

Text

Session 14

PMAP 8551/4551: Data Visualization with R
Andrew Young School of Policy Studies
Spring 2026

Plan for today

Qualitative text-based data

**Crash course in
computational linguistics**

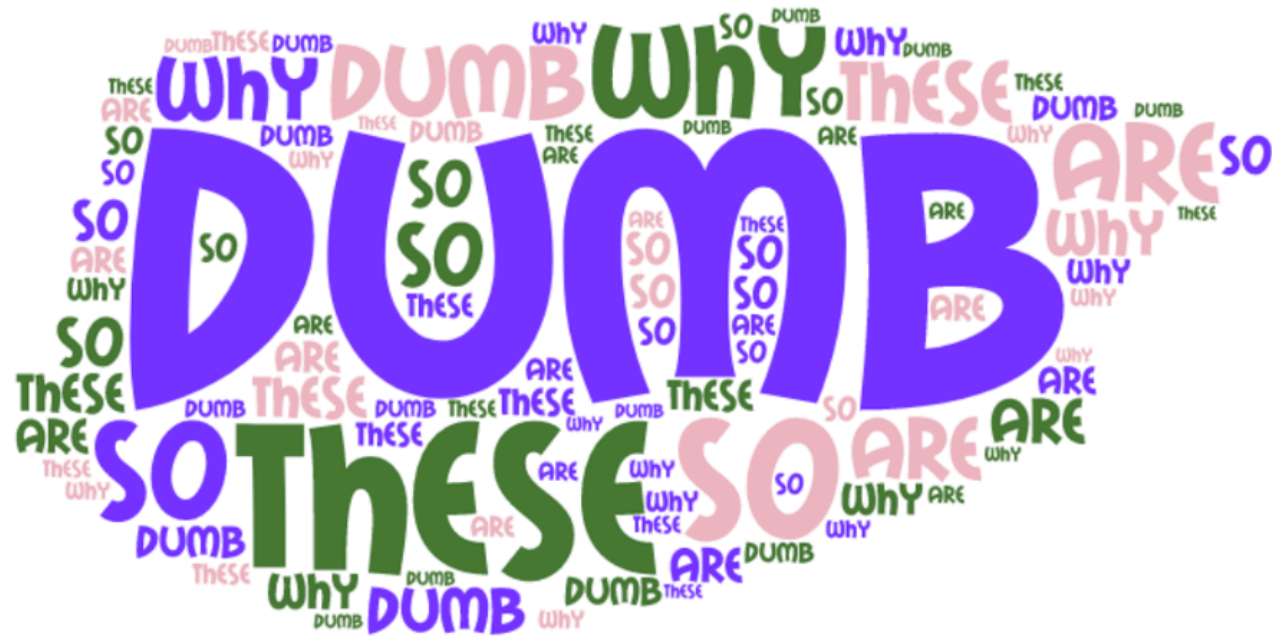
Qualitative text-based data

Free responses

N	O	P	
donate_likely	amount_donate	amount_keep	amount_why
Somewhat unlikely	0	100	I am poor
Somewhat unlikely	0	100	I really feel like I deserve to treat myself recently. I have been wor
Somewhat likely	10	90	I donate the amount that I usually would
Somewhat unlikely	0	100	i'm poor
Neither likely nor unlikely	10	90	It is not a cause that is very important to me. i have other things tl
Extremely likely	29	71	I want to contribute to the cause, but also keep some of the mone
Somewhat likely	20	80	It's a reasonable amount of money for an individual to donate to a
Extremely unlikely	0	100	I don't fully agree with their mission
Somewhat likely	10	90	I am pretty poor so I need to keep some for myself, but I also war
Extremely likely	5	95	I think it would be a good amount to give from the money I have a
Neither likely nor unlikely	69	31	to help with their cause
Somewhat unlikely	0	100	My dad always told me to give until it hurts, and right now I am hu
Neither likely nor unlikely	0	100	I would rather keep the money for myself and find a charity that I
Extremely unlikely	0	100	I want the most for myself.
Neither likely nor unlikely	5	95	Can afford to give a little
Extremely unlikely	0	100	Because I would then have 100\$ more dollars.
Extremely unlikely	0	100	I'm a broke boi. If anyone need humanitarian aid, it's me.
Somewhat likely	10	90	I'm in a position where I would need the extra money, but I also w
Somewhat unlikely	90	10	I think it is a worthy cause and I think donating 90% of the amoun
Extremely likely	50	50	I feel splitting it 50/50 would be a fair deal. I get to help make a di
Extremely likely	20	80	I feel that my contribution is enough. I would gladly donate again
Somewhat likely	9	91	give a little
Somewhat likely	1	99	I like money
Somewhat unlikely	0	100	I do not really know what they will do with the money.

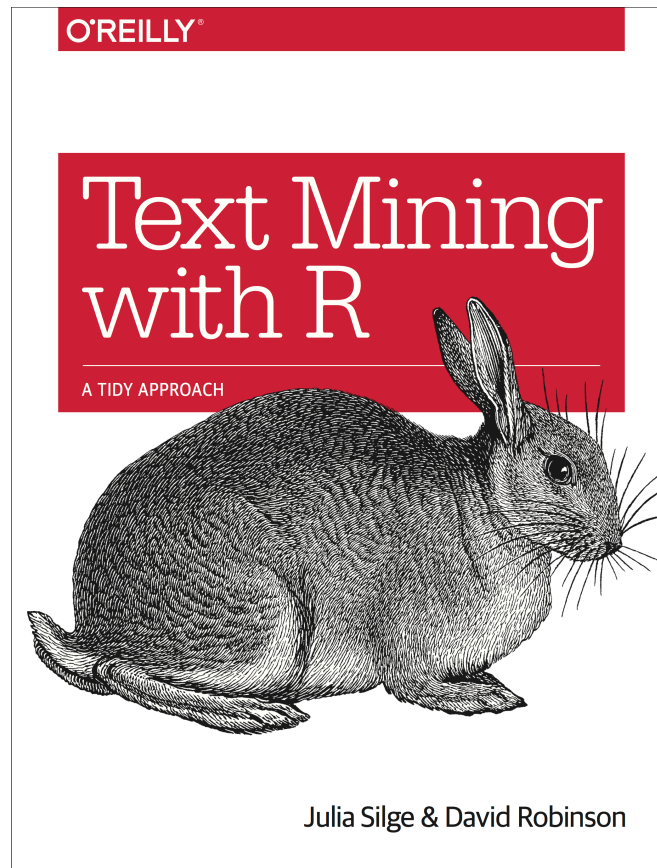
Typical free responses from a survey

y tho?



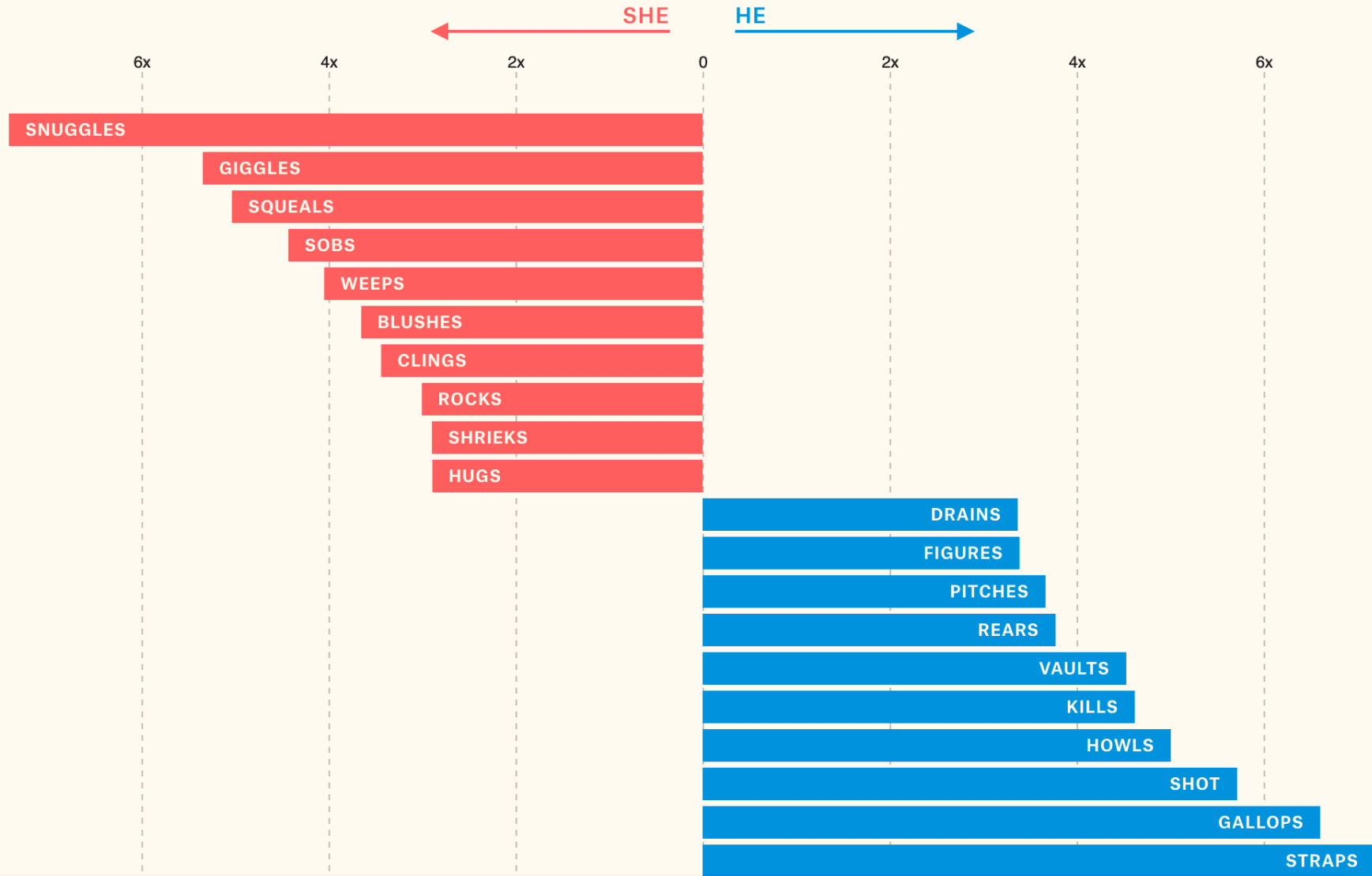
Word clouds for grownups

Count words, but in fancier ways

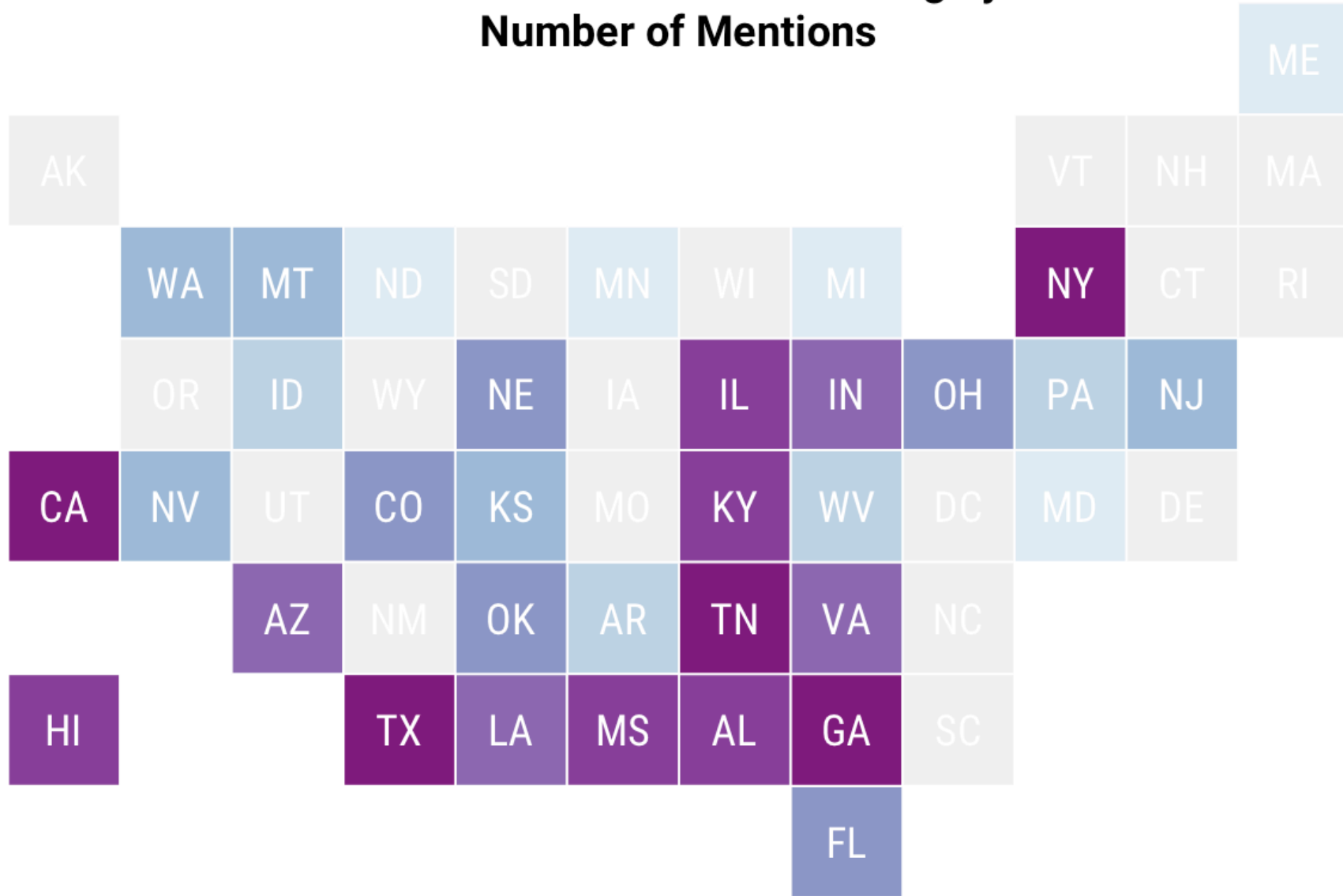


The most used words for women vs. men

Likelihood that certain words appear after "she" vs. "he" in screen direction.



What States Are Mentioned in Song Lyrics? Number of Mentions



Crash course in computational linguistics

Core concepts and techniques

Tokens, lemmas, and parts of speech

Sentiment analysis

tf-idf

Topics and LDA

Fingerprinting

Regular text

THE BOY WHO LIVED Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense. Mr. Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large mustache. Mrs. Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbors. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere. The Dursleys had everything they wanted, but they also had a secret, and their greatest fear was that somebody would discover it. They didn't think they could bear it if anyone found out about the Potters. Mrs. Potter was Mrs. Dursley's sister, but they hadn't met for several years; in fact, Mrs. Dursley pretended she didn't have a sister, because her sister and her good-for-nothing husband were as unDursleyish as it was possible to be. The Dursleys shuddered to think what the neighbors would say if the Potters a...

Tidy text

One row for each text element

Can be chapter, page, verse, etc.

`.small-code[`

```
[38;5;246m# A tibble: 6 × 3 [39m
```

```
  chapter book
```

```
      text
```

```
    [3m [38;5;246m<int> [39m [23m    [3m [38;5;246m<chr> [39m [23m
```

```
[38;5;250m1 [39m
```

```
1 Harry Potter and the Philosopher's Stone
```

```
[38;5;250m2 [39m
```

```
2 Harry Potter and the Philosopher's Stone
```

```
[38;5;250m3 [39m
```

```
3 Harry Potter and the Philosopher's Stone
```

```
[38;5;250m4 [39m
```

```
4 Harry Potter and the Philosopher's Stone
```

Tokens

Split the text into even smaller parts

Paragraph, line, verse, sentence, n-gram, word, letter, etc.

.pull-left.small-code[

```
[38;5;246m# A tibble: 6 × 3 [39m
```

```
word chapter book
```

```
[3m [38;5;246m<chr> [39m [23m      [3m [38;5;246m<int> [39m [23m [3m
```

```
[38;5;250m1 [39m the          1 Harry Potter...
```

```
[38;5;250m2 [39m boy          1 Harry Potter...
```

```
[38;5;250m3 [39m who          1 Harry Potter...
```

```
[38;5;250m4 [39m lived        1 Harry Potter...
```

Stop words

Common words that we can generally ignore

.center.small-code[

[38;5;246m# A tibble: 1,149 × 2 [39m

word lexicon

[3m [38;5;246m<chr> [39m [23m

[3m [38;5;246m<chr> [39m [23

[38;5;250m 1 [39m a SMART

[38;5;250m 2 [39m a's SMART

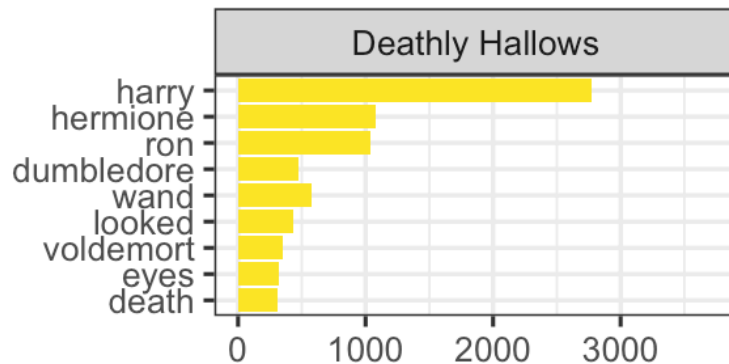
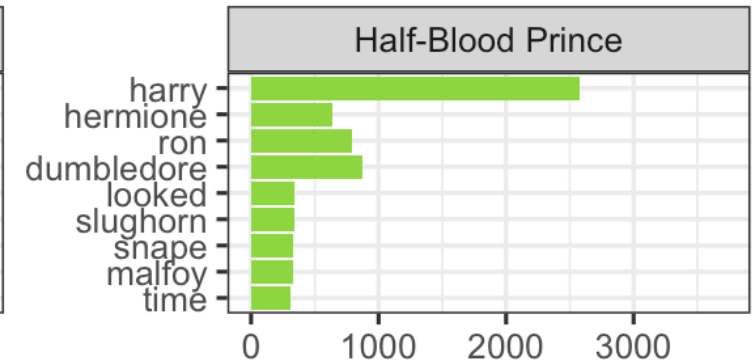
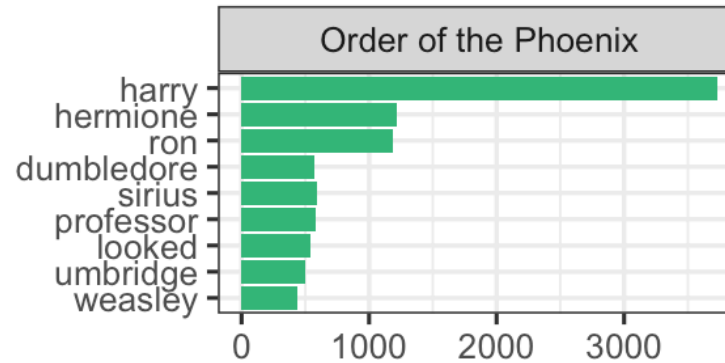
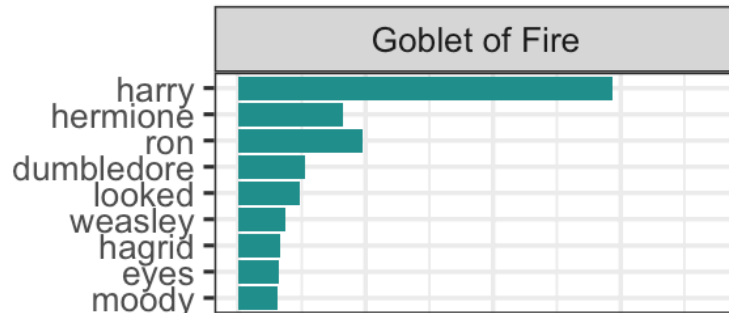
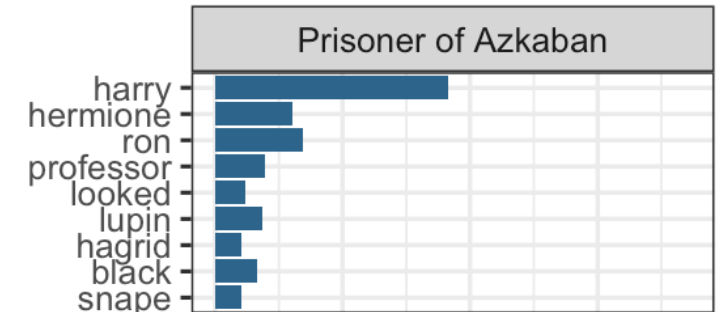
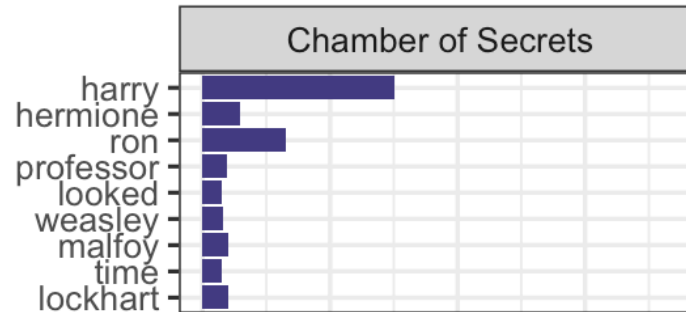
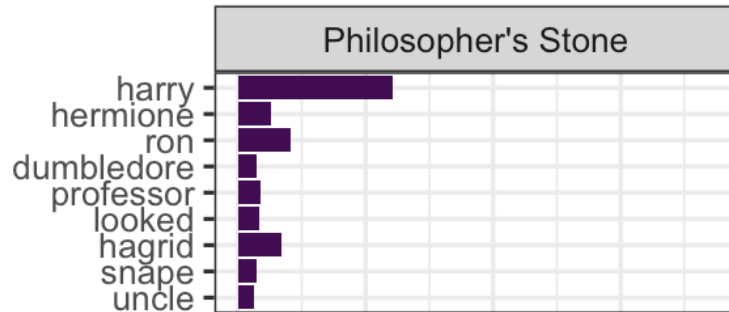
[38;5;250m 3 [39m able SMART

[38;5;250m 4 [39m about SMART

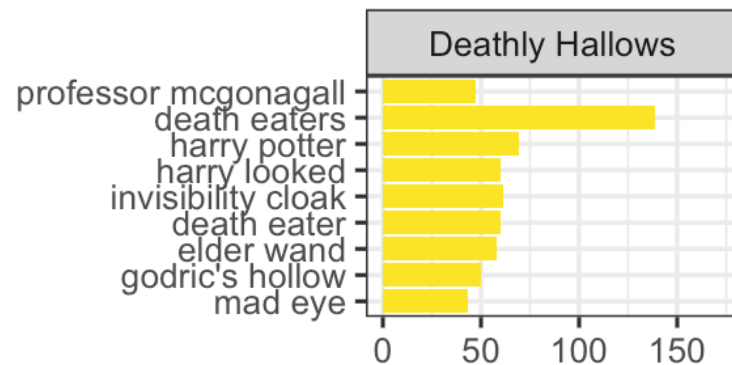
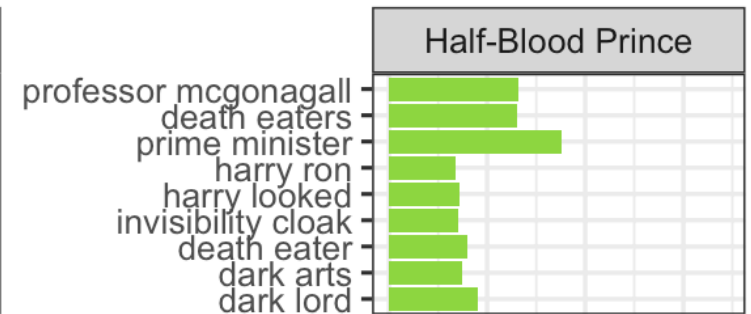
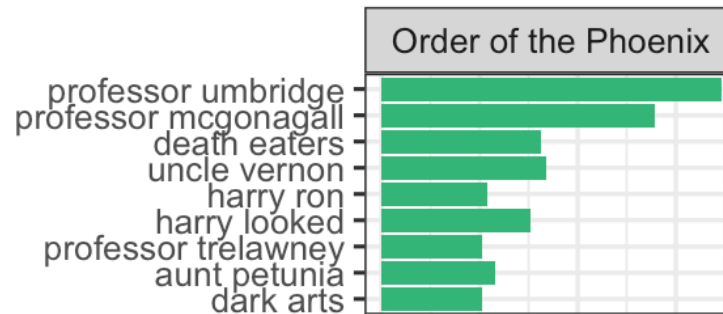
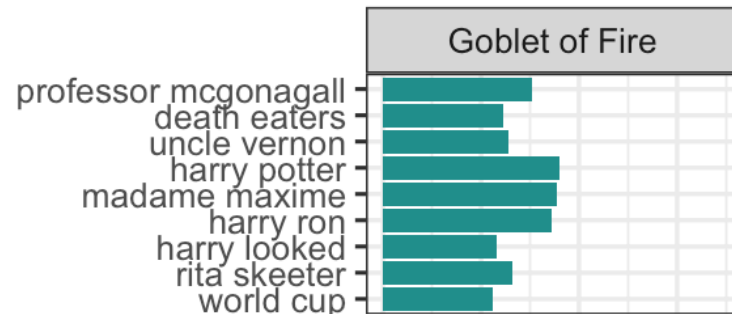
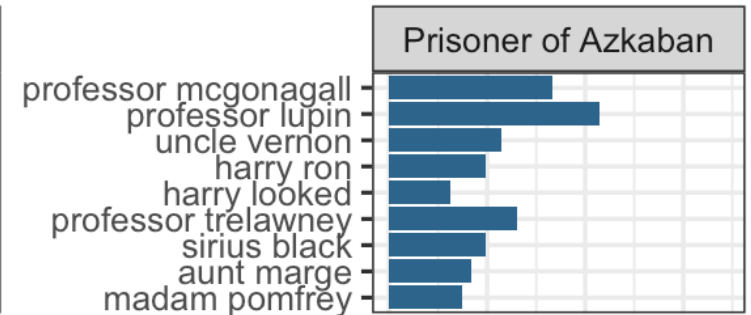
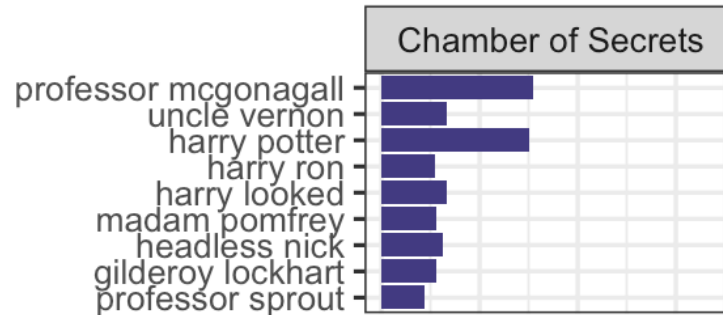
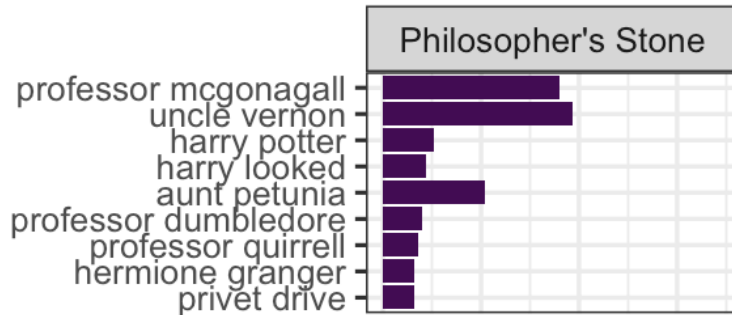
[38;5;250m 5 [39m above SMART

[38;5;250m 6 [39m according SMART

Token frequency: words

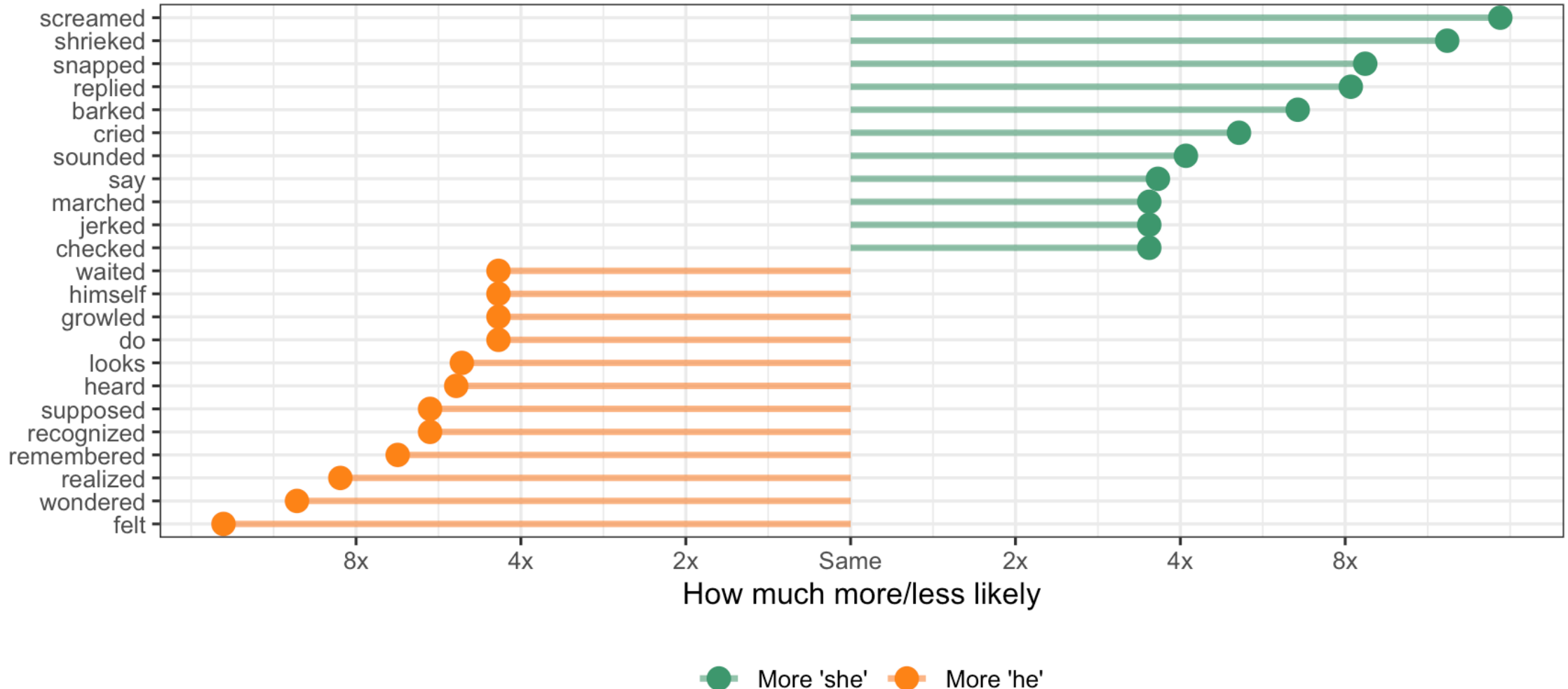


Token frequency: n-grams



0 50 100 150

Token frequency: n-gram ratios



Parts of speech

.small-code[

[38;5;246m# A tibble: 50 × 11 [39m

doc_id	sid	tid	token	token_with_ws	lemma	upos	xpos	f
[3m [38;5;246m<dbl> [39m	[23m	[3m [38;5;246m<dbl> [39m	[23m	[3m	[38;5;246m<dbl> [39m	[23m	[3	
[38;5;250m 1 [39m			1	1	1 THE	THE	the	
[38;5;250m 2 [39m			1	1	2 BOY	BOY	Boy	
[38;5;250m 3 [39m			1	1	3 WHO	WHO	who	
[38;5;250m 4 [39m			1	1	4 LIVED	LIVED	live	
[38;5;250m 5 [39m			1	1	5 Mr.	Mr.	Mr.	
[38;5;250m 6 [39m			1	1	6 and	and	and	
[38;5;250m 7 [39m			1	1	7 Mrs.	Mrs.	Mrs.	

Parts of speech frequency

.pull-left-3.small-code[

Verbs

[38;5;246m# A tibble: 1,557 × 2 [39m

lemma n

[3m [38;5;246m<chr> [39m [23m [3m [38;5;246m<dbl> [39m [23m

[38;5;250m 1 [39m say 920

[38;5;250m 2 [39m get 440

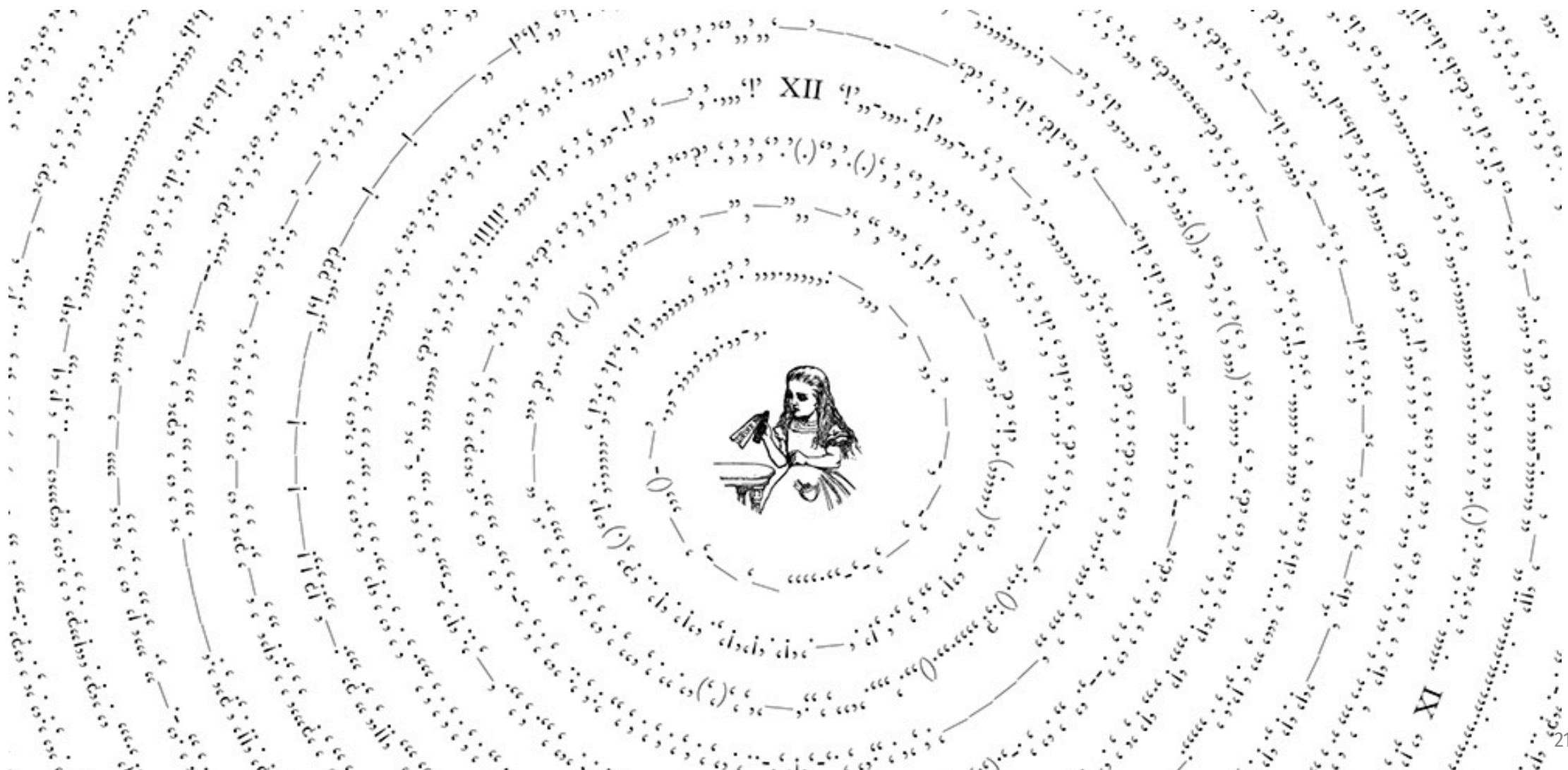
[38;5;250m 3 [39m have 417

[38;5;250m 4 [39m go 384

[38;5;250m 5 [39m look 380

[38;5;250m 6 [39m be 310

Artsy stuff



Sentiment analysis

```
.pull-left-3.small-code[
```

```
get_sentiments("bing")
```

```
[38;5;246m# A tibble: 6,786 × 2 [39m
```

```
  word          sentiment
```

```
  [3m [38;5;246m<chr> [39m [23m
```

```
[3m [38;5;246m<chr> [39m [23
```

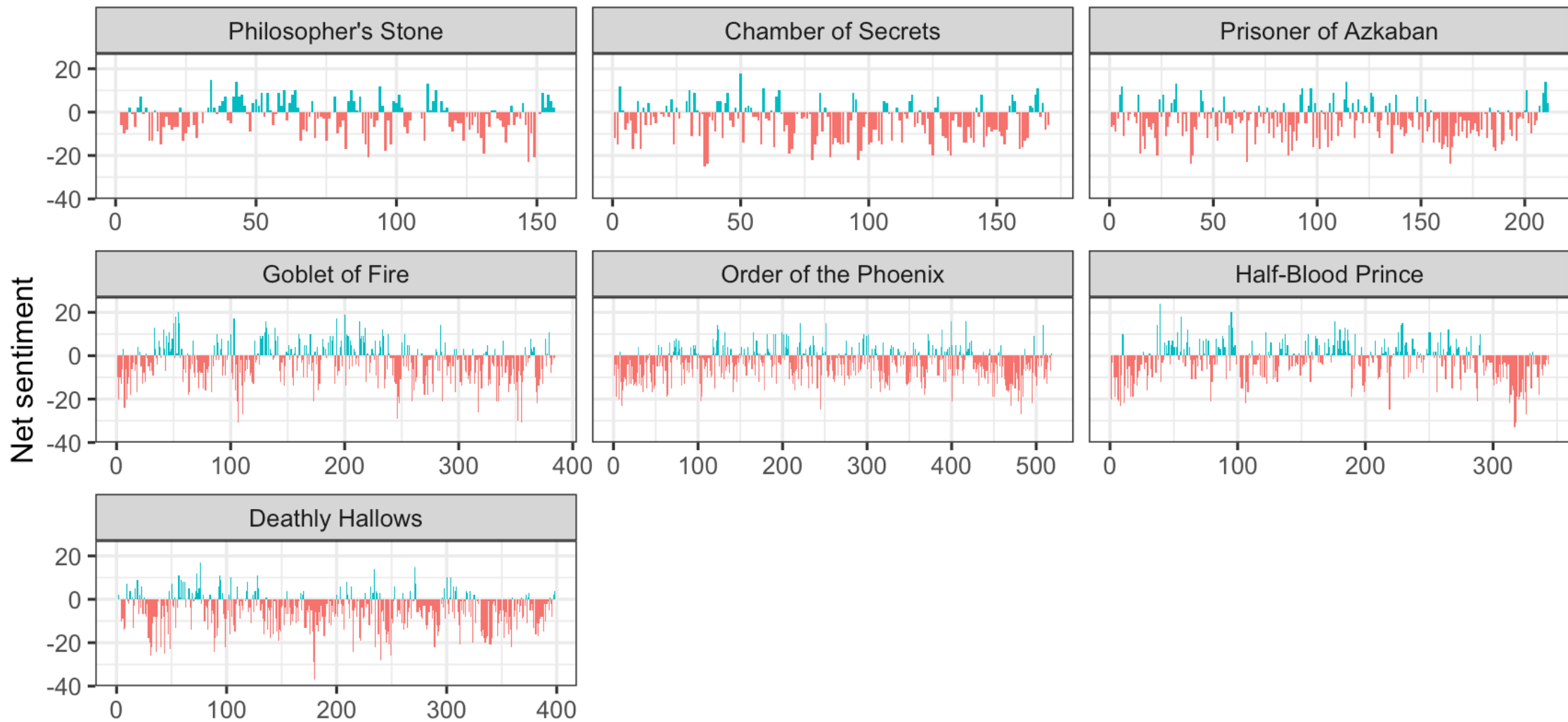
```
[38;5;250m 1 [39m 2-faces      negative
```

```
[38;5;250m 2 [39m abnormal    negative
```

```
[38;5;250m 3 [39m abolish     negative
```

```
[38;5;250m 4 [39m abominable  negative
```

```
[38;5;250m 5 [39m abominably  negative
```



tf-idf

Term frequency-inverse document frequency

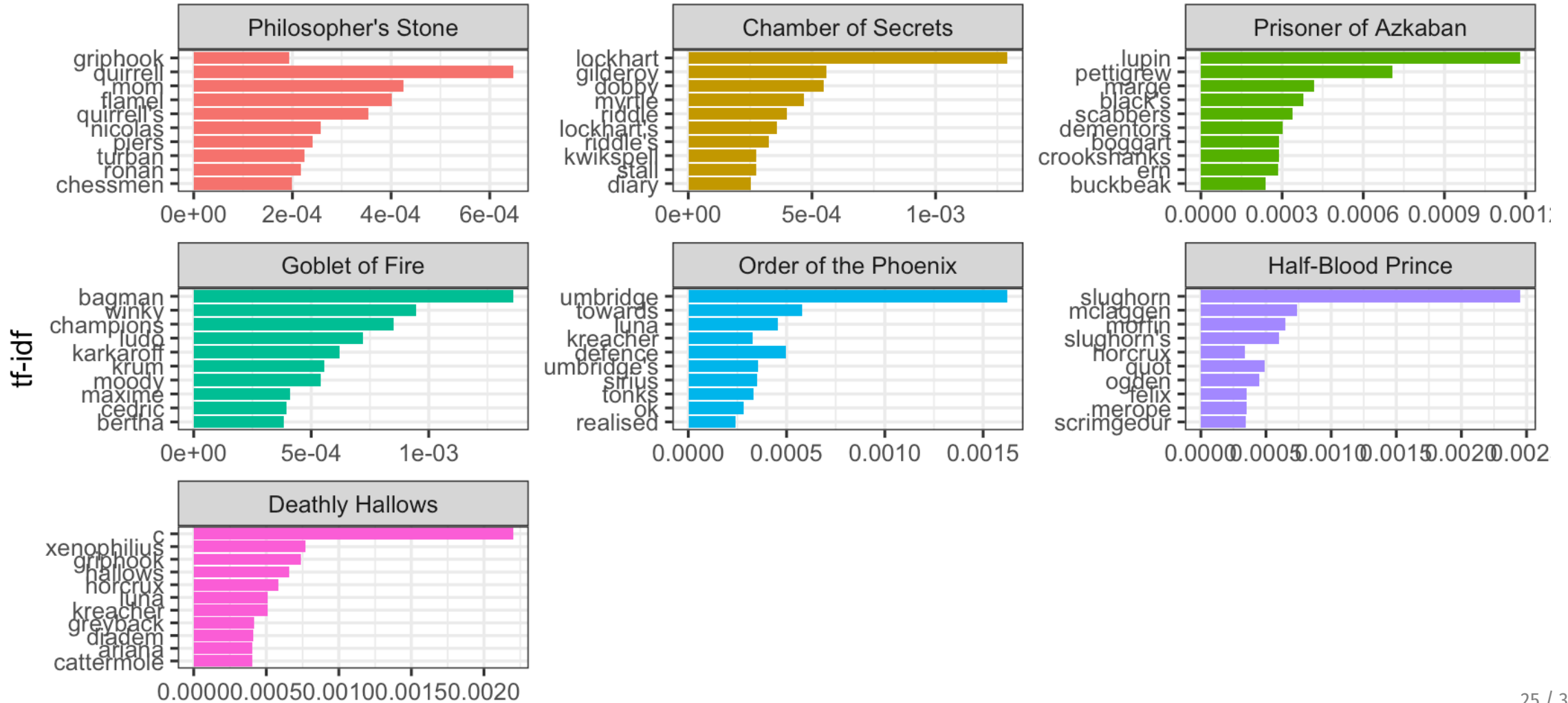
How important a term is compared to the rest of the documents

$$tf = \frac{n_{\text{term}}}{n_{\text{terms in document}}}$$

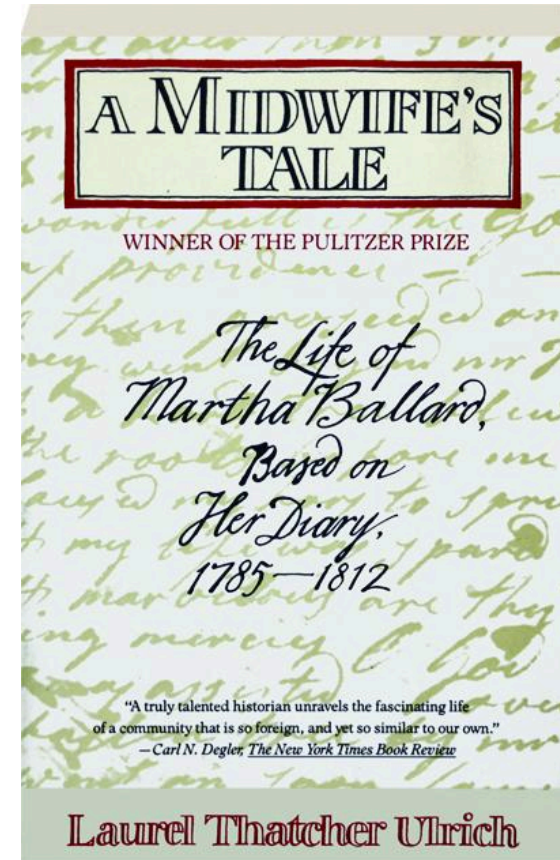
$$idf(\text{term}) = \ln \left(\frac{n_{\text{documents}}}{n_{\text{documents containing term}}} \right)$$

$$tf-idf(\text{term}) = tf(\text{term}) \times idf(\text{term})$$

tf-idf



Topic modeling



Latent Dirichlet Allocation (LDA)

Topics

egypt (p_w)
peopl (p_w)
egyptian (p_w)
...

protest (p_w)
tahrir_squar (p_w)
...

court (p_w)
right (p_w)
case (p_w)
...

constitu (p_w)
brotherhood (p_w)
...

militar (p_w)
scaf (p_w)
...

politic (p_w)
mubarak (p_w)
brotherhood (p_w)
...

Documents

Maspero interrogation continues, virginity checks case adjourned

December 0 / 2011 ,13 Comments / 3 Views

CAIRO: An investigations judge began Tuesday interrogating 29 defendants allegedly involved in the Maspero violence between Contic protesters and

army f

Abdel-

were d

Fattah

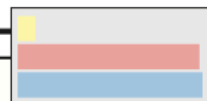
weapd

Egypt political forces call for mass 'Eyes of Freedom' rally Friday

Rejection of President Morsi's new constitutional declaration will likely take centre stage in planned Friday protests commemorating last year's clashes on Mohamed Mahmoud Street

Osman El Sharnoubi, Thursday 22 Nov 2012

President Mohamed Morsi's Thursday constitutional declaration has prompted Egyptian political forces that had been planning to commemorate last year's Mohamed Mahmoud clashes with mass protests on Friday to fine-tune their demands.

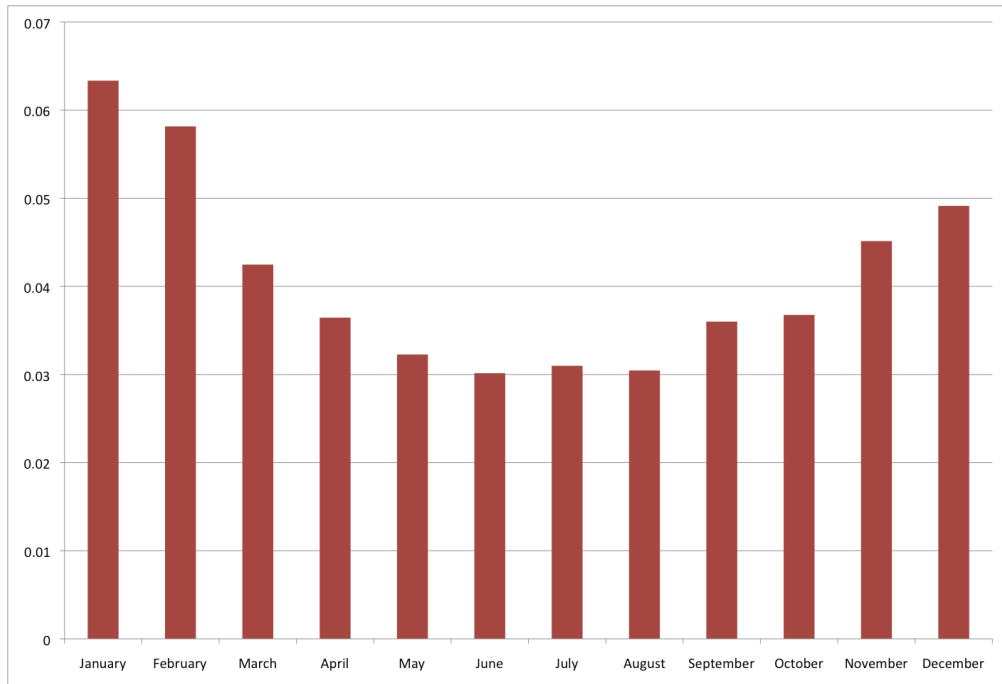


Clusters of related words

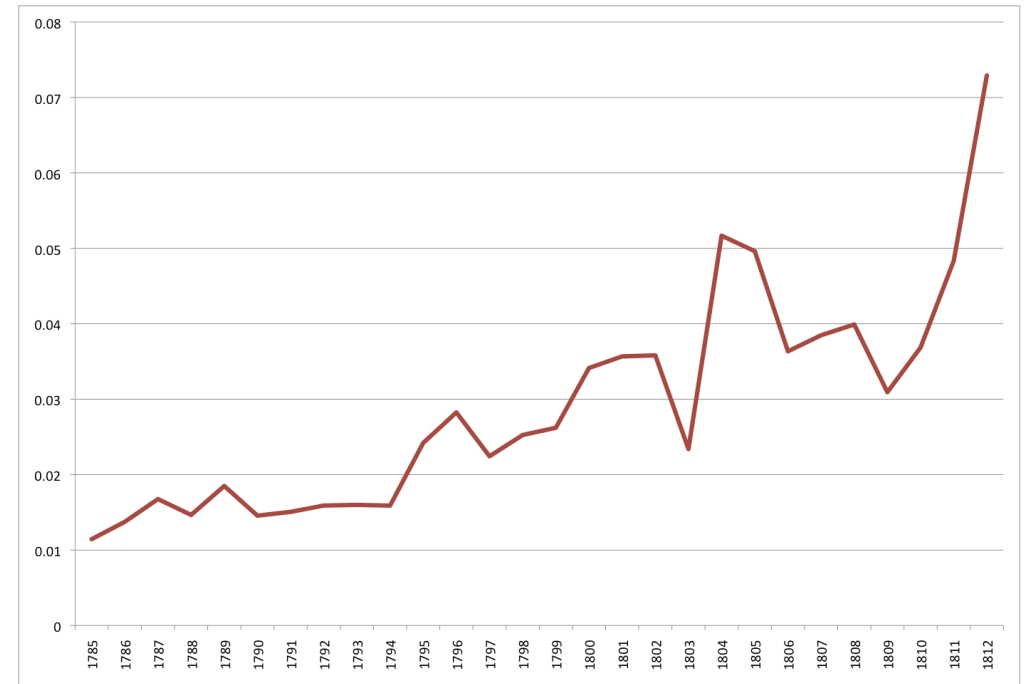
Topic label **Topic words**

Midwifery	birth safe morn receivd calld left cleverly pm labour ...
Church	meeting attended afternoon reverend worship ...
Death	day yesterday informd morn years death expired ...
Gardening	gardin sett worked clear beens corn warm planted ...
Shopping	lb made brot bot tea butter sugar carried ...
Illness	unwell sick gave dr rainy easier care head neighbor ...

Track topics over time

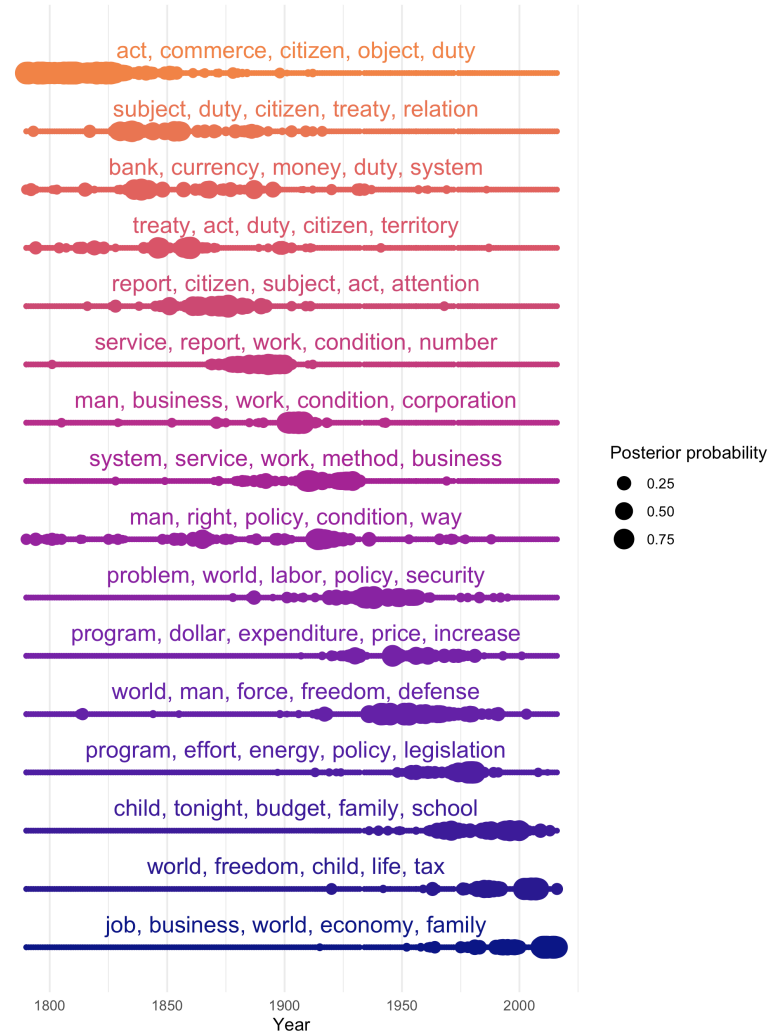


Cold weather topic by month



Emotion topic over time

State of the Union addresses



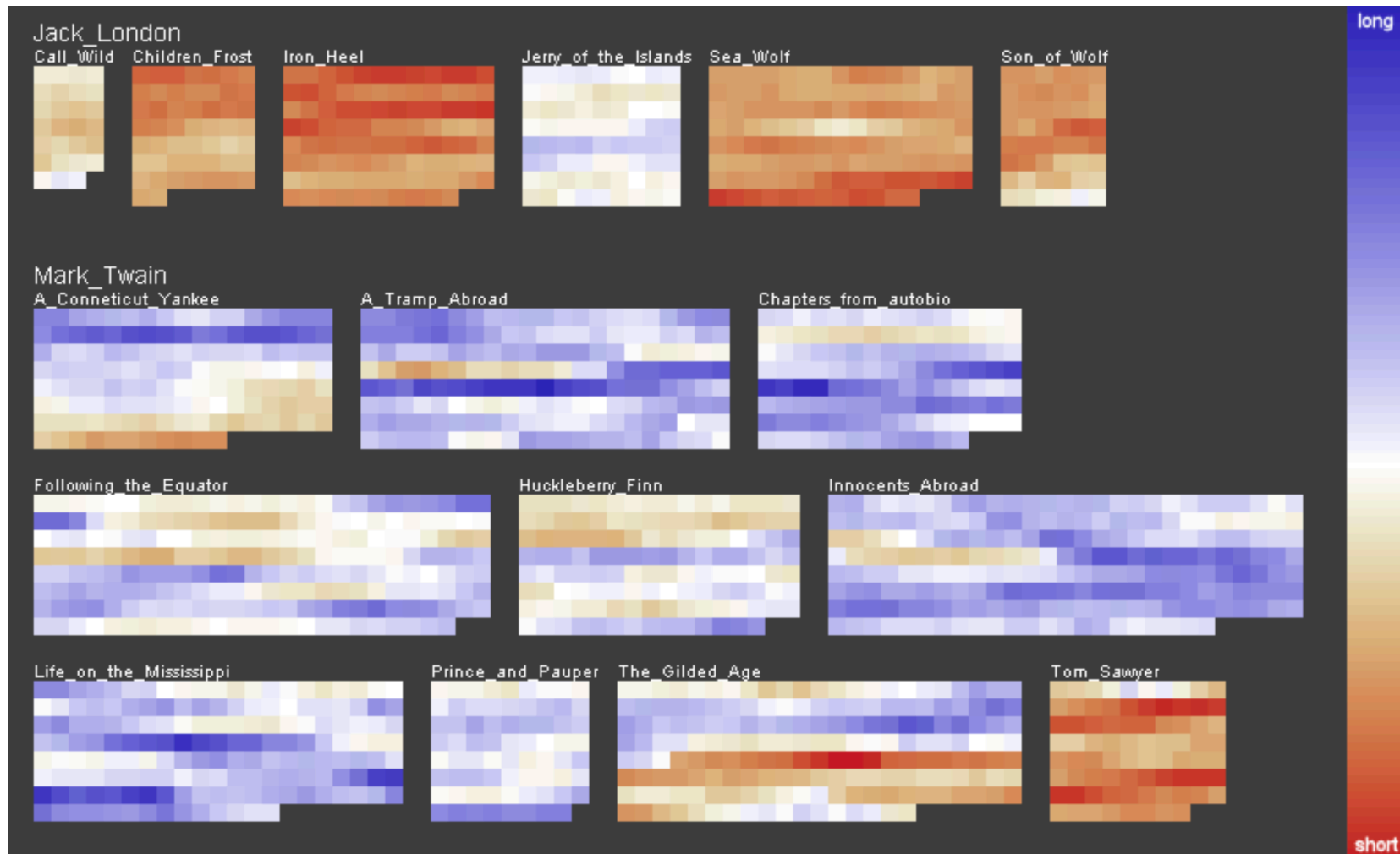
Fingerprinting

Analyze richness or uniqueness of a document

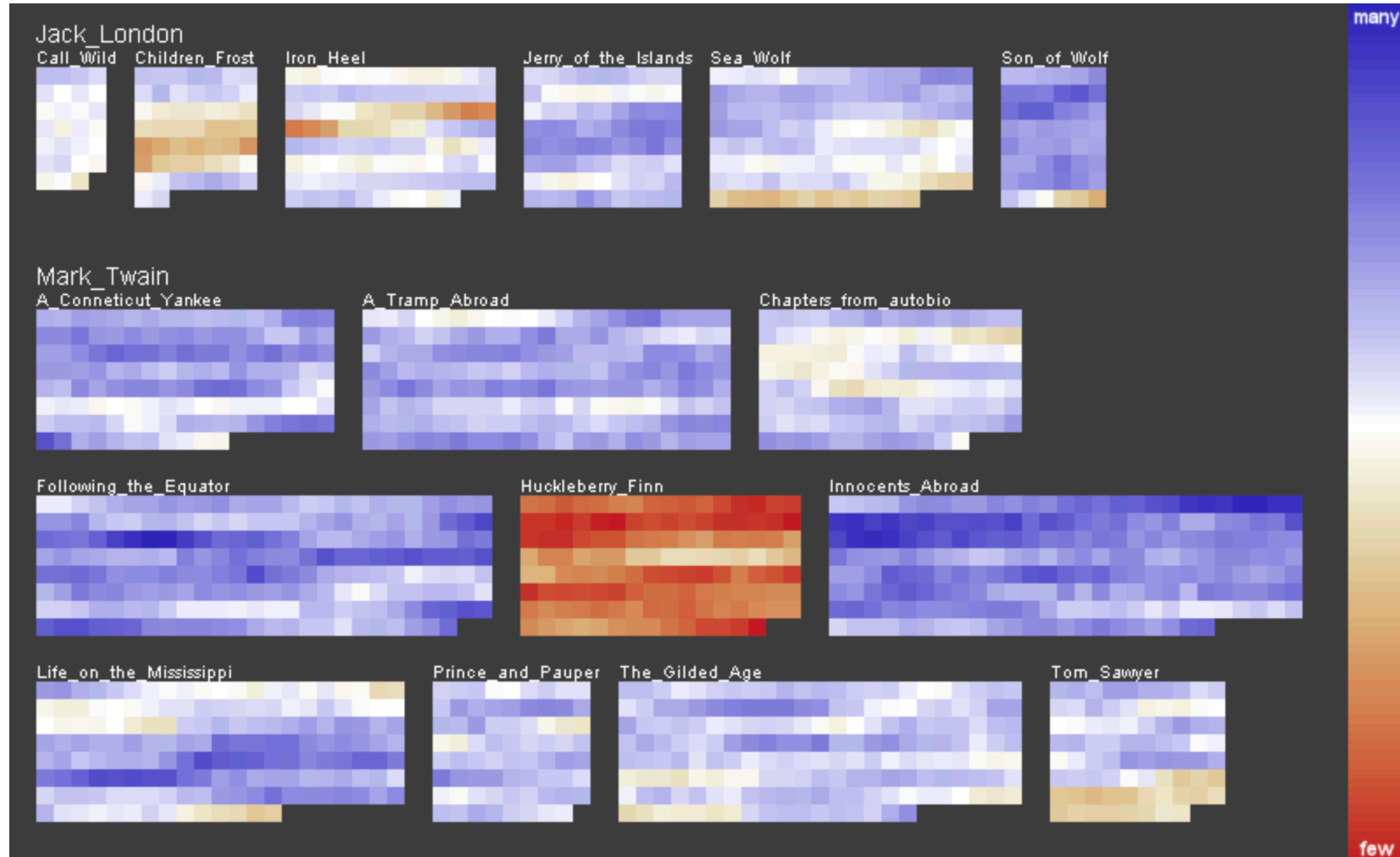
Punctuation patterns, vocabulary choices, sentence length

Hapax legomenon

Sentence length



Hapax legomena



Verse length

