

Text

Session 13

PMAP 8921: Data Visualization with R
Andrew Young School of Policy Studies
May 2020

Plan for today

Qualitative text-based data

**Crash course in
computational linguistics**

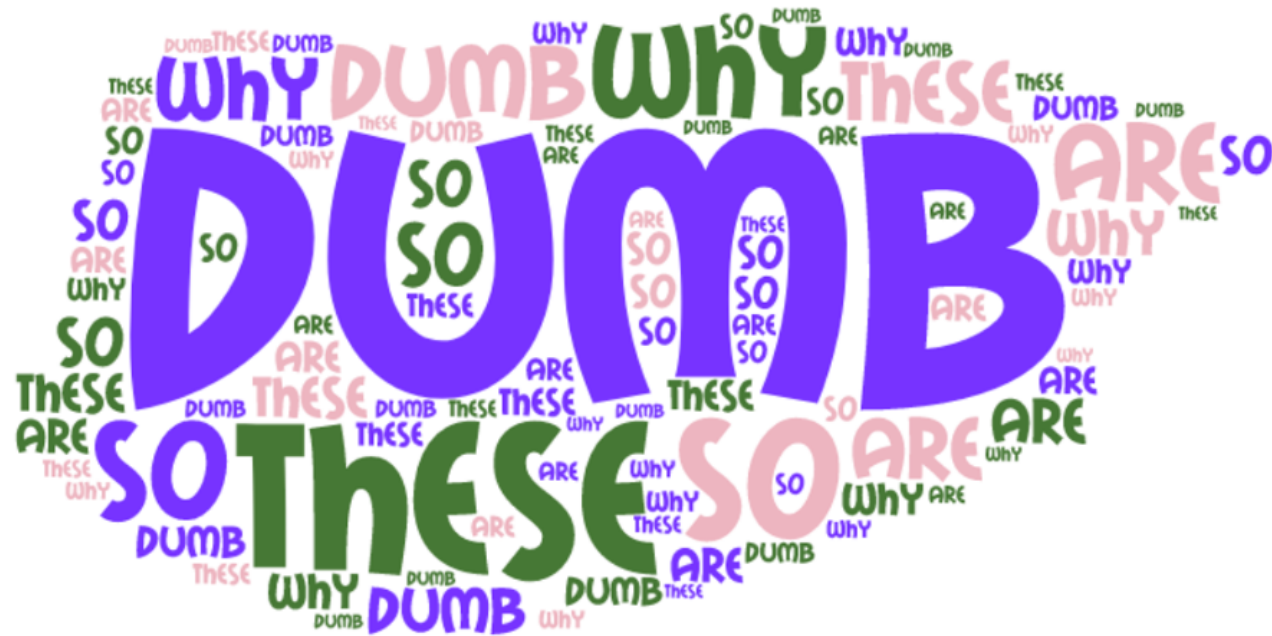
Qualitative text-based data

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?



Some cases are okay


400

What Happened

the result of a relentless barrage of political attacks and negative coverage. But I also know that it was my job to try to break through all that noise and convince the American people to vote for me. I wasn't able to do it.

What Americans Have Heard or Read About Donald Trump

What specifically do you recall reading, hearing or seeing about Donald Trump in the last day or two?




immigration speech convention make president people obama mexico wall email campaign

GALLUP DAILY TRACKING
JULY 17-SEPT 18, 2016

What Americans Have Heard or Read About Hillary Clinton

What specifically do you recall reading, hearing or seeing about Hillary Clinton in the last day or two?

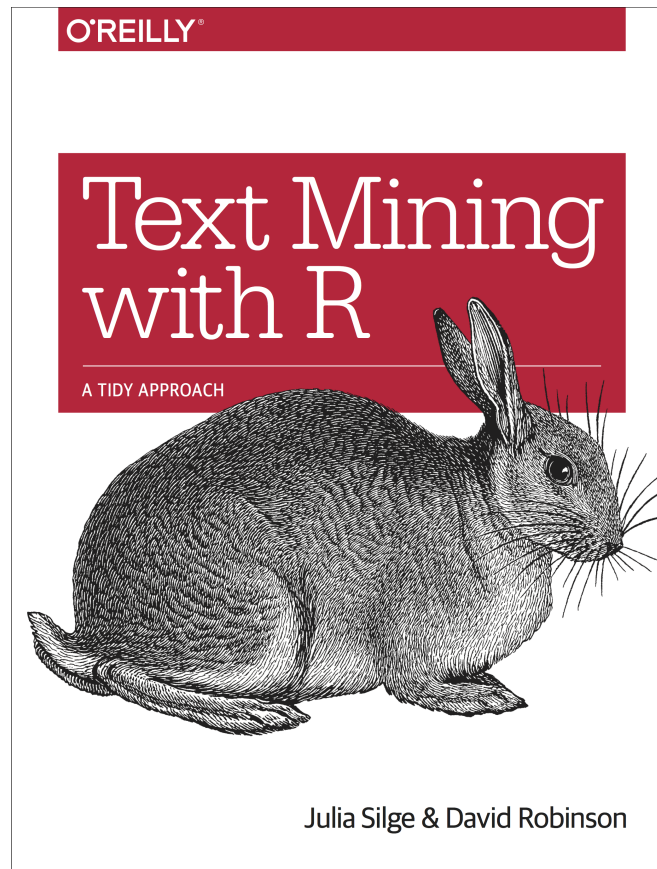


email foundation health convention scandal speech president campaign

GALLUP DAILY TRACKING
JULY 17-SEPT 18, 2016

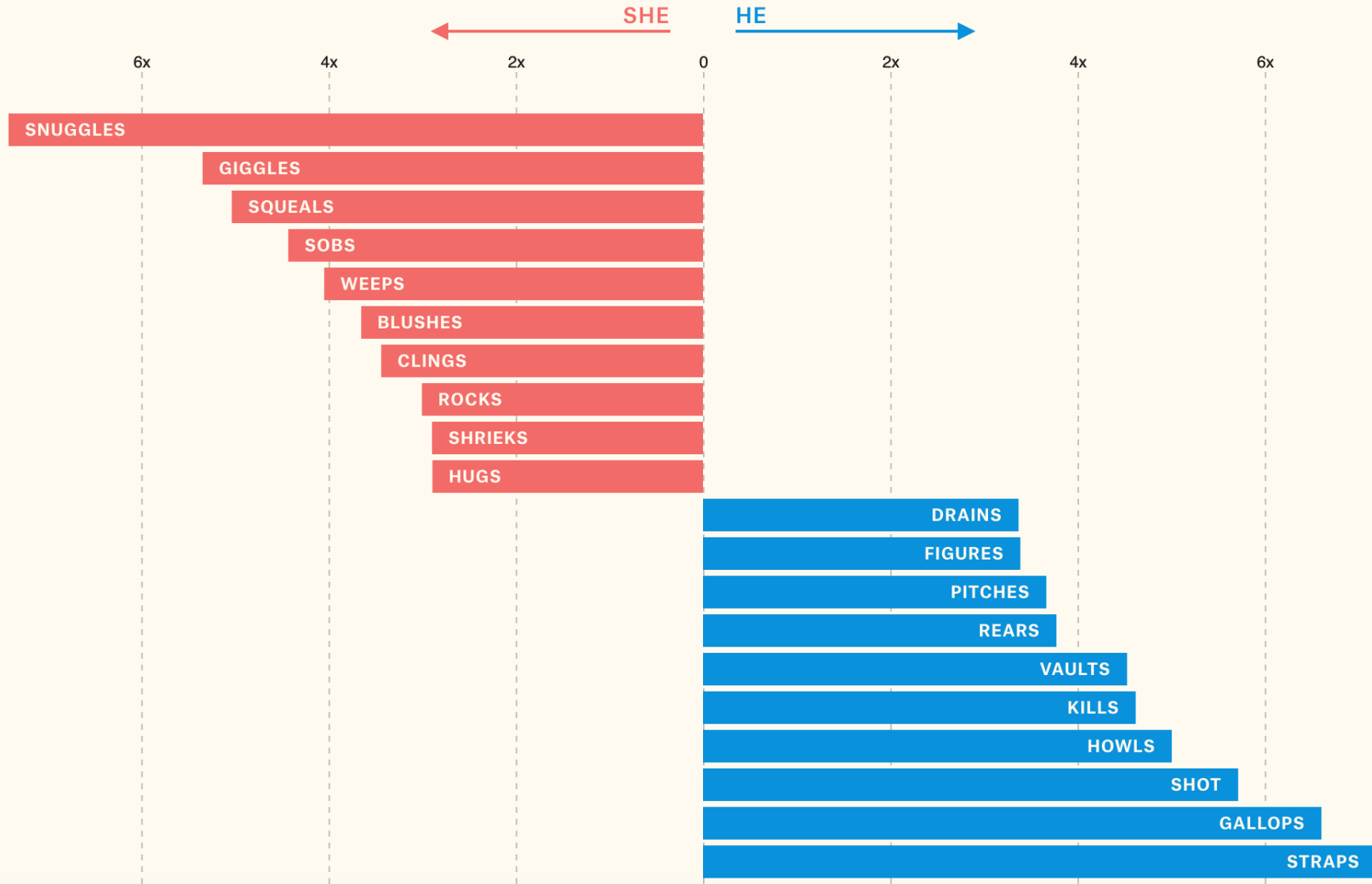
Word clouds for grownups

Count words, but in fancier ways

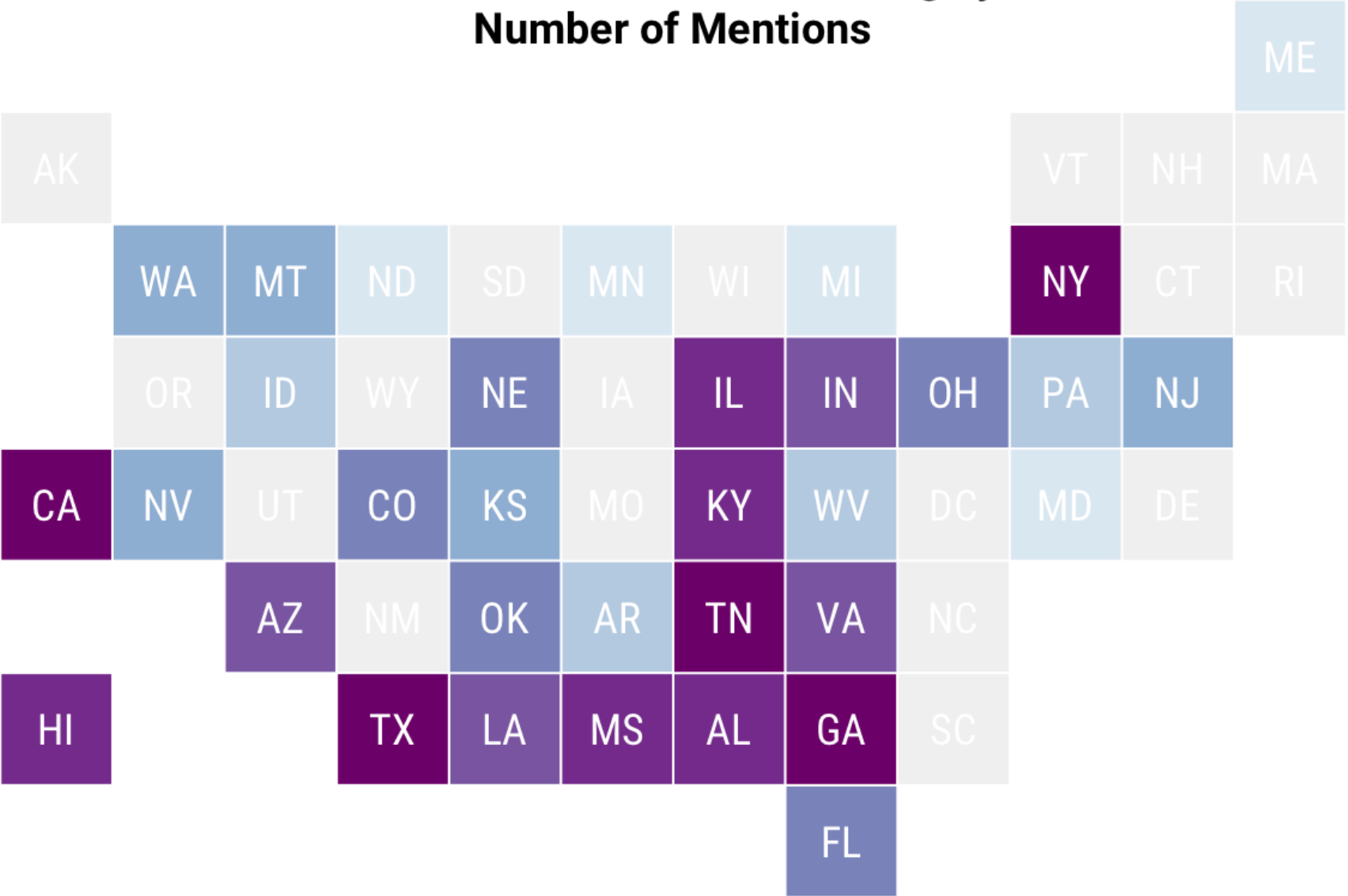


The most used words for women vs. men

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.

```
# A tibble: 6 x 3
  chapter book          text
  <int> <chr>          <chr>
1     1 Harry Potter and the Phil... "THE BOY WHO LIVED    Mr. and Mrs. Dursley, of number ...
2     2 Harry Potter and the Phil... "THE VANISHING GLASS  Nearly ten years had passed si...
3     3 Harry Potter and the Phil... "THE LETTERS FROM NO ONE    The escape of the Brazilia...
4     4 Harry Potter and the Phil... "THE KEEPER OF THE KEYS    BOOM. They knocked again. D...
5     5 Harry Potter and the Phil... "DIAGON ALLEY    Harry woke early the next morning. Al...
6     6 Harry Potter and the Phil... "THE JOURNEY FROM PLATFORM NINE AND THREE-QUARTERS    ...
```

Tokens

Split the text into even smaller parts

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

A tibble: 6 x 3

word	chapter	book
<chr>	<int>	<chr>
1 the	1	Harry Potter...
2 boy	1	Harry Potter...
3 who	1	Harry Potter...
4 lived	1	Harry Potter...
5 mr	1	Harry Potter...
6 and	1	Harry Potter...

A tibble: 6 x 3

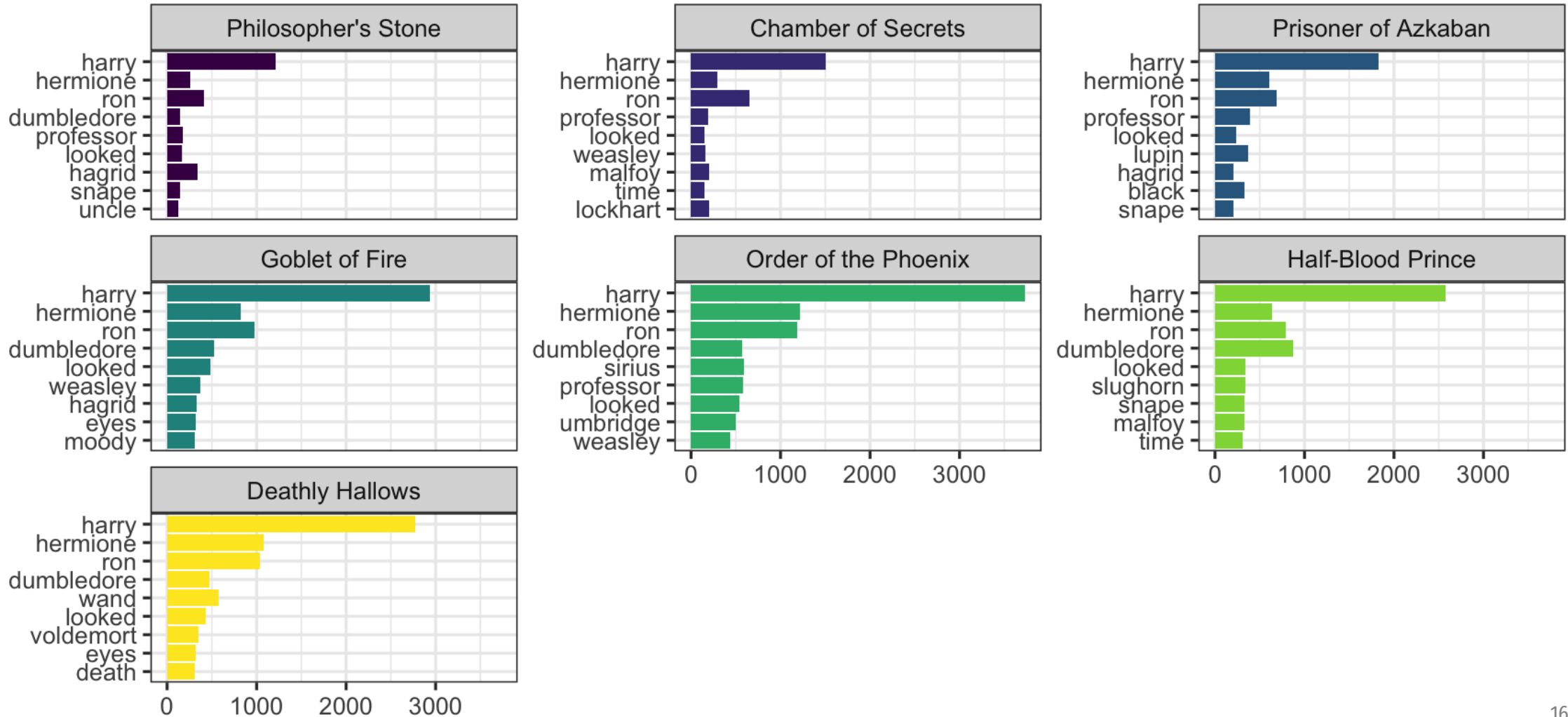
bigram	chapter	book
<chr>	<int>	<chr>
1 the boy	1	Harry Potter...
2 boy who	1	Harry Potter...
3 who lived	1	Harry Potter...
4 lived mr	1	Harry Potter...
5 mr and	1	Harry Potter...
6 and mrs	1	Harry Potter...

Stop words

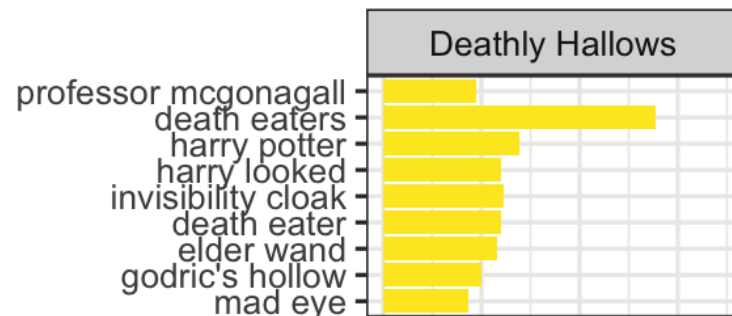
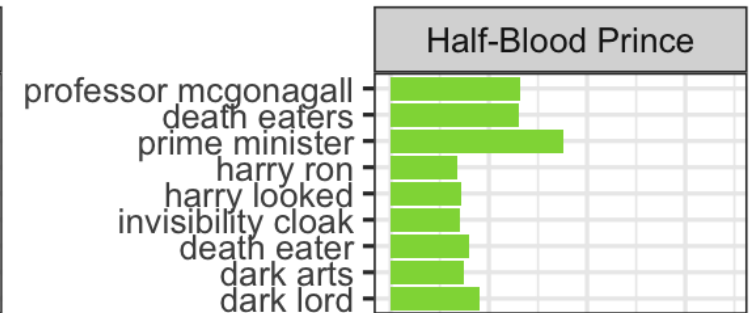
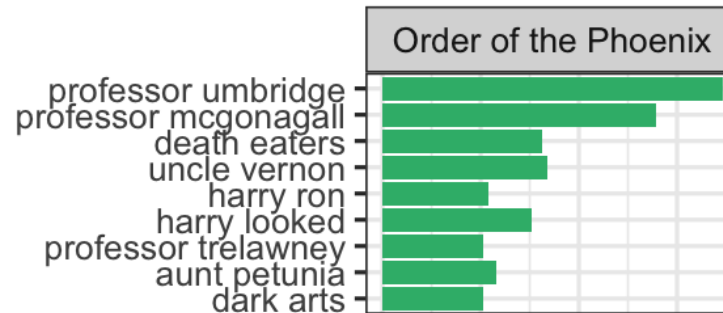
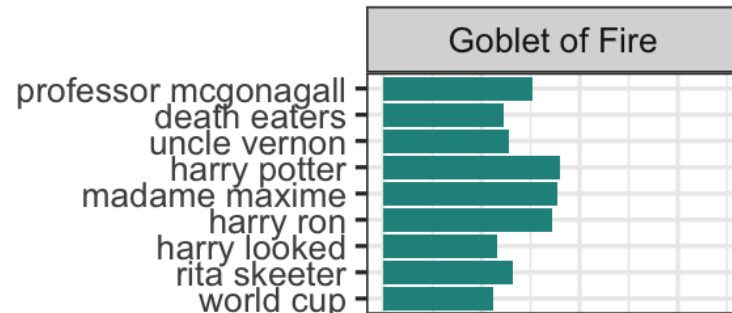
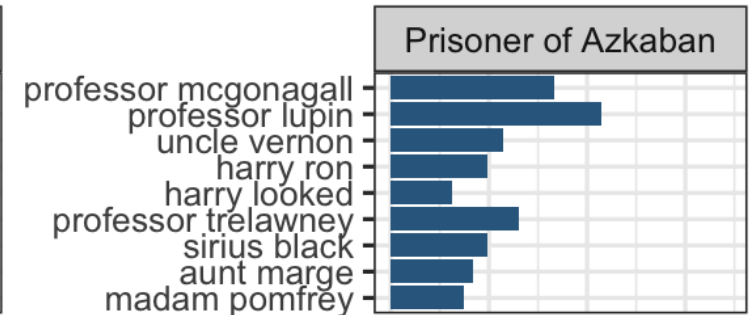
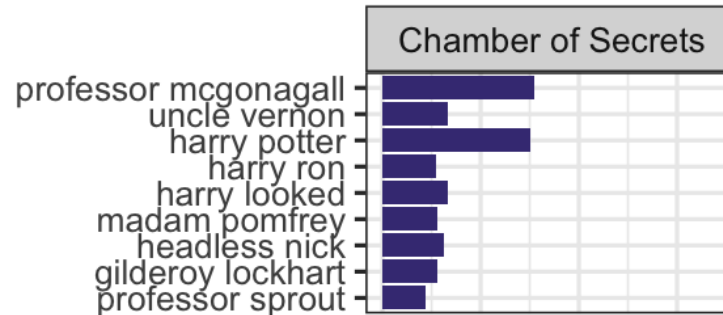
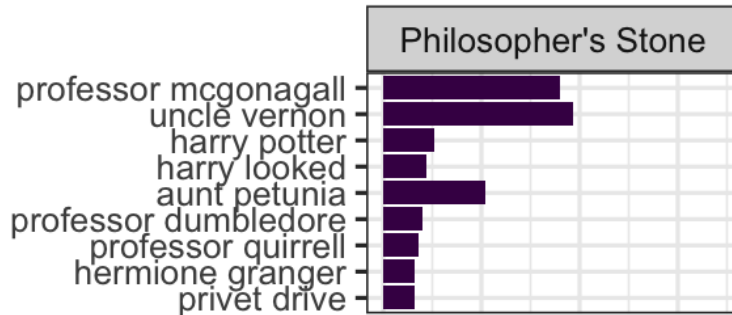
Common words that we can generally ignore

```
# A tibble: 1,149 x 2
  word      lexicon
  <chr>    <chr>
1 a        SMART
2 a's     SMART
3 able    SMART
4 about   SMART
5 above   SMART
6 according SMART
7 accordingly SMART
8 across  SMART
9 actually SMART
10 after   SMART
# ... with 1,139 more rows
```

Token frequency: words



Token frequency: n-grams

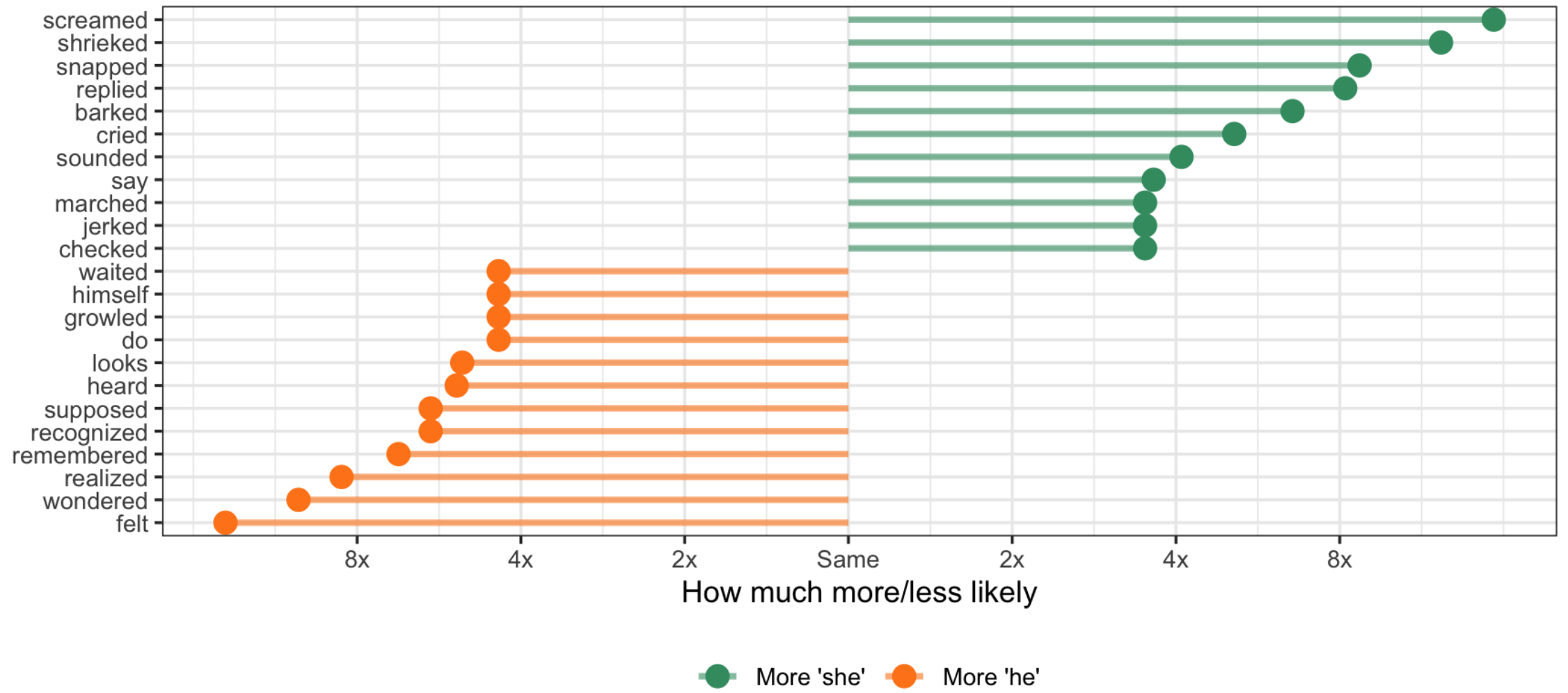


0 50 100 150

0 50 100 150

0 50 100 150

Token frequency: n-gram ratios



Parts of speech

```
# A tibble: 50 x 11
```

```
  doc_id  sid  tid token  token_with_ws lemma  upos  xpos  feats  tid_source relation
  <dbl> <dbl> <dbl> <chr>  <chr>          <chr> <chr> <chr> <chr>  <dbl> <chr>
1     1    1    1  THE    THE            the   DET  DT    Definite...  2 det
2     1    1    2  BOY    BOY            Boy   NOUN  NN    Number=S... 18 nsubj
3     1    1    3  WHO    WHO            who   PRON  WP    PronType...  4 nsubj
4     1    1    4  LIVED  LIVED          live  VERB  VBD    Mood=Ind...  2 acl:rel...
5     1    1    5  Mr.    Mr.            Mr.   PROPN NNP    Number=S...  4 xcomp
6     1    1    6  and    and            and   CCONJ CC    <NA>         7 cc
7     1    1    7  Mrs.   Mrs.           Mrs.  PROPN NNP    Number=S...  5 conj
8     1    1    8  Dursl... Dursley       Durs... PROPN NNP    Number=S...  7 flat
9     1    1    9  ,      ,              ,     PUNCT ,      <NA>         5 punct
10    1    1   10  of     of             of    ADP   IN    <NA>         11 case
```

```
# ... with 40 more rows
```

These use the **Penn part of speech tags**

Parts of speech frequency

Verbs

```
# A tibble: 1,557 x 2
  lemma      n
  <chr> <dbl>
1 say      920
2 get      440
3 have     417
4 go       384
5 look     380
6 be       310
7 know     310
8 see      303
9 think    230
10 do      227
# ... with 1,547 more rows
```

Nouns

```
# A tibble: 2,852 x 2
  lemma      n
  <chr>      <dbl>
1 Harry     1315
2 Ron       423
3 Hagrid    258
4 Professor 167
5 Snape     154
6 Hermione  153
7 Dumbledore 144
8 time      138
9 Dudley    136
10 uncle     122
# ... with 2,842 more rows
```

Adjectives & adverbs

```
# A tibble: 1,240 x 2
  lemma      n
  <chr> <dbl>
1 back    223
2 so      215
3 just    180
4 when    178
5 very    171
6 now     166
7 then    165
8 all     147
9 how     136
10 there   123
# ... with 1,230 more rows
```

Artsy stuff

Sentiment analysis

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

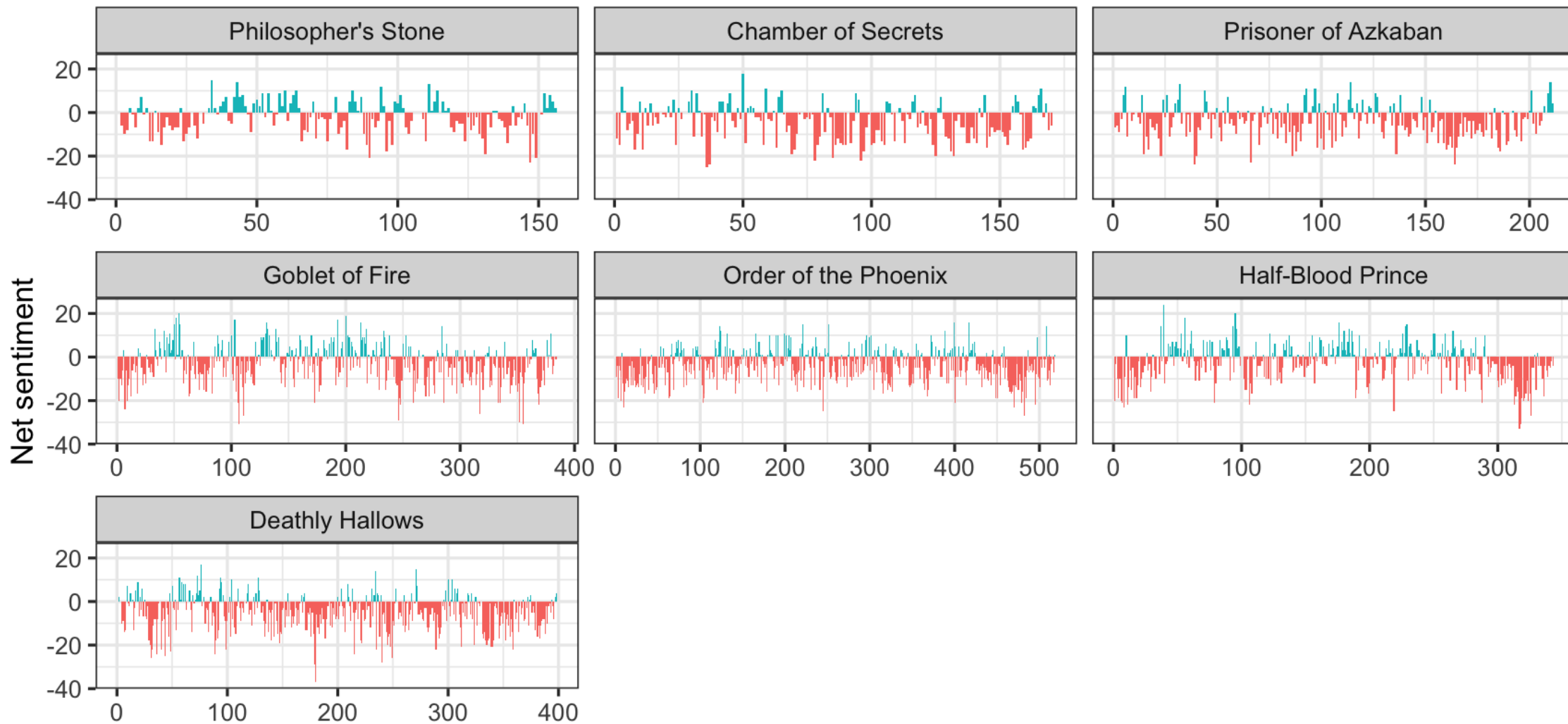
```
# A tibble: 6,786 x 2
  word      sentiment
  <chr>    <chr>
1 2-faces  negative
2 abnormal negative
3 abolish negative
4 abominable negative
5 abominably negative
6 abominate negative
7 abomination negative
8 abort    negative
9 aborted  negative
10 aborts  negative
# ... with 6,776 more rows
```

```
get_sentiments("afinn")
```

```
# A tibble: 2,477 x 2
  word      value
  <chr>    <dbl>
1 abandon  -2
2 abandoned -2
3 abandons -2
4 abducted -2
5 abduction -2
6 abductions -2
7 abhor    -3
8 abhorred -3
9 abhorrent -3
10 abhors  -3
# ... with 2,467 more rows
```

```
get_sentiments("nrc")
```

```
# A tibble: 13,901 x 2
  word      sentiment
  <chr>    <chr>
1 abacus   trust
2 abandon  fear
3 abandon  negative
4 abandon  sadness
5 abandoned anger
6 abandoned fear
7 abandoned negative
8 abandoned sadness
9 abandonment anger
10 abandonment fear
# ... with 13,891 more rows
```



tf-idf

Term frequency-inverse document frequency

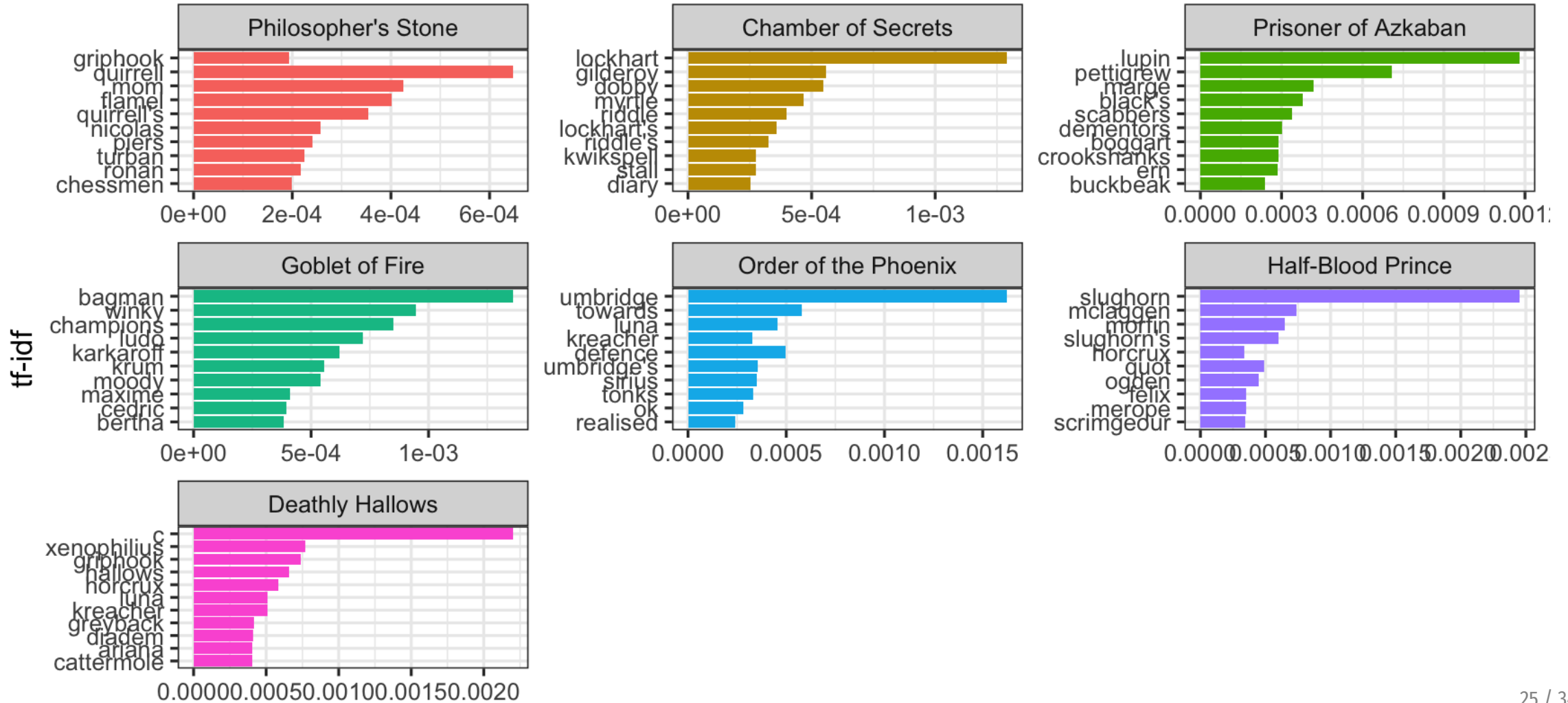
How important a term is compared to the rest of the documents

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

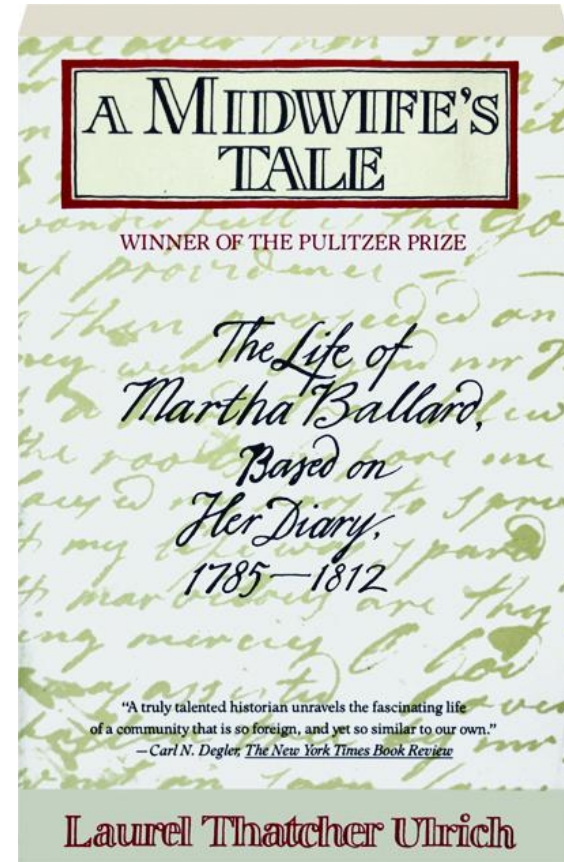
$$idf(\text{term}) = \ln \left(\frac{n_{\text{documents}}}{n_{\text{documents containing term}}} \right)$$

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

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 Coptic 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

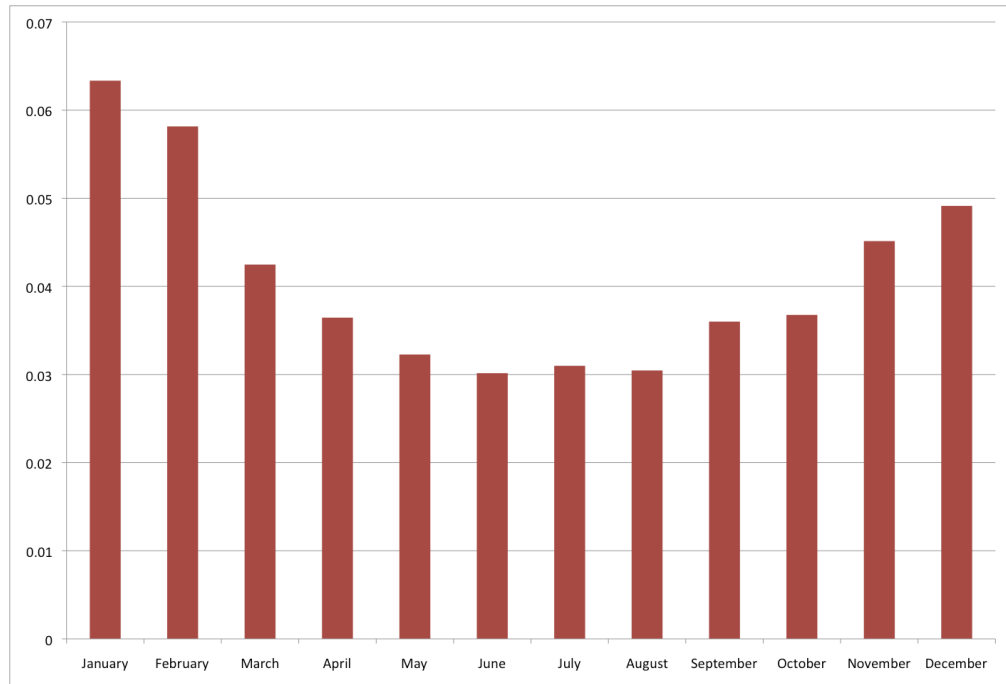
President Mohamed Morsi's Thursday constitutional declaration has prompted Egyptian political forces that had been planning to commemorate last year's Mohamed Mahmoud clashes with mass protests on Friday to fine-tune their demands.

Clusters of related words

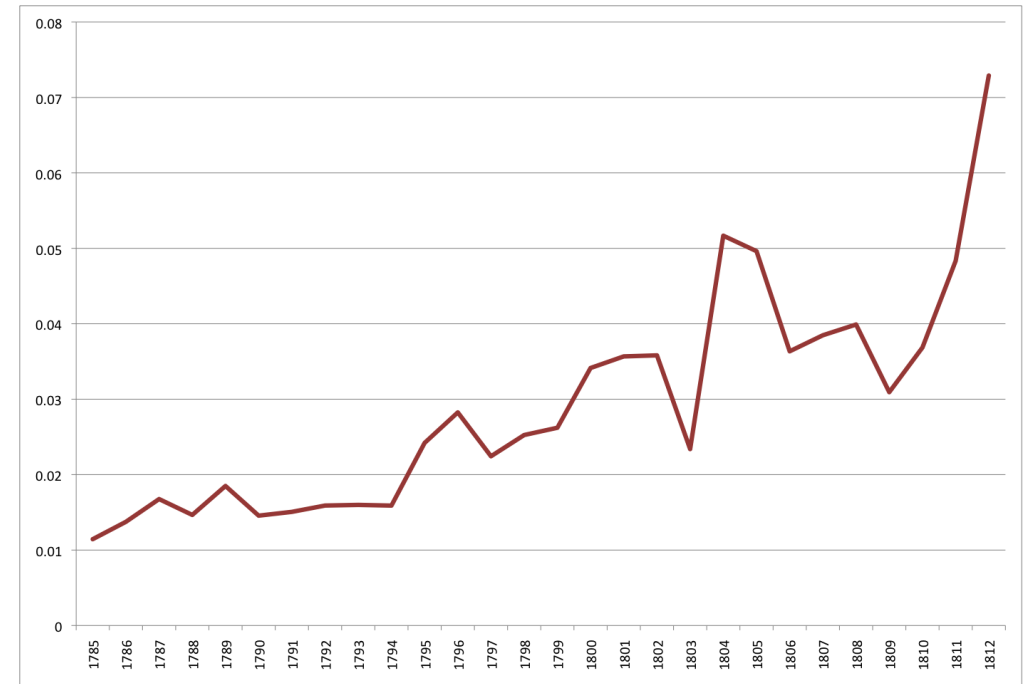
Topic label **Topic words**

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

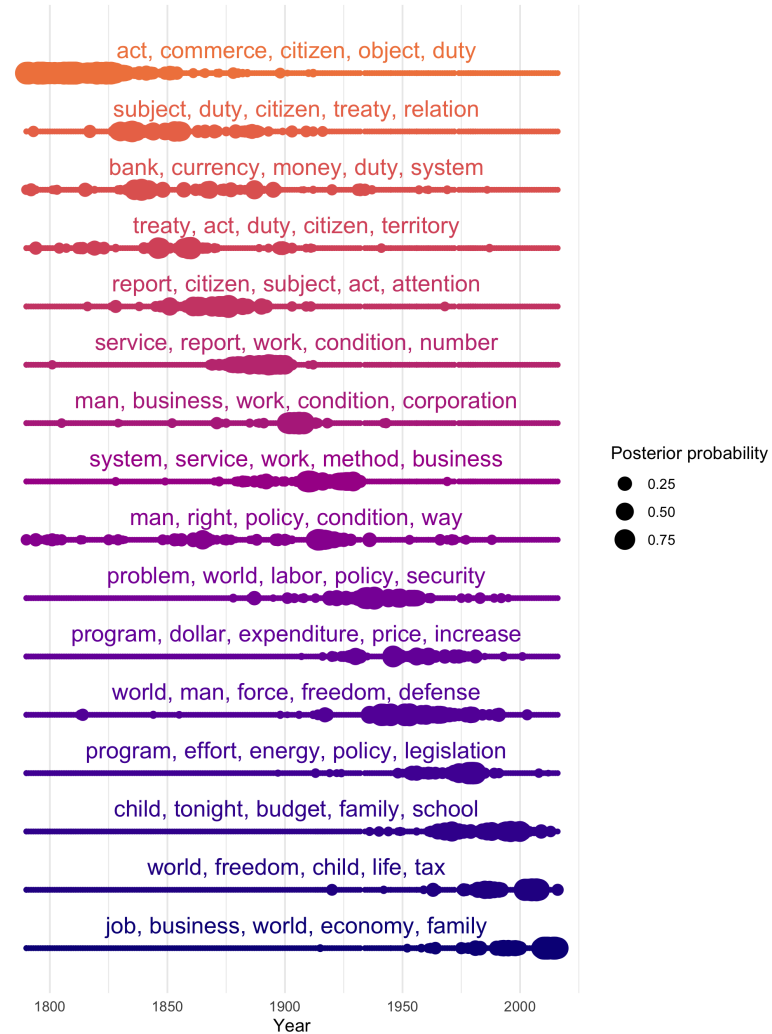


Cold weather topic by month



Emotion topic over time

State of the Union addresses



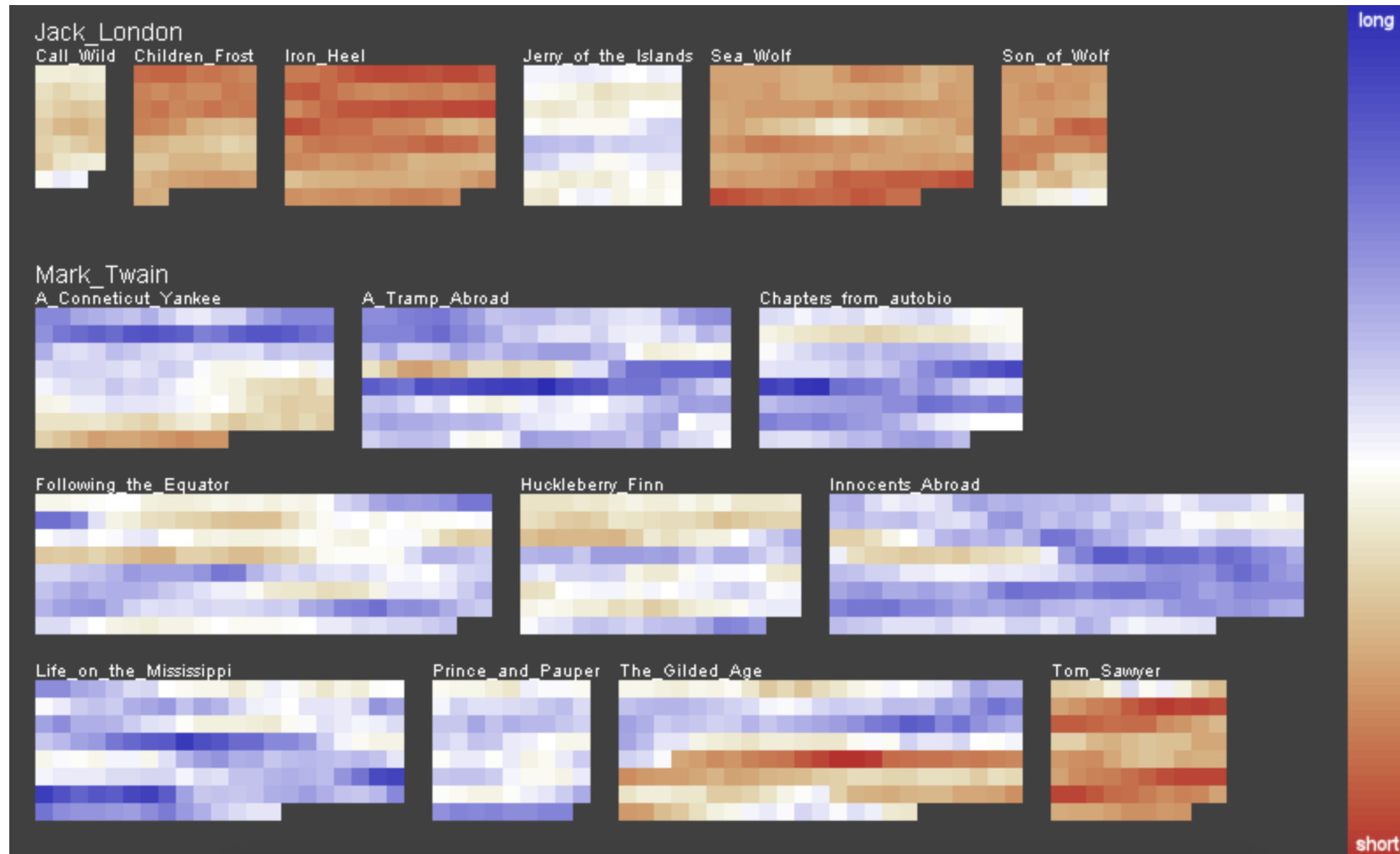
Fingerprinting

Analyze richness or uniqueness of a document

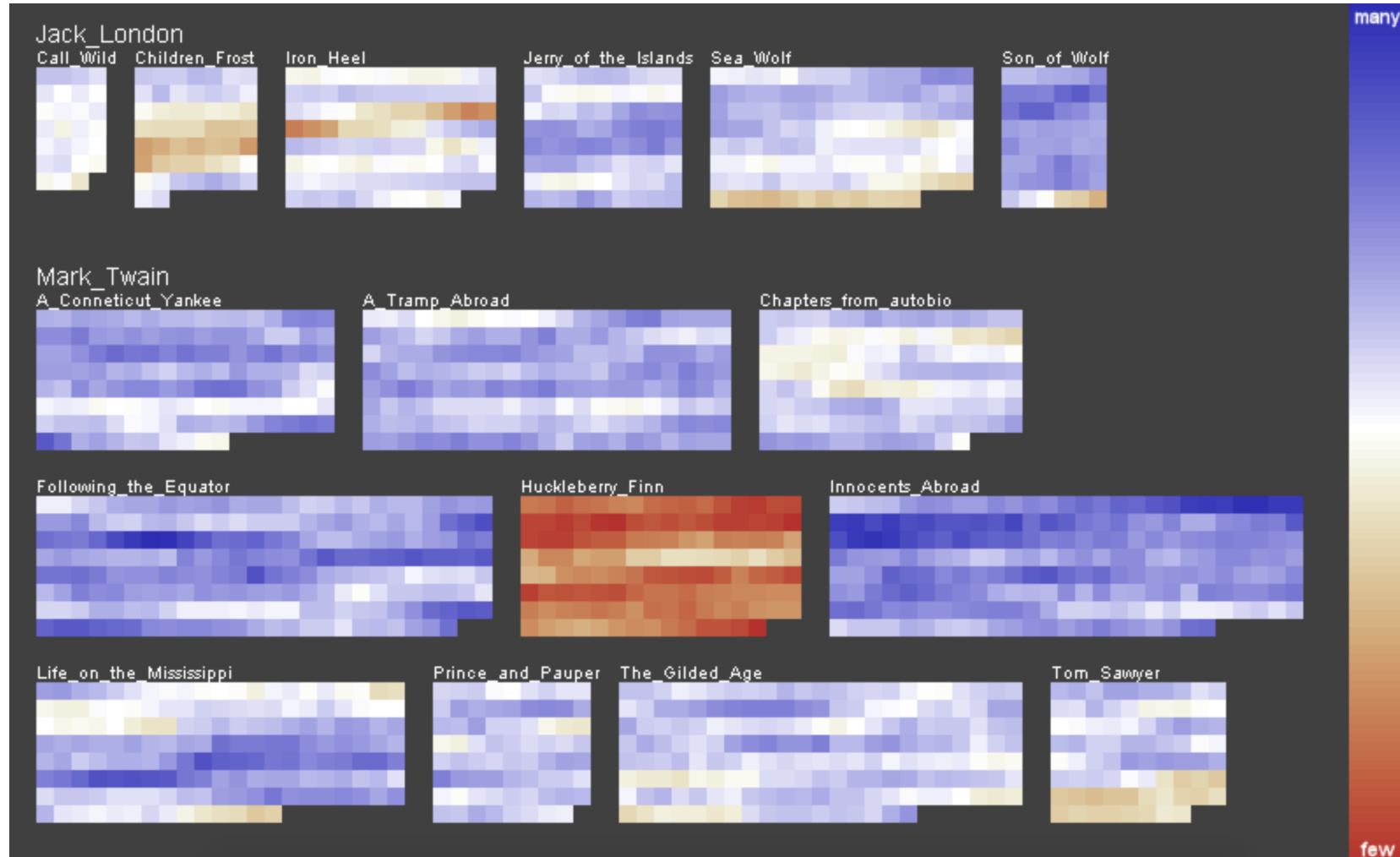
Punctuation patterns, vocabulary choices, sentence length

Hapax legomenon

Sentence length



Hapax legomena



Verse length

