CS544 Homework 6

Phillip Escandon - [escandon@bu.edu](mailto:escandon@bu.edu)

24 February, 2021

Table of Contents

## 1. Stringr and strings

# Read in the lincoln text and place into the words variable.  
file <- "http://people.bu.edu/kalathur/datasets/lincoln.txt"  
 words <- scan(file, what=character())

### 1 a. Detect and show all words with punctuation.

# Words with punctuation  
punctWords <- words[str\_detect(words, "[:punct:]")]  
punctWords

## [1] "States:" "itself," "office."   
## [4] "excitement." "endangered." "apprehension."   
## [7] "Indeed," "inspection." "you."   
## [10] "--" "purpose," "indirectly,"   
## [13] "exists." "so," "so."   
## [16] "them;" "this," "acceptance,"   
## [19] "me," "read:" "Resolved,"   
## [22] "States," "exclusively," "depend;"   
## [25] "Territory," "pretext," "crimes."   
## [28] "sentiments," "property," "peace,"   
## [31] "Administration." "add," "too,"   
## [34] "which," "laws," "demanded,"   
## [37] "--" "another." "labor."   
## [40] "provisions:" "State," "thereof,"   
## [43] "another," "labor," "due."   
## [46] "slaves;" "law." "--"   
## [49] "other." "proposition," "then,"   
## [52] "unanimous." "Now," "temper,"   
## [55] "oath?" "authority," "one."   
## [58] "surrendered," "done." "kept?"   
## [61] "Again:" "introduced," "slave?"   
## [64] "States?" "rules;" "enforced,"   
## [67] "all," "stations," "unconstitutional."  
## [70] "seventy-two" "Constitution." "Government."   
## [73] "perils," "success." "Yet,"   
## [76] "precedent," "difficulty." "Union,"   
## [79] "menaced," "attempted." "perpetual."   
## [82] "implied," "expressed," "governments."   
## [85] "termination." "Constitution," "forever,"   
## [88] "itself." "Again:" "proper,"   
## [91] "merely," "it," "contract,"   
## [94] "it?" "--" "it,"   
## [97] "--" "it?" "principles,"   
## [100] "itself." "Constitution." "formed,"   
## [103] "fact," "1774." "1776."   
## [106] "matured," "perpetual," "1778."   
## [109] "finally," "1787," "Union."   
## [112] "possible," "Constitution," "perpetuity."   
## [115] "Union;" "void," "revolutionary,"   
## [118] "circumstances." "unbroken," "ability,"   
## [121] "care," "me," "States."   
## [124] "part," "masters," "people,"   
## [127] "contrary." "menace," "itself."   
## [130] "violence," "authority." "hold,"   
## [133] "occupy," "imposts;" "objects,"   
## [136] "invasion," "anywhere." "offices,"   
## [139] "object." "offices," "offices."   
## [142] "mails," "repelled," "Union."   
## [145] "reflection." "proper," "exercised,"   
## [148] "affections." "deny;" "such,"   
## [151] "them." "those," "however,"   
## [154] "speak?" "fabric," "benefits,"   
## [157] "memories," "hopes," "it?"   
## [160] "existence?" "you," "from,"   
## [163] "mistake?" "maintained." "true,"   
## [166] "then," "denied?" "not."   
## [169] "Happily," "this." "Think,"   
## [172] "can," "denied." "right,"   
## [175] "revolution;" "one." "case."   
## [178] "negations," "prohibitions," "them."   
## [181] "administration." "questions." "authority?"   
## [184] "say." "Territories?" "say."   
## [187] "Territories?" "say." "controversies,"   
## [190] "minorities." "acquiesce," "must,"   
## [193] "cease." "alternative," "other."   
## [196] "acquiesce," "them," "minority."   
## [199] "instance," "again," "it?"   
## [202] "this." "secession?" "anarchy."   
## [205] "limitations," "sentiments," "people."   
## [208] "despotism." "impossible." "minority,"   
## [211] "arrangement," "inadmissible;" "that,"   
## [214] "principle," "left." "Court,"   
## [217] "suit," "Government." "case,"   
## [220] "it," "case," "cases,"   
## [223] "practice." "time," "Court,"   
## [226] "rulers," "tribunal." "judges."   
## [229] "them," "purposes." "extended,"   
## [232] "extended." "dispute." "enforced,"   
## [235] "perhaps," "itself." "cases,"   
## [238] "each." "This," "think,"   
## [241] "cured," "before." "trade,"   
## [244] "suppressed," "section," "slaves,"   
## [247] "surrendered," "other." "speaking,"   
## [250] "separate." "them." "other,"   
## [253] "this." "face," "intercourse,"   
## [256] "hostile," "them." "possible,"   
## [259] "then," "?" "laws?"   
## [262] "friends?" "war," "always;"   
## [265] "when," "either," "fighting,"   
## [268] "questions," "intercourse," "you."   
## [271] "country," "institutions," "it."   
## [274] "Government," "it." "amended."   
## [277] "amendments," "subject," "itself;"   
## [280] "should," "circumstances," "it."   
## [283] "preferable," "themselves," "others,"   
## [286] "purpose," "refuse." "--"   
## [289] "amendment," "however," "--"   
## [292] "Congress," "States," "service."   
## [295] "said," "that," "law,"   
## [298] "irrevocable." "people," "States."   
## [301] "choose," "it." "successor."   
## [304] "people?" "world?" "differences,"   
## [307] "right?" "Nations," "justice,"   
## [310] "North," "South," "people."   
## [313] "mischief," "intervals." "years."   
## [316] "countrymen," "all," "subject."   
## [319] "time." "," "time;"   
## [322] "it." "unimpaired," "and,"   
## [325] "point," "it;" "power,"   
## [328] "would," "either." "dispute,"   
## [331] "action." "Intelligence," "patriotism,"   
## [334] "Christianity," "difficulty." "hands,"   
## [337] "countrymen," "," "war."   
## [340] "." "aggressors." "Government,"   
## [343] "preserve," "protect," "it."   
## [346] "close." "enemies," "friends."   
## [349] "enemies." "affection." "memory,"   
## [352] "land," "Union," "touched,"   
## [355] "be," "nature."

### 1 b. Replace punctuation with space.

Remove any empty words and convert to lowercase

# replace punctuation with space  
  
new\_words<- str\_remove\_all(str\_replace\_all(punctWords,"[:punct:]","")," ")  
# Could not get rid of "" and had difficulty with '--'  
new\_words <- new\_words[new\_words!='']  
new\_words <- str\_to\_lower(new\_words)  
new\_words

## [1] "states" "itself" "office"   
## [4] "excitement" "endangered" "apprehension"   
## [7] "indeed" "inspection" "you"   
## [10] "purpose" "indirectly" "exists"   
## [13] "so" "so" "them"   
## [16] "this" "acceptance" "me"   
## [19] "read" "resolved" "states"   
## [22] "exclusively" "depend" "territory"   
## [25] "pretext" "crimes" "sentiments"   
## [28] "property" "peace" "administration"   
## [31] "add" "too" "which"   
## [34] "laws" "demanded" "another"   
## [37] "labor" "provisions" "state"   
## [40] "thereof" "another" "labor"   
## [43] "due" "slaves" "law"   
## [46] "other" "proposition" "then"   
## [49] "unanimous" "now" "temper"   
## [52] "oath" "authority" "one"   
## [55] "surrendered" "done" "kept"   
## [58] "again" "introduced" "slave"   
## [61] "states" "rules" "enforced"   
## [64] "all" "stations" "unconstitutional"  
## [67] "seventytwo" "constitution" "government"   
## [70] "perils" "success" "yet"   
## [73] "precedent" "difficulty" "union"   
## [76] "menaced" "attempted" "perpetual"   
## [79] "implied" "expressed" "governments"   
## [82] "termination" "constitution" "forever"   
## [85] "itself" "again" "proper"   
## [88] "merely" "it" "contract"   
## [91] "it" "it" "it"   
## [94] "principles" "itself" "constitution"   
## [97] "formed" "fact" "1774"   
## [100] "1776" "matured" "perpetual"   
## [103] "1778" "finally" "1787"   
## [106] "union" "possible" "constitution"   
## [109] "perpetuity" "union" "void"   
## [112] "revolutionary" "circumstances" "unbroken"   
## [115] "ability" "care" "me"   
## [118] "states" "part" "masters"   
## [121] "people" "contrary" "menace"   
## [124] "itself" "violence" "authority"   
## [127] "hold" "occupy" "imposts"   
## [130] "objects" "invasion" "anywhere"   
## [133] "offices" "object" "offices"   
## [136] "offices" "mails" "repelled"   
## [139] "union" "reflection" "proper"   
## [142] "exercised" "affections" "deny"   
## [145] "such" "them" "those"   
## [148] "however" "speak" "fabric"   
## [151] "benefits" "memories" "hopes"   
## [154] "it" "existence" "you"   
## [157] "from" "mistake" "maintained"   
## [160] "true" "then" "denied"   
## [163] "not" "happily" "this"   
## [166] "think" "can" "denied"   
## [169] "right" "revolution" "one"   
## [172] "case" "negations" "prohibitions"   
## [175] "them" "administration" "questions"   
## [178] "authority" "say" "territories"   
## [181] "say" "territories" "say"   
## [184] "controversies" "minorities" "acquiesce"   
## [187] "must" "cease" "alternative"   
## [190] "other" "acquiesce" "them"   
## [193] "minority" "instance" "again"   
## [196] "it" "this" "secession"   
## [199] "anarchy" "limitations" "sentiments"   
## [202] "people" "despotism" "impossible"   
## [205] "minority" "arrangement" "inadmissible"   
## [208] "that" "principle" "left"   
## [211] "court" "suit" "government"   
## [214] "case" "it" "case"   
## [217] "cases" "practice" "time"   
## [220] "court" "rulers" "tribunal"   
## [223] "judges" "them" "purposes"   
## [226] "extended" "extended" "dispute"   
## [229] "enforced" "perhaps" "itself"   
## [232] "cases" "each" "this"   
## [235] "think" "cured" "before"   
## [238] "trade" "suppressed" "section"   
## [241] "slaves" "surrendered" "other"   
## [244] "speaking" "separate" "them"   
## [247] "other" "this" "face"   
## [250] "intercourse" "hostile" "them"   
## [253] "possible" "then" "laws"   
## [256] "friends" "war" "always"   
## [259] "when" "either" "fighting"   
## [262] "questions" "intercourse" "you"   
## [265] "country" "institutions" "it"   
## [268] "government" "it" "amended"   
## [271] "amendments" "subject" "itself"   
## [274] "should" "circumstances" "it"   
## [277] "preferable" "themselves" "others"   
## [280] "purpose" "refuse" "amendment"   
## [283] "however" "congress" "states"   
## [286] "service" "said" "that"   
## [289] "law" "irrevocable" "people"   
## [292] "states" "choose" "it"   
## [295] "successor" "people" "world"   
## [298] "differences" "right" "nations"   
## [301] "justice" "north" "south"   
## [304] "people" "mischief" "intervals"   
## [307] "years" "countrymen" "all"   
## [310] "subject" "time" "time"   
## [313] "it" "unimpaired" "and"   
## [316] "point" "it" "power"   
## [319] "would" "either" "dispute"   
## [322] "action" "intelligence" "patriotism"   
## [325] "christianity" "difficulty" "hands"   
## [328] "countrymen" "war" "aggressors"   
## [331] "government" "preserve" "protect"   
## [334] "it" "close" "enemies"   
## [337] "friends" "enemies" "affection"   
## [340] "memory" "land" "union"   
## [343] "touched" "be" "nature"

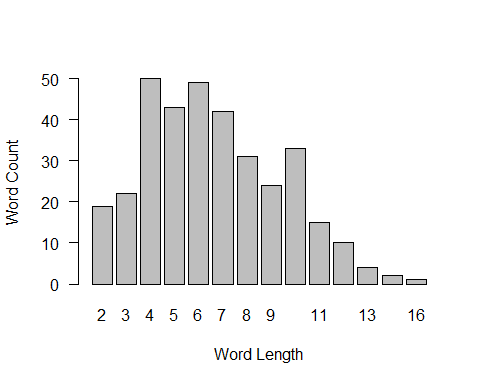
### 1 c. Top 5 frequent words

# Find the top 5 frequently used words.  
top5 <- sort(table(new\_words),decreasing = TRUE)[1:5]  
top5

## new\_words  
## it them itself states people   
## 14 7 6 6 5

### 1 d. Plot the distribution of the word frequencies of the word lengths.

# Should sort the barplot  
# Barplot seems more applicable   
barplot(table(str\_length(new\_words)),horiz = FALSE,las = 1,xlab = 'Word Length',ylab = 'Word Count')



#hist(table(str\_length(new\_words)))

### 1 e. What are the words in the new\_words data with the longest length

lenlongestWord <- str\_length(new\_words) %>% max()  
  
longestWord <- new\_words[str\_length(new\_words) == lenlongestWord]  
  
glue::glue("The longest word is ", longestWord)

## The longest word is unconstitutional

### 1 f. Show all words that begin with letter p

pWords <- new\_words[str\_detect(new\_words,"^[pP]")]  
pWords

## [1] "purpose" "pretext" "property" "peace" "provisions"   
## [6] "proposition" "perils" "precedent" "perpetual" "proper"   
## [11] "principles" "perpetual" "possible" "perpetuity" "part"   
## [16] "people" "proper" "prohibitions" "people" "principle"   
## [21] "practice" "purposes" "perhaps" "possible" "preferable"   
## [26] "purpose" "people" "people" "people" "point"   
## [31] "power" "patriotism" "preserve" "protect"

### 1 g. Show all words that end with a letter r

rWords <- new\_words[str\_detect(new\_words,"[rR]$")]  
rWords

## [1] "another" "labor" "another" "labor" "other" "temper"   
## [7] "forever" "proper" "proper" "however" "other" "other"   
## [13] "other" "war" "either" "however" "successor" "power"   
## [19] "either" "war"

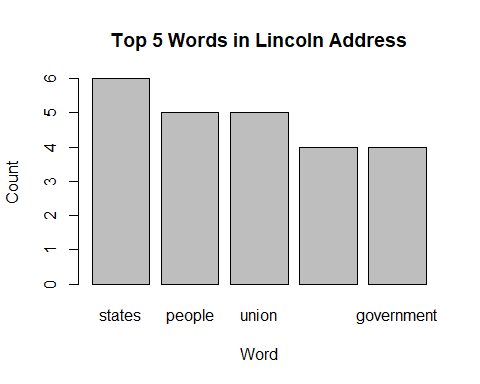
## Stopwords Bonus

stopfile <- "http://people.bu.edu/kalathur/datasets/stopwords.txt"  
stopwords <- scan(stopfile, what=character())

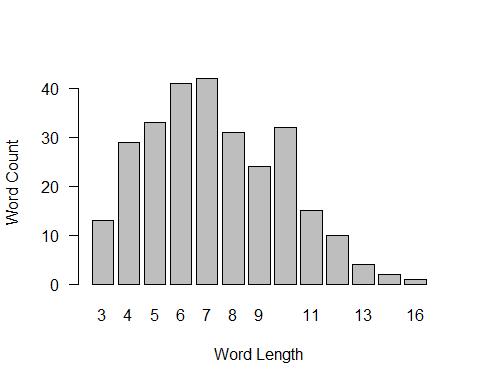
Filter the stopwords from the new\_words data. Use the %in% operator.  
Repeat c) and d) for this data without the stop words.

### Bonus c.

#new\_words[stopwords %in% new\_words]  
  
# this is the one we want  
#new\_words  
new\_words\_wo\_stops <- new\_words[!new\_words %in% stopwords]  
  
# Top 5 most frequent words  
top5Main <- sort(table(new\_words\_wo\_stops),decreasing = TRUE)[1:5]  
barplot(top5Main,xlab = 'Word',ylab = "Count",main = "Top 5 Words in Lincoln Address")



# Frequency of the word lengths  
barplot(table(str\_length(new\_words\_wo\_stops)),horiz = FALSE,las = 1,xlab = 'Word Length',ylab = 'Word Count')



## 2. Data Wrangling and the Tidyverse

Download the following csv file, <http://people.bu.edu/kalathur/usa_daily_avg_temps.csv> locally first and use read.csv to load the data into a data frame.

dailyTemps <- read\_csv("usa\_daily\_avg\_temps.csv")

##   
## -- Column specification --------------------------------------------------------  
## cols(  
## state = col\_character(),  
## city = col\_character(),  
## month = col\_double(),  
## day = col\_double(),  
## year = col\_double(),  
## avgtemp = col\_double()  
## )

### 2 a. Convert the data frame into a tibble and assign it to the variable

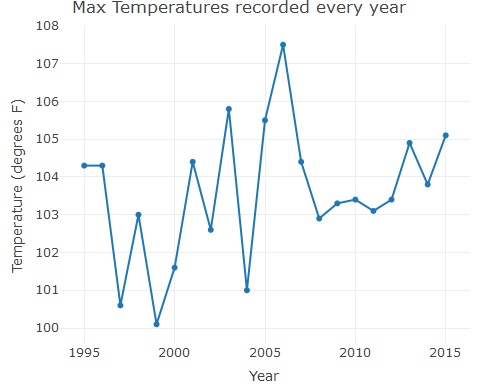
usaDailyTemps.

usaDailyTemps <- as\_tibble(dailyTemps)

### 2 b. What are the maximum temperatures recorded for each year? Show the

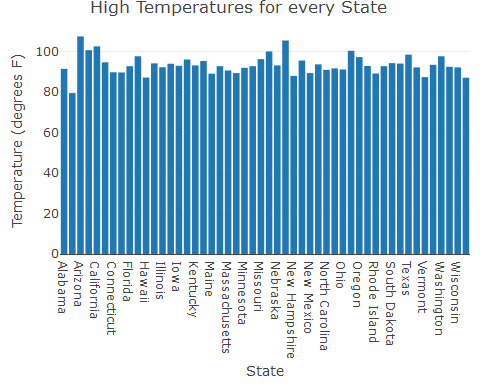
values and also the appropriate plot for the results.

usaDailyTemps %>%   
 group\_by(year) %>%  
 summarize(tmp = max(avgtemp))%>%  
 plot\_ly(x=~year,y=~tmp,type = "scatter",mode = 'lines+markers')%>%  
 layout(title = "Max Temperatures recorded every year",  
 xaxis = list(title = "Year"),  
 yaxis = list (title = "Temperature (degrees F)"))



### 2 c. What are the maximum temperatures recorded for each state? Show the values and also the appropriate plot for the results.

usaDailyTemps %>%   
 group\_by(state) %>%  
 summarize(tmp = max(avgtemp)) %>%  
 arrange(desc(tmp)) %>% # This portion doesn't quite work- I'd like the barplot to be ordered by temp.  
 plot\_ly(x = ~state)%>%  
 add\_trace(y=~tmp,name = 'Temp', type = 'bar')%>%  
 layout(title = "High Temperatures for every State",  
 xaxis = list(title = "State"),  
 yaxis = list (title = "Temperature (degrees F)"))



### 2 d. Filter the Boston data from usaDailyTemps and assign it to the variable bostonDailyTemps.

bostonDailyTemps <- usaDailyTemps %>%   
 filter(city=="Boston")

### 2 e. What are the average monthly temperatures for Boston?

Show the values and also the appropriate plot for the results. Use the bostonDailyTemps.

bostonDailyTemps %>%  
 group\_by(month) %>%  
 summarize(avg = mean(avgtemp)) %>%  
 plot\_ly(x = ~month, y = ~avg,type = 'scatter', mode = 'lines+markers')%>%  
 layout(title = "Average Temperatures in Boston",  
 xaxis = list(title = "Month"),  
 yaxis = list (title = "Temperature (degrees F)"))

