Tidy Text

Bilal Gilani

4/1/2020

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.3
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## Warning: package 'tidyr' was built under R version 3.6.3
```

```
## Warning: package 'purrr' was built under R version 3.6.3
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidytext)
```

```
## Warning: package 'tidytext' was built under R version 3.6.3
```

```
library(stringr)
library(janeaustenr)
```

```
## Warning: package 'janeaustenr' was built under R version 3.6.3
```

```
library(gutenbergr)
```

```
## Warning: package 'gutenbergr' was built under R version 3.6.3
```

```
library(ggplot2)
library(scales)
```

```
## Warning: package 'scales' was built under R version 3.6.3
```

```
##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard

## The following object is masked from 'package:readr':
##
##     col_factor
```

Exercise 1

```
orig_books <- austen_books() %>%
group_by(book) %>%
mutate(linenumber = row_number(),
chapter = cumsum(str_detect(text,
regex("^chapter [\\divxlc]",
ignore_case = TRUE)))) %>%
ungroup() %>%
select(chapter, linenumber, everything())
orig_books
```

```
## # A tibble: 73,422 x 4
##    chapter linenumber text                  book
##      <int>      <int> <chr>                 <fct>
##  1       0          1 SENSE AND SENSIBILITY Sense & Sensibility
##  2       0          2 ""                    Sense & Sensibility
##  3       0          3 by Jane Austen        Sense & Sensibility
##  4       0          4 ""                    Sense & Sensibility
##  5       0          5 (1811)                Sense & Sensibility
##  6       0          6 ""                    Sense & Sensibility
##  7       0          7 ""                    Sense & Sensibility
##  8       0          8 ""                    Sense & Sensibility
##  9       0          9 ""                    Sense & Sensibility
## 10       1         10 CHAPTER 1             Sense & Sensibility
## # ... with 73,412 more rows
```

```
tidy_books <- orig_books %>%
unnest_tokens(word, text) %>%

mutate(word = str_extract(word, "[a-z']+")) %>%
anti_join(stop_words)
```

```
## Joining, by = "word"
```

```
tidy_books %>%
  count(word, sort = TRUE)
```

```
## # A tibble: 13,464 x 2
##    word       n
##    <chr>  <int>
##  1 miss    1860
##  2 time    1339
##  3 fanny    862
##  4 dear     822
##  5 lady     819
##  6 sir      807
##  7 day      797
##  8 emma     787
##  9 sister   727
## 10 house    699
## # ... with 13,454 more rows
```

```
vec1 <- gutenberg_works(author == "Wells, H. G. (Herbert George)")
hgwells_vec <- vec1$gutenberg_id
hgwells <- gutenberg_download(hgwells_vec)
```

```
## Determining mirror for Project Gutenberg from http://www.gutenberg.org/robot/harvest
```

```
## Using mirror http://aleph.gutenberg.org
```

```
gutenberg_authors %>%
  filter(gutenberg_author_id %in% c(404, 408, 405))
```

```
## # A tibble: 3 x 7
##   gutenberg_author~ author  alias birthdate deathdate wikipedia   aliases
##               <int> <chr>   <chr>     <int>     <int> <chr>       <chr>
## 1               404 Brontë~ <NA>       1820      1849 http://en.~ Bronte, ~
## 2               405 Brontë~ <NA>       1818      1848 http://en.~ Bronte, ~
## 3               408 Brontë~ <NA>       1816      1855 http://en.~ Bell, Cu~
```

```
bronte2 <- gutenberg_works(gutenberg_author_id %in% c(404, 408, 405))
bronte_vec <- bronte2$gutenberg_id
bronte <- gutenberg_download(bronte_vec)
```

```
tidy_hgwells <- hgwells %>%
unnest_tokens(word, text) %>%
mutate(word = str_extract(word, "[a-z']+")) %>%
anti_join(stop_words) %>%
  drop_na()
```

```
## Joining, by = "word"
```

```
tidy_bronte <- bronte %>%
unnest_tokens(word, text) %>%
mutate(word = str_extract(word, "[a-z']+")) %>%
anti_join(stop_words) %>%
  drop_na()
```

```
## Joining, by = "word"
```

```
tidy_hgwells %>%
count(word, sort = TRUE)
```

```
## # A tibble: 54,987 x 2
##    word       n
##    <chr>  <int>
##  1 world   7728
##  2 time    7327
##  3 people  6758
##  4 life    6186
##  5 mind    4212
##  6 day     4107
##  7 sort    3385
##  8 hand    3197
##  9 found   3172
## 10 war     3156
## # ... with 54,977 more rows
```

```
tidy_bronte %>%
count(word, sort = TRUE)
```

```
## # A tibble: 26,641 x 2
##    word       n
##    <chr>  <int>
##  1 time    1464
##  2 miss    1344
##  3 day     1228
##  4 hand    1071
##  5 eyes     989
##  6 night    937
##  7 heart    909
##  8 looked   874
##  9 house    851
## 10 love     805
## # ... with 26,631 more rows
```

```
tidy_hgwells
```

```
## # A tibble: 1,562,274 x 2
##    gutenberg_id word
##           <int> <chr>
##  1           35 time
##  2           35 machine
##  3           35 time
##  4           35 traveller
##  5           35 convenient
##  6           35 speak
##  7           35 expounding
##  8           35 recondite
##  9           35 matter
## 10           35 grey
## # ... with 1,562,264 more rows
```

```
tidy_bronte
```

```
## # A tibble: 360,948 x 2
##    gutenberg_id word
##           <int> <chr>
##  1          767 facsimile
##  2          767 title
##  3          767 page
##  4          767 edition
##  5          767 issued
##  6          767 wuthering
##  7          767 heights
##  8          767 volumes
##  9          767 wuthering
## 10          767 heights
## # ... with 360,938 more rows
```

```
frequency_by_word_across_authors <- bind_rows(mutate(tidy_bronte,
author = "Bronte"),
mutate(tidy_hgwells, author = "Wells"),
mutate(tidy_books, author = "Austen")) %>%
mutate(word = str_extract(word, "[a-z']+")) %>%
count(author, word) %>%
group_by(author) %>%
mutate(proportion = n / sum(n)) %>%
select(-n) %>%
spread(author, proportion)
frequency_by_word_across_authors
```

```
## # A tibble: 61,913 x 4
##    word           Austen     Bronte        Wells
##    <chr>           <dbl>      <dbl>        <dbl>
##  1 '          NA          NA          0.00000128
##  2 a'ch       NA          NA          0.000000640
##  3 a'chitect  NA          NA          0.00000128
##  4 a'eplane   NA          NA          0.000000640
##  5 a'hm       NA          NA          0.000000640
##  6 a'll       NA          NA          0.00000192
##  7 a'most     NA           0.0000111  0.00000256
##  8 a'n't       0.00000462 NA         NA
##  9 a'penny    NA          NA          0.000000640
## 10 aa         NA          NA          0.000000640
## # ... with 61,903 more rows
```

```
frequency <- frequency_by_word_across_authors %>%
gather(author, proportion, `Bronte`:`Wells`)
frequency
```

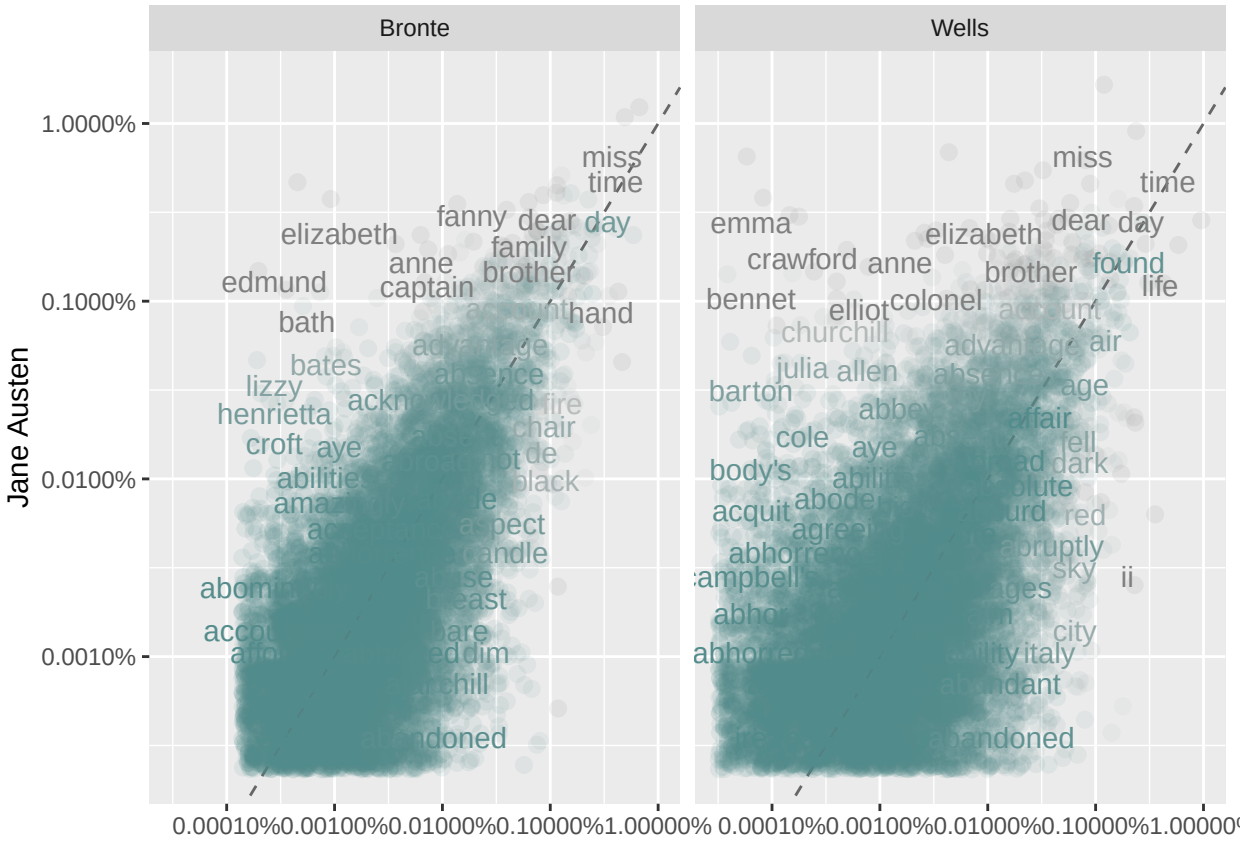
```
## # A tibble: 123,826 x 4
##    word           Austen author proportion
##    <chr>           <dbl> <chr>       <dbl>
##  1 '          NA          Bronte NA
```

```
##  2 a'ch      NA          Bronte NA
##  3 a'chitect NA          Bronte NA
##  4 a'eplane  NA          Bronte NA
##  5 a'hm      NA          Bronte NA
##  6 a'll      NA          Bronte NA
##  7 a'most    NA          Bronte  0.0000111
##  8 a'n't      0.00000462 Bronte NA
##  9 a'penny   NA          Bronte NA
## 10 aa        NA          Bronte NA
## # ... with 123,816 more rows
```

```
frequency %>% ggplot(aes(x = proportion,
y = `Austen`,
color = abs(`Austen` - proportion))) +
geom_abline(color = "gray40", lty = 2) +
geom_jitter(alpha = 0.1, size = 2.5,
width = 0.3, height = 0.3) +
geom_text(aes(label = word),
check_overlap = TRUE, vjust = 1.5) +
scale_x_log10(labels = percent_format()) +
scale_y_log10(labels = percent_format()) +
scale_color_gradient(limits = c(0, 0.001),
low = "darkslategray4",
high = "gray75") +
facet_wrap(~author, ncol = 2) +
theme(legend.position="none") +
labs(y = "Jane Austen", x = NULL)
```

```
## Warning: Removed 101059 rows containing missing values (geom_point).
```

```
## Warning: Removed 101059 rows containing missing values (geom_text).
```




```
df_Bronte <- frequency[frequency$author == "Bronte",]
df_Bronte
```

```
## # A tibble: 61,913 x 4
##    word           Austen author proportion
##    <chr>           <dbl> <chr>       <dbl>
##  1 '          NA         Bronte NA
##  2 a'ch       NA         Bronte NA
##  3 a'chitect  NA         Bronte NA
##  4 a'eplane   NA         Bronte NA
##  5 a'hm       NA         Bronte NA
##  6 a'll       NA         Bronte NA
##  7 a'most     NA         Bronte  0.0000111
##  8 a'n't       0.00000462 Bronte NA
##  9 a'penny    NA         Bronte NA
## 10 aa         NA         Bronte NA
## # ... with 61,903 more rows
```

```
cor.test(data = df_Bronte, ~ proportion + `Austen`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  proportion and Austen
## t = 118.49, df = 10913, p-value < 2.2e-16
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7417693 0.7581829
## sample estimates:
##       cor
## 0.7500916
```

```
df_Wells <- frequency[frequency$author == "Wells",]
df_Wells
```

```
## # A tibble: 61,913 x 4
##    word           Austen author   proportion
##    <chr>           <dbl> <chr>         <dbl>
##  1 '          NA         Wells   0.00000128
##  2 a'ch       NA         Wells   0.000000640
##  3 a'chitect  NA         Wells   0.00000128
##  4 a'eplane   NA         Wells   0.000000640
##  5 a'hm       NA         Wells   0.000000640
##  6 a'll       NA         Wells   0.00000192
##  7 a'most     NA         Wells   0.00000256
##  8 a'n't       0.00000462 Wells  NA
##  9 a'penny    NA         Wells   0.000000640
## 10 aa         NA         Wells   0.000000640
## # ... with 61,903 more rows
```

```
cor.test(data = df_Wells, ~ proportion + `Austen`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  proportion and Austen
## t = 68.947, df = 11850, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5220976 0.5477983
## sample estimates:
##       cor
## 0.5350717
```

Exercise 2

DICKENS

```
gutenberg_authors %>%
  filter(gutenberg_author_id == 37)
```

```
## # A tibble: 1 x 7
##   gutenberg_author~ author  alias birthdate deathdate wikipedia     aliases
##               <int> <chr>   <chr>     <int>     <int> <chr>         <chr>
## 1                37 Dicken~ Boz        1812      1870 http://en.wi~ <NA>
```

```
dickens2 <- gutenberg_works(gutenberg_author_id == 37)
dickens_vec <- dickens2$gutenberg_id
dickens <- gutenberg_download(dickens_vec)
```

TWAIN

```
gutenberg_authors %>%
  filter(gutenberg_author_id == 53)
```

```
## # A tibble: 1 x 7
##   gutenberg_author~ author alias birthdate deathdate wikipedia   aliases
##               <int> <chr>  <chr>     <int>     <int> <chr>       <chr>
## 1                53 Twain~ <NA>       1835      1910 http://en.~ Twain, Ma~
```

```
twain2 <- gutenberg_works(gutenberg_author_id == 53)
twain_vec <- twain2$gutenberg_id
twain <- gutenberg_download(twain_vec)
```

```
## Warning in .f(.x[[i]], ...): Could not download a book at http://
## aleph.gutenberg.org/1/9/6/8/19682/19682.zip
```

```
## Warning in .f(.x[[i]], ...): Could not download a book at http://
## aleph.gutenberg.org/1/9/8/4/19841/19841.zip
```

VERNE

```
gutenberg_authors %>%
  filter(gutenberg_author_id == 60)
```

```
## # A tibble: 1 x 7
##   gutenberg_author~ author alias birthdate deathdate wikipedia    aliases
##               <int> <chr>  <chr>     <int>     <int> <chr>        <chr>
## 1                60 Verne~ <NA>       1828      1905 http://fr.w~ Verne, J~
```

```
verne2 <- gutenberg_works(gutenberg_author_id == 60)
verne_vec <- verne2$gutenberg_id
verne <- gutenberg_download(verne_vec)
```

```
## Warning in .f(.x[[i]], ...): Could not download a book at http://
## aleph.gutenberg.org/1/9/5/1/19513/19513.zip
```

```
tidy_dickens <- dickens %>%
unnest_tokens(word, text) %>%
mutate(word = str_extract(word, "[a-z']+")) %>%
anti_join(stop_words) %>%
  drop_na()
```

```
## Joining, by = "word"
```

```
tidy_twain <- twain %>%
unnest_tokens(word, text) %>%
mutate(word = str_extract(word, "[a-z']+")) %>%
anti_join(stop_words) %>%
  drop_na()
```

```
## Joining, by = "word"
```

```
tidy_verne <- verne %>%
unnest_tokens(word, text) %>%
mutate(word = str_extract(word, "[a-z']+")) %>%
anti_join(stop_words) %>%
  drop_na()
```

```
## Joining, by = "word"
```

```
tidy_dickens %>%
count(word, sort = TRUE)
```

```
## # A tibble: 48,729 x 2
##    word          n
##    <chr>     <int>
##  1 time      13115
##  2 sir       12993
##  3 dear       9336
##  4 hand       8492
##  5 miss       8348
##  6 head       8327
##  7 day        8322
##  8 night      8081
##  9 house      7716
## 10 gentleman  7084
## # ... with 48,719 more rows
```

```
tidy_twain %>%
count(word, sort = TRUE)
```

```
## # A tibble: 46,963 x 2
##    word        n
##    <chr>   <int>
##  1 time    10504
##  2 day      6651
##  3 people   4861
##  4 night    3815
##  5 hundred  3650
##  6 tom      3543
##  7 life     3394
##  8 head     3361
##  9 house    3230
## 10 half     3126
## # ... with 46,953 more rows
```

```
tidy_verne %>%
count(word, sort = TRUE)
```

```
## # A tibble: 45,392 x 2
##    word        n
##    <chr>   <int>
##  1 captain  6518
##  2 time     5732
##  3 replied  4806
##  4 day      4276
##  5 sea      4159
##  6 island   3789
##  7 water    3377
##  8 feet     2967
##  9 left     2900
## 10 found    2875
## # ... with 45,382 more rows
```

```
tidy_dickens
```

```
## # A tibble: 2,373,161 x 2
##    gutenberg_id word
##           <int> <chr>
##  1           46 christmas
##  2           46 carol
##  3           46 prose
##  4           46 ghost
##  5           46 story
##  6           46 christmas
##  7           46 charles
##  8           46 dickens
##  9           46 preface
## 10           46 endeavoured
## # ... with 2,373,151 more rows
```

```
tidy_twain
```

```
## # A tibble: 1,595,401 x 2
##    gutenberg_id word
##           <int> <chr>
##  1           70 essays
##  2           70 mark
##  3           70 twain
##  4           70 samuel
##  5           70 langhorne
##  6           70 clemens
##  7           70 contents
##  8           70 death
##  9           70 jean
## 10           70 life
## # ... with 1,595,391 more rows
```

```
tidy_verne
```

```
## # A tibble: 1,366,438 x 2
##    gutenberg_id word
##           <int> <chr>
##  1           83 ebook
##  2           83 project
##  3           83 gutenberg's
##  4           83 files
##  5           83 produced
##  6           83 time
##  7           83 proofing
##  8           83 methods
##  9           83 tools
## 10           83 developed
## # ... with 1,366,428 more rows
```

```
frequency_by_word_across_authors2 <- bind_rows(mutate(tidy_dickens,
author = "Dickens"),
mutate(tidy_twain, author = "Twain"),
mutate(tidy_verne, author = "Verne")) %>%
mutate(word = str_extract(word, "[a-z']+")) %>%
count(author, word) %>%
group_by(author) %>%
mutate(proportion = n / sum(n)) %>%
select(-n) %>%
spread(author, proportion)
frequency_by_word_across_authors2
```

```
## # A tibble: 83,402 x 4
##    word            Dickens        Twain Verne
##    <chr>             <dbl>        <dbl> <dbl>
##  1 a'beckett   0.00000126  NA              NA
##  2 a'exposer  NA            0.000000627    NA
##  3 a'hoy      NA            0.000000627    NA
##  4 a'int       0.000000421 NA              NA
##  5 a'mighty's  0.000000421 NA              NA
##  6 a'most      0.0000265   NA              NA
##  7 a'ms       NA            0.000000627    NA
##  8 a'n't       0.00000253  NA              NA
##  9 a'nt        0.00000169   0.000000627    NA
## 10 a'purpose   0.000000843 NA              NA
## # ... with 83,392 more rows
```

```
frequency2 <- frequency_by_word_across_authors2 %>%
gather(author, proportion, `Dickens`:`Twain`)
frequency2
```

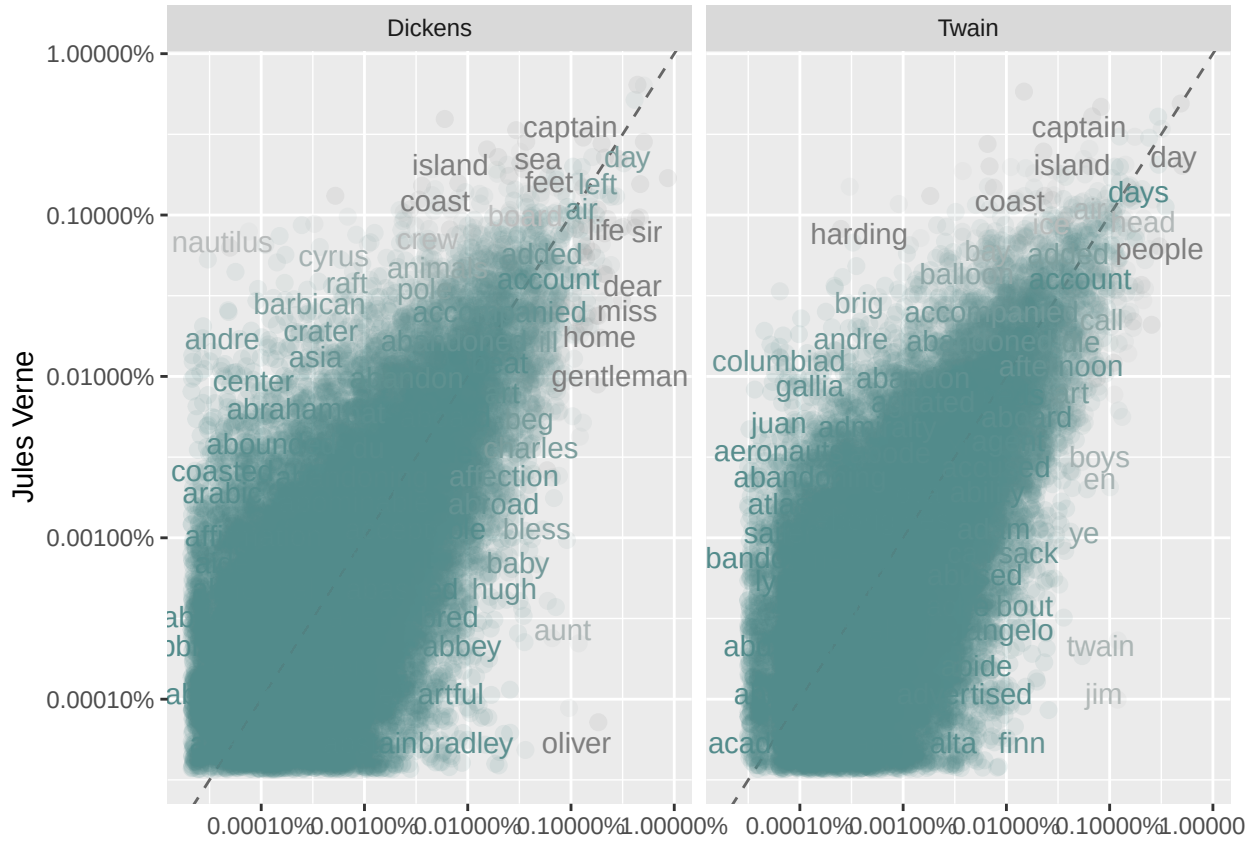
```
## # A tibble: 166,804 x 4
##    word       Verne author    proportion
##    <chr>      <dbl> <chr>          <dbl>
##  1 a'beckett     NA Dickens  0.00000126
```

```
##  2 a'exposer     NA Dickens NA
##  3 a'hoy         NA Dickens NA
##  4 a'int         NA Dickens  0.000000421
##  5 a'mighty's    NA Dickens  0.000000421
##  6 a'most        NA Dickens  0.0000265
##  7 a'ms          NA Dickens NA
##  8 a'n't         NA Dickens  0.00000253
##  9 a'nt          NA Dickens  0.00000169
## 10 a'purpose     NA Dickens  0.000000843
## # ... with 166,794 more rows
```

```
frequency2 %>% ggplot(aes(x = proportion,
y = `Verne`,
color = abs(`Verne` - proportion))) +
geom_abline(color = "gray40", lty = 2) +
geom_jitter(alpha = 0.1, size = 2.5,
width = 0.3, height = 0.3) +
geom_text(aes(label = word),
check_overlap = TRUE, vjust = 1.5) +
scale_x_log10(labels = percent_format()) +
scale_y_log10(labels = percent_format()) +
scale_color_gradient(limits = c(0, 0.001),
low = "darkslategray4",
high = "gray75") +
facet_wrap(~author, ncol = 2) +
theme(legend.position="none") +
labs(y = "Jules Verne", x = NULL)
```

```
## Warning: Removed 115608 rows containing missing values (geom_point).
```

```
## Warning: Removed 115608 rows containing missing values (geom_text).
```




```
df_Dickens <- frequency2[frequency2$author == "Dickens",]
df_Dickens
```

```
## # A tibble: 83,402 x 4
##    word       Verne author    proportion
##    <chr>      <dbl> <chr>          <dbl>
##  1 a'beckett     NA Dickens  0.00000126
##  2 a'exposer     NA Dickens NA
##  3 a'hoy         NA Dickens NA
##  4 a'int         NA Dickens  0.000000421
##  5 a'mighty's    NA Dickens  0.000000421
##  6 a'most        NA Dickens  0.0000265
##  7 a'ms          NA Dickens NA
##  8 a'n't         NA Dickens  0.00000253
##  9 a'nt          NA Dickens  0.00000169
## 10 a'purpose     NA Dickens  0.000000843
## # ... with 83,392 more rows
```

```
cor.test(data = df_Dickens, ~ proportion + `Verne`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  proportion and Verne
## t = 132.16, df = 25849, p-value < 2.2e-16
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6276673 0.6422179
## sample estimates:
##       cor
## 0.6349989
```

```
df_Twain <- frequency2[frequency2$author == "Twain",]
df_Twain
```

```
## # A tibble: 83,402 x 4
##    word       Verne author   proportion
##    <chr>      <dbl> <chr>         <dbl>
##  1 a'beckett     NA Twain  NA
##  2 a'exposer     NA Twain   0.000000627
##  3 a'hoy         NA Twain   0.000000627
##  4 a'int         NA Twain  NA
##  5 a'mighty's    NA Twain  NA
##  6 a'most        NA Twain  NA
##  7 a'ms          NA Twain   0.000000627
##  8 a'n't         NA Twain  NA
##  9 a'nt          NA Twain   0.000000627
## 10 a'purpose     NA Twain  NA
## # ... with 83,392 more rows
```

```
cor.test(data = df_Twain, ~ proportion + `Verne`)
```

```
##
##  Pearson's product-moment correlation
##
## data:  proportion and Verne
## t = 156.24, df = 25343, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6941169 0.7066602
## sample estimates:
##       cor
## 0.7004426
```