hw6-Rohit-Thakur

Rohit Thakur

3/31/2020

```
library(textdata)
```

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

```
library(readr)
library(ggplot2)
library(tidytext)
```

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

```
library(tokenizers)
```

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

```
library(dplyr)
```

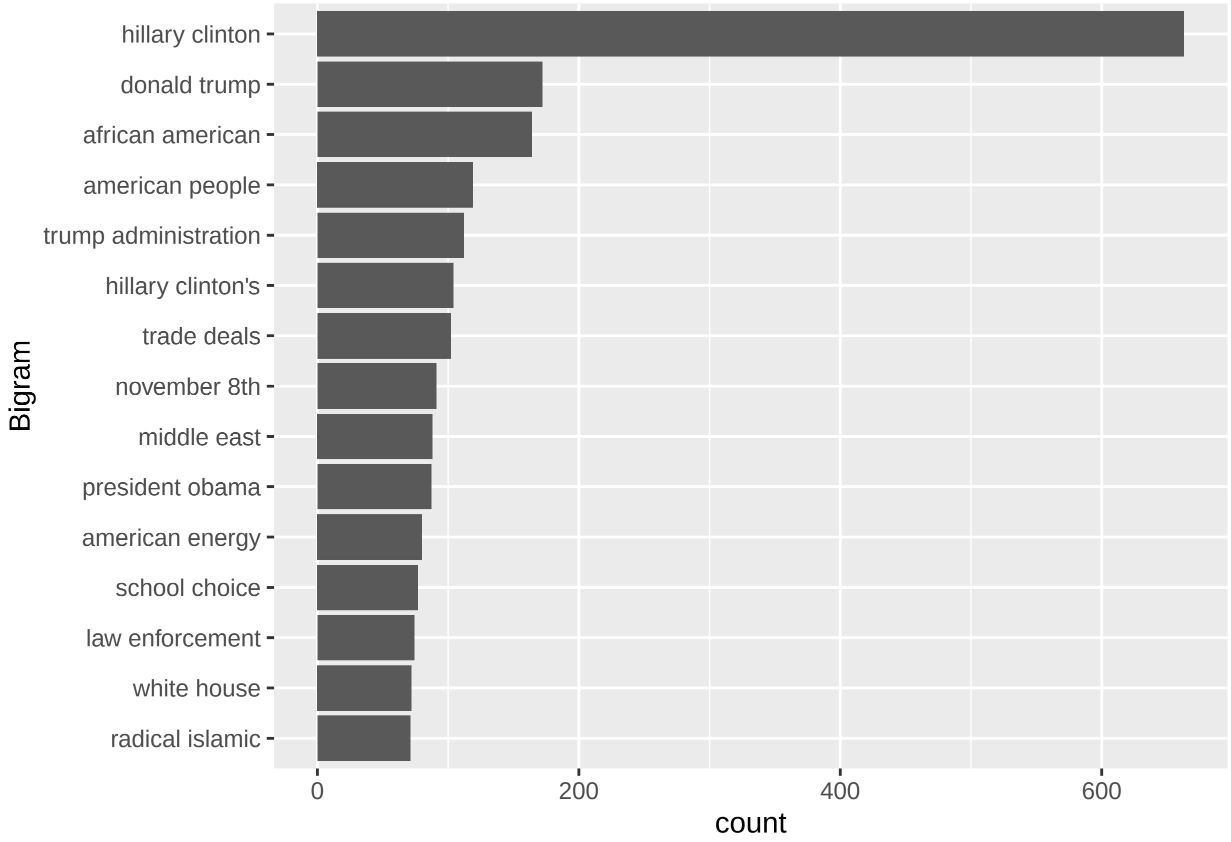
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
```

PROBLEM 1

```
speech<-read_lines("D:/Spring 20 Sem 2/DMP/full_speech.txt")
tidy_speech<-tibble(line=1:length(speech), text=speech)
tidy_speech<-tidy_speech%>%unnest_tokens(bigram,text,token="ngrams",n=2)
tidy_bigram<-tidy_speech%>%separate(bigram, c("word1", "word2"), sep = " ")
tidy_bigram1<-tidy_bigram%>%
  filter(!word1 %in% c(stop_words$word, "applause"))%>%
  filter(!word1 %in% c("not", "no", "never", "without"))%>%
  filter(!word2 %in% c(stop_words$word, "applause"))%>%
  unite(bigram, word1, word2, sep = " ")
tidy_bigram1%>%count(bigram,sort = TRUE)%>%
  top_n(15)%>%
  ggplot()+geom_bar(aes(x=reorder(bigram,n),y=n),stat="identity")+
  coord_flip()+xlab("Bigram")+ylab("count")
```

```
## Selecting by n
```

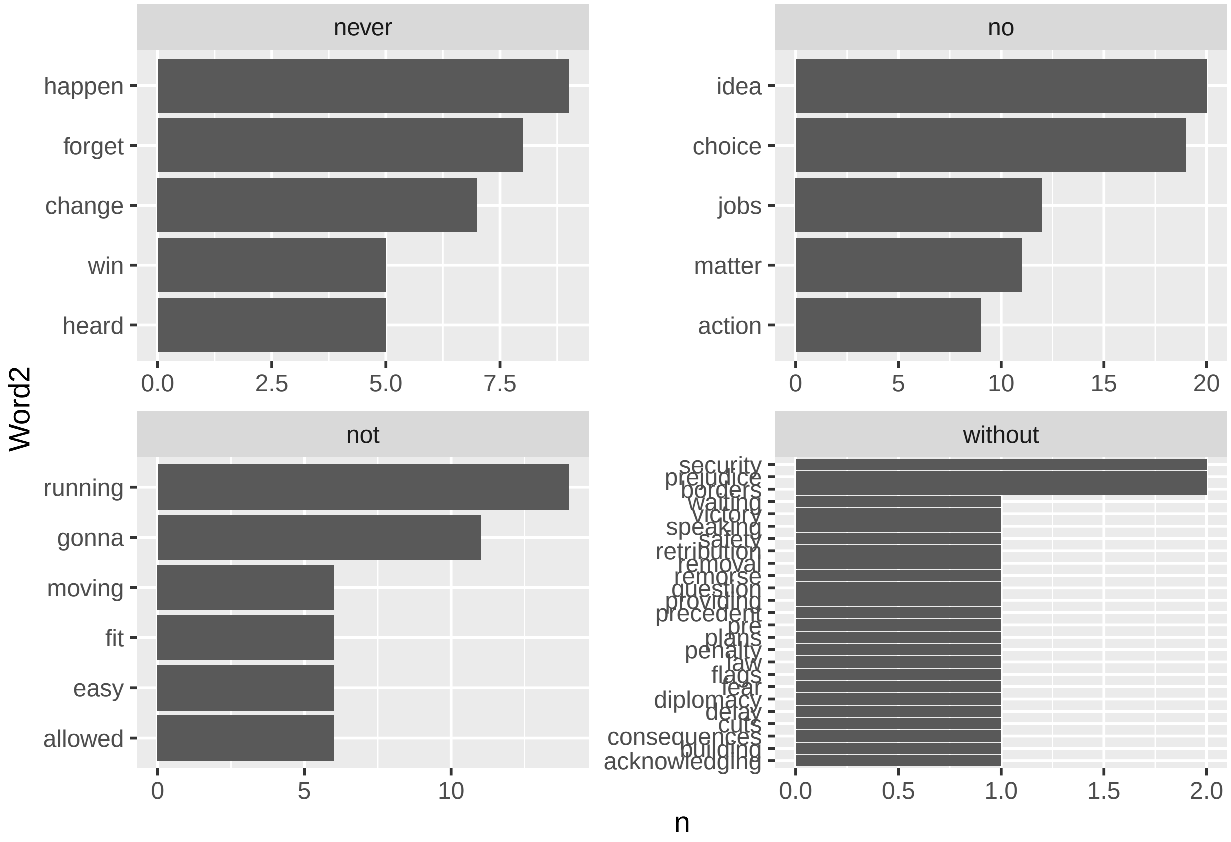


PROBLEM 2

```
tidy_bigram2<-tidy_speech%>%separate(bigram, c("word1", "word2"), sep = " ")
tidy_bigram2<-tidy_bigram2%>%
  filter(word1 %in% c("not", "no", "never", "without"))%>%
  filter(!word2 %in% c(stop_words$word, "applause"))
problem_2<-tidy_bigram2%>%count(word1,word2,sort=TRUE)%>%
  group_by(word1)%>%
  top_n(5)%>%
  ggplot()+geom_bar(aes(x=reorder(word2,n),y=n),stat="identity")+
  facet_wrap(~word1,scales="free")+
  coord_flip()+xlab("Word2")
```

```
## Selecting by n
```

```
print(problem_2)
```

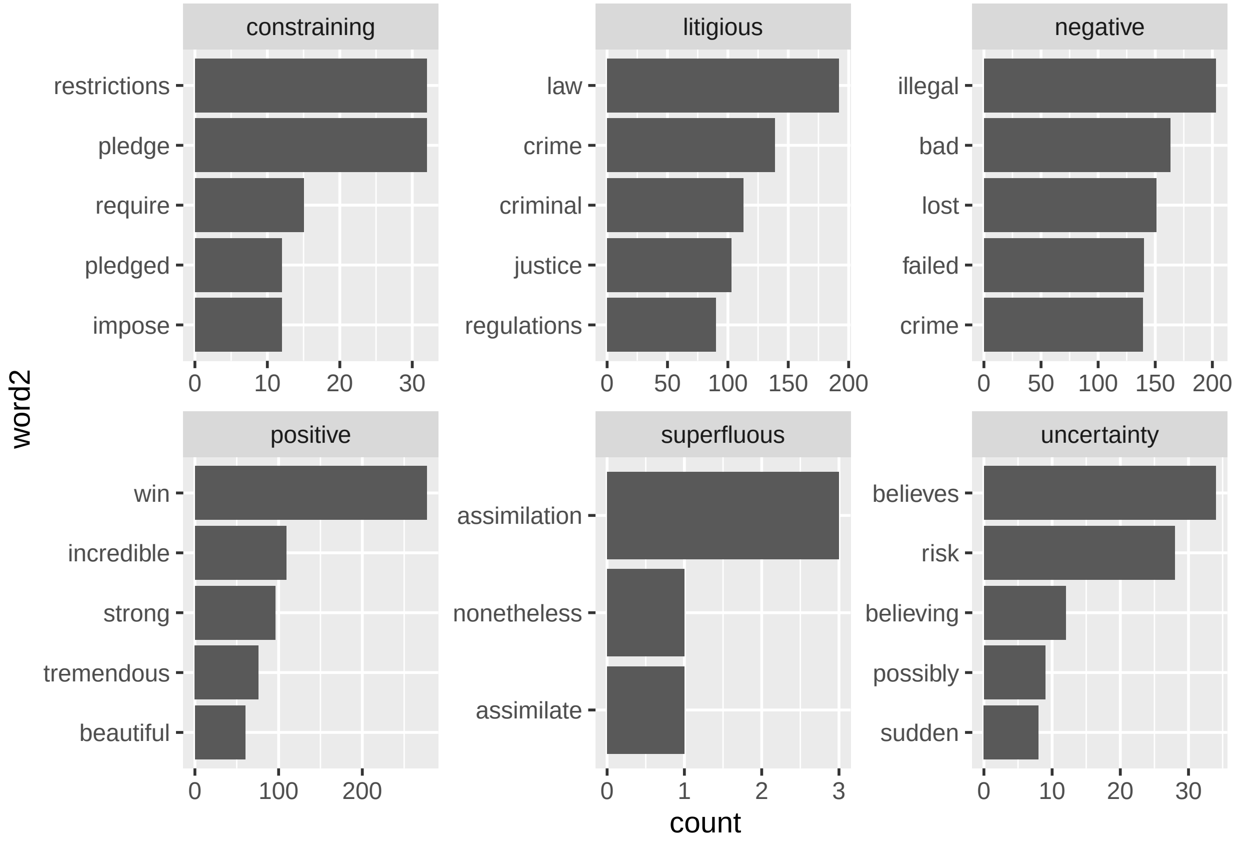


PROBLEM 3

```
problem3<-tidy_bigram%>%
  filter(!word1 %in% c("not", "no", "never", "without"))%>%
  filter(!word2 %in% c(stop_words$word,"applause"))
problem3<-problem3%>%
  inner_join(get_sentiments("loughran"), by=c("word2"="word"))%>%
  count(word2, sentiment, sort=TRUE)%>%
  group_by(sentiment)%>%
  top_n(5)
```

```
## Selecting by n
```

```
problem3_plot<-problem3%>%
  ggplot()+geom_bar(aes(x=reorder(word2,n),y=n),stat="identity")+
  facet_wrap(~sentiment,scales="free")+
  coord_flip()+xlab("word2")+ylab("count")
print(problem3_plot)
```



PROBLEM 4

```
library(gutenbergr)
```

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

```
titles<-c("Pride and Prejudice","The War of the Worlds")
books<-gutenberg_works(title %in% titles)%>%
  gutenberg_download(meta_fields = c("title","author"))
```

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

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

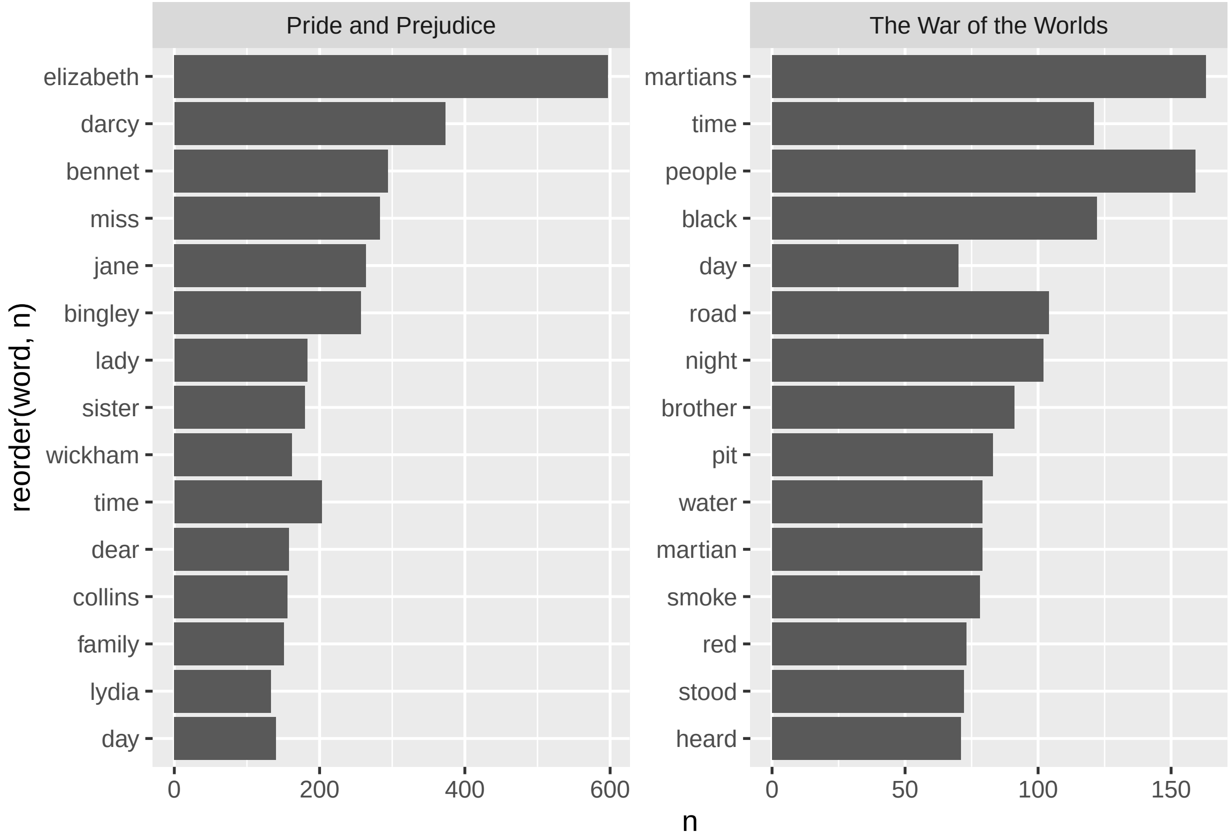
```
books<-mutate(books,document=row_number())
tidy_book<-books%>%
  unnest_tokens(word,text)%>%
  group_by(word)%>%
  filter(!n()<10)
tidy_book<-tidy_book%>%
  anti_join(stop_words)%>%
  count(title,word,sort=TRUE)%>%
  group_by(title)%>%
  top_n(15)
```

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

## Selecting by n
```

```
tidy_book
```

```
## # A tibble: 30 x 3
## # Groups:   title [2]
##    word      title                    n
##    <chr>     <chr>                <int>
##  1 elizabeth Pride and Prejudice    597
##  2 darcy     Pride and Prejudice    373
##  3 bennet    Pride and Prejudice    294
##  4 miss      Pride and Prejudice    283
##  5 jane      Pride and Prejudice    264
##  6 bingley   Pride and Prejudice    257
##  7 time      Pride and Prejudice    203
##  8 lady      Pride and Prejudice    183
##  9 sister    Pride and Prejudice    180
## 10 martians  The War of the Worlds   163
## # ... with 20 more rows
```

```
ggplot(tidy_book)+geom_bar(aes(x=reorder(word,n),y=n),stat="identity")+
  facet_wrap(~title,scales="free")+
  coord_flip()
```



PROBLEM 5

```
tidy_book_1<-books%>%
  mutate(document=row_number())%>%
  unnest_tokens(word,text)%>%
  group_by(word)%>%
  filter(!n()<10)
head(tidy_book_1)
```

```
## # A tibble: 6 x 5
## # Groups:   word [5]
##   gutenberg_id title                author                       document word
##          <int> <chr>                <chr>                           <int> <chr>
## 1           36 The War of the Worlds Wells, H. G. (Herbert Georg~        1 the
## 2           36 The War of the Worlds Wells, H. G. (Herbert Georg~        1 war
## 3           36 The War of the Worlds Wells, H. G. (Herbert Georg~        1 of
## 4           36 The War of the Worlds Wells, H. G. (Herbert Georg~        1 the
## 5           36 The War of the Worlds Wells, H. G. (Herbert Georg~        3 by
## 6           36 The War of the Worlds Wells, H. G. (Herbert Georg~        6 but
```

```
doc<-tidy_book_1%>%
  count(document,word)
head(doc)
```

```
## # A tibble: 6 x 3
## # Groups:   word [1]
##   word  document     n
##   <chr>    <int> <int>
## 1 _he_      8780     1
## 2 _he_      9168     1
## 3 _he_      9924     1
## 4 _he_     10978     1
## 5 _he_     11677     1
## 6 _he_     12770     1
```

Creating Document term matrix

```
doc_mat<-doc%>%
  cast_dtm(document,word,n)%>%
  as.matrix()
```

```
doc_id<-data.frame(document=as.integer(rownames(doc_mat)))
doc_id<-doc_id%>%left_join(books)%>%
  select(document,author)
```

```
## Joining, by = "document"
```

```
doc_id<-mutate(doc_id,author=as.factor(author))
```

```r
library(caret)
```

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

```
## Loading required package: lattice
```

```r
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 3.6.3
```

```r
set.seed(0)
partition<- createDataPartition(doc_id$author,p=0.75, list=FALSE)
test_data<-doc_mat[-partition,]
x=doc_mat[partition,]
y=doc_id$author[partition]
model<-train(x=x,y=y,method="svmLinear",
             trControl =trainControl(method="none"))
prediction<-predict(model,test_data)
confusionMatrix(prediction,doc_id$author[-partition])
```

```
## Confusion Matrix and Statistics
##
##                                 Reference
## Prediction                      Austen, Jane Wells, H. G. (Herbert George)
##   Austen, Jane                          2403                          202
##   Wells, H. G. (Herbert George)          260                         1148
##
##                Accuracy : 0.8849
##                  95% CI : (0.8746, 0.8946)
##     No Information Rate : 0.6636
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7448
##
##  Mcnemar's Test P-Value : 0.008004
##
##             Sensitivity : 0.9024
##             Specificity : 0.8504
##          Pos Pred Value : 0.9225
##          Neg Pred Value : 0.8153
##              Prevalence : 0.6636
##          Detection Rate : 0.5988
##    Detection Prevalence : 0.6491
##       Balanced Accuracy : 0.8764
##
##        'Positive' Class : Austen, Jane
##
```