



SENTIMENT ANALYSIS IN R: THE TIDY WAY

Tidying Shakespeare

Julia Silge

Data Scientist at Stack Overflow



Six Shakespearean plays

shakespeare

- title, the title of a Shakespearean play
- type, the type of play, either tragedy or comedy
- text, a line from that play

Texts available from [Project Gutenberg](#)



Six Shakespearean plays

```
> shakespeare %>%  
  count(type, title)
```

```
# A tibble: 6 x 3
```

	type	title	n
	<chr>	<chr>	<int>
1	Comedy	A Midsummer Night's Dream	3459
2	Comedy	Much Ado about Nothing	3799
3	Comedy	The Merchant of Venice	4225
4	Tragedy	Hamlet, Prince of Denmark	6776
5	Tragedy	The Tragedy of Macbeth	3188
6	Tragedy	The Tragedy of Romeo and Juliet	4441



Learning how to tidy text

```
> hamlet
# A tibble: 6,776 x 3
  title    type    text
  <chr>   <chr>   <chr>
1 Hamlet, Prince of Denmark Tragedy HAMLET, PRINCE OF DENMARK
2 Hamlet, Prince of Denmark Tragedy
3 Hamlet, Prince of Denmark Tragedy by William Shakespeare
4 Hamlet, Prince of Denmark Tragedy
5 Hamlet, Prince of Denmark Tragedy
6 Hamlet, Prince of Denmark Tragedy
7 Hamlet, Prince of Denmark Tragedy
8 Hamlet, Prince of Denmark Tragedy PERSONS REPRESENTED.
9 Hamlet, Prince of Denmark Tragedy
10 Hamlet, Prince of Denmark Tragedy Claudius, King of Denmark.
# ... with 6,766 more rows
```



Learning how to tidy text

```
> library(tidytext)
>
> hamlet %>%
  unnest_tokens(word, text)
# A tibble: 32,068 x 3
   title      type      word
   <chr>    <chr>    <chr>
1 Hamlet, Prince of Denmark Tragedy hamlet
2 Hamlet, Prince of Denmark Tragedy prince
3 Hamlet, Prince of Denmark Tragedy of
4 Hamlet, Prince of Denmark Tragedy denmark
5 Hamlet, Prince of Denmark Tragedy by
6 Hamlet, Prince of Denmark Tragedy william
7 Hamlet, Prince of Denmark Tragedy shakespeare
8 Hamlet, Prince of Denmark Tragedy persons
9 Hamlet, Prince of Denmark Tragedy represented
10 Hamlet, Prince of Denmark Tragedy claudius
# ... with 32,058 more rows
```



Tokenization

In our case,

TOKEN = SINGLE WORD



What did `unnest_tokens()` do?

- Other columns have been retained
- Punctuation has been stripped
- Words have been converted to lower-case

```
> library(tidytext)
>
> hamlet %>%
  unnest_tokens(word, text)
# A tibble: 32,068 x 3
   title      type      word
   <chr>    <chr>    <chr>
1 Hamlet, Prince of Denmark Tragedy hamlet
2 Hamlet, Prince of Denmark Tragedy prince
3 Hamlet, Prince of Denmark Tragedy of
4 Hamlet, Prince of Denmark Tragedy denmark
5 Hamlet, Prince of Denmark Tragedy by
6 Hamlet, Prince of Denmark Tragedy william
7 Hamlet, Prince of Denmark Tragedy shakespeare
8 Hamlet, Prince of Denmark Tragedy persons
9 Hamlet, Prince of Denmark Tragedy represented
10 Hamlet, Prince of Denmark Tragedy claudius
# ... with 32,058 more rows
```



SENTIMENT ANALYSIS IN R: THE TIDY WAY

The game's afoot!



SENTIMENT ANALYSIS IN R: THE TIDY WAY

Using count and mutate

Julia Silge

Data Scientist at Stack Overflow

Positive and negative words

```
> shakespeare_sentiment %>%  
  count(title, sentiment)
```

```
# A tibble: 12 x 3
```

```
      title sentiment      n  
    <chr>    <chr> <int>
```

1	A Midsummer Night's Dream	negative	681
2	A Midsummer Night's Dream	positive	773
3	Hamlet, Prince of Denmark	negative	1323
4	Hamlet, Prince of Denmark	positive	1223
5	Much Ado about Nothing	negative	767
6	Much Ado about Nothing	positive	1127
7	The Merchant of Venice	negative	740
8	The Merchant of Venice	positive	962
9	The Tragedy of Macbeth	negative	914
10	The Tragedy of Macbeth	positive	749
11	The Tragedy of Romeo and Juliet	negative	1235
12	The Tragedy of Romeo and Juliet	positive	1090



Defining a new column

Using `mutate()`:

```
mutate(total = sum(n),  
       percent = n / total)
```

Looking at contributions by word

```
> hamlet_sentiment %>%  
  count(word, sentiment, sort = TRUE)
```

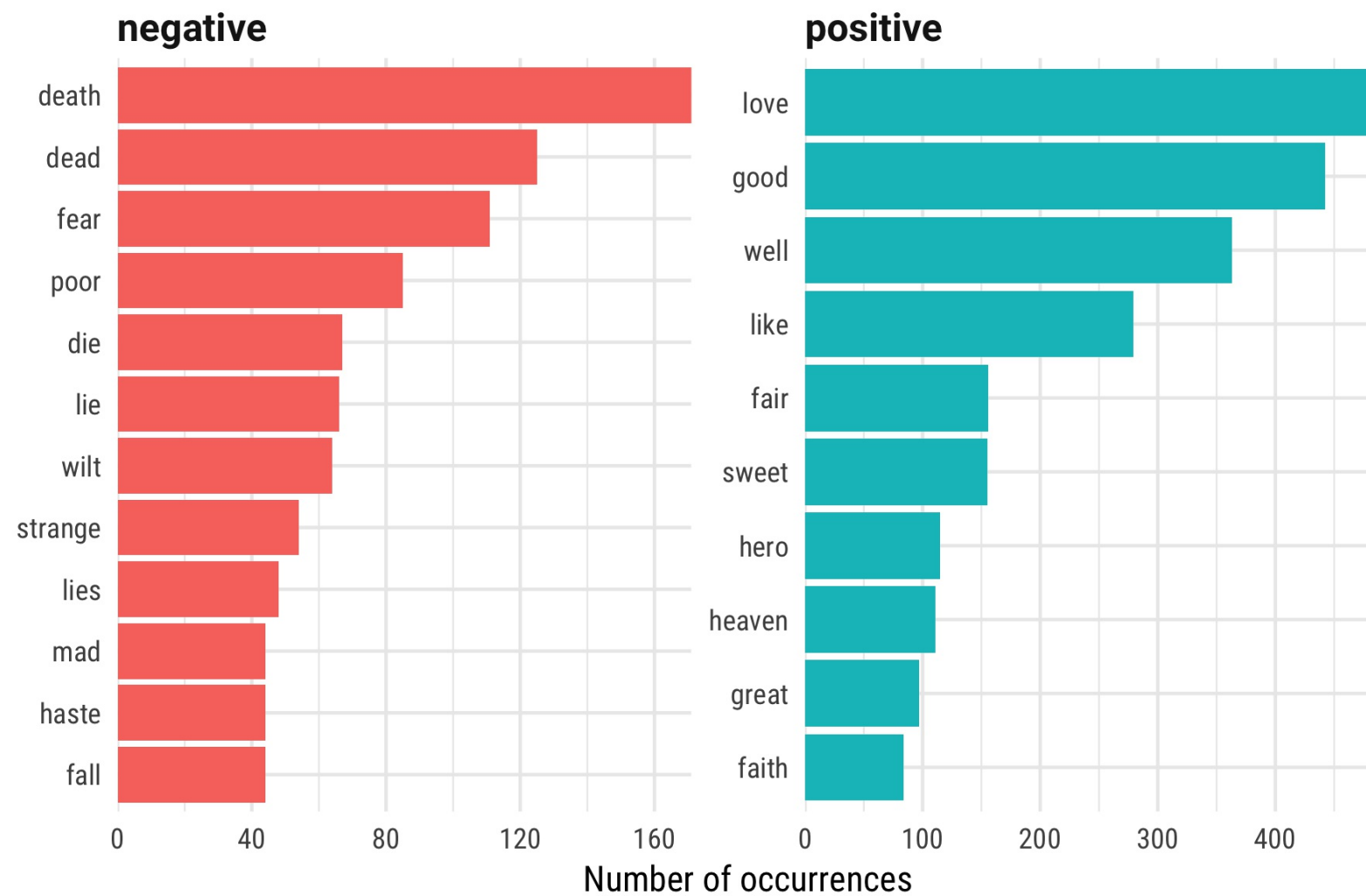
```
# A tibble: 823 x 3
```

	word	sentiment	n
	<chr>	<chr>	<int>
1	good	positive	109
2	like	positive	84
3	well	positive	78
4	love	positive	68
5	heaven	positive	44
6	death	negative	38
7	mar	negative	35
8	dead	negative	33
9	great	positive	26
10	sweet	positive	26

```
# ... with 813 more rows
```



Visualizing word contributions





SENTIMENT ANALYSIS IN R: THE TIDY WAY

**Once more unto the
breach!**



SENTIMENT ANALYSIS IN R: THE TIDY WAY

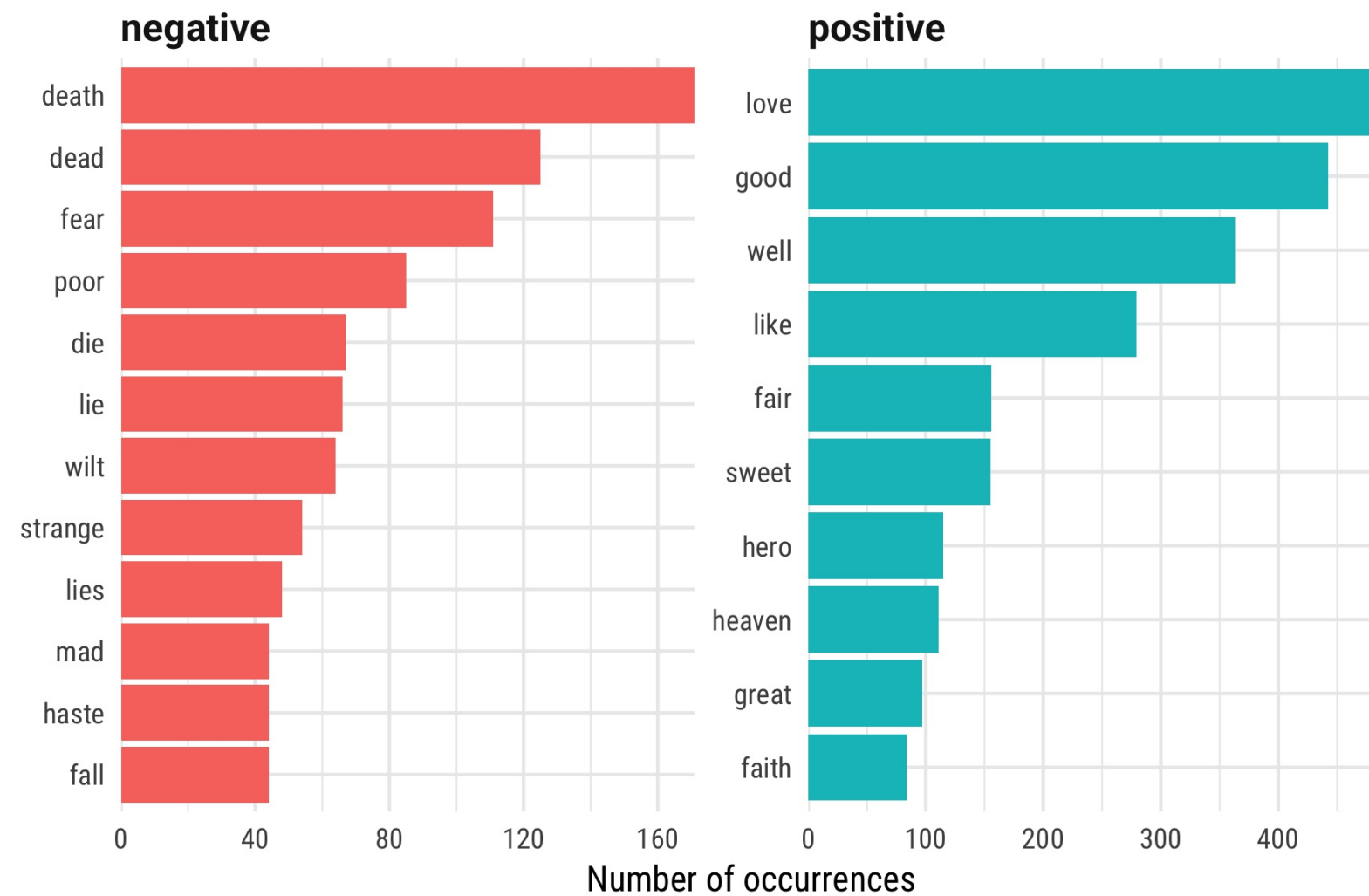
Sentiment contributions by individual words

Julia Silge

Data Scientist at Stack Overflow



What words contribute to sentiment scores?





Back to the sentiment lexicons

```
> get_sentiments("bing") %>%  
  filter(word == "wilt")
```

```
# A tibble: 1 x 2  
  word sentiment  
  <chr>    <chr>  
1 wilt negative
```

Think what thou wilt

```
> library(stringr)
>
> shakespeare %>%
  filter(str_detect(text, "wilt")) %>%
  select(text)

# A tibble: 52 x 1
      text
  <chr>
1      Take it in what sense thou wilt.
2      Which thou wilt propagate, to have it prest
3      Thou wilt fall backward when thou hast more wit;
4      Thou wilt fall backward when thou comest to age;
5      Ben. An if he hear thee, thou wilt anger him.
6      Or, if thou wilt not, be but sworn my love,
7      Dost thou love me, I know thou wilt say 'Ay';
8      So thou wilt woo; but else, not for the world.
9      Or if thou wilt, swear by thy gracious self,
10     Rom. O, wilt thou leave me so unsatisfied?
# ... with 42 more rows
```



What next?

anti_join()

```
> tidy_shakespeare %>%  
  anti_join(data_frame(word = "wilt"))  
Joining, by = "word"
```

```
# A tibble: 141,003 x 4
```

	title	type	linenumber	word
	<chr>	<chr>	<int>	<chr>
1	Hamlet, Prince of Denmark	Tragedy	6776	multiple
2	Hamlet, Prince of Denmark	Tragedy	6761	loudly
3	Hamlet, Prince of Denmark	Tragedy	6737	upshot
4	Hamlet, Prince of Denmark	Tragedy	6736	deaths
5	Hamlet, Prince of Denmark	Tragedy	6735	slaughters
6	Hamlet, Prince of Denmark	Tragedy	6735	casual
7	Hamlet, Prince of Denmark	Tragedy	6734	carnal
8	Hamlet, Prince of Denmark	Tragedy	6732	unknowing
9	Hamlet, Prince of Denmark	Tragedy	6681	volley
10	Hamlet, Prince of Denmark	Tragedy	6666	have't

```
# ... with 140,993 more rows
```



SENTIMENT ANALYSIS IN R: THE TIDY WAY

Let's practice!



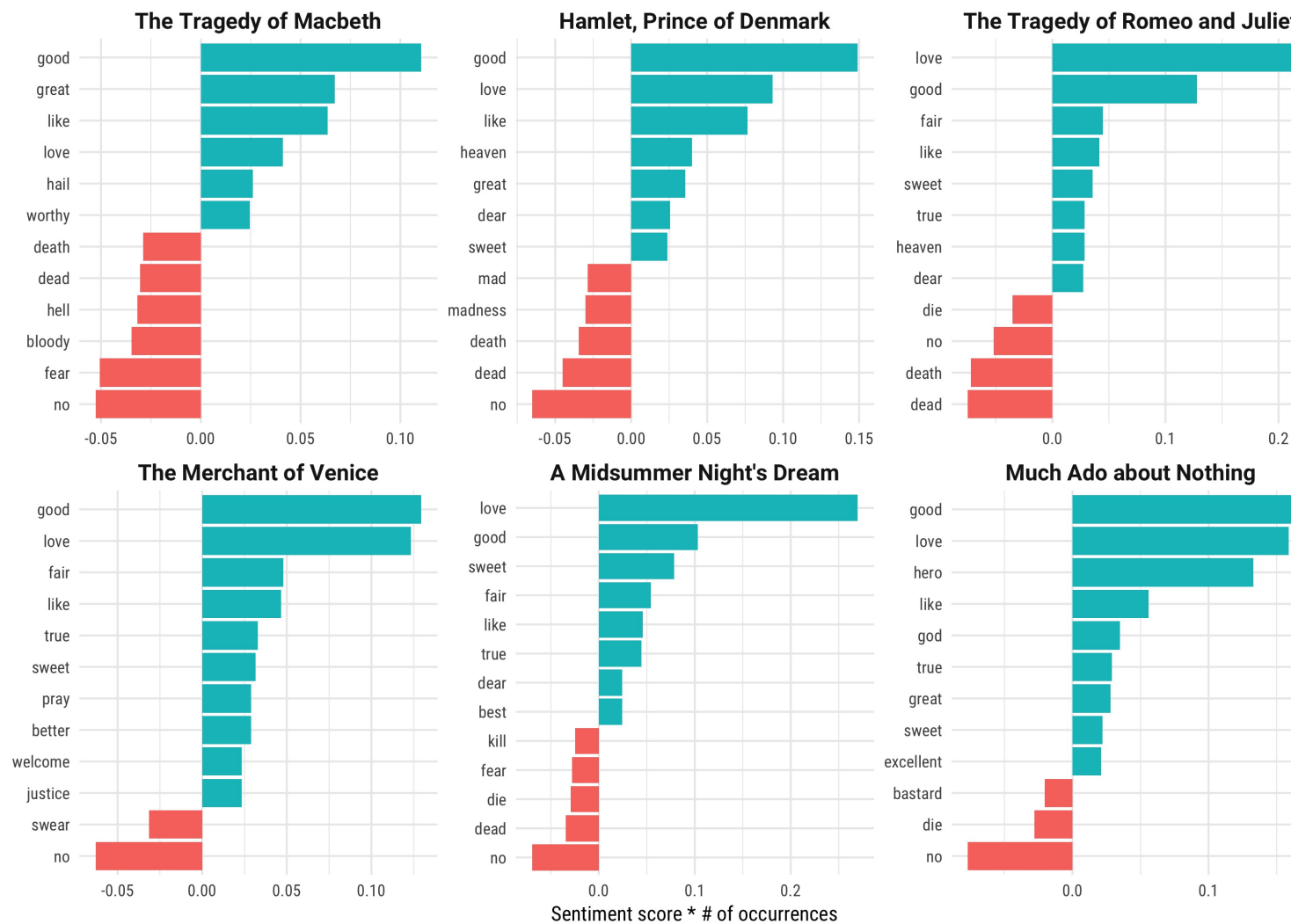
SENTIMENT ANALYSIS IN R: THE TIDY WAY

**Which words are
important in each play?**

Julia Silge

Data Scientist at Stack Overflow

Word contributions



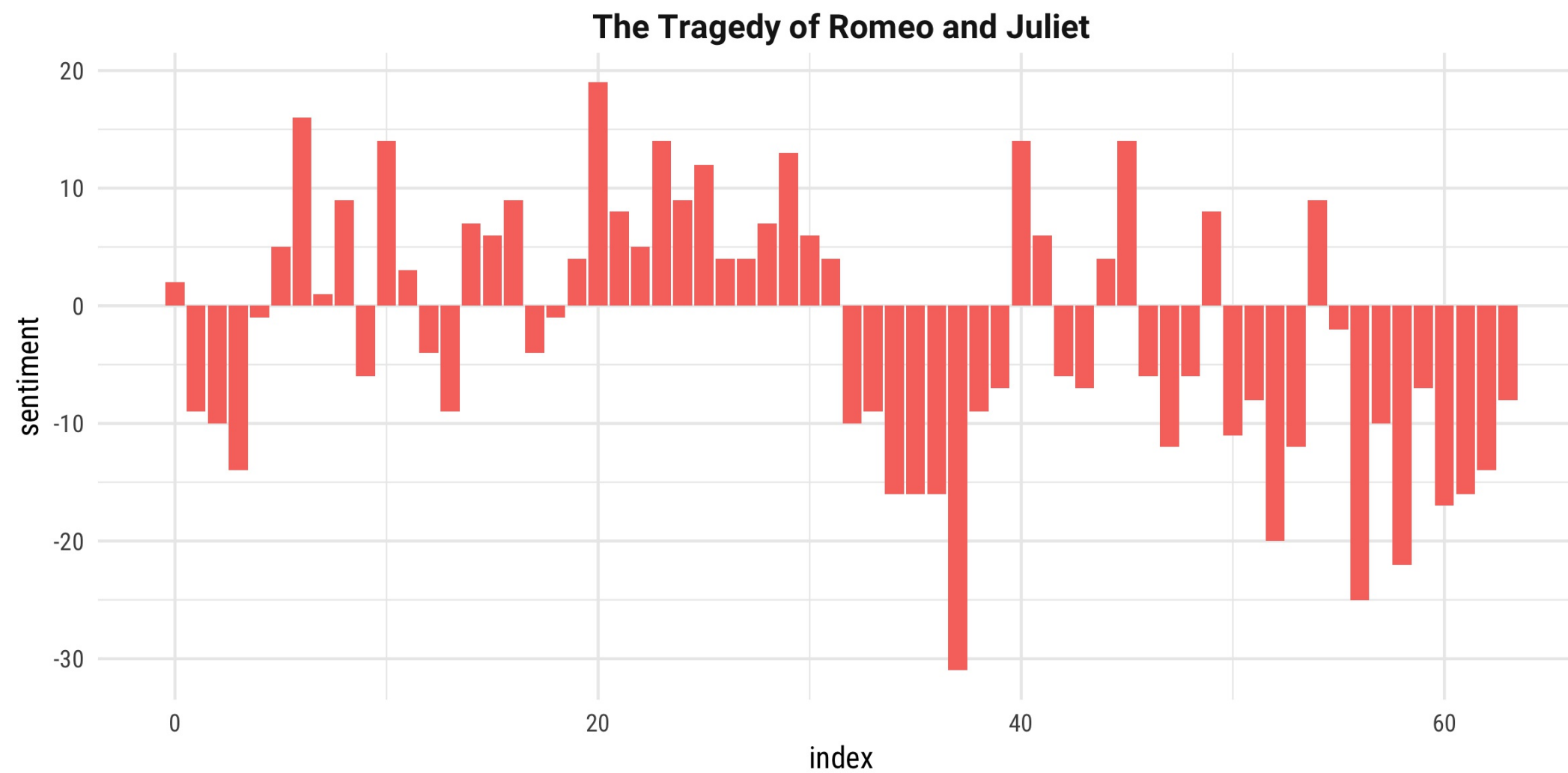


Which words are important?

Tidy data principles make sentiment analysis easier and more effective



Narrative arcs





SENTIMENT ANALYSIS IN R: THE TIDY WAY

Let's practice!