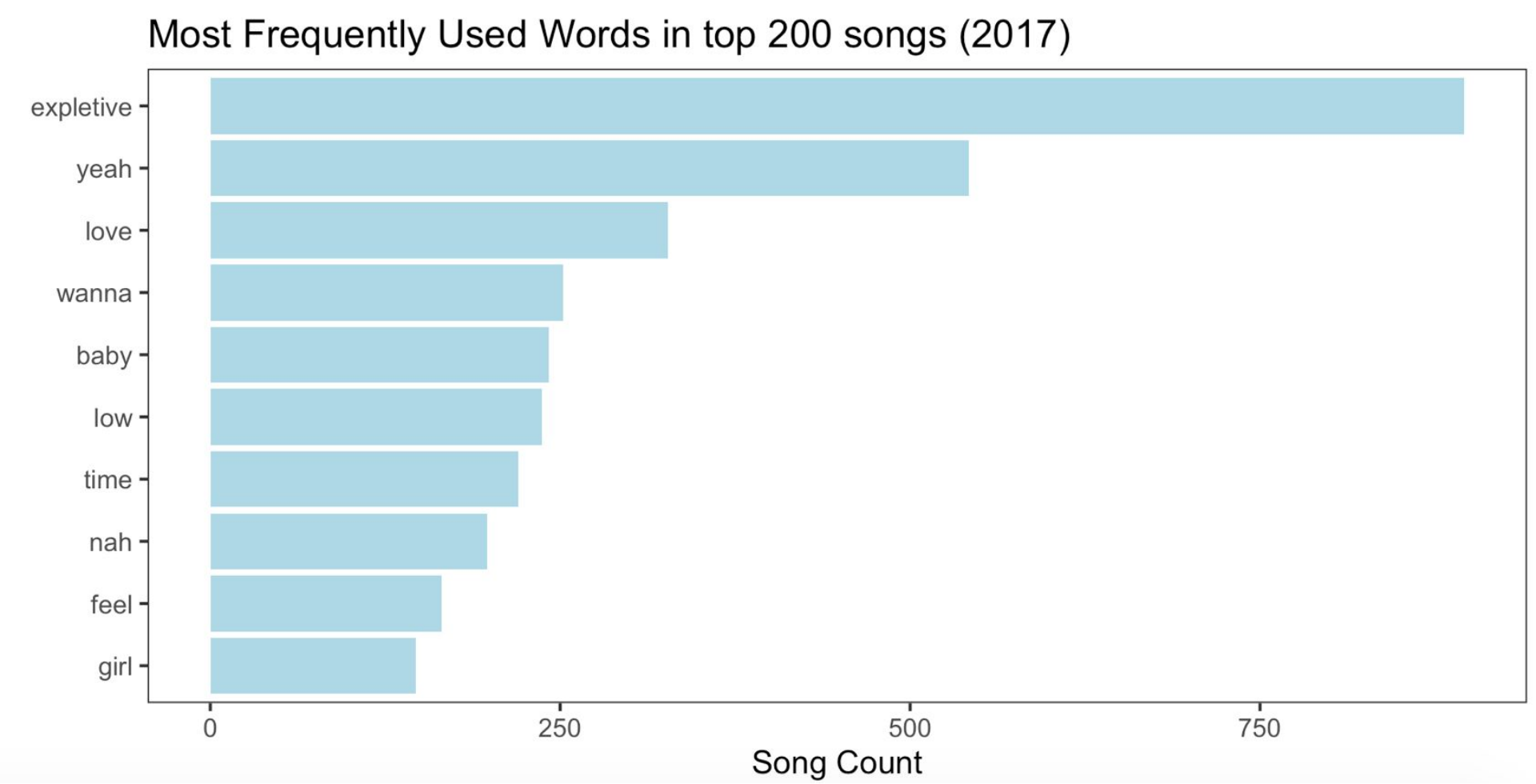

Lexical Diversity
800
600
400
200
0
2017
2018
2019

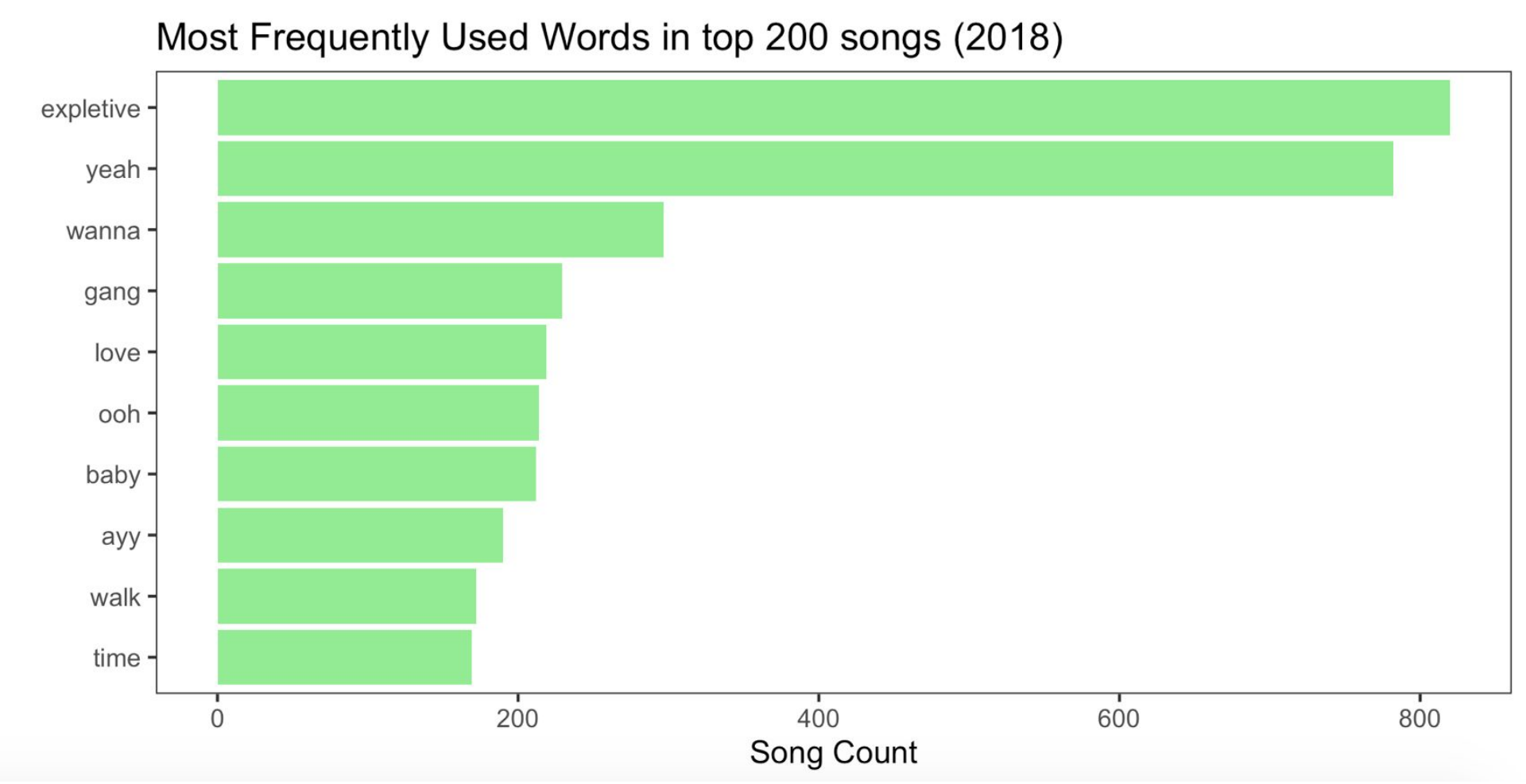

Lexical Density
1.00
0.75
0.50
0.25
2017
2018
2019

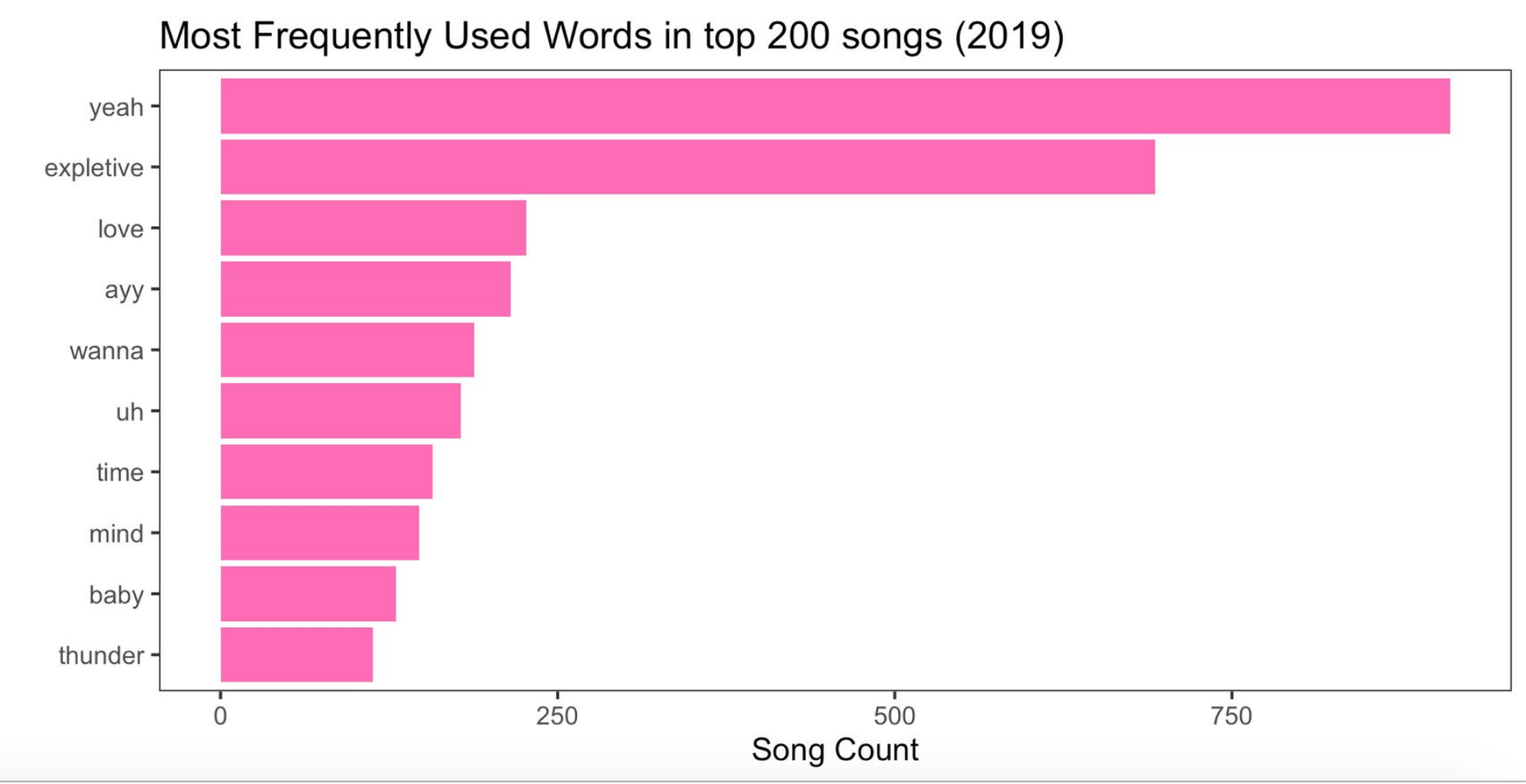
Questions we can ask...

1. Does the total number of words change?
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?

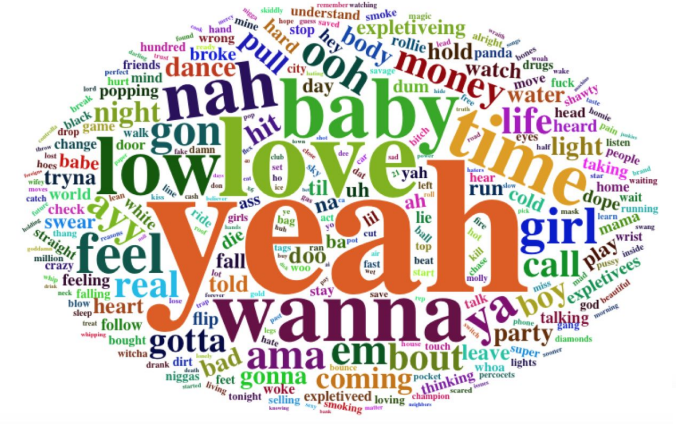



Most Frequently Used Words in top 200 songs (2017)
expletive
yeah
love
wanna
baby
low
time
nah
feel
girl
0
250
500
750
Song Count


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

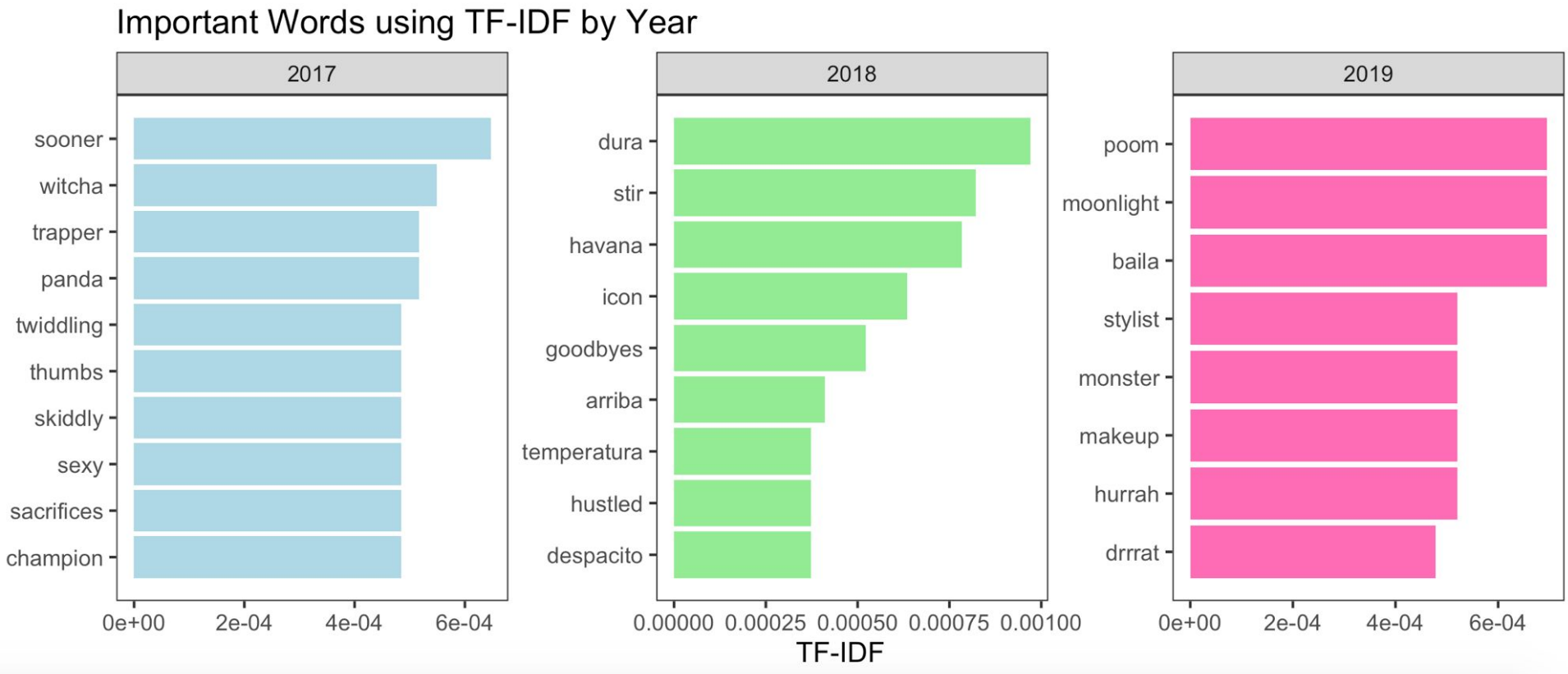
Word clouds display the word size proportional to their frequency within the textual dataset



2017

2018

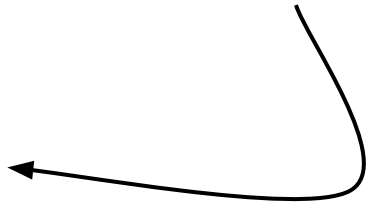
2019

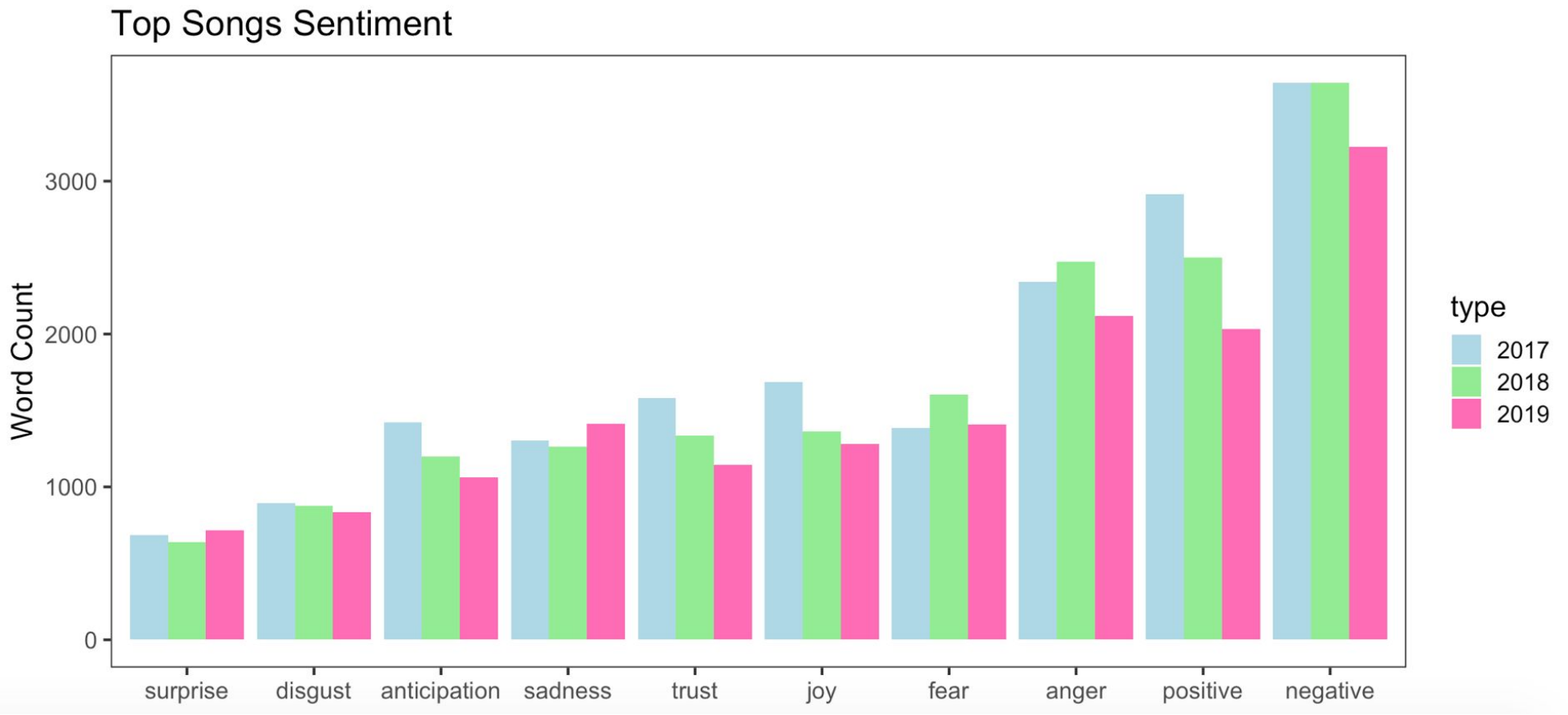

Important Words using TF-IDF by Year
2017
sooner
witcha
trapper
panda
twiddling
thumbs
skiddly
sexy
sacrifices
champion
0e+00
2e-04
4e-04
6e-04
2018
dura
stir
havana
icon
goodbyes
arriba
temperatura
hustled
despacito
0.00000
0.00025
0.00050
0.00075
0.00100
2019
poom
moonlight
baila
stylist
monster
makeup
hurrah
drrrat
0e+00
2e-04
4e-04
6e-04
TF-IDF

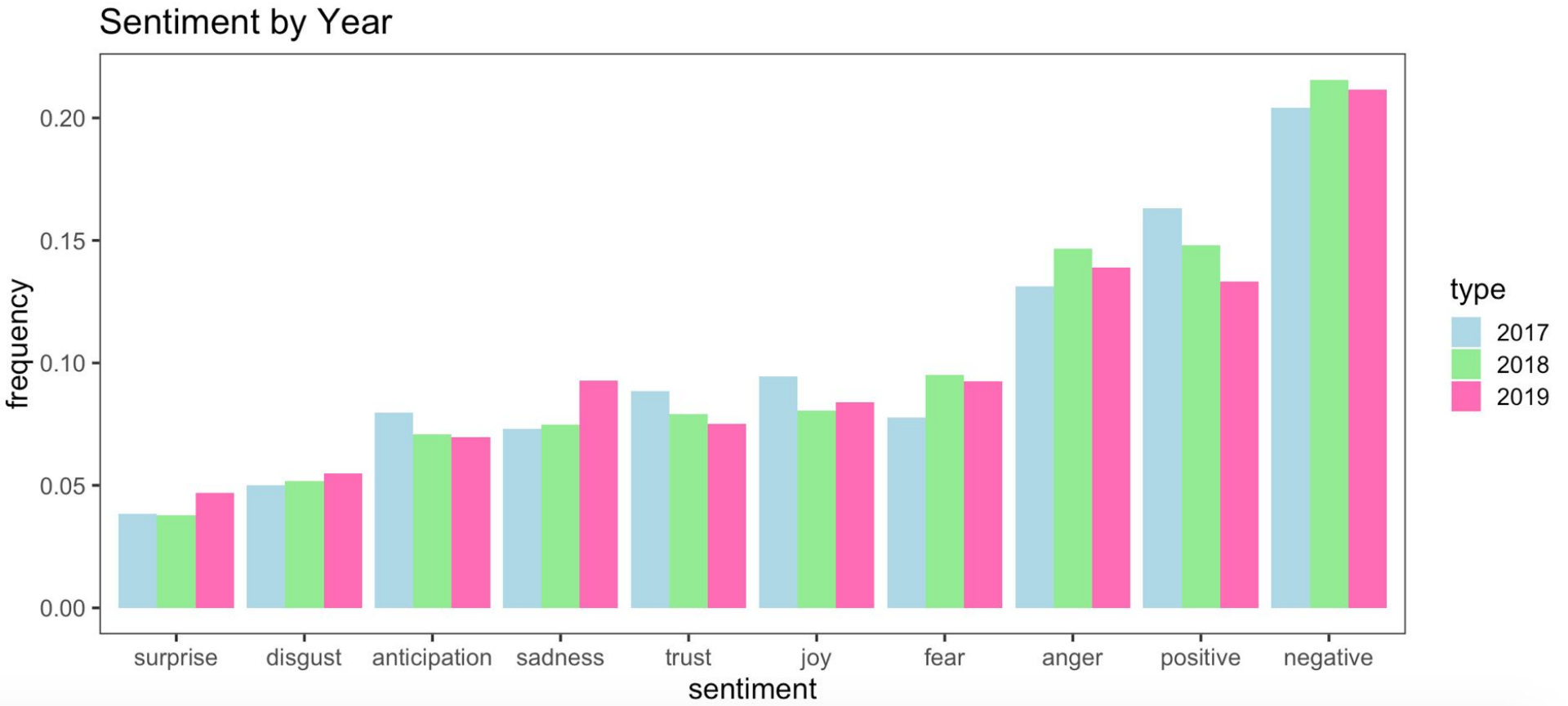
Questions we can ask...

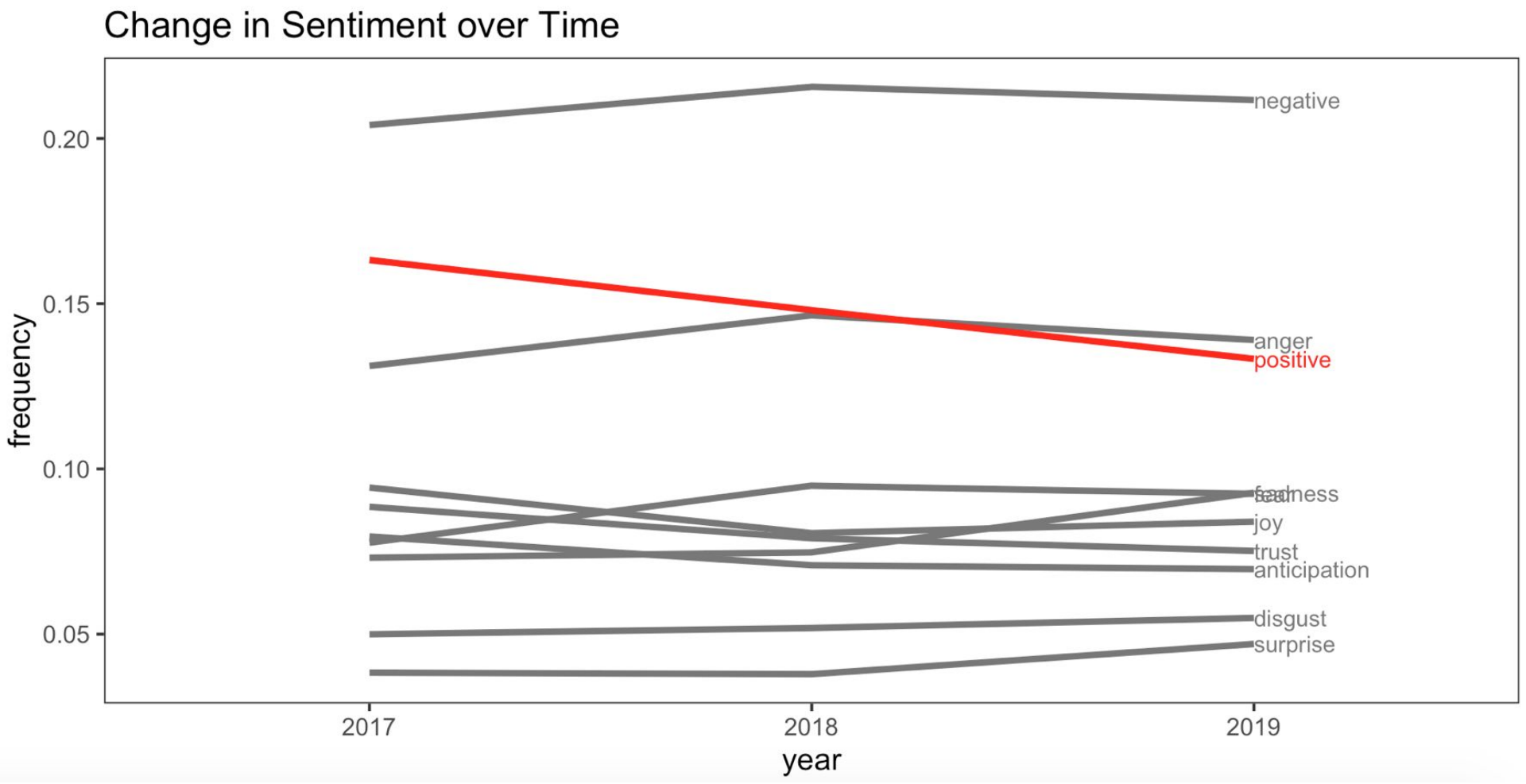
1. Does the total number of words change?
2. Does uniqueness change over time?
3. Does the diversity or density change?
4. What words are most common?
5. What words are most unique to each year?
6. What sentiment do songs convey most frequently?
7. Has sentiment changed over time?
8. What are the sentiment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

Sentiment Analysis




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


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


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