HW text mining

Lin Pin Tzu (Ruby)

2022-07-14

```
text1 <- c("The Gettysburg Address is a speech that US President Abraham",
           "Lincoln delivered during the American Civil War at the dedication of the",
           "Soldiers' National Cemetery in Gettysburg, Pennsylvania, on the afternoon of",
           "November 19,1863, four and a half months after the Union armies defeated those",
           "of the Confederacy at the Battle of Gettysburg. It is one of the best-known",
           "speeches in American history. Lincoln's carefully crafted address, not",
           "even that day's primary speech, came to be seen as one of the greatest and",
           "most influential statements of American national purpose. In just 271 words,",
           "beginning with the now famous phrase Four score and seven years ago",
           "referring to the signing of the Declaration of Independence 87 years earlier",
           " Lincoln described the US as a nation conceived in Liberty, and dedicated",
           "to the proposition that all men are created equal and represented the Civil",
           "War as a test that would determine whether such a nation, the Union sundered by",
           "the secession crisis, could endure. He extolled the sacrifices of those who",
           "died at Gettysburg in defense of those principles.")

text1
```

```
##  [1] "The Gettysburg Address is a speech that US President Abraham"
##  [2] "Lincoln delivered during the American Civil War at the dedication of the"
##  [3] "Soldiers' National Cemetery in Gettysburg, Pennsylvania, on the afternoon of"
##  [4] "November 19,1863, four and a half months after the Union armies defeated those"
##  [5] "of the Confederacy at the Battle of Gettysburg. It is one of the best-known"
##  [6] "speeches in American history. Lincoln's carefully crafted address, not"
##  [7] "even that day's primary speech, came to be seen as one of the greatest and"
##  [8] "most influential statements of American national purpose. In just 271 words,"
##  [9] "beginning with the now famous phrase Four score and seven years ago"
## [10] "referring to the signing of the Declaration of Independence 87 years earlier"
## [11] " Lincoln described the US as a nation conceived in Liberty, and dedicated"
## [12] "to the proposition that all men are created equal and represented the Civil"
## [13] "War as a test that would determine whether such a nation, the Union sundered by"
## [14] "the secession crisis, could endure. He extolled the sacrifices of those who"
## [15] "died at Gettysburg in defense of those principles."
```

1 Use and show R code to convert text1 above to a tibble

```
text_tibble <- tibble(line = 16:30, text1 = text1)

text_tibble
```

```
## # A tibble: 15 x 2
##     line text1
```

```
##    <int> <chr>
##  1    16 "The Gettysburg Address is a speech that US President Abraham"
##  2    17 "Lincoln delivered during the American Civil War at the dedication of ~
##  3    18 "Soldiers' National Cemetery in Gettysburg, Pennsylvania, on the after~
##  4    19 "November 19,1863, four and a half months after the Union armies defea~
##  5    20 "of the Confederacy at the Battle of Gettysburg. It is one of the best~
##  6    21 "speeches in American history. Lincoln's carefully crafted address, no~
##  7    22 "even that day's primary speech, came to be seen as one of the greates~
##  8    23 "most influential statements of American national purpose. In just 271~
##  9    24 "beginning with the now famous phrase Four score and seven years ago"
## 10    25 "referring to the signing of the Declaration of Independence 87 years ~
## 11    26 " Lincoln described the US as a nation conceived in Liberty, and dedic~
## 12    27 "to the proposition that all men are created equal and represented the~
## 13    28 "War as a test that would determine whether such a nation, the Union s~
## 14    29 "the secession crisis, could endure. He extolled the sacrifices of tho~
## 15    30 "died at Gettysburg in defense of those principles."
```

2 Use and show R code to produce a table that shows the line location for every word in the text.

```
text_tibble %>%
  unnest_tokens(word, text1) -> tibble2
tibble2
```

```
## # A tibble: 179 x 2
##     line word
##    <int> <chr>
##  1    16 the
##  2    16 gettysburg
##  3    16 address
##  4    16 is
##  5    16 a
##  6    16 speech
##  7    16 that
##  8    16 us
##  9    16 president
## 10    16 abraham
## # ... with 169 more rows
```

3 Use and show R code to Find frequencies >= 3 for each word in text1.

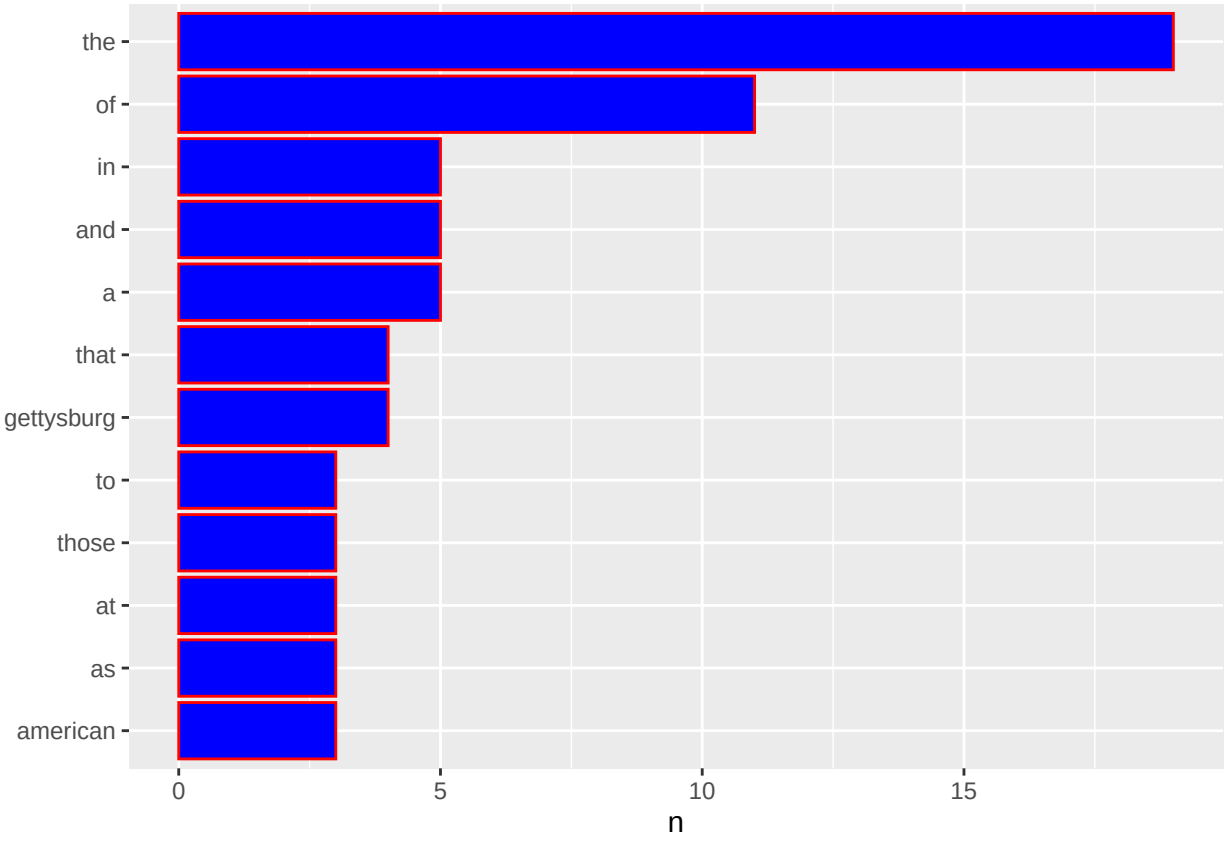
```
tibble2%>%
  count(word, sort =TRUE) %>%
 filter(n >= 3)
```

```
## # A tibble: 12 x 2
##    word           n
##    <chr>      <int>
##  1 the           19
##  2 of            11
##  3 a              5
```

```
##  4 and            5
##  5 in             5
##  6 gettysburg     4
##  7 that           4
##  8 american       3
##  9 as             3
## 10 at             3
## 11 those          3
## 12 to             3
```

4 Create a Data visual (Bar Graph) showing and comparing word frequencies that are found in the table produced for # 3

```
tibble2%>%
  count(word, sort =TRUE) %>%
  filter(n >= 3) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(n, word)) +
  geom_col(fill = "blue", color = "red") +
  labs(y = NULL)
```



5 Use and show R code to find the total number of words in the text1.

```
total <- tibble2 %>%
  count(word, sort =TRUE)
sum(total$n)
```

```
## [1] 179
```

```
nrow(tibble2)
```

```
## [1] 179
```