Automatisierte Inhaltsanalyse mit R

Wort- und Textmetriken

Cornelius Puschmann

inhaltsanalyse-mit-r.de

In diesem zweiten Kapitel steht nun die Analyse vor Wörtern und Texten im Mittelpunkt. Auf den ersten Blick erscheinen die Metriken, die hier vorgestellt werden, möglicherweise nicht als besonders relevant für sozialwissenschaftliche Fragestellung. Das liegt zum einen daran, das wir uns an dieser Stelle noch nicht mit abstrakten Konzepten wie Themen oder Sentiment beschäftigen, die in den folgenden Kapitel im Mittelpunkt stehen werden, sondern mit Aspekten wie der Frequenz von Begriffen und der Ähnlichkeit von Texten, die augenscheinlich vielleicht der Linguistik näher sind. Wort- und Textmetriken sind aber aus zwei Gründen von Bedeutung: erstens bilden sie die Grundlage der höherstufigen Verfahren, egal ob Lexikon-, Themen- oder Sentimentanalyse, und zu anderen lassen sich auch schon mit ihnen interessante sozialwissenschaftliche Fragestellungen bearbeiten.

Einige Beispiele:

- Welche Begriffe sind besonders distinktiv für einen politische Partei?
- Wie sprachlich komplex sind Nachrichtenbeiträge in unterschiedlichen Medien?
- Welche anderen Begriffe sind mit einem gesellschaftlichen Schlüsselbegriff ('Klima', 'Migration', 'Gerechtigkeit', 'Digitalisierung') verknüpft und wie verändern sich diese über die Zeit?
- Wie ähnlich sind sich Online-Kommentare zu unterschiedlichen Themen?

Diese und ähnliche Fragen werden in den folgenden Kapiteln aufgegriffen – zunächst werden aber die Funktionen vorgestellt, welche die Arbeit mit Wörtern und Texten in quanteda ermöglichen.

Installation und Laden der benötigten R-Bibliotheken, Erstellen des Korpus, Berechnen einer DFM

Zunächst werden wieder die notwendigen Bibliotheken geladen. Dann wird in einem zweiten Schritt das Sherlock-Korpus eingelesen und aufbereitet und dann Korpus-Statistiken berechnet. Schließlich erstellen wir wieder eine DFM (vgl. Kapitel 1), da wir diese später noch benötigen.

```
# Installation und Laden der Bibliotheken
if(!require("quanteda")) install.packages("quanteda")
if(!require("tidyverse")) install.packages("tidyverse")
theme_set(theme_bw())

# Laden der Sherlock Holmes-Daten (bereits als RData-File gespeichert)
load("daten/sherlock/sherlock.korpus.RData")

# Berechnen einer DFM
meine.dfm <- dfm(korpus, remove_numbers = TRUE, remove_punct = TRUE, remove_symbols = TRUE, remove = st
```

Korkordanzen erstellen

Zu den einfachsten Funktionen von quanteda gehört die Möglichkeit, Konkordanzen (auch KWIC genannt) zu erstellen, also die Textstelle eines Suchterms sowie dessen umgebenden Satzekontext zu extrahieren. Konkordanzen lassen sich in quanteda für einzelen Wörter, aber auch für Phrasen erzeugen. Oftmals ist der Export einer Konkodanz (etwa als CSV-Datei, die mit Excel geöffnet werden kann) neben der Darstellung innerhalb von R besonders nützlich. Dies geschieht hier mit der Funktion write_delim().

Anmerkung: Die Konkordanz kann mit dem kleinen Pfeil rechts oben gescrollt werden.

```
konkordanz <- kwic(korpus, "data")
konkordanz
```

```
##
##                  [A Scandal in Bohemia, 1630]
##                  [A Scandal in Bohemia, 1643]
##         [The Man with the Twisted Lip, 7776]
##   [The Adventure of the Speckled Band, 5718]
##  [The Adventure of the Speckled Band, 11126]
##  [The Adventure of the Noble Bachelor, 4959]
##  [The Adventure of the Copper Beeches, 3413]
##  [The Adventure of the Copper Beeches, 4044]
##  [The Adventure of the Copper Beeches, 4046]
##  [The Adventure of the Copper Beeches, 4048]
##
##                     "" I have no | data | yet. It is a
##      to theorize before one has | data | . Insensibly one begins to
##   or convinced himself that his | data | were insufficient. It was
##              I hope to get some | data | which may help us in
##  is to reason from insufficient | data | . The presence of the
##            I have nearly all my | data | . May I ask whether
##                  Ah, I have no | data | . I cannot tell.
##                I mentioned it." | Data | ! data! data!
##                       it." Data! | data | ! data!" he
##                   " Data! data! | data | !" he cried impatiently
```

```
konkordanz <- kwic(korpus, phrase("John|Mary [A-Z]+"), valuetype = "regex", case_insensitive = FALSE)
konkordanz
```

```
##
##                    [A Scandal in Bohemia, 979:980]
##                  [A Scandal in Bohemia, 4247:4248]
##                  [A Scandal in Bohemia, 7006:7007]
##                   [The Red-headed League, 332:333]
##                 [The Red-headed League, 8031:8032]
##                 [The Red-headed League, 8086:8087]
##                 [The Red-headed League, 8199:8200]
##                 [The Red-headed League, 9837:9838]
##               [The Red-headed League, 10091:10092]
##               [The Red-headed League, 10205:10206]
##                    [A Case of Identity, 1200:1201]
##                    [A Case of Identity, 1558:1559]
##                    [A Case of Identity, 5831:5832]
##             [The Boscombe Valley Mystery, 656:657]
##           [The Boscombe Valley Mystery, 2664:2665]
##           [The Boscombe Valley Mystery, 9429:9430]
##                  [The Five Orange Pips, 1307:1308]
##                  [The Five Orange Pips, 4540:4541]
##                  [The Five Orange Pips, 4766:4767]
##                  [The Five Orange Pips, 4781:4782]
##                  [The Five Orange Pips, 5426:5427]
##                  [The Five Orange Pips, 7640:7641]
##                  [The Five Orange Pips, 8763:8764]
##    [The Adventure of the Blue Carbuncle, 453:454]
```

```
##  [The Adventure of the Blue Carbuncle, 2831:2832]
##  [The Adventure of the Blue Carbuncle, 2913:2914]
##  [The Adventure of the Blue Carbuncle, 7186:7187]
##   [The Adventure of the Beryl Coronet, 6453:6454]
##    [The Adventure of the Copper Beeches, 656:657]
##
##                      deduce it. As to |    Mary Jane    |
##               Serpentine Avenue, St. |   John's Wood   |
##                      were such as Mr. |    John Hare    |
##    simple problem presented by Miss | Mary Sutherland |
##                         your hands."" |    John Clay    |
##             remarkable man, is young |    John Clay    |
##                  turns also with Mr. |    John Clay    |
##                       " It's no use, |    John Clay    |
##                     is better," said |    John Clay    |
##                   to settle with Mr. |    John Clay    |
##           . entered to announce Miss | Mary Sutherland |
##        somewhat vacuous face of Miss | Mary Sutherland |
##  the disappearing bridegroom of Miss | Mary Sutherland |
##                         part is a Mr. |   John Turner   |
##             driven over to Ross with |    John Cobb    |
##                                -"" Mr. |   John Turner   |
##                          said he," is |  John Openshaw  |
##                     one thing," said |  John Openshaw  |
##              McCauley, Paramore, and |   John Swain    |
##              McCauley cleared. 1Oth. |   John Swain    |
##                   that. And yet this |  John Openshaw  |
##                   in his pocket, was |  John Openshaw  |
##                , and the murderers of |  John Openshaw  |
##            the singular case of Miss | Mary Sutherland |
##                   just five days ago. |   John Horner   |
##    Hotel Cosmopolitan Jewel Robbery. |   John Horner   |
##                         ." My name is |   John Robin    |
##        the honour of addressing Miss |   Mary Holder   |
##      the singular experience of Miss | Mary Sutherland |
##
##  , she is incorrigible,
##  ." Holmes took a
##  alone could have equalled.
##  , that for strange effects
##  , the murderer, thief
##  . His grandfather was a
##  , and I agree with
##  ," said Holmes blandly
##  serenely. He made a
##  ," said Holmes.
##  , while the lady her
##  ." Yes, I
##  . A professional case of
##  , who made his money
##  , the groom. Shortly
##  ," cried the hotel
##  , but my own affairs
##  . He rummaged in his
```

```
##  , of St. Augustine
##  cleared. 12th. Visited
##  seems to me to be
##  , and whose residence is
##  were never to receive the
##  , and to the adventure
##  , a plumber, was
##  , 26, plumber,
##  - son," he
##  . Might I ask you
##  , the problem connected with
```

```
konkordanz <- kwic(korpus, c("log*", "emot*"), window = 10, case_insensitive = FALSE)
konkordanz
```

```
##
##                   [A Scandal in Bohemia, 46]
##                   [A Scandal in Bohemia, 55]
##                  [A Scandal in Bohemia, 205]
##          [The Boscombe Valley Mystery, 7002]
##                  [The Five Orange Pips, 147]
##  [The Adventure of the Blue Carbuncle, 3210]
##  [The Adventure of the Blue Carbuncle, 7499]
##    [The Adventure of the Speckled Band, 553]
##    [The Adventure of the Beryl Coronet, 645]
##   [The Adventure of the Copper Beeches, 151]
##   [The Adventure of the Copper Beeches, 401]
##
##                      her sex. It was not that he felt any | emotion  |
##            any emotion akin to love for Irene Adler. All | emotions |
##       lenses, would not be more disturbing than a strong | emotion  |
##           . Men who had only known the quiet thinker and | logician |
##      upon conjecture and sur- mise than on that absolute | logical  |
##          Assizes. Horner, who had shown signs of intense | emotion  |
##          black bar across the tail." Ryder quivered with | emotion  |
##         swift as intuitions, and yet always founded on a | logical  |
##       or more with a heaving chest, fighting against his | emotion  |
##  have given room for those faculties of deduction and of | logical  |
##                . Logic is rare. Therefore it is upon the |  logic   |
##
##  akin to love for Irene Adler. All emotions,
##  , and that one particularly, were abhorrent to his
##  in a nature such as his. And yet there
##  of Baker Street would have failed to recognize him.
##  proof which was so dear to him. There is
##  during the proceedings, fainted away at the conclusion and
##  ." Oh, sir," he cried,
##  basis wlth which he unravelled the problems which were submitted
##  . Then he passed his handkerchief over his brow,
##  synthesis which I have made my special province."
##  rather than upon the crime that you should dwell.
```

```
write_delim(konkordanz, path = "konkordanz.csv", delim = ";") # Datei ist Excel-kompatibel
```

Die Konkordanzen bestehen aus den Metadaten (Textname und Position), dem linken Kontext, dem Suchterm, sowie dem rechten Kontext. Die erste Konkordanz enthält alle Vorkommnisse des Begriffs ‘data’, die zweite

alle Vorkommnisse der Namen 'John' und 'Mary' gefolgt von einem weiteren Wort in Großschreibung (i.d.R. der Nachname). Die dritte Konkordanz enthält schließlich die Wortfragmente 'log' und 'emot', also Wörter wie 'logical' und 'emotional', aber auch die Pluralform 'emotions'. Strenggenommen handelt es sich hierbei nicht um Wortstämme, weil die Flexionsform bei unregelmäßigen Wörtern ganz vom Lemma abweicht (vgl. 'go' und 'went'). In den meisten sozialwissenschaftlichen Anwendungsszenarien ist es aber bereits ausreichend, durch die Verwendung von Platzhaltern (*) verschieden Wortvarianten zu identifizieren. Hier bringt quanteda eine Reihe nützlicher Eigenschaften mit, die in der Dokumentation von kwic() genau beschrieben werden.

Als nächstes berechnen wir die Häufigkeit und Dispersion von Tokens pro Erzählung, welche die Begriffe 'dark' und 'light' enthalten.

```
term1 <- kwic(korpus, "dark", valuetype = "regex", case_insensitive = FALSE) %>% group_by(docname) %>% s
term2 <- kwic(korpus, "light", valuetype = "regex", case_insensitive = FALSE) %>% group_by(docname) %>%
term1
```

```
## # A tibble: 12 x 4
##    docname                                Treffer Prozentanteil Suchterm
##    <chr>                                    <int>         <dbl> <chr>
##  1 The Adventure of the Speckled Band          12       0.102   dark
##  2 The Adventure of the Beryl Coronet           7       0.0823  dark
##  3 The Adventure of the Copper Beeches          6       0.0676  dark
##  4 The Man with the Twisted Lip                 7       0.0600  dark
##  5 The Red-headed League                        7       0.0583  dark
##  6 A Scandal in Bohemia                         5       0.0450  dark
##  7 The Adventure of the Engineer's Thumb        4       0.0358  dark
##  8 The Five Orange Pips                         3       0.0300  dark
##  9 The Adventure of the Blue Carbuncle          2       0.0174  dark
## 10 The Adventure of the Noble Bachelor          1       0.0104  dark
## 11 The Boscombe Valley Mystery                  1       0.0100  dark
## 12 A Case of Identity                           1       0.00949 dark
```

```
term2
```

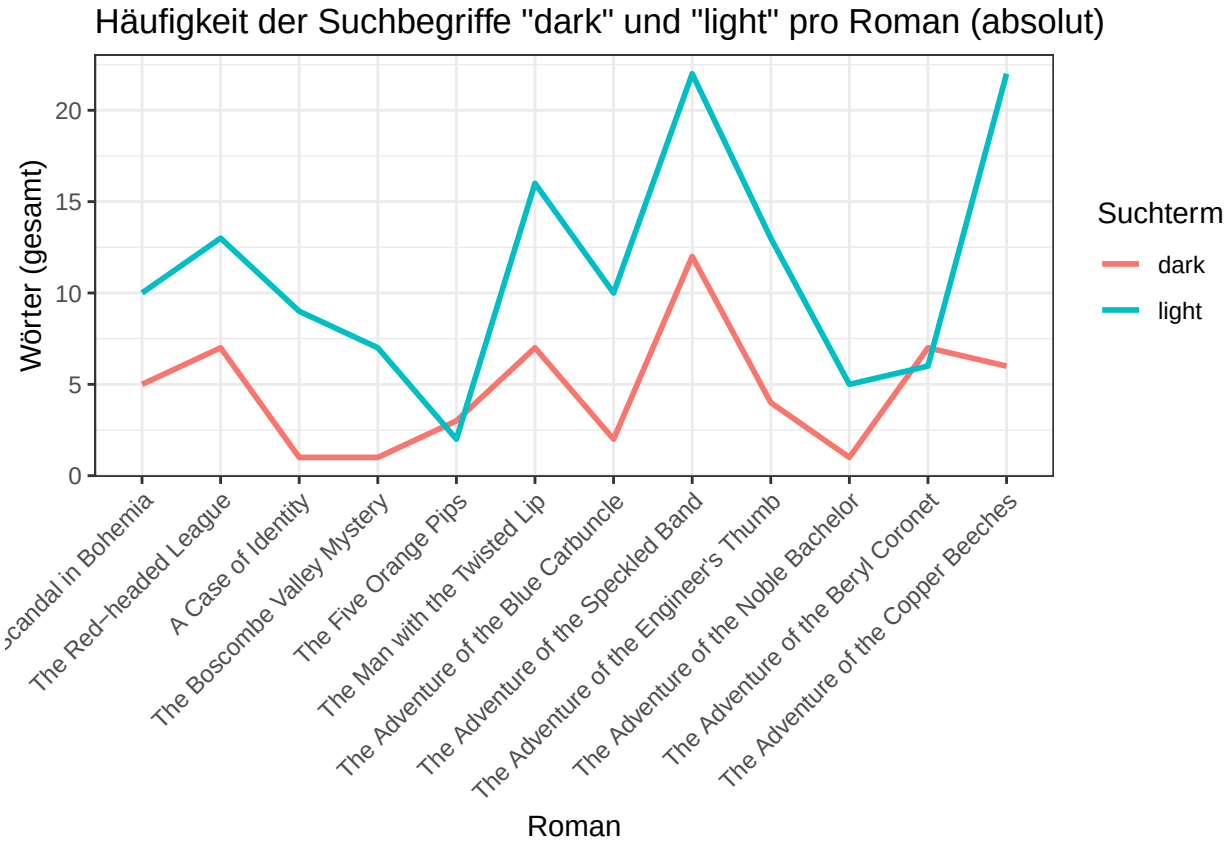
```
## # A tibble: 12 x 4
##    docname                                Treffer Prozentanteil Suchterm
##    <chr>                                    <int>         <dbl> <chr>
##  1 The Adventure of the Copper Beeches         22        0.248  light
##  2 The Adventure of the Speckled Band          22        0.187  light
##  3 The Man with the Twisted Lip                16        0.137  light
##  4 The Adventure of the Engineer's Thumb       13        0.116  light
##  5 The Red-headed League                       13        0.108  light
##  6 A Scandal in Bohemia                        10        0.0899 light
##  7 The Adventure of the Blue Carbuncle         10        0.0870 light
##  8 A Case of Identity                           9        0.0854 light
##  9 The Adventure of the Beryl Coronet           6        0.0705 light
## 10 The Boscombe Valley Mystery                  7        0.0700 light
## 11 The Adventure of the Noble Bachelor          5        0.0518 light
## 12 The Five Orange Pips                         2        0.0200 light
```

Wieder wenden wir zunächst die Funktion kwic() an, allerdings hier in Kombination mit mehreren Funktionen aus dem Paket dplyr (tidyverse). Diese Funktionen haben nichts mit quanteda zu tun, sondern sind für die Umformung jeglicher Daten in R nützlich (wer mehr wissen möchte, sollte sich dieses Buch anschauen). Während zuvor einfach die resultierende Konkordanz ausgegeben wurde, wird das Ergebnis jetzt mit Hilfe der Funktionen group_by(), summarise(), mutate() und arrange() weiter verarbeitet. Dabei machen wir uns die Tatsache zunutze, dass in einem KWIC-Ergebnis bereits alle Informationen vorliegen, um die absolute und relative Frequenz eines Begriffs (hier 'light' und 'dark') in einer Reihe von Dokumenten zu berechnen. Den Prozentanteil haben wir dabei einfach mittels Dreisatz abgeleitet (mit *Treffer/(korpus.stats$Tokens/100)*).

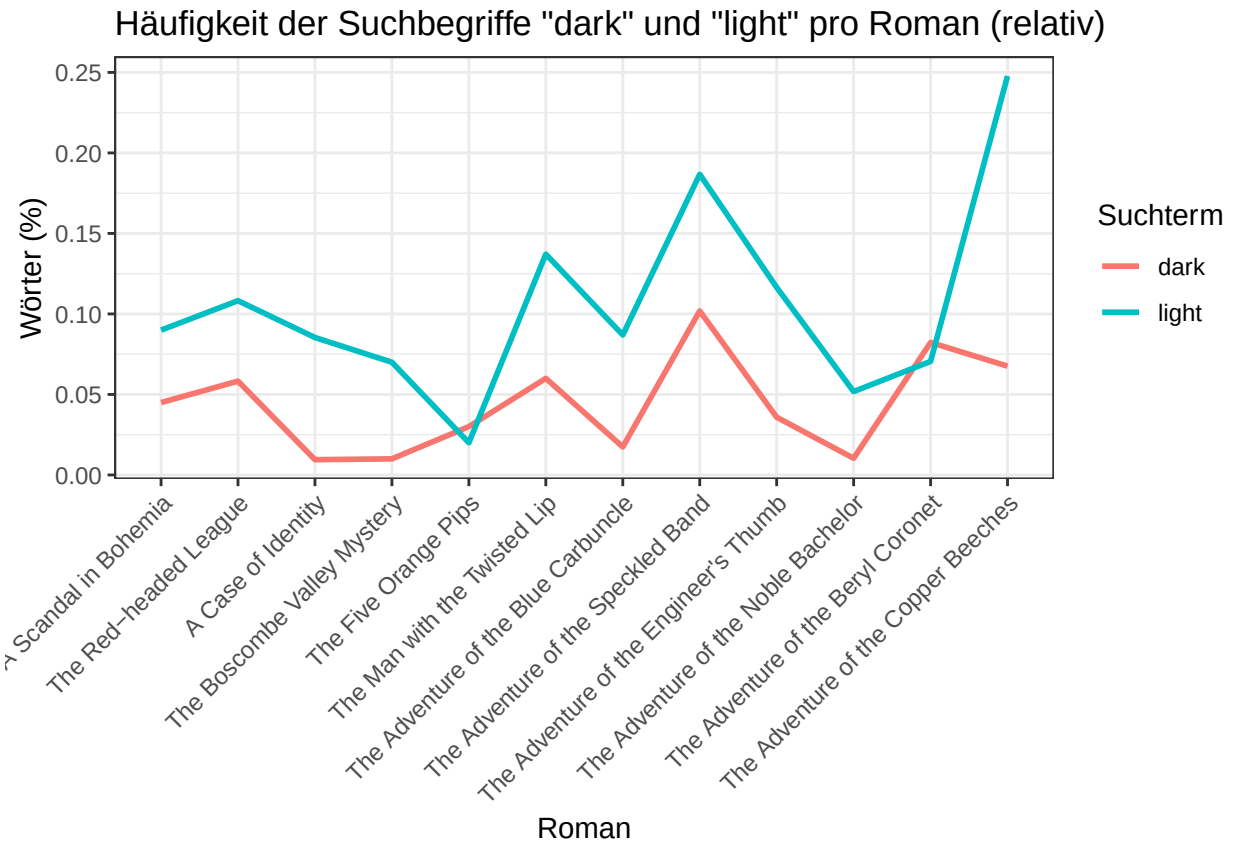
Wortfrequenzen lassen sich allerdings wesentlich einfacher durch die quanteda-eigenen Funktion textstat_frequency() umsetzen, die wir folgend auch konsequent nutzen werden – auch dazu gleich noch etwas mehr.

Zunächst plotten wie die absolute und relativen Häufigkeit der beiden Begriffe.

```
terme.kombiniert <- bind_rows(term1, term2)
terme.kombiniert$docname <- factor(terme.kombiniert$docname, levels = levels(korpus.stats$Text))
ggplot(terme.kombiniert, aes(docname, Treffer, group = Suchterm, col = Suchterm)) + geom_line(size = 1)
```



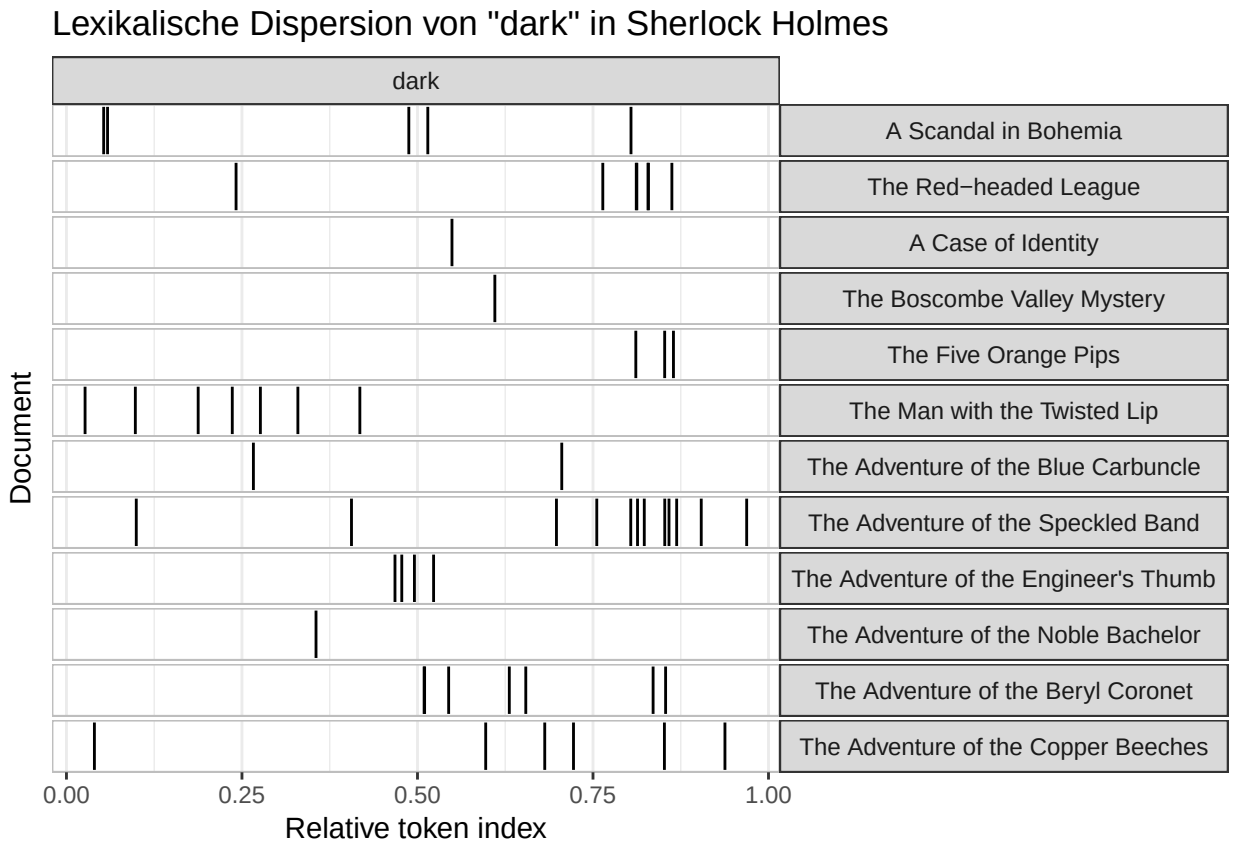

```
ggplot(terme.kombiniert, aes(docname, Prozentanteil, group = Suchterm, col = Suchterm)) + geom_line(siz
```



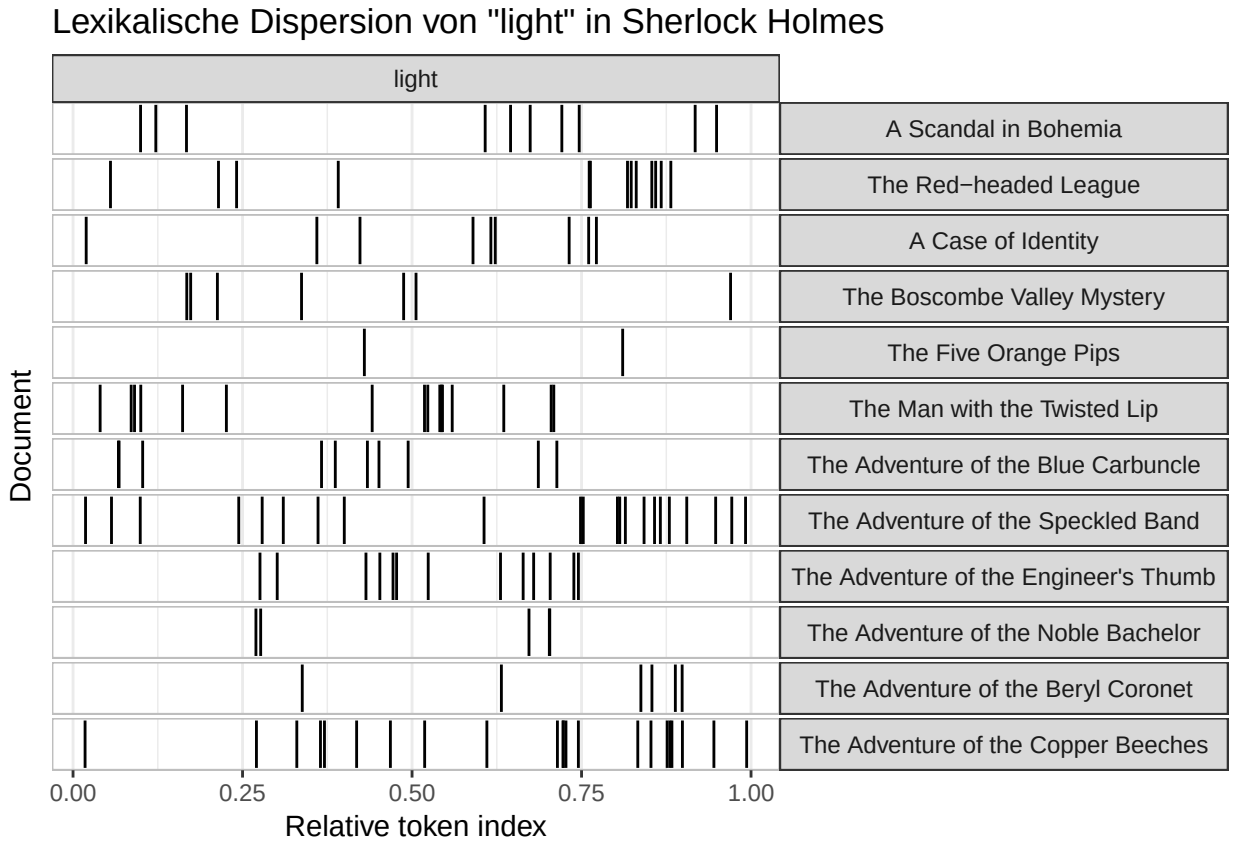
Wir sehen zwei unterschiedliche Berechnungen: das erste Plot zeigt die absolute Häufigkeit der beiden Begriffe, das zweite hingegen den relativen Prozenzanteil des Begriffs an der Gesamtwortzahl des jeweiligen Romans. Wieso sind die beiden Plots nahezu identisch? Dies hat mit der im Vergleich relativ ähnlichen Wortanzahl der Romane untereinander zu tun (zwischen 8,500 und 12,000 Tokens). Sind zwei Teilkorpora von sehr unterschiedlicher Größe, ist eine Normalisierung der Wortfrequenzen extrem wichtig, da sonst die Ergebnisse massiv verzerrt werden. Auch so ergeben sich durchaus Unterschiede, wenn man etwa den Anteil von 'The Adventure of the Speckled Band' und 'The Adventure of the Copper Beeches' vergleicht. Während die Anszahl der absoluten Treffer auf 'light' in beiden Romanen identisch ist, fällt der relative Anteil bei 'Speckled Band' im Vergleich ab.

Was tun, wenn man sich weniger für die Häufigkeit als für die Position der Suchterme interessiert? Dazu kann das Plotten der Begiffsdispersion als 'xray-plot' nützlich sein, wozu die Funktion textplot_xray existiert. Die X-Achse stellt hierbei die Position innerhalb des Textes dar, an dem der Suchbegriff vorkommt.

```
textplot_xray(kwic(korpus, "dark", valuetype = "regex", case_insensitive = FALSE)) + ggtitle("Lexikalis
```




```
textplot_xray(kwic(korpus, "light", valuetype = "regex", case_insensitive = FALSE)) + ggtitle("Lexikalis
```



Kollokationen

Wir gehen nun zu den sogenannten Textstatistken über. Dabei handelt es sich um Funktionen anhand derer sich Wörter und Texte mit Blick auf ihre Ähnlichkeit mit oder Distanz zu anderen Wörtern oder Texte analysieren lassen. Eine wichtige Funktion in diesem Zusammenhang ist die Extraktion von Kollokationen. Die Kollokate eines Begriffs sind solche Begriffe, die häufig gemeinsam mit dem Term vorkommen. Der Prozess ist dabei induktiv.

Folgend wenden wir die Funktion textstat_collocations() an, die häufige Kollokate im Sherlock Holmes-Korpus ermittelt. Anhand von write_delim wird das Ergebnis dann noch als Excel-kompatible CSV-Datei gespeichert.

```
meine.tokens <- tokens(korpus)
kollokationen <- textstat_collocations(meine.tokens, min_count = 10)
arrange(kollokationen, desc(count))
```

```
##          collocation count count_nested length      lambda          z
## 1              of the   707            0      2  1.93512589 41.48444370
## 2              in the   504            0      2  2.01246505 36.54144840
## 3               it is   334            0      2  3.61356047 49.64404560
## 4              to the   302            0      2  0.78489570 12.48617241
## 5              i have   299            0      2  2.81388870 38.17767146
## 6              it was   278            0      2  3.00566665 40.77948755
## 7              that i   249            0      2  2.44788414 33.04409207
## 8              at the   236            0      2  2.04164971 25.64398309
## 9               and i   207            0      2  1.56096182 20.25911391
## 10              to be   199            0      2  2.79790897 31.71857091
## 11           and the   198            0      2  0.18486146  2.46611290
```

```
## 12          upon the   196          0      2  2.53972500 26.66473930
## 13             i was   186          0      2  1.61564530 19.87948723
## 14            with a   184          0      2  2.41013272 27.92474703
## 15              i am   182          0      2  7.40635244 13.71613613
## 16             i had   168          0      2  2.11241726 23.79015438
## 17              of a   168          0      2  0.94126070 11.41684163
## 18             was a   160          0      2  1.60394595 18.51521036
## 19          that he   153          0      2  2.40233266 26.14213737
## 20            he was   151          0      2  2.23996944 24.58040073
## 21          that the   149          0      2  0.49719336  5.71763061
## 22             is a   143          0      2  1.75535452 19.04958772
## 23            on the   141          0      2  2.35070121 21.77172619
## 24           said he   140          0      2  3.90647705 36.24068656
## 25          from the   136          0      2  1.86738944 18.35990344
## 26          with the   136          0      2  1.21110799 12.76557296
## 27             in a   135          0      2  1.15477861 12.55600685
## 28            to me   134          0      2  2.21636051 22.33756325
## 29          you have   133          0      2  2.61439171 26.52649721
## 30            of his   131          0      2  1.57441642 16.51656969
## 31         have been   130          0      2  4.03991688 35.67714003
## 32            he had   130          0      2  2.64944797 26.49850170
## 33          into the   125          0      2  2.67296650 21.81922836
## 34            by the   123          0      2  2.25130166 19.84752115
## 35             and a   122          0      2  0.45927731  4.85087369
## 36            in his   120          0      2  1.92702467 19.27904966
## 37              as i   118          0      2  2.27891008 22.22761975
## 38          had been   116          0      2  3.98547315 34.05719824
## 39           that it   116          0      2  1.83469097 18.16160700
## 40           for the   110          0      2  1.12649488 10.77606128
## 41       said holmes   109          0      2  5.01770711 38.06652349
## 42          that you   108          0      2  1.78815664 17.15617806
## 43         there was   106          0      2  3.11434196 27.10750725
## 44          there is   104          0      2  3.33991072 28.65184832
## 45           which i   103          0      2  2.24686200 20.60507983
## 46             but i   100          0      2  2.42532576 21.66472464
## 47          i should    99          0      2  3.36268362 23.83984865
## 48          a little    99          0      2  3.01939619 23.50690378
## 49           i shall    97          0      2  3.75946992 23.77142127
## 50            one of    93          0      2  2.66595353 21.43225912
## 51           i could    92          0      2  2.70589835 21.07687598
## 52            a very    90          0      2  2.45863961 19.90034732
## 53          the door    86          0      2  3.16132029 18.32161643
## 54          and that    86          0      2  0.60286181  5.36345002
## 55   sherlock holmes    85          0      2  8.46440557 21.87511087
## 56             of my    84          0      2  1.31469210 11.29644539
## 57            when i    82          0      2  3.04404792 23.27092300
## 58        the matter    82          0      2  3.38028347 17.85441563
## 59             as to    81          0      2  1.29065228 10.91485266
## 60             a man    80          0      2  2.59289208 19.46762012
## 61         the other    80          0      2  2.89245688 17.46429672
## 62           that is    80          0      2  1.48810525 12.54365040
## 63           i think    78          0      2  3.22842716 20.97740738
## 64           you are    77          0      2  3.15481393 23.61394579
## 65             in my    77          0      2  1.66301764 13.66285520
```

```
## 66        think that    76           0      2  4.37638997 25.01958705
## 67            out of    75           0      2  2.58523034 18.91777631
## 68             to my    74           0      2  1.15768107  9.41814716
## 69          you will    73           0      2  3.40266479 24.06143052
## 70            is the    73           0      2  0.19404001  1.59422305
## 71          that she    70           0      2  2.94470503 20.94842895
## 72            if you    69           0      2  3.42878548 23.86960750
## 73          has been    68           0      2  4.49859081 30.00357371
## 74          and then    68           0      2  3.54119022 19.86898315
## 75          with his    68           0      2  2.07442751 15.93809620
## 76             he is    68           0      2  1.55372719 12.13070066
## 77          would be    66           0      2  3.72851605 25.85271219
## 78         could not    64           0      2  3.89208470 26.26456756
## 79             me to    64           0      2  1.83988522 13.43108996
## 80            do you    63           0      2  3.42116427 22.82042728
## 81         the house    63           0      2  2.94252170 15.56418365
## 82            to his    63           0      2  0.73456821  5.61055089
## 83            to you    63           0      2  0.68824580  5.26359413
## 84             a few    61           0      2  4.89326104 17.17445662
## 85            to see    61           0      2  2.53834208 16.79260762
## 86          which he    61           0      2  2.16638767 15.78429717
## 87          the room    61           0      2  2.52104442 14.58554093
## 88            and he    61           0      2  0.72000764  5.40472581
## 89           was the    61           0      2 -0.26363804 -2.00845816
## 90             as he    60           0      2  2.02102264 14.68766726
## 91            do not    59           0      2  4.01485429 25.74325441
## 92              to a    59           0      2 -0.20559511 -1.54857952
## 93             of it    57           0      2  0.55909452  4.08779814
## 94          that we     55           0      2  2.28374434 15.42416299
## 95             to do    55           0      2  2.28916400 14.84600117
## 96              as a    55           0      2  0.92932041  6.61939665
## 97            he has    54           0      2  2.87336522 18.60430945
## 98          we have     54           0      2  2.49915142 16.97290742
## 99        that there    53           0      2  2.48995752 16.24358280
## 100         the same    53           0      2  5.77561597  8.92383396
## 101          all the    53           0      2  1.18382168  7.87956821
## 102          to have    53           0      2  0.74575482  5.23603805
## 103          and his    53           0      2  0.44250493  3.12615372
## 104          but the    53           0      2  0.42910435  2.98726189
## 105         it would    52           0      2  2.57051904 16.71870799
## 106        seemed to    52           0      2  4.81838413 16.20417273
## 107           such a    51           0      2  3.43640941 18.01681261
## 108        of course    51           0      2  5.31717645 14.34118528
## 109         no doubt    50           0      2  7.12270475 23.38309038
## 110        there are    50           0      2  3.74043104 23.17162673
## 111          she had    50           0      2  2.81416311 18.17695847
## 112        which was    50           0      2  1.57192775 10.59731190
## 113         down the    50           0      2  1.68148401 10.34838981
## 114        should be    49           0      2  3.86372270 22.99975471
## 115           said i    49           0      2  1.97039856 12.88492612
## 116         we shall    48           0      2  4.32870326 24.52103303
## 117             if i    48           0      2  2.58970355 16.04750930
## 118           is not    48           0      2  2.02449111 13.22423076
## 119          in this    48           0      2  2.00528189 12.89770571
```

```
## 120          i will    48           0      2  1.95822821 12.18066886
## 121         when he    47           0      2  2.85280276 17.72206904
## 122     through the    47           0      2  3.02344846 13.91110487
## 123          to him    47           0      2  1.39423576  8.99575858
## 124           for a    47           0      2  0.95200432  6.27841954
## 125          have a    47           0      2  0.68504311  4.55264799
## 126         and yet    46           0      2  3.95539643 16.20218043
## 127           i can    46           0      2  2.49264007 14.26723686
## 128         she was    46           0      2  2.17150425 13.71042214
## 129         was not    46           0      2  1.71530038 11.05064337
## 130           be a    46           0      2  1.11649424  7.24467171
## 131          as we    45           0      2  2.75804402 16.97395417
## 132         a small    45           0      2  3.81482225 16.69276157
## 133        from his    45           0      2  2.16574440 13.64208870
## 134          and we    45           0      2  1.45293746  9.09784758
## 135         must be    44           0      2  4.06322789 22.43899167
## 136          my own    44           0      2  4.26213181 21.07084335
## 137         you can    44           0      2  3.30677163 18.53899539
## 138           i saw    43           0      2  3.40224812 15.80525538
## 139        the case    43           0      2  2.28521592 11.68941545
## 140           i do    43           0      2  1.88241968 11.17277985
## 141         the man    43           0      2  0.99761093  6.03356114
## 142         did you    42           0      2  3.52695787 19.06320152
## 143        upon his    42           0      2  2.15953613 13.16675348
## 144      the window    42           0      2  3.10953210 12.82250469
## 145     the morning    42           0      2  2.46092407 11.99808794
## 146       and there    42           0      2  1.61515041  9.67347574
## 147         did not    41           0      2  4.16168583 22.06324916
## 148        his eyes    41           0      2  4.25801160 19.74316276
## 149          into a    41           0      2  1.87674527 11.04619776
## 150         i would    41           0      2  1.52063339  9.06496733
## 151         of this    41           0      2  1.38567503  8.37401695
## 152        his face    40           0      2  3.60006720 18.70802468
## 153         we were    40           0      2  3.22630628 18.57108253
## 154         you see    40           0      2  2.76825155 15.72238916
## 155         this is    40           0      2  2.11123304 12.60144093
## 156       the first    40           0      2  2.68502509 12.13222953
## 157        the most    40           0      2  2.59295828 11.97811661
## 158        in which    40           0      2  1.27243233  7.74303510
## 159        his hand    39           0      2  3.74867928 18.78284659
## 160        you know    39           0      2  3.07075668 16.71498020
## 161         you may    39           0      2  2.94226196 16.23730888
## 162       i thought    39           0      2  3.37588863 15.04226499
## 163          i know    39           0      2  2.19712724 12.04368422
## 164          to her    39           0      2  1.27079223  7.53367838
## 165         man who    38           0      2  5.06904447 25.45347037
## 166         me that    38           0      2  1.81469803 10.59624208
## 167          and it    38           0      2  0.01302870  0.07895709
## 168       my friend    37           0      2  4.65455080 19.24735122
## 169          may be    37           0      2  3.58447581 19.22312548
## 170         he said    37           0      2  3.81268390 17.94239347
## 171          we are    37           0      2  3.17100705 17.66036283
## 172        see that    37           0      2  2.69401006 14.62689410
## 173          then i    37           0      2  2.56625414 14.04661851
```

```
## 174         i cannot    37           0      2  4.35068809 14.04240347
## 175         he would    37           0      2  2.22583810 12.62458279
## 176           i must    37           0      2  2.22368047 11.84275612
## 177          so that    37           0      2  1.96833717 11.26230721
## 178           what i    37           0      2  1.86065311 10.68478129
## 179         the time    37           0      2  1.59176836  8.43729326
## 180            for i    37           0      2  1.15690853  6.81475922
## 181           and to    37           0      2 -0.85783175 -5.18087543
## 182           i knew    36           0      2  3.92932622 14.41839204
## 183           we had    36           0      2  2.14404450 12.22586567
## 184        the table    36           0      2  4.29151357 10.69799323
## 185           but he    36           0      2  1.77491636 10.21555013
## 186            up to    36           0      2  1.76149866  9.76994427
## 187            had a    36           0      2  0.52149794  3.05652657
## 188           was in    36           0      2  0.39427780  2.32383458
## 189        they were    35           0      2  4.18192363 21.52841359
## 190         in front    35           0      2  4.77586544 15.63638835
## 191          who had    35           0      2  2.86237864 15.57327860
## 192          have no    35           0      2  2.76030542 15.02389325
## 193          i found    35           0      2  2.71307068 13.12754129
## 194          came to    35           0      2  2.41599321 12.41417012
## 195        the whole    35           0      2  3.92704609 11.17014542
## 196       the police    35           0      2  4.02714152 11.01678415
## 197         over the    35           0      2  1.51576707  7.97762633
## 198            as it    35           0      2  1.22456809  7.03450834
## 199          at once    34           0      2  4.43395098 19.71759591
## 200          my dear    34           0      2  5.60720462 16.39394939
## 201         all that    34           0      2  1.95910254 10.76599095
## 202         have you    34           0      2  1.15490616  6.55277755
## 203           as the    34           0      2 -0.36133333 -2.07165136
## 204         his head    33           0      2  4.36474017 17.72953304
## 205        must have    33           0      2  3.29932765 16.73352409
## 206        would not    33           0      2  2.91643147 15.55594382
## 207           who is    33           0      2  2.49361486 13.31157942
## 208            to go    33           0      2  2.82369095 13.23613730
## 209         which is    33           0      2  1.36138661  7.58563314
## 210           you to    33           0      2 -0.09189014 -0.52132485
## 211         front of    32           0      2  4.42031568 13.62494273
## 212           to say    32           0      2  2.83992295 13.07711398
## 213       would have    32           0      2  2.45642448 13.04839581
## 214           was no    32           0      2  2.22411400 11.73847542
## 215        about the    32           0      2  1.40394336  7.15273173
## 216         the very    32           0      2  0.41598890  2.25333341
## 217           be the    32           0      2 -0.05235785 -0.28933013
## 218        more than    31           0      2  5.21251175 23.80269743
## 219        his hands    31           0      2  4.06333479 17.14330475
## 220          will be    31           0      2  3.07012795 15.74114573
## 221        that this    31           0      2  1.54761413  8.25194733
## 222             by a    31           0      2  1.29476707  6.88300187
## 223           and so    31           0      2  1.20917106  6.39917645
## 224           as you    31           0      2  1.13143957  6.14486119
## 225          and was    31           0      2 -0.35604332 -1.96400869
## 226        could see    30           0      2  4.01478006 19.60465618
## 227         might be    30           0      2  3.88182448 18.26352323
```

```
## 228       there were    30           0      2  3.07776663 15.67760142
## 229          side of    30           0      2  3.89054933 13.78029362
## 230          i heard    30           0      2  2.41441055 11.32947450
## 231        you would    30           0      2  2.03845349 10.57597852
## 232           for me    30           0      2  1.87516890  9.89334892
## 233          and now    30           0      2  1.94042793  9.59902658
## 234         the last    30           0      2  1.98992769  8.99155466
## 235           for it    30           0      2  1.24957652  6.66409908
## 236        which you    30           0      2  1.21854475  6.50432025
## 237           from a    30           0      2  0.86509197  4.59932405
## 238       the little    30           0      2  0.66557592  3.44886396
## 239         have the    30           0      2 -0.56397408 -3.05270066
## 240           and in    30           0      2 -0.58160937 -3.16480752
## 241          at last    29           0      2  4.02133484 17.80375398
## 242     baker street    29           0      2  9.27899246 17.71108500
## 243        his chair    29           0      2  4.43144281 16.61171322
## 244         say that    29           0      2  3.72749135 15.46119405
## 245           up and    29           0      2  2.18277844 11.02196913
## 246      against the    29           0      2  3.08087768 10.99246007
## 247          come to    29           0      2  2.31603763 10.98479691
## 248         he could    29           0      2  2.07110705 10.53369616
## 249          to come    29           0      2  2.17152653 10.43086131
## 250            i did    29           0      2  2.18096365 10.36125608
## 251           up in     29           0      2  2.00081807 10.11910411
## 252         down to    29           0      2  1.72620933  8.63805148
## 253          him to    29           0      2  1.48630388  7.55180089
## 254        and what    29           0      2  1.42477523  7.21217506
## 255          but it    29           0      2  1.33965464  7.01561729
## 256         lord st    28           0      2  8.54484490 22.80325305
## 257         tell me    28           0      2  4.18282706 18.25559973
## 258        shall be    28           0      2  3.39576203 16.21769019
## 259        may have    28           0      2  2.87056581 13.95637081
## 260         then he    28           0      2  2.74934934 13.42868037
## 261        you must    28           0      2  2.74304740 13.14093629
## 262       enough to    28           0      2  4.26950968 12.89860521
## 263         able to    28           0      2  4.62095379 12.31690652
## 264          a long    28           0      2  2.76573636 12.09489852
## 265          i came    28           0      2  1.99683204  9.50829040
## 266         to know    28           0      2  1.88196291  9.12916746
## 267            i see   28           0      2  1.47475249  7.32020856
## 268         but you    28           0      2  1.33744832  6.89002395
## 269         for you    28           0      2  1.21185787  6.25920859
## 270          of her    28           0      2  0.92415021  4.73159309
## 271          up the    28           0      2  0.76145213  3.81525986
## 272         that my    28           0      2  0.60432190  3.14492447
## 273        when the    28           0      2  0.49882361  2.53279441
## 274          and of    28           0      2 -1.12110441 -5.92182527
## 275        at least    27           0      2  5.84877448 15.46699345
## 276       know that    27           0      2  2.70873487 12.60349722
## 277          had no    27           0      2  2.57690068 12.58509823
## 278         i asked    27           0      2  3.00960656 12.14448234
## 279         back to    27           0      2  2.45957003 11.07229672
## 280          not be    27           0      2  1.94701391  9.76192116
## 281         upon my    27           0      2  1.90977276  9.56191938
```

```
## 282         the lady    27           0      2  2.27175046  9.26785781
## 283            of us    27           0      2  1.76578857  8.50054785
## 284            i may    27           0      2  1.62499834  7.84430735
## 285          the way    27           0      2  1.52645282  6.98707570
## 286         that was    27           0      2  0.09266439  0.47696377
## 287          and you    27           0      2 -0.30236091 -1.55850166
## 288        very much    26           0      2  4.25774651 18.74148579
## 289           let me    26           0      2  4.48527519 17.95809710
## 290       might have    26           0      2  3.32537278 14.99736971
## 291        knew that    26           0      2  4.05368593 14.82241147
## 292          to tell    26           0      2  2.61766000 11.27762688
## 293           a good    26           0      2  2.48073862 10.90644029
## 294            is no    26           0      2  2.24668163 10.83894073
## 295         when you    26           0      2  2.00282638  9.75735599
## 296         with him    26           0      2  1.96846699  9.66889781
## 297        the name     26           0      2  2.41458494  9.39191036
## 298         of them     26           0      2  1.99018516  9.23608664
## 299         what is     26           0      2  1.86368167  9.13746005
## 300          she is     26           0      2  1.77473886  8.72715493
## 301         with me     26           0      2  1.57179666  7.79570560
## 302           at my     26           0      2  1.33762616  6.66201487
## 303         with my     26           0      2  1.26389438  6.30225960
## 304            at a     26           0      2  0.25166104  1.26594761
## 305           it to     26           0      2 -0.44304044 -2.24728265
## 306         to find     25           0      2  2.73175713 11.34733366
## 307          it may     25           0      2  2.29545808 10.69445707
## 308         is very     25           0      2  1.87217097  9.00410591
## 309          it all     25           0      2  1.63319908  7.86559389
## 310         had not     25           0      2  1.60523013  7.80940943
## 311          in one     25           0      2  1.50518756  7.23692968
## 312          at his     25           0      2  1.06168470  5.20912326
## 313         to your     25           0      2  0.99529251  4.80819601
## 314        with you     25           0      2  0.94336933  4.63624393
## 315         to this     25           0      2  0.83336298  4.04880225
## 316           not a     25           0      2  0.45708116  2.24725453
## 317        that his     25           0      2  0.25266956  1.25096408
## 318         what do     24           0      2  3.35694754 15.27416278
## 319        tell you     24           0      2  3.37760080 14.20208876
## 320          to get     24           0      2  3.17827023 11.92293614
## 321          like a     24           0      2  2.58756149 10.90701111
## 322          at all     24           0      2  2.21652410 10.41095280
## 323       among the     24           0      2  3.56547177  9.96648799
## 324        the fire     24           0      2  2.98772003  9.69029531
## 325        with her     24           0      2  1.96390477  9.28705643
## 326          it has     24           0      2  1.85963472  8.70110385
## 327           in an     24           0      2  1.44408078  6.82345970
## 328         of your     24           0      2  0.96544505  4.57954205
## 329            on a     24           0      2  0.93654540  4.45777731
## 330        you that     24           0      2  0.15328096  0.74512419
## 331           is to     24           0      2 -0.25063637 -1.22051631
## 332          far as     23           0      2  4.77132878 17.18541806
## 333          we may     23           0      2  3.21851887 14.30850780
## 334        he asked     23           0      2  3.57340812 13.93002038
## 335       it seemed     23           0      2  3.49277438 13.67157288
```

```
## 336         will not    23           0      2  2.78080403 12.59430989
## 337          wish to    23           0      2  3.75951253 12.07048546
## 338        that they    23           0      2  2.47978940 10.79785758
## 339            if he    23           0      2  2.25054315 10.28379796
## 340          who was    23           0      2  1.85267947  8.50852070
## 341         to think    23           0      2  1.63912407  7.35741367
## 342           for he    23           0      2  1.17403961  5.52828564
## 343           in her    23           0      2  1.15141580  5.38131664
## 344         have had    23           0      2  1.08623344  5.12218434
## 345         the only    23           0      2  1.01421631  4.50929070
## 346           it had    23           0      2  0.64213441  3.04155372
## 347            is in    23           0      2  0.19249411  0.91619392
## 348           and as    23           0      2  0.17705349  0.83800519
## 349           of you    23           0      2 -0.34514758 -1.64792655
## 350           let us    22           0      2  5.56355350 20.77743144
## 351        know what    22           0      2  4.38924972 18.25556598
## 352         they are    22           0      2  3.64176388 15.74099914
## 353        have done    22           0      2  3.91445104 15.00550671
## 354       doubt that    22           0      2  3.85160128 13.59843552
## 355     my companion    22           0      2  5.46978658 13.54303161
## 356         for some    22           0      2  2.67582820 11.85504481
## 357          a large    22           0      2  3.75130861 11.74685171
## 358         he might    22           0      2  2.74418710 11.65046191
## 359          a woman    22           0      2  2.93695440 11.08834392
## 360          to make    22           0      2  2.70174092 10.59404666
## 361        you think    22           0      2  2.35898838 10.35409755
## 362           i took    22           0      2  2.47408022  9.89623171
## 363         heard of    22           0      2  2.23770490  9.37481699
## 364          upon me    22           0      2  1.99771928  9.08369234
## 365            if it    22           0      2  2.00020983  8.98396821
## 366          him in     22           0      2  1.66785367  7.50520398
## 367           at me     22           0      2  1.46988730  6.74871389
## 368          of our     22           0      2  1.50717391  6.68155572
## 369        not have     22           0      2  1.36411683  6.27036079
## 370        which had    22           0      2  1.23357224  5.68418658
## 371        said that    22           0      2  1.14369876  5.24839068
## 372           of him    22           0      2  0.58861121  2.70761151
## 373            is it    22           0      2  0.49692995  2.30966599
## 374          and my     22           0      2 -0.23061872 -1.07519888
## 375        once more    21           0      2  5.61673474 20.54819028
## 376          an hour    21           0      2  6.27830771 16.98144824
## 377          so much    21           0      2  4.11169812 16.76769379
## 378          my wife    21           0      2  4.08186059 14.58654261
## 379        my father    21           0      2  3.64092300 14.03037669
## 380           no one    21           0      2  3.14453734 13.59385518
## 381         not know    21           0      2  3.08108478 13.14667236
## 382          here is    21           0      2  3.25661719 13.09312797
## 383       have heard    21           0      2  3.15748258 13.03280678
## 384     i understand    21           0      2  3.11293341 10.87370031
## 385           i went    21           0      2  2.68990251 10.17723435
## 386       during the    21           0      2  3.07318644  9.38068354
## 387        the facts    21           0      2  3.10023180  9.11563198
## 388          time to    21           0      2  2.18622363  8.98658564
## 389         the king    21           0      2  4.27042278  8.23484052
```

```
## 390           was so    21           0      2  1.57242973  6.98079208
## 391        the night    21           0      2  1.72208052  6.80211665
## 392          in your    21           0      2  1.25895495  5.61279868
## 393            of an    21           0      2  0.86663070  3.86617116
## 394           was at    21           0      2  0.72425068  3.28001972
## 395          are the    21           0      2  0.22261903  0.99324248
## 396         said the    21           0      2 -0.18395685 -0.83047637
## 397           that a    21           0      2 -0.77253633 -3.54290885
## 398       young lady    20           0      2  6.49853388 22.03060418
## 399        very well    20           0      2  4.41056648 16.84763069
## 400       an instant    20           0      2  6.22747453 16.70649652
## 401     this morning    20           0      2  3.99391636 15.76894791
## 402           as far    20           0      2  4.23754887 14.82137107
## 403          as well    20           0      2  3.61820367 13.96881707
## 404         when she    20           0      2  3.19522101 13.48449489
## 405         is quite    20           0      2  3.23148882 12.71864337
## 406        thank you    20           0      2  5.53670664 12.27194200
## 407         said she    20           0      2  2.82104287 12.01898758
## 408          his own    20           0      2  2.97246042 11.94293031
## 409         over his    20           0      2  2.47812711 10.41010806
## 410         began to    20           0      2  5.03562115  9.66595218
## 411          it must    20           0      2  2.23980419  9.40326761
## 412       the corner    20           0      2  3.37524576  8.88808072
## 413         which we    20           0      2  1.97358459  8.56571832
## 414    the inspector    20           0      2  4.22261546  8.10851905
## 415      between the    20           0      2  2.20413053  8.01335699
## 416         think of    20           0      2  1.80696659  7.51293852
## 417          the old    20           0      2  1.88673759  7.10701913
## 418         what you    20           0      2  1.54237146  6.72817197
## 419         you were    20           0      2  1.53179581  6.66645062
## 420        the light    20           0      2  1.72789338  6.65953270
## 421           it out    20           0      2  1.48021168  6.43567892
## 422         was very    20           0      2  1.37813369  6.01263305
## 423          upon it    20           0      2  1.27878138  5.61102761
## 424             so i    20           0      2  1.25466930  5.47798995
## 425         and down    20           0      2  1.07704872  4.63485449
## 426            me in    20           0      2  1.04292228  4.57555933
## 427            be in    20           0      2  0.65973025  2.91703790
## 428            it in    20           0      2 -0.22104167 -0.98711553
## 429         quite so    19           0      2  4.21266066 16.39192548
## 430          just as    19           0      2  3.32656287 13.31661880
## 431      he remarked    19           0      2  3.82951854 12.97227396
## 432        not think    19           0      2  2.95357515 12.10685325
## 433          she has    19           0      2  2.89759756 12.03841458
## 434          my hand    19           0      2  2.92477661 11.65228271
## 435           end of    19           0      2  3.58918020 11.03739459
## 436           i felt    19           0      2  3.76199192 10.61028860
## 437        which has    19           0      2  2.24227770  9.41753377
## 438          into my    19           0      2  2.10118782  8.84317217
## 439        which are    19           0      2  2.04098272  8.62818405
## 440      the coronet    19           0      2  3.53638914  8.59928147
## 441       which were    19           0      2  2.00513973  8.48545849
## 442            is an    19           0      2  1.68122603  7.13276337
## 443       the street    19           0      2  1.50911939  5.83123193
```

```
## 444         what was    19           0      2  1.29599692  5.53188846
## 445         have not    19           0      2  1.20394340  5.17354920
## 446          you not    19           0      2  0.86908780  3.74554695
## 447          you had    19           0      2  0.55513281  2.40290080
## 448            in it    19           0      2 -0.13689609 -0.59596885
## 449          in that    19           0      2 -0.40337876 -1.75948691
## 450            to it    19           0      2 -0.58531811 -2.55318182
## 451          that of    19           0      2 -0.92029576 -4.02429494
## 452     few minutes     18           0      2  6.93125856 21.13788519
## 453          am sure    18           0      2  5.97782226 18.96076853
## 454           or two    18           0      2  4.41356407 16.84662610
## 455        after all    18           0      2  4.19444823 15.82772205
## 456          look at    18           0      2  3.93709203 14.22462308
## 457    have already     18           0      2  4.19571560 13.86695422
## 458          who has    18           0      2  3.27493104 13.14211013
## 459          my mind    18           0      2  3.70838000 13.12995419
## 460         he cried    18           0      2  3.88736833 12.67602063
## 461          we must    18           0      2  3.13343446 12.46874664
## 462          as much    18           0      2  3.00187355 11.72710197
## 463            as if    18           0      2  2.78153947 11.03833344
## 464       should not    18           0      2  2.70191249 10.93503393
## 465        i believe    18           0      2  3.57082910 10.35448129
## 466       i answered    18           0      2  2.96161422  9.89964769
## 467          a great    18           0      2  2.73967931  9.70052290
## 468         of those    18           0      2  2.67078299  9.56422583
## 469          to look    18           0      2  2.56703483  9.31468710
## 470      should have    18           0      2  2.28590007  9.29620695
## 471         of these    18           0      2  2.45935824  9.07165292
## 472        which she    18           0      2  2.16030927  8.86768102
## 473       the papers    18           0      2  2.30921885  7.67149061
## 474        round the    18           0      2  1.80329462  6.62612451
## 475            it up    18           0      2  1.42124878  5.88476575
## 476        the young    18           0      2  1.48781907  5.61360310
## 477          the day    18           0      2  1.42573576  5.41659358
## 478           for my    18           0      2  1.03343426  4.33960591
## 479          the two    18           0      2  0.90837383  3.61292134
## 480          and she    18           0      2  0.75928538  3.14364083
## 481         which it    18           0      2  0.65651957  2.76685501
## 482           been a    18           0      2  0.59116120  2.47172021
## 483           not to    18           0      2  0.05534419  0.23353747
## 484          is that    18           0      2  0.02756501  0.11680732
## 485         of which    18           0      2  0.00437913  0.01848940
## 486          the one    18           0      2 -0.07524793 -0.31351506
## 487          of that    18           0      2 -0.89179697 -3.79820339
## 488          had the    18           0      2 -0.96514721 -4.09618548
## 489           an old    17           0      2  4.59964260 16.20902109
## 490        away from    17           0      2  3.95997552 14.51393157
## 491        have seen    17           0      2  3.45470629 12.48066955
## 492         it seems    17           0      2  4.76045959 11.78605224
## 493    believe that     17           0      2  4.69571924 11.62735601
## 494      miss hunter    17           0      2  9.64151093 11.21837593
## 495          he took    17           0      2  2.95026388 10.83223253
## 496           a word    17           0      2  3.20321249 10.12490209
## 497       i remarked    17           0      2  2.83785315  9.44447906
```

```
## 498        they have    17           0      2  2.28878269  9.05612641
## 499         could be    17           0      2  2.25529578  9.01738757
## 500     hosmer angel    17           0      2 13.44889189  8.90796472
## 501        but there    17           0      2  2.21208430  8.85443488
## 502         upon her    17           0      2  2.19642746  8.82168925
## 503       the letter    17           0      2  3.04210679  8.20890880
## 504         the lamp    17           0      2  3.69610509  8.04299443
## 505       the ground    17           0      2  3.86317115  7.89963058
## 506         not been    17           0      2  1.95563335  7.88279159
## 507      which would    17           0      2  1.95355286  7.85259002
## 508       the second    17           0      2  4.06385382  7.68605525
## 509         the city    17           0      2  4.06385382  7.68605525
## 510           made a    17           0      2  1.99664369  7.58650118
## 511           in our    17           0      2  1.65825301  6.57510311
## 512          from my    17           0      2  1.35507972  5.51113880
## 513        after the    17           0      2  1.31890142  4.97591626
## 514        the right    17           0      2  1.21677894  4.59121892
## 515        heard the    17           0      2  1.19742146  4.56577167
## 516          of some    17           0      2  1.04680361  4.18881479
## 517            to an    17           0      2  0.63217887  2.56323221
## 518            is my    17           0      2  0.54817669  2.25111436
## 519      our visitor    16           0      2  6.59251658 16.42608301
## 520          told me    16           0      2  4.25649249 14.02656817
## 521       neville st    16           0      2  9.34257026 13.48435034
## 522         had left    16           0      2  3.29515823 11.89944642
## 523        we should    16           0      2  2.72065771 10.44207036
## 524           am not    16           0      2  2.72499686 10.41995056
## 525        had come    16           0      2  2.69040990 10.17131151
## 526        seems to    16           0      2  4.06972532  9.98062672
## 527        out upon    16           0      2  2.55723579  9.89634950
## 528       she would    16           0      2  2.54158083  9.85059311
## 529     appeared to    16           0      2  4.30613721  9.76801417
## 530        but what    16           0      2  2.36587622  9.15802478
## 531       holmes had    16           0      2  2.33109997  8.97518227
## 532       thought of    16           0      2  2.45679693  8.59880011
## 533     the windows     16           0      2  3.25755991  7.99391402
## 534        the stone    16           0      2  2.90611413  7.91321548
## 535          to take    16           0      2  2.20519262  7.90584979
## 536      across the     16           0      2  2.64041771  7.88633582
## 537        was only     16           0      2  2.06546710  7.87530002
## 538       i suppose     16           0      2  5.28042811  7.77743063
## 539        the road     16           0      2  2.70540751  7.76457301
## 540        before i     16           0      2  2.03291779  7.71523532
## 541        the hall     16           0      2  4.25616057  7.23034815
## 542         me with     16           0      2  1.77619595  6.98045207
## 543       with your     16           0      2  1.73767982  6.81521473
## 544           i say     16           0      2  1.83273833  6.75089746
## 545           on my     16           0      2  1.59291204  6.26506577
## 546         are you     16           0      2  1.51641704  5.94866167
## 547       you could     16           0      2  1.47882586  5.79148087
## 548        him that     16           0      2  1.41050660  5.51180561
## 549       where the     16           0      2  1.00972197  3.79301596
## 550          out to     16           0      2  0.76323995  2.99654208
## 551           at it     16           0      2  0.52678160  2.10151529
```

```
## 552         the more    16           0      2  0.46692006  1.80025018
## 553          see the    16           0      2  0.45282009  1.75492085
## 554           her to    16           0      2  0.41562858  1.64622323
## 555          out the    16           0      2  0.05308802  0.20873691
## 556          and had    16           0      2 -0.49159302 -1.97064224
## 557     miss stoner    15           0      2  8.64363060 14.31429355
## 558        he spoke    15           0      2  4.50463318 11.54471568
## 559        his wife    15           0      2  3.43890024 11.37602728
## 560          whom i    15           0      2  3.62199139 11.03524204
## 561      his father    15           0      2  3.05511314 10.59765483
## 562          can be    15           0      2  2.70337304 10.01849641
## 563         used to    15           0      2  4.11805275  9.63794057
## 564          cry of    15           0      2  4.26101480  9.56970687
## 565        man with    15           0      2  2.50650561  9.36173307
## 566    your majesty    15           0      2  7.78729278  9.09056935
## 567         to give    15           0      2  2.55220081  8.49683555
## 568         has not    15           0      2  2.19455467  8.29160919
## 569         back in    15           0      2  2.22496053  8.06646498
## 570         seem to    15           0      2  5.34189751  7.83443768
## 571        about it    15           0      2  2.08099909  7.75024058
## 572         case of    15           0      2  2.20630187  7.70390327
## 573     the country    15           0      2  2.64270804  7.46535849
## 574         a young    15           0      2  2.05555201  7.25945780
## 575          a most    15           0      2  1.98193159  7.04720500
## 576         the end    15           0      2  1.76949232  5.89753255
## 577      the centre    15           0      2  5.04078290  5.89424623
## 578         down in    15           0      2  1.48526720  5.59883273
## 579          by his    15           0      2  1.37348365  5.24658656
## 580          was an    15           0      2  1.17987394  4.50693855
## 581          man of    15           0      2  1.07818699  4.06172471
## 582        and your    15           0      2  0.35626237  1.36611145
## 583          upon a    15           0      2  0.21735315  0.83856789
## 584       which the    15           0      2 -1.06919037 -4.15949007
## 585       am afraid    14           0      2  6.93400765 15.19358566
## 586         saw him    14           0      2  3.76689107 12.87995928
## 587          one or    14           0      2  3.58609816 12.60989755
## 588        had gone    14           0      2  4.09889541 12.47483150
## 589        what did    14           0      2  3.44683901 12.14793589
## 590        only one    14           0      2  3.37102446 11.98450725
## 591         one who    14           0      2  3.25656725 11.62697793
## 592     this matter    14           0      2  3.26591041 11.52648448
## 593         name is    14           0      2  3.71344986 11.52475268
## 594     irene adler    14           0      2 13.26089577 11.19826743
## 595       chair and    14           0      2  3.74137670 11.07290411
## 596       sure that    14           0      2  3.83905285 10.91504533
## 597  understand that    14           0      2  4.50563449 10.77308446
## 598           do so    14           0      2  2.97583532 10.73345585
## 599     he answered    14           0      2  3.39804559 10.71827799
## 600        hand and    14           0      2  2.99387737  9.93690005
## 601         find it    14           0      2  2.90268885  9.89140808
## 602       have made    14           0      2  2.80099910  9.80234956
## 603           a cab    14           0      2  3.64541332  9.43259789
## 604        face and    14           0      2  2.67235173  9.16665550
## 605         when we    14           0      2  2.51124252  9.15813099
```

```
## 606        was quite    14           0      2  2.53976330  8.79895375
## 607        have come    14           0      2  2.42839515  8.69175153
## 608        before he    14           0      2  2.39666997  8.57704458
## 609          my room    14           0      2  2.36117911  8.44264574
## 610          i cried    14           0      2  2.68211338  8.34822797
## 611         we could    14           0      2  2.24036945  8.19794926
## 612       found that    14           0      2  2.31931852  8.12481277
## 613        time that    14           0      2  2.27302182  7.98889315
## 614         it might    14           0      2  2.16172716  7.68248620
## 615     the passage    14           0      2  3.23922758  7.49460629
## 616   the photograph    14           0      2  3.12798998  7.48975423
## 617          the air    14           0      2  2.85351721  7.38855178
## 618        the floor    14           0      2  2.77654420  7.33646141
## 619         the best    14           0      2  2.51698511  7.08239225
## 620           took a    14           0      2  1.97242601  6.83877273
## 621           as she    14           0      2  1.77658171  6.52898121
## 622        the paper    14           0      2  2.14041174  6.50187010
## 623           than i    14           0      2  1.74394614  6.30141919
## 624          it over    14           0      2  1.71685400  6.23683239
## 625       could have    14           0      2  1.68987441  6.21915373
## 626        the money    14           0      2  1.99465203  6.21345123
## 627          of such    14           0      2  1.70297100  5.99143810
## 628          i never    14           0      2  1.71566864  5.98315878
## 629        the front    14           0      2  1.78164620  5.73372333
## 630            of mr    14           0      2  1.56430163  5.55462319
## 631          i might    14           0      2  1.40691053  5.01188551
## 632           is all    14           0      2  1.29722441  4.80215037
## 633          in some    14           0      2  1.27646544  4.67959501
## 634          what it    14           0      2  1.13921823  4.22648559
## 635           out in    14           0      2  1.10944569  4.09391885
## 636         and when    14           0      2  0.80996392  2.96557252
## 637         than the    14           0      2  0.51341427  1.86059750
## 638           one to    14           0      2  0.45584379  1.69170428
## 639           if the    14           0      2 -0.02067053 -0.07636683
## 640            be of    14           0      2 -0.16688874 -0.62593932
## 641            but a    14           0      2 -0.17434480 -0.65423967
## 642            is of    14           0      2 -0.77085566 -2.90702223
## 643            you a    14           0      2 -0.88653979 -3.34617961
## 644        years ago    13           0      2  7.66167596 18.80491820
## 645         too much    13           0      2  5.61941666 17.14936700
## 646       each other    13           0      2  6.12670371 16.81218126
## 647        less than    13           0      2  5.75502275 16.69773086
## 648       why should    13           0      2  5.11861025 15.72362062
## 649        will find    13           0      2  4.33810455 14.19705184
## 650      waiting for    13           0      2  5.24729268 13.28230796
## 651          made up    13           0      2  3.84071113 12.89317245
## 652        very good    13           0      2  3.58302459 12.06146985
## 653        down into    13           0      2  3.17515390 11.01304352
## 654      some little    13           0      2  3.03165396 10.56006258
## 655          at first    13           0      2  3.14226313 10.47336030
## 656    possible that    13           0      2  3.66424031 10.42499953
## 657        have ever    13           0      2  3.15467067 10.35027811
## 658         out into    13           0      2  2.87322991 10.04736864
## 659      afraid that    13           0      2  4.98008829  9.83163208
```

```
## 660        in london    13           0      2  3.18685681  9.61931297
## 661           we can    13           0      2  2.75308151  9.56931837
## 662        corner of    13           0      2  3.69458232  9.19522860
## 663          part of    13           0      2  3.39604313  9.10897714
## 664           a year    13           0      2  3.57357267  9.09252325
## 665          i don't    13           0      2  3.74363996  8.83136076
## 666         door and    13           0      2  2.64706540  8.78312713
## 667           to ask    13           0      2  3.02174987  8.68780284
## 668         where he    13           0      2  2.51883306  8.64402139
## 669            yet i    13           0      2  2.60492891  8.56811941
## 670       account of    13           0      2  4.63293281  8.50539636
## 671      a gentleman    13           0      2  2.74921221  8.30600584
## 672            i met    13           0      2  2.71758981  8.11432674
## 673         this man    13           0      2  2.29513437  8.11290489
## 674           for us    13           0      2  2.30438652  8.07765455
## 675       and looked    13           0      2  2.33364847  7.36488602
## 676         the next    13           0      2  3.05635196  7.21334956
## 677       the office    13           0      2  2.95625653  7.18047536
## 678          went to    13           0      2  2.19673143  7.15442779
## 679           to put    13           0      2  2.18750620  7.10972364
## 680           is now    13           0      2  2.03134751  7.10015911
## 681          where i    13           0      2  2.00878100  6.91088111
## 682           way to    13           0      2  1.93648119  6.46421501
## 683      the evening    13           0      2  2.23944714  6.44390629
## 684      the coroner    13           0      2  4.39143683  6.37670117
## 685           but we    13           0      2  1.71308084  6.11103545
## 686       nothing of    13           0      2  1.78588883  6.04138918
## 687          that if    13           0      2  1.72811277  6.01306592
## 688          the bed    13           0      2  1.99324222  5.99274963
## 689          and how    13           0      2  1.70032542  5.72463600
## 690          heard a    13           0      2  1.66758054  5.70536578
## 691        the world    13           0      2  4.90227443  5.69519394
## 692       you should    13           0      2  1.59953576  5.65138637
## 693       holmes was    13           0      2  1.58399178  5.58919040
## 694           you do    13           0      2  1.42246527  5.05250710
## 695          for him    13           0      2  1.39407277  4.99523314
## 696            is so    13           0      2  1.32800774  4.74255996
## 697          said my    13           0      2  1.18141873  4.24156583
## 698       the stairs    13           0      2  6.00089869  4.16662262
## 699       before the    13           0      2  0.57600202  2.00841258
## 700          were to    13           0      2  0.32094879  1.15274195
## 701           of all    13           0      2  0.30543815  1.09707398
## 702           was it    13           0      2 -0.28483062 -1.03525981
## 703          and not    13           0      2 -0.38589420 -1.39840718
## 704         and with    13           0      2 -0.45299735 -1.64277444
## 705           was to    13           0      2 -1.11589067 -4.06904047
## 706       can hardly    12           0      2  5.29031502 15.67384385
## 707       other side    12           0      2  4.94529104 15.38896486
## 708     asked holmes    12           0      2  4.75058097 14.11014080
## 709      should like    12           0      2  4.44264225 14.03919120
## 710        young man    12           0      2  4.16469396 13.20678166
## 711         one side    12           0      2  3.95418266 12.64211586
## 712           be able    12           0      2  4.57376966 12.54140985
## 713        came down    12           0      2  3.68458457 12.12415305
```

```
## 714       at present    12           0      2  4.55056608 11.91523886
## 715        some time    12           0      2  3.54098625 11.70498046
## 716        how could    12           0      2  3.50188590 11.55997369
## 717       have given    12           0      2  4.09082128 11.35010787
## 718        my sister    12           0      2  4.33772267 11.20053335
## 719        felt that    12           0      2  3.68928012 10.05058061
## 720          his son    12           0      2  3.33089706 10.04876852
## 721        in order     12           0      2  4.20814068  9.76108715
## 722        give you     12           0      2  3.06661634  9.58202533
## 723        until he     12           0      2  2.93595276  9.42287696
## 724         see him     12           0      2  2.68917137  9.06486650
## 725        pounds a     12           0      2  3.43328382  8.88273437
## 726          but if     12           0      2  2.63964252  8.82476846
## 727       likely to     12           0      2  3.90177405  8.76753367
## 728         sign of     12           0      2  4.04471577  8.74872099
## 729        not only     12           0      2  2.60000460  8.71918942
## 730         take it     12           0      2  2.70484998  8.71205925
## 731          a week     12           0      2  3.41925749  8.69989105
## 732        a single     12           0      2  4.00711353  8.66767475
## 733       was still     12           0      2  2.72659035  8.63126800
## 734    thought that     12           0      2  2.64584958  8.37802176
## 735        his long     12           0      2  2.54725742  8.30232268
## 736        a strong     12           0      2  2.95380174  8.29168069
## 737       of events     12           0      2  4.52230591  8.22078676
## 738    confess that     12           0      2  5.69094271  8.20829361
## 739         while i     12           0      2  2.56338383  8.14669412
## 740          i hope     12           0      2  2.94709542  8.12274588
## 741        were all     12           0      2  2.32088734  7.91043083
## 742         name of     12           0      2  2.60935402  7.79160823
## 743          gave a     12           0      2  2.56624241  7.75829141
## 744         i fancy     12           0      2  5.00141635  7.21624441
## 745         but she     12           0      2  1.92364991  6.58001695
## 746         are not     12           0      2  1.86610668  6.39549691
## 747           be so     12           0      2  1.84900256  6.33859873
## 748       the scene     12           0      2  4.31429677  6.22591260
## 749      the church     12           0      2  4.31429677  6.22591260
## 750         door of     12           0      2  1.90793191  6.15285104
## 751        from her     12           0      2  1.76881655  6.07963606
## 752           us in     12           0      2  1.79226095  5.99268751
## 753        open the     12           0      2  2.05518398  5.98427235
## 754        not very     12           0      2  1.69817181  5.83719449
## 755          at one     12           0      2  1.57275946  5.40613538
## 756          had an     12           0      2  1.49020693  5.12654630
## 757     the strange     12           0      2  1.68924501  5.11243841
## 758          up his     12           0      2  1.48209976  5.07794678
## 759           do it     12           0      2  1.45581791  4.98336043
## 760        will you     12           0      2  1.43795832  4.92788038
## 761           me at     12           0      2  1.39516485  4.81458300
## 762       the place     12           0      2  1.50442707  4.65859860
## 763           to mr     12           0      2  1.38166352  4.60182228
## 764           i beg     12           0      2  6.61087746  4.58342616
## 765       and found     12           0      2  1.36840421  4.53501597
## 766       the great     12           0      2  1.41236027  4.41861247
## 767           us to     12           0      2  1.30882332  4.38199394
```

```
## 768        while the    12           0      2  1.33379469  4.24858219
## 769          what he    12           0      2  1.18302911  4.08519590
## 770          it will    12           0      2  1.18215393  4.05780097
## 771        was there    12           0      2  1.04822665  3.61010683
## 772     the business    12           0      2  1.11084375  3.57683809
## 773         took the    12           0      2  1.02095811  3.33654010
## 774         the left    12           0      2  0.96006367  3.12843844
## 775          it were    12           0      2  0.89823775  3.10195288
## 776            to us    12           0      2  0.86614386  2.94974711
## 777       with which    12           0      2  0.79301637  2.75462104
## 778        that your    12           0      2  0.71555739  2.47295034
## 779             up a    12           0      2  0.63892020  2.19594695
## 780          it from    12           0      2  0.55850729  1.93911703
## 781          but his    12           0      2  0.51639191  1.79701216
## 782          for his    12           0      2  0.39383866  1.37207357
## 783           it not    12           0      2  0.29951868  1.04309769
## 784          at that    12           0      2 -0.02476237 -0.08646533
## 785          and for    12           0      2 -0.32198734 -1.12213746
## 786           you in    12           0      2 -0.61951750 -2.17093375
## 787          have to    12           0      2 -0.77034808 -2.69820834
## 788          not the    12           0      2 -1.05878165 -3.69975540
## 789           was of    12           0      2 -1.17620935 -4.12923239
## 790       last night    11           0      2  5.42925744 15.88479161
## 791         two days    11           0      2  5.58035950 15.61834896
## 792       don't know    11           0      2  5.90385885 15.22441507
## 793       shall soon    11           0      2  5.42233253 15.13433891
## 794       our client    11           0      2  5.53656651 14.75304445
## 795     could hardly    11           0      2  4.60462765 13.44428735
## 796          how did    11           0      2  4.23861724 13.17850718
## 797        come back    11           0      2  4.22351865 13.13873970
## 798      her husband    11           0      2  5.47859568 13.03602455
## 799           so far    11           0      2  4.30409033 12.71442240
## 800         your son    11           0      2  4.31389136 12.57221017
## 801         far from    11           0      2  3.97468235 11.83880232
## 802        which led    11           0      2  4.61479385 11.45200769
## 803         your own    11           0      2  3.37745394 10.64758054
## 804       his pocket    11           0      2  4.34551203 10.39366531
## 805          ask you    11           0      2  3.70853707 10.19508684
## 806      as possible    11           0      2  3.42227577 10.15924048
## 807          no more    11           0      2  3.09831311  9.96331802
## 808         this way    11           0      2  3.07372823  9.78447240
## 809          but how    11           0      2  3.09256219  9.68295740
## 810        hope that    11           0      2  3.66080410  9.62083676
## 811          will do    11           0      2  2.92440806  9.46233383
## 812       in england    11           0      2  4.12417462  9.39217512
## 813         had seen    11           0      2  3.03747400  9.38548517
## 814          such as    11           0      2  2.84288680  9.05293347
## 815        he looked    11           0      2  2.94380262  8.77979996
## 816      stoke moran    11           0      2 13.61691449  8.63462804
## 817     briony lodge    11           0      2 13.28043103  8.62168994
## 818           air of    11           0      2  4.10405098  8.35632549
## 819          to meet    11           0      2  4.05707547  8.26097958
## 820            if we    11           0      2  2.46545622  8.04702134
## 821        down upon    11           0      2  2.46490626  8.02501823
```

```
## 822        i confess    11           0      2  4.12919788  7.95946489
## 823         though i    11           0      2  2.57564097  7.84410453
## 824          gone to    11           0      2  2.84439271  7.83483385
## 825       what could    11           0      2  2.34893717  7.68380527
## 826          and why    11           0      2  2.85308740  7.66384847
## 827   copper beeches    11           0      2 15.22637484  7.53175028
## 828      of interest    11           0      2  2.62816678  7.47209472
## 829         within a    11           0      2  2.51187181  7.33621870
## 830         signs of    11           0      2  5.05960862  7.24753625
## 831          until i    11           0      2  2.32641376  7.23612599
## 832          you say    11           0      2  2.26886466  7.18399435
## 833         in every    11           0      2  2.27223936  7.04316685
## 834         out from    11           0      2  2.14763637  7.04303387
## 835         but when    11           0      2  2.14192983  6.99451175
## 836          my face    11           0      2  2.16843310  6.97833210
## 837          be some    11           0      2  2.11222344  6.90969960
## 838           i wish    11           0      2  2.35239265  6.84570146
## 839          and let    11           0      2  2.34211091  6.81587885
## 840       the garden    11           0      2  3.27517685  6.67163047
## 841        which may    11           0      2  2.03665519  6.64485017
## 842      the colonel    11           0      2  2.89565131  6.60133450
## 843        the truth    11           0      2  3.64292557  6.54954936
## 844       beside the    11           0      2  2.57474427  6.51612838
## 845          do with    11           0      2  1.97846715  6.48734014
## 846            i got    11           0      2  2.18290488  6.48552528
## 847            be no    11           0      2  1.93703529  6.36271431
## 848         they had    11           0      2  1.93080975  6.31035701
## 849          you did    11           0      2  1.92688558  6.21000379
## 850       opened the    11           0      2  2.27841895  6.12337824
## 851        cried the    11           0      2  2.13528240  5.88587610
## 852          he came    11           0      2  1.75747317  5.69370108
## 853         were not    11           0      2  1.66185846  5.48763149
## 854        under the    11           0      2  1.86432413  5.35814449
## 855         from him    11           0      2  1.60205543  5.29911637
## 856         and took    11           0      2  1.61081708  5.04059636
## 857          for all    11           0      2  1.47749456  4.88649790
## 858         of their    11           0      2  1.54112516  4.88272838
## 859        the point    11           0      2  1.66542761  4.84995939
## 860            me as    11           0      2  1.44943574  4.80107410
## 861          quite a    11           0      2  1.45314711  4.64795686
## 862          with an    11           0      2  1.39608131  4.61696198
## 863         into his    11           0      2  1.30084234  4.29449323
## 864          out his    11           0      2  1.28935766  4.25759279
## 865         and made    11           0      2  1.32845057  4.23317027
## 866          at this    11           0      2  1.27168410  4.21672558
## 867           just a    11           0      2  1.28190744  4.13484045
## 868    the contrary    11           0      2  5.84019803  4.04250910
## 869           of any    11           0      2  1.18343186  3.82080941
## 870           is one    11           0      2  1.13567970  3.76078864
## 871           on his    11           0      2  0.98039980  3.25527901
## 872          saw the    11           0      2  0.95090513  2.99550266
## 873            am a     11           0      2  0.88310005  2.89195329
## 874         that all    11           0      2  0.60385988  2.00622290
## 875          to some    11           0      2  0.58180344  1.91690729
```

```
## 876            has a    11            0      2  0.43159424  1.42977273
## 877            it as    11            0      2  0.20166158  0.67451745
## 878           and on    11            0      2  0.09822757  0.32637036
## 879           and an    11            0      2  0.08517268  0.28306036
## 880          and all    11            0      2  0.01927956  0.06414588
## 881             me a    11            0      2  0.01896273  0.06325804
## 882       that which    11            0      2 -0.01960712 -0.06564996
## 883          have it    11            0      2 -0.02524840 -0.08465408
## 884         and this    11            0      2 -0.11156773 -0.37196383
## 885            of me    11            0      2 -0.49829470 -1.67068648
## 886         and which    11            0      2 -0.60423592 -2.02612310
## 887          was that    11            0      2 -0.71326148 -2.39929966
## 888            had to    11            0      2 -0.74139184 -2.49075573
## 889           to that    11            0      2 -1.38845134 -4.67989440
## 890      either side    10            0      2  6.42443267 16.26853940
## 891    boscombe pool    10            0      2  9.79292293 15.83772416
## 892      good enough    10            0      2  5.44123849 15.21800397
## 893        two years    10            0      2  4.88257627 14.03001541
## 894  remarked holmes    10            0      2  5.01289866 13.15421578
## 895          tell us    10            0      2  4.18948516 12.44739102
## 896       would give    10            0      2  4.21795136 12.12762391
## 897        came back    10            0      2  3.99592348 12.01986584
## 898        she cried    10            0      2  4.25770191 11.91734335
## 899       do nothing    10            0      2  3.98568981 11.87057302
## 900           put on    10            0      2  3.93382187 11.58416592
## 901         be happy    10            0      2  5.22647877 11.28225135
## 902         does not    10            0      2  4.17456805 11.24654741
## 903        all right    10            0      2  3.55372257 10.70256328
## 904        as though    10            0      2  4.86669785 10.51813057
## 905      will excuse    10            0      2  7.35576326 10.40842742
## 906         my heart    10            0      2  4.09465347 10.21205527
## 907          her own    10            0      2  3.31003461 10.03178398
## 908         his lips    10            0      2  4.55597881  9.85384472
## 909       had better    10            0      2  3.42382299  9.75620181
## 910    his shoulders    10            0      2  4.82426553  9.64907805
## 911          read it    10            0      2  3.58237050  9.55485843
## 912        all about    10            0      2  3.01651809  9.29416962
## 913        this case    10            0      2  3.03914036  9.26781393
## 914          my life    10            0      2  3.23794287  9.20806109
## 915        away with    10            0      2  3.00864545  9.05255257
## 916          his hat    10            0      2  3.12620531  8.86730210
## 917         all over    10            0      2  2.86259112  8.86146368
## 918        come from    10            0      2  2.86371243  8.79686580
## 919         told you    10            0      2  3.05368010  8.76163638
## 920         what can    10            0      2  2.83251066  8.75733352
## 921         at night    10            0      2  2.88696463  8.65846954
## 922        is always    10            0      2  2.92592354  8.52878154
## 923       friend and    10            0      2  2.85264690  8.19032637
## 924         sight of    10            0      2  3.74439671  8.10764140
## 925         order to    10            0      2  3.72664321  8.06932799
## 926            a low    10            0      2  3.97510992  7.95854914
## 927       appears to    10            0      2  4.16199595  7.90140410
## 928         beg that    10            0      2  5.51523843  7.83170844
## 929          to keep    10            0      2  2.97519998  7.61196513
```

```
## 930         became a    10           0      2  2.86411672  7.59588616
## 931           a note    10           0      2  2.77868572  7.37017024
## 932           i fear    10           0      2  2.82470807  7.30193066
## 933          to hear    10           0      2  2.72382771  7.27786919
## 934       a question    10           0      2  2.64803084  7.16870642
## 935           a pair    10           0      2  4.93066761  7.00368446
## 936         are very    10           0      2  2.21700614  6.95696934
## 937         i passed    10           0      2  2.32461290  6.48783915
## 938         amid the    10           0      2  3.65376577  6.48265893
## 939         all this    10           0      2  2.05359560  6.45963720
## 940       closed the    10           0      2  2.80639651  6.42673990
## 941         saw that    10           0      2  2.11990813  6.41081076
## 942         to leave    10           0      2  2.23986881  6.38596496
## 943       the cellar    10           0      2  3.55177484  6.30188040
## 944     the boscombe    10           0      2  3.55177484  6.30188040
## 945         the bell    10           0      2  3.55177484  6.30188040
## 946          we will    10           0      2  2.00032396  6.28498255
## 947        number of    10           0      2  5.47907860  6.27539769
## 948    the direction    10           0      2  2.80450057  6.26023050
## 949          he made    10           0      2  2.04745150  6.25747179
## 950           in its    10           0      2  2.07111909  6.22633806
## 951     the carriage    10           0      2  2.70440514  6.19346760
## 952        above the    10           0      2  2.48357552  6.12505519
## 953        i presume    10           0      2  5.33721089  6.11311470
## 954           i told    10           0      2  2.11791743  6.05710688
## 955       the copper    10           0      2  4.13958545  5.88120255
## 956         his room    10           0      2  1.86240835  5.77248483
## 957         near the    10           0      2  2.13722788  5.62764761
## 958           he did    10           0      2  1.77756913  5.50013085
## 959          in such    10           0      2  1.76663289  5.40675941
## 960        without a    10           0      2  1.79653001  5.37623537
## 961           a lady    10           0      2  1.76032282  5.24972578
## 962         save the    10           0      2  1.91891480  5.22916259
## 963           take a    10           0      2  1.73444846  5.21296550
## 964            now i    10           0      2  1.67126264  5.16097376
## 965         upon him    10           0      2  1.57340159  4.98015534
## 966        have some    10           0      2  1.58302960  4.97844204
## 967            go to    10           0      2  1.61037997  4.85939762
## 968          he must    10           0      2  1.55155496  4.84046591
## 969         in spite    10           0      2  6.97716068  4.81948173
## 970          away to    10           0      2  1.56867258  4.74607012
## 971           see it    10           0      2  1.50392406  4.71598901
## 972           for an    10           0      2  1.44774398  4.58051407
## 973         spite of    10           0      2  6.57770244  4.54375983
## 974         on which    10           0      2  1.43098162  4.53296970
## 975          like to    10           0      2  1.47761985  4.49499319
## 976         who have    10           0      2  1.40821725  4.44711762
## 977        that time    10           0      2  1.36275867  4.24785639
## 978          and saw    10           0      2  1.39423647  4.22956927
## 979       the latter    10           0      2  5.74904730  3.97154425
## 980          the top    10           0      2  5.74904730  3.97154425
## 981       upon which    10           0      2  1.20207616  3.81873296
## 982       from which    10           0      2  1.13531849  3.60925093
## 983         that our    10           0      2  1.13567542  3.56492583
```

```
## 984          he will    10            0      2  1.05593865  3.33720923
## 985          and two    10            0      2  0.94771138  2.93930715
## 986         take the    10            0      2  0.95990320  2.88786359
## 987            be at    10            0      2  0.83757909  2.66978889
## 988          over to    10            0      2  0.82482676  2.58123197
## 989          that no    10            0      2  0.79035043  2.50125258
## 990            it on    10            0      2  0.76726967  2.43569889
## 991         the side    10            0      2  0.79944906  2.41467646
## 992            see a    10            0      2  0.74017011  2.32630430
## 993          was all    10            0      2  0.70477345  2.23995266
## 994             or a    10            0      2  0.70183317  2.20809551
## 995           to our    10            0      2  0.65657789  2.06331112
## 996        found the    10            0      2  0.67384059  2.06243643
## 997          were in    10            0      2  0.54233878  1.72463276
## 998          from it    10            0      2  0.51166137  1.63279281
## 999          some of    10            0      2  0.43298219  1.36957917
## 1000           him a    10            0      2  0.43037852  1.36265479
## 1001         been in    10            0      2  0.41720331  1.32908049
## 1002         of what    10            0      2  0.41713505  1.31903862
## 1003         you for    10            0      2  0.36958568  1.18046013
## 1004      which have    10            0      2  0.33826819  1.08205683
## 1005        that her    10            0      2  0.33708726  1.07462862
## 1006          it for    10            0      2  0.26267279  0.83932133
## 1007          when a    10            0      2  0.21022855  0.66832607
## 1008        the back    10            0      2  0.19067374  0.59368228
## 1009         and has    10            0      2  0.18230610  0.57797971
## 1010          are to    10            0      2  0.17471862  0.55522953
## 1011          and by    10            0      2  0.07837562  0.24896108
## 1012          what a    10            0      2  0.03797768  0.12104991
## 1013         with it    10            0      2 -0.01076477 -0.03449974
## 1014           on to    10            0      2 -0.02206900 -0.07034629
## 1015        and were    10            0      2 -0.04743671 -0.15099909
## 1016         was his    10            0      2 -0.46264332 -1.48601837
## 1017        were the    10            0      2 -0.64968326 -2.07044580
## 1018          it and    10            0      2 -0.72398927 -2.32804743
## 1019          me the    10            0      2 -0.84957960 -2.71500214
## 1020        and have    10            0      2 -1.05489495 -3.39177790
## 1021         that in    10            0      2 -1.08303989 -3.48687042
```

```
arrange(kollokationen, desc(lambda))
```

```
##          collocation count count_nested length      lambda           z
## 1     copper beeches    11            0      2 15.22637484  7.53175028
## 2        stoke moran    11            0      2 13.61691449  8.63462804
## 3       hosmer angel    17            0      2 13.44889189  8.90796472
## 4       briony lodge    11            0      2 13.28043103  8.62168994
## 5        irene adler    14            0      2 13.26089577 11.19826743
## 6      boscombe pool    10            0      2  9.79292293 15.83772416
## 7        miss hunter    17            0      2  9.64151093 11.21837593
## 8         neville st    16            0      2  9.34257026 13.48435034
## 9       baker street    29            0      2  9.27899246 17.71108500
## 10       miss stoner    15            0      2  8.64363060 14.31429355
## 11           lord st    28            0      2  8.54484490 22.80325305
## 12   sherlock holmes    85            0      2  8.46440557 21.87511087
## 13      your majesty    15            0      2  7.78729278  9.09056935
```

```
## 14         years ago    13           0      2  7.66167596 18.80491820
## 15              i am   182           0      2  7.40635244 13.71613613
## 16       will excuse    10           0      2  7.35576326 10.40842742
## 17          no doubt    50           0      2  7.12270475 23.38309038
## 18          in spite    10           0      2  6.97716068  4.81948173
## 19         am afraid    14           0      2  6.93400765 15.19358566
## 20       few minutes    18           0      2  6.93125856 21.13788519
## 21             i beg    12           0      2  6.61087746  4.58342616
## 22       our visitor    16           0      2  6.59251658 16.42608301
## 23          spite of    10           0      2  6.57770244  4.54375983
## 24        young lady    20           0      2  6.49853388 22.03060418
## 25       either side    10           0      2  6.42443267 16.26853940
## 26           an hour    21           0      2  6.27830771 16.98144824
## 27        an instant    20           0      2  6.22747453 16.70649652
## 28        each other    13           0      2  6.12670371 16.81218126
## 29        the stairs    13           0      2  6.00089869  4.16662262
## 30           am sure    18           0      2  5.97782226 18.96076853
## 31        don't know    11           0      2  5.90385885 15.22441507
## 32          at least    27           0      2  5.84877448 15.46699345
## 33     the contrary     11           0      2  5.84019803  4.04250910
## 34          the same    53           0      2  5.77561597  8.92383396
## 35         less than    13           0      2  5.75502275 16.69773086
## 36        the latter    10           0      2  5.74904730  3.97154425
## 37           the top    10           0      2  5.74904730  3.97154425
## 38     confess that     12           0      2  5.69094271  8.20829361
## 39          too much    13           0      2  5.61941666 17.14936700
## 40         once more    21           0      2  5.61673474 20.54819028
## 41           my dear    34           0      2  5.60720462 16.39394939
## 42          two days    11           0      2  5.58035950 15.61834896
## 43            let us    22           0      2  5.56355350 20.77743144
## 44         thank you    20           0      2  5.53670664 12.27194200
## 45        our client    11           0      2  5.53656651 14.75304445
## 46          beg that    10           0      2  5.51523843  7.83170844
## 47         number of    10           0      2  5.47907860  6.27539769
## 48       her husband    11           0      2  5.47859568 13.03602455
## 49      my companion    22           0      2  5.46978658 13.54303161
## 50       good enough    10           0      2  5.44123849 15.21800397
## 51        last night    11           0      2  5.42925744 15.88479161
## 52        shall soon    11           0      2  5.42233253 15.13433891
## 53           seem to    15           0      2  5.34189751  7.83443768
## 54         i presume    10           0      2  5.33721089  6.11311470
## 55         of course    51           0      2  5.31717645 14.34118528
## 56        can hardly    12           0      2  5.29031502 15.67384385
## 57         i suppose    16           0      2  5.28042811  7.77743063
## 58       waiting for    13           0      2  5.24729268 13.28230796
## 59          be happy    10           0      2  5.22647877 11.28225135
## 60         more than    31           0      2  5.21251175 23.80269743
## 61        why should    13           0      2  5.11861025 15.72362062
## 62           man who    38           0      2  5.06904447 25.45347037
## 63          signs of    11           0      2  5.05960862  7.24753625
## 64        the centre    15           0      2  5.04078290  5.89424623
## 65          began to    20           0      2  5.03562115  9.66595218
## 66       said holmes   109           0      2  5.01770711 38.06652349
## 67   remarked holmes    10           0      2  5.01289866 13.15421578
```

```
##  68          i fancy    12           0      2  5.00141635  7.21624441
##  69       afraid that    13           0      2  4.98008829  9.83163208
##  70       other side    12           0      2  4.94529104 15.38896486
##  71           a pair    10           0      2  4.93066761  7.00368446
##  72        the world    13           0      2  4.90227443  5.69519394
##  73            a few    61           0      2  4.89326104 17.17445662
##  74        two years    10           0      2  4.88257627 14.03001541
##  75        as though    10           0      2  4.86669785 10.51813057
##  76    his shoulders    10           0      2  4.82426553  9.64907805
##  77        seemed to    52           0      2  4.81838413 16.20417273
##  78         in front    35           0      2  4.77586544 15.63638835
##  79           far as    23           0      2  4.77132878 17.18541806
##  80         it seems    17           0      2  4.76045959 11.78605224
##  81     asked holmes    12           0      2  4.75058097 14.11014080
##  82     believe that    17           0      2  4.69571924 11.62735601
##  83        my friend    37           0      2  4.65455080 19.24735122
##  84       account of    13           0      2  4.63293281  8.50539636
##  85          able to    28           0      2  4.62095379 12.31690652
##  86        which led    11           0      2  4.61479385 11.45200769
##  87     could hardly    11           0      2  4.60462765 13.44428735
##  88           an old    17           0      2  4.59964260 16.20902109
##  89          be able    12           0      2  4.57376966 12.54140985
##  90         his lips    10           0      2  4.55597881  9.85384472
##  91       at present    12           0      2  4.55056608 11.91523886
##  92        of events    12           0      2  4.52230591  8.22078676
##  93  understand that    14           0      2  4.50563449 10.77308446
##  94         he spoke    15           0      2  4.50463318 11.54471568
##  95         has been    68           0      2  4.49859081 30.00357371
##  96           let me    26           0      2  4.48527519 17.95809710
##  97      should like    12           0      2  4.44264225 14.03919120
##  98          at once    34           0      2  4.43395098 19.71759591
##  99        his chair    29           0      2  4.43144281 16.61171322
## 100        front of    32           0      2  4.42031568 13.62494273
## 101          or two    18           0      2  4.41356407 16.84662610
## 102       very well    20           0      2  4.41056648 16.84763069
## 103     the coroner    13           0      2  4.39143683  6.37670117
## 104       know what    22           0      2  4.38924972 18.25556598
## 105      think that    76           0      2  4.37638997 25.01958705
## 106        his head    33           0      2  4.36474017 17.72953304
## 107        i cannot    37           0      2  4.35068809 14.04240347
## 108      his pocket    11           0      2  4.34551203 10.39366531
## 109       will find    13           0      2  4.33810455 14.19705184
## 110       my sister    12           0      2  4.33772267 11.20053335
## 111        we shall    48           0      2  4.32870326 24.52103303
## 112       the scene    12           0      2  4.31429677  6.22591260
## 113      the church    12           0      2  4.31429677  6.22591260
## 114        your son    11           0      2  4.31389136 12.57221017
## 115     appeared to    16           0      2  4.30613721  9.76801417
## 116          so far    11           0      2  4.30409033 12.71442240
## 117       the table    36           0      2  4.29151357 10.69799323
## 118        the king    21           0      2  4.27042278  8.23484052
## 119       enough to    28           0      2  4.26950968 12.89860521
## 120          my own    44           0      2  4.26213181 21.07084335
## 121          cry of    15           0      2  4.26101480  9.56970687
```

```
## 122         his eyes    41           0      2  4.25801160 19.74316276
## 123        very much    26           0      2  4.25774651 18.74148579
## 124        she cried    10           0      2  4.25770191 11.91734335
## 125          told me    16           0      2  4.25649249 14.02656817
## 126         the hall    16           0      2  4.25616057  7.23034815
## 127          how did    11           0      2  4.23861724 13.17850718
## 128           as far    20           0      2  4.23754887 14.82137107
## 129        come back    11           0      2  4.22351865 13.13873970
## 130    the inspector    20           0      2  4.22261546  8.10851905
## 131       would give    10           0      2  4.21795136 12.12762391
## 132         quite so    19           0      2  4.21266066 16.39192548
## 133         in order    12           0      2  4.20814068  9.76108715
## 134     have already    18           0      2  4.19571560 13.86695422
## 135        after all    18           0      2  4.19444823 15.82772205
## 136          tell us    10           0      2  4.18948516 12.44739102
## 137          tell me    28           0      2  4.18282706 18.25559973
## 138        they were    35           0      2  4.18192363 21.52841359
## 139         does not    10           0      2  4.17456805 11.24654741
## 140        young man    12           0      2  4.16469396 13.20678166
## 141       appears to    10           0      2  4.16199595  7.90140410
## 142          did not    41           0      2  4.16168583 22.06324916
## 143       the copper    10           0      2  4.13958545  5.88120255
## 144        i confess    11           0      2  4.12919788  7.95946489
## 145       in england    11           0      2  4.12417462  9.39217512
## 146          used to    15           0      2  4.11805275  9.63794057
## 147          so much    21           0      2  4.11169812 16.76769379
## 148           air of    11           0      2  4.10405098  8.35632549
## 149         had gone    14           0      2  4.09889541 12.47483150
## 150         my heart    10           0      2  4.09465347 10.21205527
## 151       have given    12           0      2  4.09082128 11.35010787
## 152          my wife    21           0      2  4.08186059 14.58654261
## 153         seems to    16           0      2  4.06972532  9.98062672
## 154       the second    17           0      2  4.06385382  7.68605525
## 155         the city    17           0      2  4.06385382  7.68605525
## 156        his hands    31           0      2  4.06333479 17.14330475
## 157          must be    44           0      2  4.06322789 22.43899167
## 158          to meet    11           0      2  4.05707547  8.26097958
## 159        knew that    26           0      2  4.05368593 14.82241147
## 160          sign of    12           0      2  4.04471577  8.74872099
## 161        have been   130           0      2  4.03991688 35.67714003
## 162       the police    35           0      2  4.02714152 11.01678415
## 163          at last    29           0      2  4.02133484 17.80375398
## 164           do not    59           0      2  4.01485429 25.74325441
## 165        could see    30           0      2  4.01478006 19.60465618
## 166         a single    12           0      2  4.00711353  8.66767475
## 167        came back    10           0      2  3.99592348 12.01986584
## 168     this morning    20           0      2  3.99391636 15.76894791
## 169       do nothing    10           0      2  3.98568981 11.87057302
## 170         had been   116           0      2  3.98547315 34.05719824
## 171            a low    10           0      2  3.97510992  7.95854914
## 172         far from    11           0      2  3.97468235 11.83880232
## 173        away from    17           0      2  3.95997552 14.51393157
## 174          and yet    46           0      2  3.95539643 16.20218043
## 175         one side    12           0      2  3.95418266 12.64211586
```

```
## 176          look at    18           0      2  3.93709203 14.22462308
## 177           put on    10           0      2  3.93382187 11.58416592
## 178           i knew    36           0      2  3.92932622 14.41839204
## 179        the whole    35           0      2  3.92704609 11.17014542
## 180        have done    22           0      2  3.91445104 15.00550671
## 181          said he   140           0      2  3.90647705 36.24068656
## 182        likely to    12           0      2  3.90177405  8.76753367
## 183        could not    64           0      2  3.89208470 26.26456756
## 184          side of    30           0      2  3.89054933 13.78029362
## 185         he cried    18           0      2  3.88736833 12.67602063
## 186         might be    30           0      2  3.88182448 18.26352323
## 187        should be    49           0      2  3.86372270 22.99975471
## 188       the ground    17           0      2  3.86317115  7.89963058
## 189       doubt that    22           0      2  3.85160128 13.59843552
## 190          made up    13           0      2  3.84071113 12.89317245
## 191        sure that    14           0      2  3.83905285 10.91504533
## 192      he remarked    19           0      2  3.82951854 12.97227396
## 193          a small    45           0      2  3.81482225 16.69276157
## 194          he said    37           0      2  3.81268390 17.94239347
## 195          saw him    14           0      2  3.76689107 12.87995928
## 196           i felt    19           0      2  3.76199192 10.61028860
## 197          wish to    23           0      2  3.75951253 12.07048546
## 198          i shall    97           0      2  3.75946992 23.77142127
## 199          a large    22           0      2  3.75130861 11.74685171
## 200         his hand    39           0      2  3.74867928 18.78284659
## 201         sight of    10           0      2  3.74439671  8.10764140
## 202          i don't    13           0      2  3.74363996  8.83136076
## 203        chair and    14           0      2  3.74137670 11.07290411
## 204        there are    50           0      2  3.74043104 23.17162673
## 205         would be    66           0      2  3.72851605 25.85271219
## 206         say that    29           0      2  3.72749135 15.46119405
## 207         order to    10           0      2  3.72664321  8.06932799
## 208          name is    14           0      2  3.71344986 11.52475268
## 209          ask you    11           0      2  3.70853707 10.19508684
## 210          my mind    18           0      2  3.70838000 13.12995419
## 211         the lamp    17           0      2  3.69610509  8.04299443
## 212        corner of    13           0      2  3.69458232  9.19522860
## 213        felt that    12           0      2  3.68928012 10.05058061
## 214        came down    12           0      2  3.68458457 12.12415305
## 215    possible that    13           0      2  3.66424031 10.42499953
## 216        hope that    11           0      2  3.66080410  9.62083676
## 217         amid the    10           0      2  3.65376577  6.48265893
## 218            a cab    14           0      2  3.64541332  9.43259789
## 219        the truth    11           0      2  3.64292557  6.54954936
## 220         they are    22           0      2  3.64176388 15.74099914
## 221        my father    21           0      2  3.64092300 14.03037669
## 222           whom i    15           0      2  3.62199139 11.03524204
## 223          as well    20           0      2  3.61820367 13.96881707
## 224            it is   334           0      2  3.61356047 49.64404560
## 225         his face    40           0      2  3.60006720 18.70802468
## 226           end of    19           0      2  3.58918020 11.03739459
## 227           one or    14           0      2  3.58609816 12.60989755
## 228           may be    37           0      2  3.58447581 19.22312548
## 229        very good    13           0      2  3.58302459 12.06146985
```

```
## 230          read it    10           0      2  3.58237050  9.55485843
## 231           a year    13           0      2  3.57357267  9.09252325
## 232         he asked    23           0      2  3.57340812 13.93002038
## 233        i believe    18           0      2  3.57082910 10.35448129
## 234        among the    24           0      2  3.56547177  9.96648799
## 235        all right    10           0      2  3.55372257 10.70256328
## 236       the cellar    10           0      2  3.55177484  6.30188040
## 237     the boscombe    10           0      2  3.55177484  6.30188040
## 238         the bell    10           0      2  3.55177484  6.30188040
## 239         and then    68           0      2  3.54119022 19.86898315
## 240        some time    12           0      2  3.54098625 11.70498046
## 241      the coronet    19           0      2  3.53638914  8.59928147
## 242          did you    42           0      2  3.52695787 19.06320152
## 243        how could    12           0      2  3.50188590 11.55997369
## 244        it seemed    23           0      2  3.49277438 13.67157288
## 245        have seen    17           0      2  3.45470629 12.48066955
## 246         what did    14           0      2  3.44683901 12.14793589
## 247         his wife    15           0      2  3.43890024 11.37602728
## 248           such a    51           0      2  3.43640941 18.01681261
## 249         pounds a    12           0      2  3.43328382  8.88273437
## 250           if you    69           0      2  3.42878548 23.86960750
## 251       had better    10           0      2  3.42382299  9.75620181
## 252      as possible    11           0      2  3.42227577 10.15924048
## 253           do you    63           0      2  3.42116427 22.82042728
## 254           a week    12           0      2  3.41925749  8.69989105
## 255         you will    73           0      2  3.40266479 24.06143052
## 256            i saw    43           0      2  3.40224812 15.80525538
## 257      he answered    14           0      2  3.39804559 10.71827799
## 258          part of    13           0      2  3.39604313  9.10897714
## 259         shall be    28           0      2  3.39576203 16.21769019
## 260       the matter    82           0      2  3.38028347 17.85441563
## 261         tell you    24           0      2  3.37760080 14.20208876
## 262         your own    11           0      2  3.37745394 10.64758054
## 263        i thought    39           0      2  3.37588863 15.04226499
## 264       the corner    20           0      2  3.37524576  8.88808072
## 265         only one    14           0      2  3.37102446 11.98450725
## 266         i should    99           0      2  3.36268362 23.83984865
## 267          what do    24           0      2  3.35694754 15.27416278
## 268         there is   104           0      2  3.33991072 28.65184832
## 269          his son    12           0      2  3.33089706 10.04876852
## 270          just as    19           0      2  3.32656287 13.31661880
## 271       might have    26           0      2  3.32537278 14.99736971
## 272          her own    10           0      2  3.31003461 10.03178398
## 273          you can    44           0      2  3.30677163 18.53899539
## 274        must have    33           0      2  3.29932765 16.73352409
## 275         had left    16           0      2  3.29515823 11.89944642
## 276       the garden    11           0      2  3.27517685  6.67163047
## 277          who has    18           0      2  3.27493104 13.14211013
## 278      this matter    14           0      2  3.26591041 11.52648448
## 279      the windows    16           0      2  3.25755991  7.99391402
## 280          here is    21           0      2  3.25661719 13.09312797
## 281          one who    14           0      2  3.25656725 11.62697793
## 282      the passage    14           0      2  3.23922758  7.49460629
## 283          my life    10           0      2  3.23794287  9.20806109
```

```
## 284         is quite    20           0      2  3.23148882 12.71864337
## 285          i think    78           0      2  3.22842716 20.97740738
## 286          we were    40           0      2  3.22630628 18.57108253
## 287           we may    23           0      2  3.21851887 14.30850780
## 288           a word    17           0      2  3.20321249 10.12490209
## 289         when she    20           0      2  3.19522101 13.48449489
## 290        in london    13           0      2  3.18685681  9.61931297
## 291           to get    24           0      2  3.17827023 11.92293614
## 292        down into    13           0      2  3.17515390 11.01304352
## 293           we are    37           0      2  3.17100705 17.66036283
## 294         the door    86           0      2  3.16132029 18.32161643
## 295       have heard    21           0      2  3.15748258 13.03280678
## 296          you are    77           0      2  3.15481393 23.61394579
## 297        have ever    13           0      2  3.15467067 10.35027811
## 298           no one    21           0      2  3.14453734 13.59385518
## 299         at first    13           0      2  3.14226313 10.47336030
## 300          we must    18           0      2  3.13343446 12.46874664
## 301   the photograph    14           0      2  3.12798998  7.48975423
## 302          his hat    10           0      2  3.12620531  8.86730210
## 303        there was   106           0      2  3.11434196 27.10750725
## 304     i understand    21           0      2  3.11293341 10.87370031
## 305       the window    42           0      2  3.10953210 12.82250469
## 306        the facts    21           0      2  3.10023180  9.11563198
## 307          no more    11           0      2  3.09831311  9.96331802
## 308          but how    11           0      2  3.09256219  9.68295740
## 309         not know    21           0      2  3.08108478 13.14667236
## 310      against the    29           0      2  3.08087768 10.99246007
## 311       there were    30           0      2  3.07776663 15.67760142
## 312         this way    11           0      2  3.07372823  9.78447240
## 313       during the    21           0      2  3.07318644  9.38068354
## 314         you know    39           0      2  3.07075668 16.71498020
## 315          will be    31           0      2  3.07012795 15.74114573
## 316         give you    12           0      2  3.06661634  9.58202533
## 317         the next    13           0      2  3.05635196  7.21334956
## 318       his father    15           0      2  3.05511314 10.59765483
## 319         told you    10           0      2  3.05368010  8.76163638
## 320           when i    82           0      2  3.04404792 23.27092300
## 321       the letter    17           0      2  3.04210679  8.20890880
## 322        this case    10           0      2  3.03914036  9.26781393
## 323         had seen    11           0      2  3.03747400  9.38548517
## 324      some little    13           0      2  3.03165396 10.56006258
## 325      through the    47           0      2  3.02344846 13.91110487
## 326           to ask    13           0      2  3.02174987  8.68780284
## 327         a little    99           0      2  3.01939619 23.50690378
## 328        all about    10           0      2  3.01651809  9.29416962
## 329          i asked    27           0      2  3.00960656 12.14448234
## 330        away with    10           0      2  3.00864545  9.05255257
## 331           it was   278           0      2  3.00566665 40.77948755
## 332          as much    18           0      2  3.00187355 11.72710197
## 333         hand and    14           0      2  2.99387737  9.93690005
## 334         the fire    24           0      2  2.98772003  9.69029531
## 335            do so    14           0      2  2.97583532 10.73345585
## 336          to keep    10           0      2  2.97519998  7.61196513
## 337          his own    20           0      2  2.97246042 11.94293031
```

```
## 338       i answered    18           0      2  2.96161422  9.89964769
## 339      the office     13           0      2  2.95625653  7.18047536
## 340        a strong     12           0      2  2.95380174  8.29168069
## 341       not think     19           0      2  2.95357515 12.10685325
## 342         he took     17           0      2  2.95026388 10.83223253
## 343          i hope     12           0      2  2.94709542  8.12274588
## 344        that she     70           0      2  2.94470503 20.94842895
## 345       he looked     11           0      2  2.94380262  8.77979996
## 346       the house     63           0      2  2.94252170 15.56418365
## 347         you may     39           0      2  2.94226196 16.23730888
## 348         a woman     22           0      2  2.93695440 11.08834392
## 349        until he     12           0      2  2.93595276  9.42287696
## 350       is always     10           0      2  2.92592354  8.52878154
## 351         my hand     19           0      2  2.92477661 11.65228271
## 352         will do     11           0      2  2.92440806  9.46233383
## 353       would not     33           0      2  2.91643147 15.55594382
## 354       the stone     16           0      2  2.90611413  7.91321548
## 355         find it     14           0      2  2.90268885  9.89140808
## 356         she has     19           0      2  2.89759756 12.03841458
## 357     the colonel     11           0      2  2.89565131  6.60133450
## 358       the other     80           0      2  2.89245688 17.46429672
## 359        at night     10           0      2  2.88696463  8.65846954
## 360          he has     54           0      2  2.87336522 18.60430945
## 361        out into     13           0      2  2.87322991 10.04736864
## 362        may have     28           0      2  2.87056581 13.95637081
## 363        became a     10           0      2  2.86411672  7.59588616
## 364       come from     10           0      2  2.86371243  8.79686580
## 365        all over     10           0      2  2.86259112  8.86146368
## 366         who had     35           0      2  2.86237864 15.57327860
## 367         the air     14           0      2  2.85351721  7.38855178
## 368         and why     11           0      2  2.85308740  7.66384847
## 369         when he     47           0      2  2.85280276 17.72206904
## 370      friend and     10           0      2  2.85264690  8.19032637
## 371         gone to     11           0      2  2.84439271  7.83483385
## 372         such as     11           0      2  2.84288680  9.05293347
## 373          to say     32           0      2  2.83992295 13.07711398
## 374      i remarked     17           0      2  2.83785315  9.44447906
## 375        what can     10           0      2  2.83251066  8.75733352
## 376          i fear     10           0      2  2.82470807  7.30193066
## 377           to go     33           0      2  2.82369095 13.23613730
## 378        said she     20           0      2  2.82104287 12.01898758
## 379         she had     50           0      2  2.81416311 18.17695847
## 380          i have    299           0      2  2.81388870 38.17767146
## 381      closed the     10           0      2  2.80639651  6.42673990
## 382   the direction     10           0      2  2.80450057  6.26023050
## 383       have made     14           0      2  2.80099910  9.80234956
## 384           to be    199           0      2  2.79790897 31.71857091
## 385          as if     18           0      2  2.78153947 11.03833344
## 386        will not     23           0      2  2.78080403 12.59430989
## 387          a note     10           0      2  2.77868572  7.37017024
## 388       the floor     14           0      2  2.77654420  7.33646141
## 389         you see     40           0      2  2.76825155 15.72238916
## 390          a long     28           0      2  2.76573636 12.09489852
## 391         have no     35           0      2  2.76030542 15.02389325
```

```
## 392           as we    45           0      2  2.75804402 16.97395417
## 393          we can    13           0      2  2.75308151  9.56931837
## 394         then he    28           0      2  2.74934934 13.42868037
## 395      a gentleman    13           0      2  2.74921221  8.30600584
## 396         he might    22           0      2  2.74418710 11.65046191
## 397         you must    28           0      2  2.74304740 13.14093629
## 398          a great    18           0      2  2.73967931  9.70052290
## 399          to find    25           0      2  2.73175713 11.34733366
## 400        was still    12           0      2  2.72659035  8.63126800
## 401           am not    16           0      2  2.72499686 10.41995056
## 402          to hear    10           0      2  2.72382771  7.27786919
## 403        we should    16           0      2  2.72065771 10.44207036
## 404           i met    13           0      2  2.71758981  8.11432674
## 405          i found    35           0      2  2.71307068 13.12754129
## 406        know that    27           0      2  2.70873487 12.60349722
## 407          i could    92           0      2  2.70589835 21.07687598
## 408         the road    16           0      2  2.70540751  7.76457301
## 409          take it    12           0      2  2.70484998  8.71205925
## 410     the carriage    10           0      2  2.70440514  6.19346760
## 411           can be    15           0      2  2.70337304 10.01849641
## 412       should not    18           0      2  2.70191249 10.93503393
## 413          to make    22           0      2  2.70174092 10.59404666
## 414         see that    37           0      2  2.69401006 14.62689410
## 415         had come    16           0      2  2.69040990 10.17131151
## 416          i went    21           0      2  2.68990251 10.17723435
## 417          see him    12           0      2  2.68917137  9.06486650
## 418        the first    40           0      2  2.68502509 12.13222953
## 419          i cried    14           0      2  2.68211338  8.34822797
## 420         for some    22           0      2  2.67582820 11.85504481
## 421         into the   125           0      2  2.67296650 21.81922836
## 422         face and    14           0      2  2.67235173  9.16665550
## 423         of those    18           0      2  2.67078299  9.56422583
## 424           one of    93           0      2  2.66595353 21.43225912
## 425           he had   130           0      2  2.64944797 26.49850170
## 426       a question    10           0      2  2.64803084  7.16870642
## 427         door and    13           0      2  2.64706540  8.78312713
## 428     thought that    12           0      2  2.64584958  8.37802176
## 429      the country    15           0      2  2.64270804  7.46535849
## 430       across the    16           0      2  2.64041771  7.88633582
## 431           but if    12           0      2  2.63964252  8.82476846
## 432      of interest    11           0      2  2.62816678  7.47209472
## 433          to tell    26           0      2  2.61766000 11.27762688
## 434         you have   133           0      2  2.61439171 26.52649721
## 435          name of    12           0      2  2.60935402  7.79160823
## 436            yet i    13           0      2  2.60492891  8.56811941
## 437         not only    12           0      2  2.60000460  8.71918942
## 438         the most    40           0      2  2.59295828 11.97811661
## 439            a man    80           0      2  2.59289208 19.46762012
## 440             if i    48           0      2  2.58970355 16.04750930
## 441          like a    24           0      2  2.58756149 10.90701111
## 442          out of    75           0      2  2.58523034 18.91777631
## 443          had no    27           0      2  2.57690068 12.58509823
## 444        though i    11           0      2  2.57564097  7.84410453
## 445       beside the    11           0      2  2.57474427  6.51612838
```

```
## 446         it would    52           0      2  2.57051904 16.71870799
## 447          to look    18           0      2  2.56703483  9.31468710
## 448           then i    37           0      2  2.56625414 14.04661851
## 449           gave a    12           0      2  2.56624241  7.75829141
## 450          while i    12           0      2  2.56338383  8.14669412
## 451         out upon    16           0      2  2.55723579  9.89634950
## 452          to give    15           0      2  2.55220081  8.49683555
## 453         his long    12           0      2  2.54725742  8.30232268
## 454        she would    16           0      2  2.54158083  9.85059311
## 455        was quite    14           0      2  2.53976330  8.79895375
## 456         upon the   196           0      2  2.53972500 26.66473930
## 457           to see    61           0      2  2.53834208 16.79260762
## 458         the room    61           0      2  2.52104442 14.58554093
## 459         where he    13           0      2  2.51883306  8.64402139
## 460         the best    14           0      2  2.51698511  7.08239225
## 461         within a    11           0      2  2.51187181  7.33621870
## 462          when we    14           0      2  2.51124252  9.15813099
## 463         man with    15           0      2  2.50650561  9.36173307
## 464          we have    54           0      2  2.49915142 16.97290742
## 465           who is    33           0      2  2.49361486 13.31157942
## 466            i can    46           0      2  2.49264007 14.26723686
## 467       that there    53           0      2  2.48995752 16.24358280
## 468        above the    10           0      2  2.48357552  6.12505519
## 469           a good    26           0      2  2.48073862 10.90644029
## 470        that they    23           0      2  2.47978940 10.79785758
## 471         over his    20           0      2  2.47812711 10.41010806
## 472           i took    22           0      2  2.47408022  9.89623171
## 473            if we    11           0      2  2.46545622  8.04702134
## 474        down upon    11           0      2  2.46490626  8.02501823
## 475      the morning    42           0      2  2.46092407 11.99808794
## 476          back to    27           0      2  2.45957003 11.07229672
## 477         of these    18           0      2  2.45935824  9.07165292
## 478           a very    90           0      2  2.45863961 19.90034732
## 479       thought of    16           0      2  2.45679693  8.59880011
## 480       would have    32           0      2  2.45642448 13.04839581
## 481           that i   249           0      2  2.44788414 33.04409207
## 482        have come    14           0      2  2.42839515  8.69175153
## 483            but i   100           0      2  2.42532576 21.66472464
## 484         came to    35           0      2  2.41599321 12.41417012
## 485        the name    26           0      2  2.41458494  9.39191036
## 486         i heard    30           0      2  2.41441055 11.32947450
## 487          with a   184           0      2  2.41013272 27.92474703
## 488         that he   153           0      2  2.40233266 26.14213737
## 489        before he    14           0      2  2.39666997  8.57704458
## 490         but what    16           0      2  2.36587622  9.15802478
## 491         my room     14           0      2  2.36117911  8.44264574
## 492        you think    22           0      2  2.35898838 10.35409755
## 493           i wish    11           0      2  2.35239265  6.84570146
## 494           on the   141           0      2  2.35070121 21.77172619
## 495       what could    11           0      2  2.34893717  7.68380527
## 496          and let    11           0      2  2.34211091  6.81587885
## 497       and looked    13           0      2  2.33364847  7.36488602
## 498       holmes had    16           0      2  2.33109997  8.97518227
## 499          until i    11           0      2  2.32641376  7.23612599
```

```
## 500         i passed    10           0      2  2.32461290  6.48783915
## 501        were all    12           0      2  2.32088734  7.91043083
## 502      found that    14           0      2  2.31931852  8.12481277
## 503         come to    29           0      2  2.31603763 10.98479691
## 504      the papers    18           0      2  2.30921885  7.67149061
## 505          for us    13           0      2  2.30438652  8.07765455
## 506          it may    25           0      2  2.29545808 10.69445707
## 507        this man    13           0      2  2.29513437  8.11290489
## 508           to do    55           0      2  2.28916400 14.84600117
## 509       they have    17           0      2  2.28878269  9.05612641
## 510     should have    18           0      2  2.28590007  9.29620695
## 511        the case    43           0      2  2.28521592 11.68941545
## 512         that we    55           0      2  2.28374434 15.42416299
## 513            as i   118           0      2  2.27891008 22.22761975
## 514      opened the    11           0      2  2.27841895  6.12337824
## 515       time that    14           0      2  2.27302182  7.98889315
## 516        in every    11           0      2  2.27223936  7.04316685
## 517        the lady    27           0      2  2.27175046  9.26785781
## 518         you say    11           0      2  2.26886466  7.18399435
## 519        could be    17           0      2  2.25529578  9.01738757
## 520          by the   123           0      2  2.25130166 19.84752115
## 521           if he    23           0      2  2.25054315 10.28379796
## 522         which i   103           0      2  2.24686200 20.60507983
## 523           is no    26           0      2  2.24668163 10.83894073
## 524       which has    19           0      2  2.24227770  9.41753377
## 525        we could    14           0      2  2.24036945  8.19794926
## 526          he was   151           0      2  2.23996944 24.58040073
## 527        to leave    10           0      2  2.23986881  6.38596496
## 528         it must    20           0      2  2.23980419  9.40326761
## 529     the evening    13           0      2  2.23944714  6.44390629
## 530        heard of    22           0      2  2.23770490  9.37481699
## 531        he would    37           0      2  2.22583810 12.62458279
## 532         back in    15           0      2  2.22496053  8.06646498
## 533          was no    32           0      2  2.22411400 11.73847542
## 534          i must    37           0      2  2.22368047 11.84275612
## 535        are very    10           0      2  2.21700614  6.95696934
## 536          at all    24           0      2  2.21652410 10.41095280
## 537           to me   134           0      2  2.21636051 22.33756325
## 538       but there    17           0      2  2.21208430  8.85443488
## 539         case of    15           0      2  2.20630187  7.70390327
## 540        to take    16           0      2  2.20519262  7.90584979
## 541     between the    20           0      2  2.20413053  8.01335699
## 542          i know    39           0      2  2.19712724 12.04368422
## 543         went to    13           0      2  2.19673143  7.15442779
## 544        upon her    17           0      2  2.19642746  8.82168925
## 545         has not    15           0      2  2.19455467  8.29160919
## 546          to put    13           0      2  2.18750620  7.10972364
## 547         time to    21           0      2  2.18622363  8.98658564
## 548           i got    11           0      2  2.18290488  6.48552528
## 549          up and    29           0      2  2.18277844 11.02196913
## 550           i did    29           0      2  2.18096365 10.36125608
## 551         to come    29           0      2  2.17152653 10.43086131
## 552         she was    46           0      2  2.17150425 13.71042214
## 553         my face    11           0      2  2.16843310  6.97833210
```

```
## 554         which he    61           0      2  2.16638767 15.78429717
## 555         from his    45           0      2  2.16574440 13.64208870
## 556         it might    14           0      2  2.16172716  7.68248620
## 557        which she    18           0      2  2.16030927  8.86768102
## 558         upon his    42           0      2  2.15953613 13.16675348
## 559         out from    11           0      2  2.14763637  7.04303387
## 560           we had    36           0      2  2.14404450 12.22586567
## 561         but when    11           0      2  2.14192983  6.99451175
## 562        the paper    14           0      2  2.14041174  6.50187010
## 563         near the    10           0      2  2.13722788  5.62764761
## 564        cried the    11           0      2  2.13528240  5.88587610
## 565         saw that    10           0      2  2.11990813  6.41081076
## 566           i told    10           0      2  2.11791743  6.05710688
## 567            i had   168           0      2  2.11241726 23.79015438
## 568          be some    11           0      2  2.11222344  6.90969960
## 569          this is    40           0      2  2.11123304 12.60144093
## 570          into my    19           0      2  2.10118782  8.84317217
## 571         about it    15           0      2  2.08099909  7.75024058
## 572         with his    68           0      2  2.07442751 15.93809620
## 573           in its    10           0      2  2.07111909  6.22633806
## 574         he could    29           0      2  2.07110705 10.53369616
## 575         was only    16           0      2  2.06546710  7.87530002
## 576          a young    15           0      2  2.05555201  7.25945780
## 577         open the    12           0      2  2.05518398  5.98427235
## 578         all this    10           0      2  2.05359560  6.45963720
## 579          he made    10           0      2  2.04745150  6.25747179
## 580           at the   236           0      2  2.04164971 25.64398309
## 581        which are    19           0      2  2.04098272  8.62818405
## 582        you would    30           0      2  2.03845349 10.57597852
## 583        which may    11           0      2  2.03665519  6.64485017
## 584         before i    16           0      2  2.03291779  7.71523532
## 585           is now    13           0      2  2.03134751  7.10015911
## 586           is not    48           0      2  2.02449111 13.22423076
## 587            as he    60           0      2  2.02102264 14.68766726
## 588           in the   504           0      2  2.01246505 36.54144840
## 589          where i    13           0      2  2.00878100  6.91088111
## 590          in this    48           0      2  2.00528189 12.89770571
## 591       which were    19           0      2  2.00513973  8.48545849
## 592         when you    26           0      2  2.00282638  9.75735599
## 593            up in    29           0      2  2.00081807 10.11910411
## 594          we will    10           0      2  2.00032396  6.28498255
## 595           if it    22           0      2  2.00020983  8.98396821
## 596          upon me    22           0      2  1.99771928  9.08369234
## 597           i came    28           0      2  1.99683204  9.50829040
## 598          made a    17           0      2  1.99664369  7.58650118
## 599       the money    14           0      2  1.99465203  6.21345123
## 600         the bed    13           0      2  1.99324222  5.99274963
## 601         of them    26           0      2  1.99018516  9.23608664
## 602        the last    30           0      2  1.98992769  8.99155466
## 603           a most    15           0      2  1.98193159  7.04720500
## 604          do with    11           0      2  1.97846715  6.48734014
## 605         which we    20           0      2  1.97358459  8.56571832
## 606           took a    14           0      2  1.97242601  6.83877273
## 607           said i    49           0      2  1.97039856 12.88492612
```

```
## 608         with him    26           0      2  1.96846699  9.66889781
## 609          so that    37           0      2  1.96833717 11.26230721
## 610         with her    24           0      2  1.96390477  9.28705643
## 611         all that    34           0      2  1.95910254 10.76599095
## 612           i will    48           0      2  1.95822821 12.18066886
## 613         not been    17           0      2  1.95563335  7.88279159
## 614      which would    17           0      2  1.95355286  7.85259002
## 615           not be    27           0      2  1.94701391  9.76192116
## 616         and now    30           0      2  1.94042793  9.59902658
## 617           be no    11           0      2  1.93703529  6.36271431
## 618          way to    13           0      2  1.93648119  6.46421501
## 619          of the   707           0      2  1.93512589 41.48444370
## 620         they had    11           0      2  1.93080975  6.31035701
## 621          in his   120           0      2  1.92702467 19.27904966
## 622         you did    11           0      2  1.92688558  6.21000379
## 623         but she    12           0      2  1.92364991  6.58001695
## 624        save the    10           0      2  1.91891480  5.22916259
## 625         upon my    27           0      2  1.90977276  9.56191938
## 626         door of    12           0      2  1.90793191  6.15285104
## 627         the old    20           0      2  1.88673759  7.10701913
## 628            i do    43           0      2  1.88241968 11.17277985
## 629         to know    28           0      2  1.88196291  9.12916746
## 630          into a    41           0      2  1.87674527 11.04619776
## 631          for me    30           0      2  1.87516890  9.89334892
## 632         is very    25           0      2  1.87217097  9.00410591
## 633        from the   136           0      2  1.86738944 18.35990344
## 634         are not    12           0      2  1.86610668  6.39549691
## 635       under the    11           0      2  1.86432413  5.35814449
## 636         what is    26           0      2  1.86368167  9.13746005
## 637        his room    10           0      2  1.86240835  5.77248483
## 638          what i    37           0      2  1.86065311 10.68478129
## 639          it has    24           0      2  1.85963472  8.70110385
## 640         who was    23           0      2  1.85267947  8.50852070
## 641           be so    12           0      2  1.84900256  6.33859873
## 642           me to    64           0      2  1.83988522 13.43108996
## 643         that it   116           0      2  1.83469097 18.16160700
## 644           i say    16           0      2  1.83273833  6.75089746
## 645         me that    38           0      2  1.81469803 10.59624208
## 646        think of    20           0      2  1.80696659  7.51293852
## 647       round the    18           0      2  1.80329462  6.62612451
## 648       without a    10           0      2  1.79653001  5.37623537
## 649           us in    12           0      2  1.79226095  5.99268751
## 650        that you   108           0      2  1.78815664 17.15617806
## 651      nothing of    13           0      2  1.78588883  6.04138918
## 652       the front    14           0      2  1.78164620  5.73372333
## 653          he did    10           0      2  1.77756913  5.50013085
## 654          as she    14           0      2  1.77658171  6.52898121
## 655         me with    16           0      2  1.77619595  6.98045207
## 656          but he    36           0      2  1.77491636 10.21555013
## 657          she is    26           0      2  1.77473886  8.72715493
## 658         the end    15           0      2  1.76949232  5.89753255
## 659        from her    12           0      2  1.76881655  6.07963606
## 660         in such    10           0      2  1.76663289  5.40675941
## 661           of us    27           0      2  1.76578857  8.50054785
```

```
## 662            up to    36           0      2  1.76149866  9.76994427
## 663           a lady    10           0      2  1.76032282  5.24972578
## 664         he came     11           0      2  1.75747317  5.69370108
## 665            is a    143           0      2  1.75535452 19.04958772
## 666          than i     14           0      2  1.74394614  6.30141919
## 667        with your    16           0      2  1.73767982  6.81521473
## 668          take a     10           0      2  1.73444846  5.21296550
## 669         that if     13           0      2  1.72811277  6.01306592
## 670       the light     20           0      2  1.72789338  6.65953270
## 671         down to     29           0      2  1.72620933  8.63805148
## 672       the night     21           0      2  1.72208052  6.80211665
## 673         it over     14           0      2  1.71685400  6.23683239
## 674         i never     14           0      2  1.71566864  5.98315878
## 675         was not     46           0      2  1.71530038 11.05064337
## 676          but we     13           0      2  1.71308084  6.11103545
## 677         of such     14           0      2  1.70297100  5.99143810
## 678         and how     13           0      2  1.70032542  5.72463600
## 679        not very     12           0      2  1.69817181  5.83719449
## 680      could have     14           0      2  1.68987441  6.21915373
## 681     the strange     12           0      2  1.68924501  5.11243841
## 682        down the     50           0      2  1.68148401 10.34838981
## 683           is an     19           0      2  1.68122603  7.13276337
## 684           now i     10           0      2  1.67126264  5.16097376
## 685          him in     22           0      2  1.66785367  7.50520398
## 686         heard a     13           0      2  1.66758054  5.70536578
## 687       the point     11           0      2  1.66542761  4.84995939
## 688           in my     77           0      2  1.66301764 13.66285520
## 689        were not     11           0      2  1.66185846  5.48763149
## 690          in our     17           0      2  1.65825301  6.57510311
## 691        to think     23           0      2  1.63912407  7.35741367
## 692          it all     25           0      2  1.63319908  7.86559389
## 693           i may     27           0      2  1.62499834  7.84430735
## 694           i was    186           0      2  1.61564530 19.87948723
## 695       and there     42           0      2  1.61515041  9.67347574
## 696        and took     11           0      2  1.61081708  5.04059636
## 697           go to     10           0      2  1.61037997  4.85939762
## 698         had not     25           0      2  1.60523013  7.80940943
## 699           was a    160           0      2  1.60394595 18.51521036
## 700        from him     11           0      2  1.60205543  5.29911637
## 701      you should     13           0      2  1.59953576  5.65138637
## 702           on my     16           0      2  1.59291204  6.26506577
## 703        the time     37           0      2  1.59176836  8.43729326
## 704      holmes was     13           0      2  1.58399178  5.58919040
## 705       have some     10           0      2  1.58302960  4.97844204
## 706          of his    131           0      2  1.57441642 16.51656969
## 707        upon him     10           0      2  1.57340159  4.98015534
## 708          at one     12           0      2  1.57275946  5.40613538
## 709          was so     21           0      2  1.57242973  6.98079208
## 710       which was     50           0      2  1.57192775 10.59731190
## 711         with me     26           0      2  1.57179666  7.79570560
## 712         away to     10           0      2  1.56867258  4.74607012
## 713           of mr     14           0      2  1.56430163  5.55462319
## 714           and i    207           0      2  1.56096182 20.25911391
## 715           he is     68           0      2  1.55372719 12.13070066
```

```
## 716          he must    10           0      2  1.55155496  4.84046591
## 717        that this    31           0      2  1.54761413  8.25194733
## 718         what you    20           0      2  1.54237146  6.72817197
## 719         of their    11           0      2  1.54112516  4.88272838
## 720         you were    20           0      2  1.53179581  6.66645062
## 721          the way    27           0      2  1.52645282  6.98707570
## 722          i would    41           0      2  1.52063339  9.06496733
## 723          are you    16           0      2  1.51641704  5.94866167
## 724         over the    35           0      2  1.51576707  7.97762633
## 725       the street    19           0      2  1.50911939  5.83123193
## 726           of our    22           0      2  1.50717391  6.68155572
## 727           in one    25           0      2  1.50518756  7.23692968
## 728        the place    12           0      2  1.50442707  4.65859860
## 729           see it    10           0      2  1.50392406  4.71598901
## 730           had an    12           0      2  1.49020693  5.12654630
## 731          that is    80           0      2  1.48810525 12.54365040
## 732        the young    18           0      2  1.48781907  5.61360310
## 733           him to    29           0      2  1.48630388  7.55180089
## 734          down in    15           0      2  1.48526720  5.59883273
## 735           up his    12           0      2  1.48209976  5.07794678
## 736           it out    20           0      2  1.48021168  6.43567892
## 737        you could    16           0      2  1.47882586  5.79148087
## 738          like to    10           0      2  1.47761985  4.49499319
## 739          for all    11           0      2  1.47749456  4.88649790
## 740            i see    28           0      2  1.47475249  7.32020856
## 741           at me    22           0      2  1.46988730  6.74871389
## 742           do it    12           0      2  1.45581791  4.98336043
## 743         quite a    11           0      2  1.45314711  4.64795686
## 744          and we    45           0      2  1.45293746  9.09784758
## 745           me as    11           0      2  1.44943574  4.80107410
## 746          for an    10           0      2  1.44774398  4.58051407
## 747           in an    24           0      2  1.44408078  6.82345970
## 748         will you    12           0      2  1.43795832  4.92788038
## 749         on which    10           0      2  1.43098162  4.53296970
## 750          the day    18           0      2  1.42573576  5.41659358
## 751         and what    29           0      2  1.42477523  7.21217506
## 752          you do    13           0      2  1.42246527  5.05250710
## 753           it up    18           0      2  1.42124878  5.88476575
## 754        the great    12           0      2  1.41236027  4.41861247
## 755         him that    16           0      2  1.41050660  5.51180561
## 756         who have    10           0      2  1.40821725  4.44711762
## 757          i might    14           0      2  1.40691053  5.01188551
## 758        about the    32           0      2  1.40394336  7.15273173
## 759          with an    11           0      2  1.39608131  4.61696198
## 760            me at    12           0      2  1.39516485  4.81458300
## 761          and saw    10           0      2  1.39423647  4.22956927
## 762           to him    47           0      2  1.39423576  8.99575858
## 763          for him    13           0      2  1.39407277  4.99523314
## 764          of this    41           0      2  1.38567503  8.37401695
## 765            to mr    12           0      2  1.38166352  4.60182228
## 766         was very    20           0      2  1.37813369  6.01263305
## 767           by his    15           0      2  1.37348365  5.24658656
## 768        and found    12           0      2  1.36840421  4.53501597
## 769         not have    22           0      2  1.36411683  6.27036079
```

```
## 770        that time    10           0      2  1.36275867  4.24785639
## 771         which is    33           0      2  1.36138661  7.58563314
## 772          from my    17           0      2  1.35507972  5.51113880
## 773           but it    29           0      2  1.33965464  7.01561729
## 774            at my    26           0      2  1.33762616  6.66201487
## 775          but you    28           0      2  1.33744832  6.89002395
## 776        while the    12           0      2  1.33379469  4.24858219
## 777         and made    11           0      2  1.32845057  4.23317027
## 778            is so    13           0      2  1.32800774  4.74255996
## 779        after the    17           0      2  1.31890142  4.97591626
## 780            of my    84           0      2  1.31469210 11.29644539
## 781            us to    12           0      2  1.30882332  4.38199394
## 782         into his    11           0      2  1.30084234  4.29449323
## 783           is all    14           0      2  1.29722441  4.80215037
## 784         what was    19           0      2  1.29599692  5.53188846
## 785             by a    31           0      2  1.29476707  6.88300187
## 786            as to    81           0      2  1.29065228 10.91485266
## 787          out his    11           0      2  1.28935766  4.25759279
## 788           just a    11           0      2  1.28190744  4.13484045
## 789          upon it    20           0      2  1.27878138  5.61102761
## 790          in some    14           0      2  1.27646544  4.67959501
## 791         in which    40           0      2  1.27243233  7.74303510
## 792          at this    11           0      2  1.27168410  4.21672558
## 793           to her    39           0      2  1.27079223  7.53367838
## 794          with my    26           0      2  1.26389438  6.30225960
## 795          in your    21           0      2  1.25895495  5.61279868
## 796             so i    20           0      2  1.25466930  5.47798995
## 797           for it    30           0      2  1.24957652  6.66409908
## 798        which had    22           0      2  1.23357224  5.68418658
## 799             as it    35           0      2  1.22456809  7.03450834
## 800        which you    30           0      2  1.21854475  6.50432025
## 801        the right    17           0      2  1.21677894  4.59121892
## 802          for you    28           0      2  1.21185787  6.25920859
## 803         with the   136           0      2  1.21110799 12.76557296
## 804           and so    31           0      2  1.20917106  6.39917645
## 805         have not    19           0      2  1.20394340  5.17354920
## 806       upon which    10           0      2  1.20207616  3.81873296
## 807        heard the    17           0      2  1.19742146  4.56577167
## 808          all the    53           0      2  1.18382168  7.87956821
## 809           of any    11           0      2  1.18343186  3.82080941
## 810          what he    12           0      2  1.18302911  4.08519590
## 811          it will    12           0      2  1.18215393  4.05780097
## 812          said my    13           0      2  1.18141873  4.24156583
## 813           was an    15           0      2  1.17987394  4.50693855
## 814           for he    23           0      2  1.17403961  5.52828564
## 815            to my    74           0      2  1.15768107  9.41814716
## 816            for i    37           0      2  1.15690853  6.81475922
## 817         have you    34           0      2  1.15490616  6.55277755
## 818             in a   135           0      2  1.15477861 12.55600685
## 819          in her    23           0      2  1.15141580  5.38131664
## 820        said that    22           0      2  1.14369876  5.24839068
## 821          what it    14           0      2  1.13921823  4.22648559
## 822           is one    11           0      2  1.13567970  3.76078864
## 823         that our    10           0      2  1.13567542  3.56492583
```

```
## 824       from which    10           0      2  1.13531849  3.60925093
## 825           as you    31           0      2  1.13143957  6.14486119
## 826         for the   110           0      2  1.12649488 10.77606128
## 827            be a    46           0      2  1.11649424  7.24467171
## 828    the business    12           0      2  1.11084375  3.57683809
## 829          out in    14           0      2  1.10944569  4.09391885
## 830        have had    23           0      2  1.08623344  5.12218434
## 831          man of    15           0      2  1.07818699  4.06172471
## 832        and down    20           0      2  1.07704872  4.63485449
## 833          at his    25           0      2  1.06168470  5.20912326
## 834         he will    10           0      2  1.05593865  3.33720923
## 835       was there    12           0      2  1.04822665  3.61010683
## 836         of some    17           0      2  1.04680361  4.18881479
## 837           me in    20           0      2  1.04292228  4.57555933
## 838          for my    18           0      2  1.03343426  4.33960591
## 839        took the    12           0      2  1.02095811  3.33654010
## 840        the only    23           0      2  1.01421631  4.50929070
## 841       where the    16           0      2  1.00972197  3.79301596
## 842         the man    43           0      2  0.99761093  6.03356114
## 843         to your    25           0      2  0.99529251  4.80819601
## 844          on his    11           0      2  0.98039980  3.25527901
## 845         of your    24           0      2  0.96544505  4.57954205
## 846        the left    12           0      2  0.96006367  3.12843844
## 847        take the    10           0      2  0.95990320  2.88786359
## 848           for a    47           0      2  0.95200432  6.27841954
## 849         saw the    11           0      2  0.95090513  2.99550266
## 850         and two    10           0      2  0.94771138  2.93930715
## 851        with you    25           0      2  0.94336933  4.63624393
## 852            of a   168           0      2  0.94126070 11.41684163
## 853            on a    24           0      2  0.93654540  4.45777731
## 854            as a    55           0      2  0.92932041  6.61939665
## 855          of her    28           0      2  0.92415021  4.73159309
## 856         the two    18           0      2  0.90837383  3.61292134
## 857         it were    12           0      2  0.89823775  3.10195288
## 858            am a    11           0      2  0.88310005  2.89195329
## 859         you not    19           0      2  0.86908780  3.74554695
## 860           of an    21           0      2  0.86663070  3.86617116
## 861           to us    12           0      2  0.86614386  2.94974711
## 862         from a    30           0      2  0.86509197  4.59932405
## 863           be at    10           0      2  0.83757909  2.66978889
## 864         to this    25           0      2  0.83336298  4.04880225
## 865         over to    10           0      2  0.82482676  2.58123197
## 866        and when    14           0      2  0.80996392  2.96557252
## 867        the side    10           0      2  0.79944906  2.41467646
## 868      with which    12           0      2  0.79301637  2.75462104
## 869         that no    10           0      2  0.79035043  2.50125258
## 870          to the   302           0      2  0.78489570 12.48617241
## 871           it on    10           0      2  0.76726967  2.43569889
## 872         out to    16           0      2  0.76323995  2.99654208
## 873          up the    28           0      2  0.76145213  3.81525986
## 874         and she    18           0      2  0.75928538  3.14364083
## 875         to have    53           0      2  0.74575482  5.23603805
## 876           see a    10           0      2  0.74017011  2.32630430
## 877          to his    63           0      2  0.73456821  5.61055089
```

```
## 878          was at    21           0      2  0.72425068  3.28001972
## 879          and he    61           0      2  0.72000764  5.40472581
## 880       that your    12           0      2  0.71555739  2.47295034
## 881         was all    10           0      2  0.70477345  2.23995266
## 882            or a    10           0      2  0.70183317  2.20809551
## 883          to you    63           0      2  0.68824580  5.26359413
## 884          have a    47           0      2  0.68504311  4.55264799
## 885       found the    10           0      2  0.67384059  2.06243643
## 886      the little    30           0      2  0.66557592  3.44886396
## 887           be in    20           0      2  0.65973025  2.91703790
## 888          to our    10           0      2  0.65657789  2.06331112
## 889        which it    18           0      2  0.65651957  2.76685501
## 890          it had    23           0      2  0.64213441  3.04155372
## 891           up a    12           0      2  0.63892020  2.19594695
## 892          to an    17           0      2  0.63217887  2.56323221
## 893         that my    28           0      2  0.60432190  3.14492447
## 894        that all    11           0      2  0.60385988  2.00622290
## 895        and that    86           0      2  0.60286181  5.36345002
## 896          been a    18           0      2  0.59116120  2.47172021
## 897          of him    22           0      2  0.58861121  2.70761151
## 898         to some    11           0      2  0.58180344  1.91690729
## 899      before the    13           0      2  0.57600202  2.00841258
## 900           of it    57           0      2  0.55909452  4.08779814
## 901         it from    12           0      2  0.55850729  1.93911703
## 902         you had    19           0      2  0.55513281  2.40290080
## 903          is my    17           0      2  0.54817669  2.25111436
## 904         were in    10           0      2  0.54233878  1.72463276
## 905          at it    16           0      2  0.52678160  2.10151529
## 906          had a    36           0      2  0.52149794  3.05652657
## 907         but his    12           0      2  0.51639191  1.79701216
## 908        than the    14           0      2  0.51341427  1.86059750
## 909         from it    10           0      2  0.51166137  1.63279281
## 910        when the    28           0      2  0.49882361  2.53279441
## 911        that the   149           0      2  0.49719336  5.71763061
## 912          is it    22           0      2  0.49692995  2.30966599
## 913        the more    16           0      2  0.46692006  1.80025018
## 914          and a   122           0      2  0.45927731  4.85087369
## 915          not a    25           0      2  0.45708116  2.24725453
## 916         one to    14           0      2  0.45584379  1.69170428
## 917         see the    16           0      2  0.45282009  1.75492085
## 918         and his    53           0      2  0.44250493  3.12615372
## 919         some of    10           0      2  0.43298219  1.36957917
## 920          has a    11           0      2  0.43159424  1.42977273
## 921          him a    10           0      2  0.43037852  1.36265479
## 922         but the    53           0      2  0.42910435  2.98726189
## 923         been in    10           0      2  0.41720331  1.32908049
## 924         of what    10           0      2  0.41713505  1.31903862
## 925        the very    32           0      2  0.41598890  2.25333341
## 926          her to    16           0      2  0.41562858  1.64622323
## 927          was in    36           0      2  0.39427780  2.32383458
## 928         for his    12           0      2  0.39383866  1.37207357
## 929         you for    10           0      2  0.36958568  1.18046013
## 930        and your    15           0      2  0.35626237  1.36611145
## 931      which have    10           0      2  0.33826819  1.08205683
```

```
## 932         that her    10           0      2  0.33708726  1.07462862
## 933          were to    13           0      2  0.32094879  1.15274195
## 934           of all    13           0      2  0.30543815  1.09707398
## 935           it not    12           0      2  0.29951868  1.04309769
## 936           it for    10           0      2  0.26267279  0.83932133
## 937         that his    25           0      2  0.25266956  1.25096408
## 938             at a    26           0      2  0.25166104  1.26594761
## 939          are the    21           0      2  0.22261903  0.99324248
## 940           upon a    15           0      2  0.21735315  0.83856789
## 941           when a    10           0      2  0.21022855  0.66832607
## 942            it as    11           0      2  0.20166158  0.67451745
## 943           is the    73           0      2  0.19404001  1.59422305
## 944            is in    23           0      2  0.19249411  0.91619392
## 945         the back    10           0      2  0.19067374  0.59368228
## 946          and the   198           0      2  0.18486146  2.46611290
## 947          and has    10           0      2  0.18230610  0.57797971
## 948           and as    23           0      2  0.17705349  0.83800519
## 949           are to    10           0      2  0.17471862  0.55522953
## 950         you that    24           0      2  0.15328096  0.74512419
## 951           and on    11           0      2  0.09822757  0.32637036
## 952         that was    27           0      2  0.09266439  0.47696377
## 953           and an    11           0      2  0.08517268  0.28306036
## 954           and by    10           0      2  0.07837562  0.24896108
## 955           not to    18           0      2  0.05534419  0.23353747
## 956          out the    16           0      2  0.05308802  0.20873691
## 957           what a    10           0      2  0.03797768  0.12104991
## 958          is that    18           0      2  0.02756501  0.11680732
## 959          and all    11           0      2  0.01927956  0.06414588
## 960             me a    11           0      2  0.01896273  0.06325804
## 961           and it    38           0      2  0.01302870  0.07895709
## 962         of which    18           0      2  0.00437913  0.01848940
## 963          with it    10           0      2 -0.01076477 -0.03449974
## 964       that which    11           0      2 -0.01960712 -0.06564996
## 965           if the    14           0      2 -0.02067053 -0.07636683
## 966            on to    10           0      2 -0.02206900 -0.07034629
## 967          at that    12           0      2 -0.02476237 -0.08646533
## 968          have it    11           0      2 -0.02524840 -0.08465408
## 969         and were    10           0      2 -0.04743671 -0.15099909
## 970           be the    32           0      2 -0.05235785 -0.28933013
## 971          the one    18           0      2 -0.07524793 -0.31351506
## 972           you to    33           0      2 -0.09189014 -0.52132485
## 973         and this    11           0      2 -0.11156773 -0.37196383
## 974            in it    19           0      2 -0.13689609 -0.59596885
## 975            be of    14           0      2 -0.16688874 -0.62593932
## 976            but a    14           0      2 -0.17434480 -0.65423967
## 977         said the    21           0      2 -0.18395685 -0.83047637
## 978             to a    59           0      2 -0.20559511 -1.54857952
## 979            it in    20           0      2 -0.22104167 -0.98711553
## 980          and my     22           0      2 -0.23061872 -1.07519888
## 981           is to     24           0      2 -0.25063637 -1.22051631
## 982          was the    61           0      2 -0.26363804 -2.00845816
## 983           was it    13           0      2 -0.28483062 -1.03525981
## 984          and you    27           0      2 -0.30236091 -1.55850166
## 985          and for    12           0      2 -0.32198734 -1.12213746
```

```
## 986           of you    23           0      2 -0.34514758 -1.64792655
## 987          and was    31           0      2 -0.35604332 -1.96400869
## 988           as the    34           0      2 -0.36133333 -2.07165136
## 989          and not    13           0      2 -0.38589420 -1.39840718
## 990          in that    19           0      2 -0.40337876 -1.75948691
## 991            it to    26           0      2 -0.44304044 -2.24728265
## 992         and with    13           0      2 -0.45299735 -1.64277444
## 993          was his    10           0      2 -0.46264332 -1.48601837
## 994          and had    16           0      2 -0.49159302 -1.97064224
## 995            of me    11           0      2 -0.49829470 -1.67068648
## 996         have the    30           0      2 -0.56397408 -3.05270066
## 997           and in    30           0      2 -0.58160937 -3.16480752
## 998            to it    19           0      2 -0.58531811 -2.55318182
## 999        and which    11           0      2 -0.60423592 -2.02612310
## 1000          you in    12           0      2 -0.61951750 -2.17093375
## 1001        were the    10           0      2 -0.64968326 -2.07044580
## 1002        was that    11           0      2 -0.71326148 -2.39929966
## 1003          it and    10           0      2 -0.72398927 -2.32804743
## 1004          had to    11           0      2 -0.74139184 -2.49075573
## 1005         have to    12           0      2 -0.77034808 -2.69820834
## 1006           is of    14           0      2 -0.77085566 -2.90702223
## 1007          that a    21           0      2 -0.77253633 -3.54290885
## 1008          me the    10           0      2 -0.84957960 -2.71500214
## 1009          and to    37           0      2 -0.85783175 -5.18087543
## 1010           you a    14           0      2 -0.88653979 -3.34617961
## 1011         of that    18           0      2 -0.89179697 -3.79820339
## 1012         that of    19           0      2 -0.92029576 -4.02429494
## 1013         had the    18           0      2 -0.96514721 -4.09618548
## 1014        and have    10           0      2 -1.05489495 -3.39177790
## 1015         not the    12           0      2 -1.05878165 -3.69975540
## 1016       which the    15           0      2 -1.06919037 -4.15949007
## 1017         that in    10           0      2 -1.08303989 -3.48687042
## 1018          was to    13           0      2 -1.11589067 -4.06904047
## 1019          and of    28           0      2 -1.12110441 -5.92182527
## 1020          was of    12           0      2 -1.17620935 -4.12923239
## 1021         to that    11           0      2 -1.38845134 -4.67989440
```

```
write_delim(kollokationen, path = "kollokationen.csv", delim = ";") # Datei ist Excel-kompatibel
```

In den beiden Tabellen zeigt uns *collocation* das Kollokat und *count* dessen absolute Häufigkeit an. Die Assoziationsstärker der Kollokation wird mit *lambda* und *z* gemessen (genauer ist z ein z-standardisiertes lambda). Lambda beschreibt dabei die Wahrscheinlichkeit, dass genau diese zwei Begriffe auf einander folgen, was insofern von der absoluten Häufigkeit zu differenzieren ist, als das diese nicht das auftreten eines Teilbegriffs mit allen anderen Wörtern im Korpus berücksichtigt.

Die beiden Tabellen illustrieren diesen Unterschied. Während die erste nach der absoluten Häufigkeit sortiert ist, so dass gängige Kollokate wie 'of the' oder 'it is' ganz vorne liegen, ist die zweite absteigend nach Lambda sortiert, so dass eine Reihe von Eigennamen wie 'hosmer angel' oder 'briony lodge' die Liste anführen. Praktisch betrachtet ist es meist sinnvoller, Eigennamen als Phrasen zu betrachten, statt als einzelne Begriffe, deren gemeinsames Auftreten wirklich etwas über den Text verrät. Echte Kollokate sind hingegen 'no doubt' oder 'young lady'.

Wort- und Textähnlichkeit und -Distanz

Wie sich im ersten Kapitel bereits angedeutet hat, lassen sich auf Grundlage einer DFM zahlreiche Metriken

berechnen, welche die Nähe und Distanz von Wörtern und Dokumenten zu einander relektieren. Dies geschieht mit textstat_simil(). Zunächst konstruieren wir dazu eine DFM, in der jeder Satz einem Dokument entspricht. Dies wird deshalb notwendig, weil sich Wortähnlichkeiten bei einer geringen Dokumentanzahl nicht besonders zuverlässig berechnen lassen, da Ähnlichkeit als Kookurenz innerhalb des gleichen Dokuments operationalisiert wird. Dann Berechnen wir die Wortähnlichkeit zum Begriff 'love' mittels Kosinusdistanz. Andere verfügbare Metriken sind 'correlation', 'jaccard', 'eJaccard', 'dice', 'eDice', 'simple matching', 'hamann' und 'faith', welche Wortähnlichkeit jeweils unterschiedlich operationalisieren.

```
korpus.saetze <- corpus_reshape(korpus, to = "sentences")
meine.dfm.saetze <- dfm(korpus.saetze, remove_numbers = TRUE, remove_punct = TRUE, remove_symbols = TRU
meine.dfm.saetze <- dfm_trim(meine.dfm.saetze, min_docfreq = 5)
aehnlichkeit.woerter <- textstat_simil(meine.dfm.saetze, "love", margin = "features", method = "cosine"
head(aehnlichkeit.woerter[order(aehnlichkeit.woerter[,1], decreasing = T),], 10)
```

```
##      love     loved     lover   working    ceased      wife   majesty
## 1.0000000 0.2108185 0.1781742 0.1781742 0.1666667 0.1295048 0.1178511
##   account   emotion     share
## 0.1081476 0.1054093 0.1054093
```

Analog zur Ähnlichkeit funktionert auch die Berechnung von Wortdistanzen mit der Funktion textstat_dist(). Auch hier haben wir wieder eine große Anzahl von Distanzmaßen zur Auswahl ('euclidean', 'chisquared', 'chisquared2', 'hamming', 'kullback'. 'manhattan', 'maximum', 'canberra', 'minkowski').

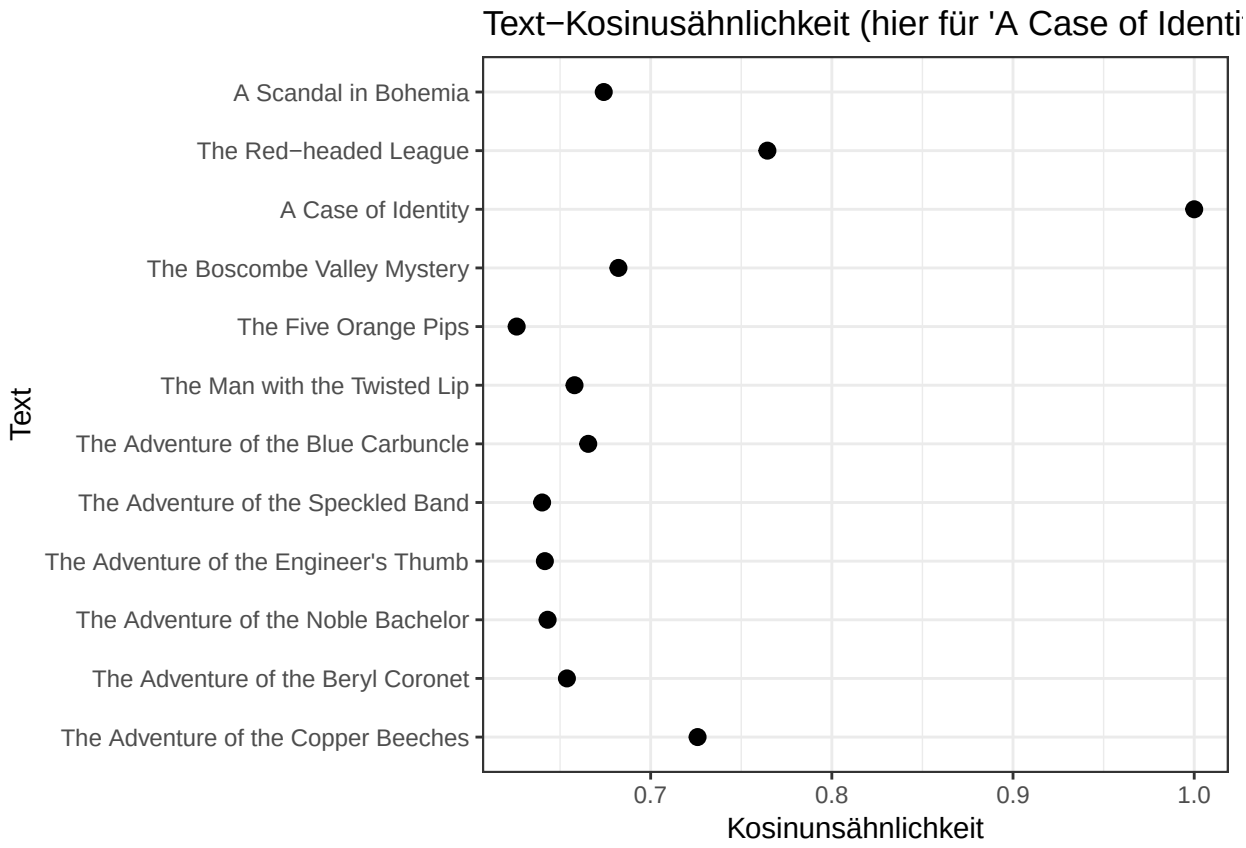
```
distanz.woerter <- textstat_dist(meine.dfm.saetze, "love", margin = "features", method = "euclidean")
head(distanz.woerter[order(distanz.woerter[,1], decreasing = T),], 10)
```

```
##     upon     said   holmes      one      man       mr   little      now
## 23.60085 22.60531 21.47091 20.97618 18.22087 17.63519 17.34935 16.12452
##      see      may
## 15.71623 15.19868
```

Was sagt das Ergebnis aus? Vor allem das (wenig überraschend) Wörter wie 'upon' und 'said' sehr weit von 'love' entfernt sind – allerdings nicht in dem Sinne, dass sie das logische Gegenteil von 'love' darstellen würden (in der Linguistik spricht man von Antonymie). Das liegt daran, dass diese Begriffe im Korpus nahezu gleich verteilt sind, also überall vorkommen. Mit dem Verfahren der Wortvektoren (welches wir hier nicht behandeln) und sehr großen Datenbeständen lassen sich allerdings auch solche und andere semantisch Beziehungen indentifizieren. Die Filterung, die wir zuvor an der DFM vorgenommen haben, schließt Begriffe aus, die vielleicht nie gemeinsam mit 'love' vorkommen, und insofern noch distanzierter wären, allerdings gibt es derer auch sehr viele. Zusammenfassend kann man sagen, dass textstatistische Nähe- und Distanzmaße immer ausreichend viele Daten benötigen, um ein zuverlässiges Resultat liefern zu können, und dass der Suchterm selbst ausreichend oft vorkommen muss.

Wer die Dokumentation von textstat_simil() und textstat_dist() anschaut, wird feststellen, dass es dort den etwas kryptischen Hinweis auf das Argument 'margin' gibt. Dieses hat zwei mögliche erinstellungen: 'documents' oder 'features'. Stellt man hier 'documents' ein, werden die besprochenen Metriken nicht auf Wörter, sondern auf Texte angewandt. Folgend plotten wir die Textnähe via Kosinusähnlichkeit (hier ausgehend vom ersten Roman, 'A Case of Identity').

```
aehnlichkeit.texte <- data.frame(Text = factor(korpus.stats$Text, levels = rev(korpus.stats$Text)), as.n
ggplot(aehnlichkeit.texte, aes(A.Case.of.Identity, Text)) + geom_point(size = 2.5) + ggtitle("Text-Kosi
```

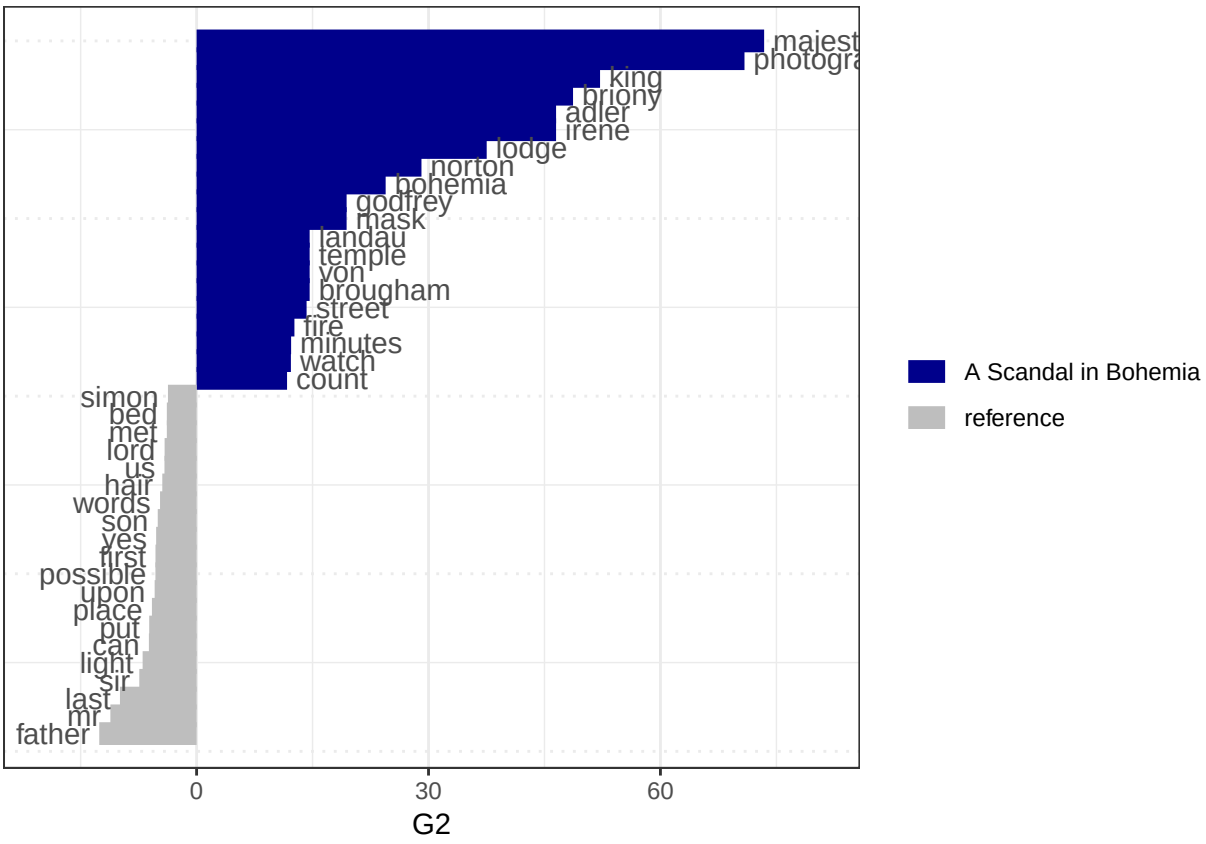


Wie wir sehen, ist die Ähnlichkeit der Romane 'The Red-headed League' und 'The Adventure of the Copper Beeches' etwas größer, als dies bei den anderen Erzählungen der Fall ist.

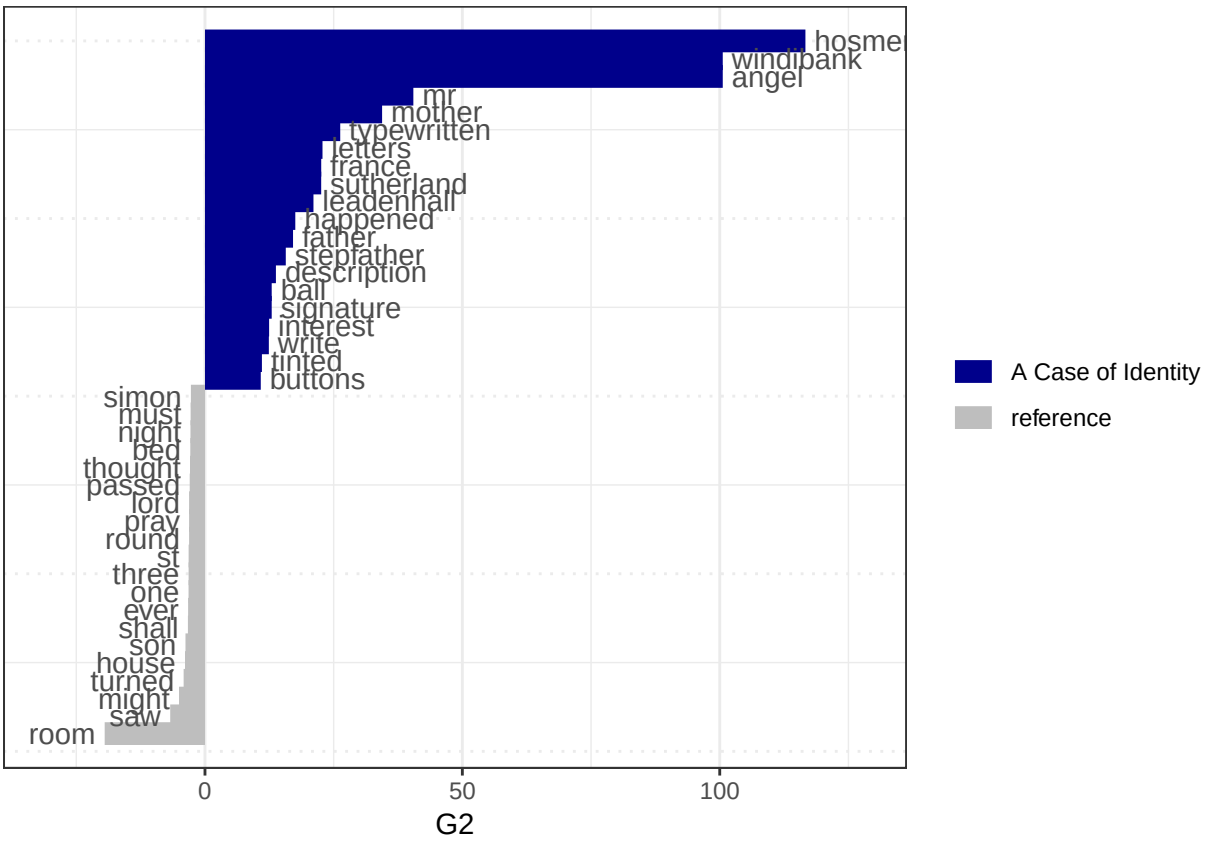
Keyness

Bei der sog. Keyness handelt es sich um ein Maß für die Distinktivität eines Begriffs für einen bestimmten Text, also wie stark der Begriff den jeweiligen Text im Vergleich zum gesamten Korpus kennzeichnet. Während wir zuvor die Distanz von Wörtern und Texten zu einander untersucht haben, macht sich die Keyness die Verteilungshäufigkeit von Wörtern auf Texte zunutze, ohne deren Position zu berücksichtigen. Keyness funktioniert daher auch mit längeren Texten gut, solange diese sich ausreichend markant unterscheiden. Folgend berechnen wir zunächst die Keyness für vier Texte mit textstat_keyness() und plotten wir diese Keyness-Statistiken dann für vier Erzählungen mit Hilfe von textplot_keyness().

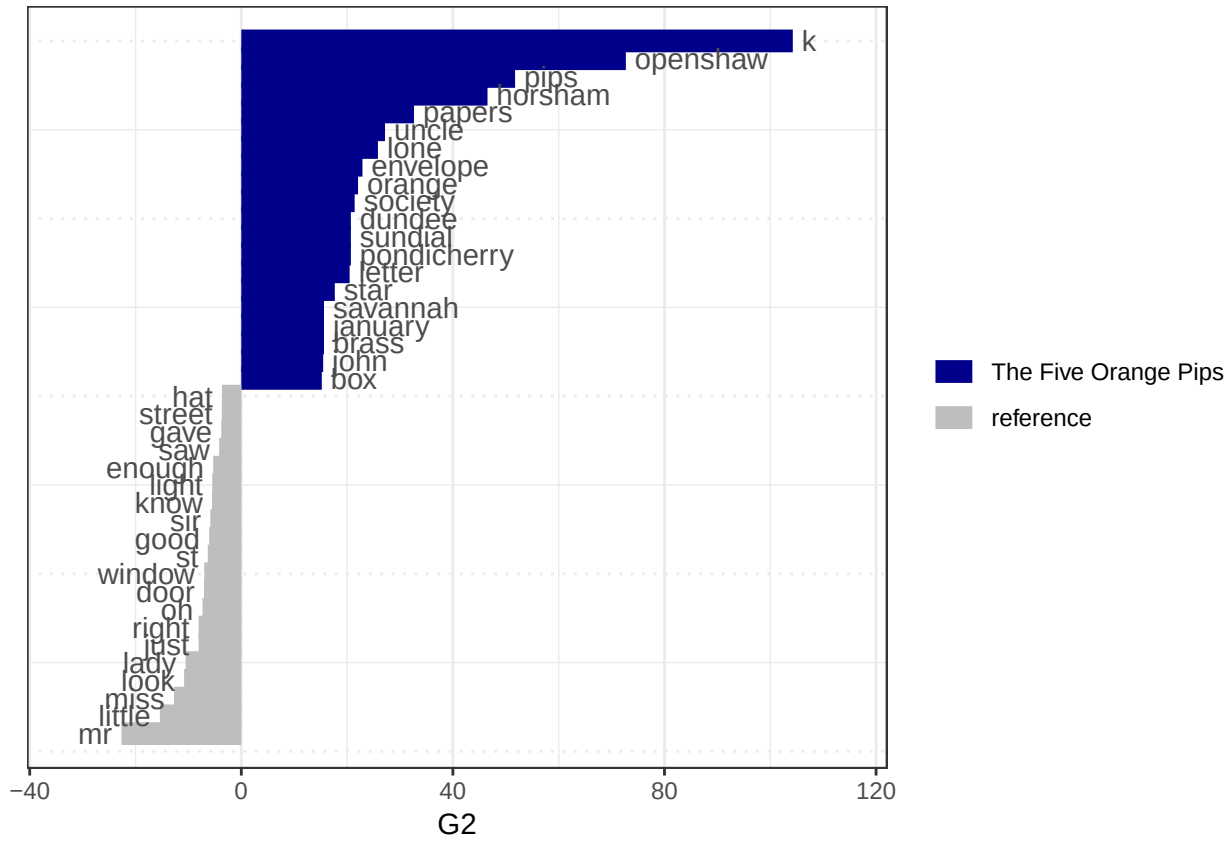
```
keyness <- textstat_keyness(meine.dfm, target = "A Scandal in Bohemia", measure = "lr")
textplot_keyness(keyness)
```



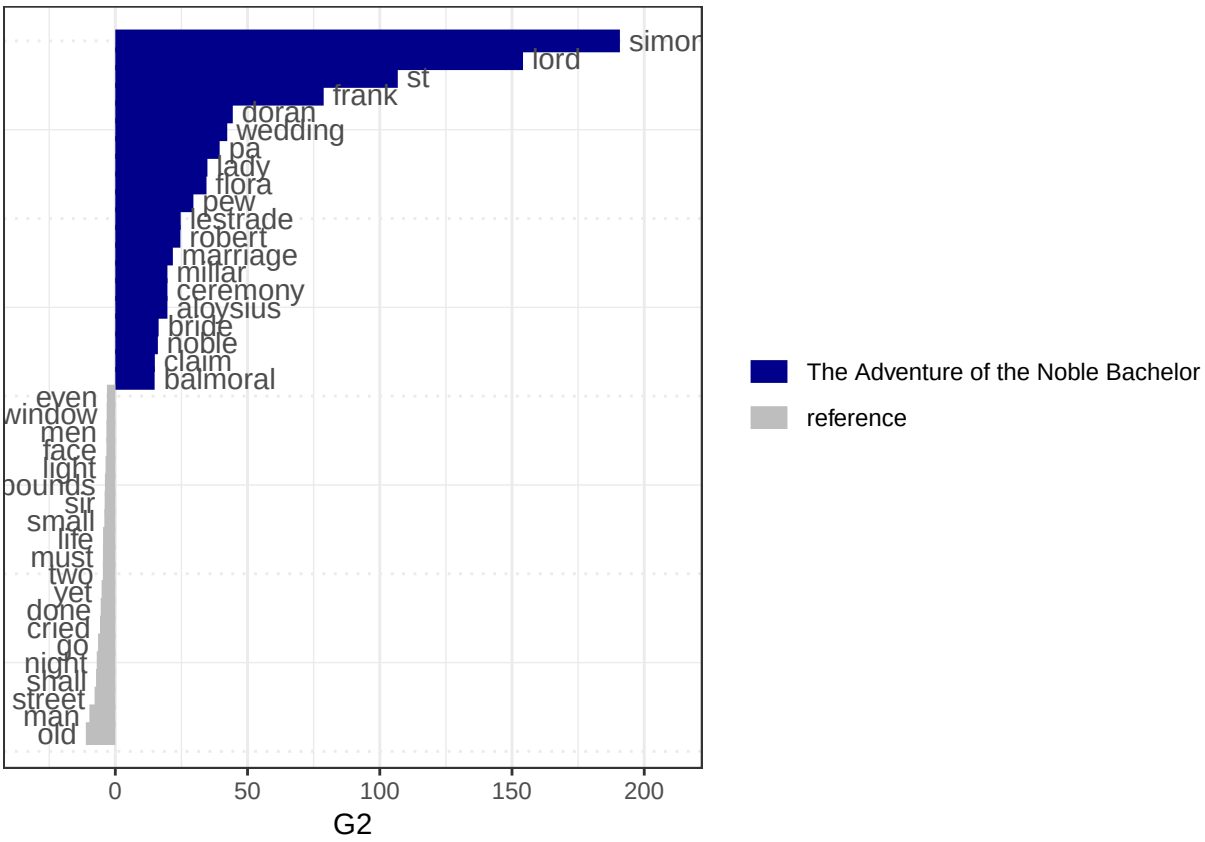

```
keyness <- textstat_keyness(meine.dfm, target = "A Case of Identity", measure = "lr")
textplot_keyness(keyness)
```




```
keyness <- textstat_keyness(meine.dfm, target = "The Five Orange Pips", measure = "lr")
textplot_keyness(keyness)
```




```
keyness <- textstat_keyness(meine.dfm, target = "The Adventure of the Noble Bachelor", measure = "lr")
textplot_keyness(keyness)
```



Schaut man sich die vier Beispieltexte einmal genauer an, so wird schnell klar, dass die Begriffe mit einem hohen Keyness-Wert tatsächlich sehr distinktiv für den jeweiligen Text sind, also Begriffe wie 'majesty' und 'photograph' tatsächlich nur in 'A Scandal in Bohemia' eine Rolle spielen. Wenig distinktive Begriffe sind hingegen solche, die zwar in anderen Texten, nicht aber dem Zieltext vorkommen. Nützlich wird diese Funktion vor allem dann, wenn man Texte nach einen Kriterium wie Medium, Sprecher, Partei, Zeitpunkt oder manuell zugeordnete Inhaltskategorie gruppiert.

Lexikalische Vielfalt

Unter Maßen lexikalischer Vielfalt versteht man Metriken, welche die Diversität eines Textes mit Blick auf den Wortgebrauch wiedergeben. Ein Beispiel ist die bereits in Kapitel 1 berechtete Typ-Token-Relation. Diese beschreibt die Wortvielfalt und geben so auch Aufschluss über die Komplexität eines Textes. Wir berechnen folgend zahlreiche Metriken für die lexikalische Diversität der zwölf Romane mit mit der Funktion textstat_lexdiv().

```
lexdiversitaet <- textstat_lexdiv(meine.dfm)
lexdiversitaet
```

```
##                                document       TTR         C        R
## 1                   A Scandal in Bohemia 0.4898924 0.9134555 30.23473
## 2                 The Red-headed League 0.4463153 0.9030169 28.57115
## 3                   A Case of Identity 0.4770497 0.9081492 26.81257
## 4            The Boscombe Valley Mystery 0.4357092 0.9004851 28.31439
## 5                  The Five Orange Pips 0.5098754 0.9168301 29.25014
## 6           The Man with the Twisted Lip 0.4543062 0.9053757 29.37222
## 7    The Adventure of the Blue Carbuncle 0.4690664 0.9074150 27.97146
## 8     The Adventure of the Speckled Band 0.4402488 0.9024741 29.53935
## 9  The Adventure of the Engineer's Thumb 0.4688943 0.9077599 28.44840
```

```
## 10   The Adventure of the Noble Bachelor 0.4589674 0.9051507 27.84234
## 11    The Adventure of the Beryl Coronet 0.4247702 0.8971770 27.31111
## 12   The Adventure of the Copper Beeches 0.4158192 0.8954733 27.66057
##        CTTR        U         S      Maas     lgV0    lgeV0
## 1  21.37919 41.37539 0.9290363 0.1554637 8.037852 18.50784
## 2  20.20285 37.24952 0.9205759 0.1638474 7.593520 17.48473
## 3  18.95935 38.10038 0.9230848 0.1620076 7.591398 17.47984
## 4  20.02130 36.43295 0.9186184 0.1656734 7.507281 17.28615
## 5  20.68298 42.29086 0.9309589 0.1537718 8.076565 18.59698
## 6  20.76930 38.26898 0.9227499 0.1616503 7.721236 17.77880
## 7  19.77881 38.35354 0.9233318 0.1614720 7.667594 17.65529
## 8  20.88747 37.46086 0.9208007 0.1633846 7.654435 17.62499
## 9  20.11606 38.65961 0.9238849 0.1608316 7.716608 17.76815
## 10 19.68751 37.59488 0.9216189 0.1630931 7.592800 17.48307
## 11 19.31187 35.17084 0.9155935 0.1686198 7.346019 16.91483
## 12 19.55898 34.88019 0.9146550 0.1693209 7.334768 16.88893
```

```
write_delim(lexdiversitaet, path = "lexdiversitaet.csv", delim = ";") # Datei ist Excel-kompatibel
```

Bei einem oberflächliche Vergleich der Metriken fällt auf, dass sich die Texte nicht sehr stark unterscheiden, was ihre jeweilige lexikalische Vielfalt betrifft, ganz unabhängig davon, welche Metrik verwendet wird. Dies ist nicht unbedingt verwunderlich, da es sich um Texte des selben Genres und Autors handelt. Interessanter werden solche Metriken dann, wenn wir sehr unterschiedliche Genres oder Autoren vergleichen wollen, etwa die Programmen von Parteien, Texte aus unterschiedlichen Medien, oder Tweets von unterschiedlichen Nutzern.

Lesbarkeitsindizes

Eine weitere Klasse von Text-Metriken, die sich für ein Dokument aufgrund seiner Wortzusammensetzung berechnen lassen, sind die sog. Lesbarkeitsindizes. Darunter versteht man Metriken, die anhand von textlichen Eigenschaften einen Zahlenwert berechnen, der die Leseschwierigkeit eines Dokumentes möglichst akkurat wiedergeben soll. Anwendung finden solche Indizes etwa im Bildungsbereich, wenn es um die Frage geht, welches Schwierigkeitsniveau eines Textes für Schüler angemessen ist, aber auch in der öffentlichen Verwaltung, wenn möglichst klare und zugängliche Sprache bspw. auf einer berhördlichen Website verwendet werden soll.

Die Kalkulation zahlreicher Lesbarkeitsindizes erfolgt in quanteda mit textstat_readability(), Auch hier wird das Ergebnis in einer Excel-kompatible CSV-Datei gespeichert.

```
lesbarkeit <- textstat_readability(korpus)
lesbarkeit
```

```
##                                document      ARI ARI.simple   Bormuth
## 1                  A Scandal in Bohemia 4.435711   50.00114 -2.137728
## 2                 The Red-headed League 5.959119   53.05488 -2.778184
## 3                    A Case of Identity 7.046479   55.20866 -3.121271
## 4            The Boscombe Valley Mystery 5.407228   51.96288 -2.608304
## 5                  The Five Orange Pips 5.685873   52.50959 -2.674306
## 6           The Man with the Twisted Lip 5.613632   52.38621 -2.735822
## 7    The Adventure of the Blue Carbuncle 4.918963   50.98740 -2.417312
## 8     The Adventure of the Speckled Band 6.077748   53.28350 -2.788627
## 9  The Adventure of the Engineer's Thumb 6.160238   53.45631 -2.854999
## 10   The Adventure of the Noble Bachelor 5.627186   52.37909 -2.594556
## 11    The Adventure of the Beryl Coronet 5.226622   51.63307 -2.670783
## 12   The Adventure of the Copper Beeches 5.805978   52.76310 -2.779212
##    Bormuth.GP  Coleman Coleman.C2 Coleman.Liau Coleman.Liau.grade
## 1    25913066 58.14334   60.36090     61.79052           6.133101
## 2    56149703 58.25812   58.19026     60.48476           6.490883
```

```
## 3    79053413 56.18532   55.44867     58.77472           6.959441
## 4    46645112 59.67333   59.96689     61.45457           6.225151
## 5    49971732 59.24405   59.38225     60.76945           6.412876
## 6    53653342 60.39290   60.21963     61.70491           6.156558
## 7    37096594 59.94674   60.85637     61.97815           6.081689
## 8    56855570 58.81974   58.67047     60.02579           6.616643
## 9    60777267 58.81911   58.48114     60.28783           6.544844
## 10   45774251 55.96133   56.68742     60.28614           6.545307
## 11   50005800 61.20065   61.12768     62.90418           5.827953
## 12   56288307 59.21849   59.04303     61.21796           6.289984
##    Coleman.Liau.short Dale.Chall Dale.Chall.old Dale.Chall.PSK
## 1            6.133532   41.36990       6.556714       5.704632
## 2            6.491506   40.08326       6.562172       5.783289
## 3            6.960147   39.21126       6.596127       5.847888
## 4            6.225727   40.85689       6.488225       5.709622
## 5            6.413474   40.01904       6.606646       5.803710
## 6            6.157164   40.90847       6.437412       5.687587
## 7            6.082209   41.34648       6.468865       5.673242
## 8            6.617272   40.17391       6.544259       5.771205
## 9            6.545487   41.47371       6.306058       5.604907
## 10           6.545887   39.58713       6.705241       5.866226
## 11           5.828536   42.92340       6.122409       5.450041
## 12           6.290603   41.55514       6.316290       5.603876
##    Danielson.Bryan Danielson.Bryan.2 Dickes.Steiwer      DRP      ELF
## 1         4.697158          89.39030      -236.7840 313.7728 3.246637
## 2         4.930815          89.81200      -275.1869 377.8184 4.036649
## 3         5.135606          89.44686      -301.0769 412.1271 4.750000
## 4         4.826566          90.02789      -262.2865 360.8304 3.658805
## 5         4.885727          89.79853      -269.7865 367.4306 3.787368
## 6         4.844153          90.33114      -269.2255 373.5822 3.725256
## 7         4.748028          89.97694      -250.9504 341.7312 3.400362
## 8         4.962003          89.60930      -277.3279 378.8627 3.977199
## 9         4.964044          89.82287      -280.5912 385.4999 4.061024
## 10        4.894529          89.45149      -265.8474 359.4556 4.074074
## 11        4.753512          90.80749      -260.6270 367.0783 3.555911
## 12        4.885682          90.16102      -272.2832 377.9212 3.921222
##    Farr.Jenkins.Paterson   Flesch Flesch.PSK Flesch.Kincaid      FOG
## 1              -43.43732 80.24222   4.905547       5.277773 8.019379
## 2              -46.68581 77.09980   5.148873       6.511516 9.155889
## 3              -48.44223 74.27932   5.340143       7.328524 9.896804
## 4              -45.81618 79.11708   5.020893       6.021593 8.908129
## 5              -46.14646 77.89787   5.093896       6.271183 9.064596
## 6              -46.46599 80.10462   4.982844       6.045078 8.742877
## 7              -44.84567 80.93373   4.901073       5.531496 8.160027
## 8              -46.72323 78.32467   5.084011       6.351534 9.006028
## 9              -47.06881 77.70789   5.125085       6.522143 9.231813
## 10             -45.76910 76.26178   5.172330       6.397064 9.012574
## 11             -46.14563 81.18312   4.917743       5.818689 8.426737
## 12             -46.69323 78.12272   5.094299       6.373569 9.027087
##     FOG.PSK  FOG.NRI  FORCAST FORCAST.RGL    Fucks Linsear.Write      LIW
## 1  4.178384 636.8048 8.768217    8.075038 53.24215      3.683109 27.95962
## 2  5.004411 844.2906 8.754870    8.060357 66.16754      4.363002 29.48063
## 3  5.477684 715.0310 8.995893    8.325482 74.05303      4.929293 32.29052
## 4  4.800191 845.3535 8.590311    7.879342 62.29403      4.270440 29.04876
```

```
## 5  4.893321 658.5691 8.640227    7.934249 63.99158      4.404211 30.02693
## 6  4.870298 830.6337 8.506640    7.787303 64.53925      3.774744 29.77109
## 7  4.453545 634.3267 8.558519    7.844371 58.37500      3.478261 27.50320
## 8  4.975008 903.0152 8.689565    7.988521 66.67915      4.104235 30.26084
## 9  5.088655 781.8095 8.689639    7.988603 67.77559      4.338583 31.37809
## 10 4.810232 713.3226 9.021938    8.354132 62.77778      4.459259 30.23405
## 11 4.740346 846.2383 8.412715    7.683986 62.52396      3.399361 28.39513
## 12 4.975975 913.2584 8.643199    7.937518 65.68810      4.147910 29.94673
##         nWS    nWS.2    nWS.3    nWS.4      RIX Scrabble     SMOG   SMOG.C
## 1  3.231183 4.016617 3.458621 3.694565 1.943199 1.727150 8.610888 8.583186
## 2  3.546336 4.326640 3.961731 4.446034 2.153578 1.736873 9.095435 9.012767
## 3  4.099415 4.840897 4.330158 4.939297 2.578283 1.720559 9.470823 9.348833
## 4  3.374569 4.189259 3.867021 4.283458 2.103774 1.735940 9.031808 8.956069
## 5  3.448772 4.253180 3.949051 4.387977 2.250526 1.735708 9.123545 9.037841
## 6  3.249285 4.074629 3.675947 4.164384 2.204778 1.733961 8.678662 8.642952
## 7  2.958476 3.768300 3.413673 3.778645 1.887681 1.727723 8.456263 8.447262
## 8  3.536682 4.332838 3.846095 4.342845 2.278502 1.720349 8.915801 8.852911
## 9  3.630806 4.428539 3.977324 4.494737 2.454724 1.722515 9.078715 8.997860
## 10 3.685929 4.418602 3.954092 4.355914 2.285185 1.710321 9.160891 9.071178
## 11 2.946110 3.780059 3.474115 3.950203 1.995208 1.726705 8.395497 8.394015
## 12 3.427765 4.228136 3.864306 4.357508 2.228296 1.716580 8.946509 8.880191
##    SMOG.simple  SMOG.de   Spache Spache.old   Strain Traenkle.Bailer
## 1     8.255789 3.255789 4.025714   4.552137 5.200448       -285.9539
## 2     8.720360 3.720360 4.389533   4.978843 6.482723       -323.7721
## 3     9.080271 4.080271 4.632814   5.258030 7.237121       -347.6011
## 4     8.659355 3.659355 4.225042   4.794493 6.082075       -310.2642
## 5     8.747311 3.747311 4.373642   4.954855 6.258947       -316.9964
## 6     8.320769 3.320769 4.377665   4.963711 6.247270       -317.8938
## 7     8.107539 3.107539 4.214289   4.769783 5.659783       -299.2696
## 8     8.548131 3.548131 4.363199   4.951840 6.427524       -325.5035
## 9     8.704329 3.704329 4.341274   4.933646 6.578740       -328.9065
## 10    8.783117 3.783117 4.372385   4.947730 6.204444       -313.5708
## 11    8.049278 3.049278 4.102323   4.670627 6.084824       -307.7318
## 12    8.577573 3.577573 4.287294   4.871885 6.430707       -320.7484
##    Traenkle.Bailer.2 Wheeler.Smith meanSentenceLength meanWordSyllables
## 1          -204.2129      32.46637           12.92377          1.341314
## 2          -203.0648      40.36649           16.12565          1.340043
## 3          -206.3125      47.50000           17.83081          1.352925
## 4          -198.3939      36.58805           15.28616          1.326270
## 5          -202.0149      37.87368           15.60632          1.336841
## 6          -198.4990      37.25256           15.93515          1.306811
## 7          -199.2214      34.00362           14.33333          1.316229
## 8          -203.6765      39.77199           16.16938          1.325040
## 9          -203.3150      40.61024           16.50984          1.328246
## 10         -203.5291      40.74074           15.19444          1.361121
## 11         -189.7193      35.55911           15.62939          1.297731
## 12         -198.5672      39.21222           16.14469          1.327724
```

```
write_delim(lesbarkeit, path = "lesbarkeit.csv", delim = ";") # Datei ist Excel-kompatibel
```

Ein schönes Beispiel für den Nutzen solcher Metriken findet sich in der Dokumentation der Funktion textstat_readability(). Hatte die Antrittsrede von George Washington im Jahr 1789 noch einen Flesh-Kincaid-Index von 28, so betrug der Wert bei der Antrittsrede von Donald Trump in 2017 nur noch 9 (was allerdings dem Trend seit Mitte des 20. Jhd. entspricht).