Case Study 2

Hadley Dixon, Haleigh Cole, Daniel Baltrons, Katherine Dhuey, Nathan Garrick, Piyush Antal

2022-10-10

1. Using the tweets.csv data that is available on the GitHub site, provide code to do the following:

Load the twitter data

```
twitter <- read.csv("https://raw.githubusercontent.com/jdwilson4/Intro-Data-Science/master/Data/tweets.csv", header = FALSE)
twitter <- twitter$V1
```

(a) Identify all tweets with the word ‘flight’ in them

The tweets with the word flight in them are:

```
grep("(flight)", twitter, value = TRUE)
```

```
## [1] "@VirginAmerica seriously would pay $30 a flight for seats that didn't have this playing."
## [2] "@VirginAmerica So excited for my first cross country flight LAX to MCO I've heard nothing but great things about Virgin America. #29DaysToGo"
## [3] "@VirginAmerica amazing to me that we can't get any cold air from the vents. #VX358 #noair #worstflightever #roasted #SFOtoBOS"
```

(b) How many tweets end in a question mark?

There are 5 tweets that end in a question mark.

```
length(grep("[?]", twitter))
```

```
## [1] 5
```

(c) How many tweets have airport codes in them (assume any three subsequent capital letters are airport codes)

There are 7 tweets that have airport codes in them.

```
length(grep("[# ][A-Z]{3}", twitter))
```

```
## [1] 7
```

(d) Identify all tweets with URLs in them

The tweets with URLs in them are:

```
grep("(http)", twitter, value = TRUE)
```

```
## [1] "@VirginAmerica Really missed a prime opportunity for Men Without Hats parody, th
ere. https://t.co/mWpG7grEZP"
## [2] "@VirginAmerica @virginmedia I'm flying your #fabulous #Seductive skies again! U
take all the #stress away from travel http://t.co/ahlXHhKiyn"
## [3] "@VirginAmerica I love this graphic. http://t.co/UT5GrRwAaA"
```

(e) Replace all instances of repeated exclamation points with a single exclamation point

```
library(stringr)
str_replace_all(twitter, "[!]{2,}", "!")
```

```
##  [1] "@VirginAmerica plus you've added commercials to the experience... tacky."
##  [2] "@VirginAmerica I didn't today... Must mean I need to take another trip!"
##  [3] "@VirginAmerica it's really aggressive to blast obnoxious \"entertainment\" in y
our guests' faces &amp; they have little recourse"
##  [4] "@VirginAmerica and it's a really big bad thing about it"
##  [5] "@VirginAmerica seriously would pay $30 a flight for seats that didn't have this
playing."
##  [6] "@VirginAmerica it's really the only bad thing about flying VA"
##  [7] "@VirginAmerica Really missed a prime opportunity for Men Without Hats parody, t
here. https://t.co/mWpG7grEZP"
##  [8] "@VirginAmerica it was amazing, and arrived an hour early. You're too good to m
e."
##  [9] "@VirginAmerica did you know that suicide is the second leading cause of death a
mong teens 10-24"
## [10] "@VirginAmerica I &lt;3 pretty graphics. so much better than minimal iconograph
y. :D"
## [11] "@VirginAmerica This is such a great deal! Already thinking about my 2nd trip to
@Australia &amp; I haven't even gone on my 1st trip yet! ;p"
## [12] "@VirginAmerica @virginmedia I'm flying your #fabulous #Seductive skies again! U
take all the #stress away from travel http://t.co/ahlXHhKiyn"
## [13] "@VirginAmerica SFO-PDX schedule is still MIA."
## [14] "@VirginAmerica So excited for my first cross country flight LAX to MCO I've hea
rd nothing but great things about Virgin America. #29DaysToGo"
## [15] "@VirginAmerica  I flew from NYC to SFO last week and couldn't fully sit in my s
eat due to two large gentleman on either side of me. HELP!"
## [16] "@VirginAmerica you know what would be amazingly awesome? BOS-FLL PLEASE! I want
to fly with only you."
## [17] "@VirginAmerica why are your first fares in May over three times more than other
carriers when all seats are available to select???"
## [18] "@VirginAmerica I love this graphic. http://t.co/UT5GrRwAaA"
## [19] "@VirginAmerica I love the hipster innovation. You are a feel good brand."
## [20] "@VirginAmerica will you be making BOS&gt;LAS non stop permanently anytime soo
n?"
## [21] "@VirginAmerica status match program.  I applied and it's been three weeks.  Cal
led and emailed with no response."
## [22] "@VirginAmerica What happened 2 ur vegan food options?! At least say on ur site
so i know I won't be able 2 eat anything for next 6 hrs #fail"
## [23] "@VirginAmerica do you miss me? Don't worry we'll be together very soon."
## [24] "@VirginAmerica amazing to me that we can't get any cold air from the vents. #VX
358 #noair #worstflightever #roasted #SFOtoBOS"
## [25] "@VirginAmerica LAX to EWR - Middle seat on a red eye. Such a noob maneuver. #se
ndambien #andchexmix"
```

(f) Replace consecutive exclamation points, question marks, and periods with a single period, split the tweet on periods, and create a list where each element is a vector of the split strings from each tweet

```
punctuation_rep <- str_replace_all(twitter, "[!?.]{2,}", ".")
strsplit(punctuation_rep, "[\\.]")
```

```
## [[1]]
## [1] "@VirginAmerica plus you've added commercials to the experience"
## [2] " tacky"
##
## [[2]]
## [1] "@VirginAmerica I didn't today"
## [2] " Must mean I need to take another trip!"
##
## [[3]]
## [1] "@VirginAmerica it's really aggressive to blast obnoxious \"entertainment\" in yo
ur guests' faces &amp; they have little recourse"
##
## [[4]]
## [1] "@VirginAmerica and it's a really big bad thing about it"
##
## [[5]]
## [1] "@VirginAmerica seriously would pay $30 a flight for seats that didn't have this
playing"
##
## [[6]]
## [1] "@VirginAmerica it's really the only bad thing about flying VA"
##
## [[7]]
## [1] "@VirginAmerica Really missed a prime opportunity for Men Without Hats parody, th
ere"
## [2] " https://t"
## [3] "co/mWpG7grEZP"
##
## [[8]]
## [1] "@VirginAmerica it was amazing, and arrived an hour early"
## [2] " You're too good to me"
##
## [[9]]
## [1] "@VirginAmerica did you know that suicide is the second leading cause of death am
ong teens 10-24"
##
## [[10]]
## [1] "@VirginAmerica I &lt;3 pretty graphics"
## [2] " so much better than minimal iconography"
## [3] " :D"
##
## [[11]]
## [1] "@VirginAmerica This is such a great deal! Already thinking about my 2nd trip to
@Australia &amp; I haven't even gone on my 1st trip yet! ;p"
##
## [[12]]
## [1] "@VirginAmerica @virginmedia I'm flying your #fabulous #Seductive skies again! U
take all the #stress away from travel http://t"
## [2] "co/ahlXHhKiyn"
##
## [[13]]
## [1] "@VirginAmerica SFO-PDX schedule is still MIA"
```

```
##
## [[14]]
## [1] "@VirginAmerica So excited for my first cross country flight LAX to MCO I've hear
d nothing but great things about Virgin America"
## [2] " #29DaysToGo"
##
## [[15]]
## [1] "@VirginAmerica  I flew from NYC to SFO last week and couldn't fully sit in my se
at due to two large gentleman on either side of me"
## [2] " HELP!"
##
## [[16]]
## [1] "@VirginAmerica you know what would be amazingly awesome? BOS-FLL PLEASE"
## [2] " I want to fly with only you"
##
## [[17]]
## [1] "@VirginAmerica why are your first fares in May over three times more than other
carriers when all seats are available to select"
##
## [[18]]
## [1] "@VirginAmerica I love this graphic" " http://t"
## [3] "co/UT5GrRwAaA"
##
## [[19]]
## [1] "@VirginAmerica I love the hipster innovation"
## [2] " You are a feel good brand"
##
## [[20]]
## [1] "@VirginAmerica will you be making BOS&gt;LAS non stop permanently anytime soon?"
##
## [[21]]
## [1] "@VirginAmerica status match program"
## [2] "  I applied and it's been three weeks"
## [3] "  Called and emailed with no response"
##
## [[22]]
## [1] "@VirginAmerica What happened 2 ur vegan food options"
## [2] " At least say on ur site so i know I won't be able 2 eat anything for next 6 hrs
#fail"
##
## [[23]]
## [1] "@VirginAmerica do you miss me? Don't worry we'll be together very soon"
##
## [[24]]
## [1] "@VirginAmerica amazing to me that we can't get any cold air from the vents"
## [2] " #VX358 #noair #worstflightever #roasted #SFOtoBOS"
##
## [[25]]
## [1] "@VirginAmerica LAX to EWR - Middle seat on a red eye"
## [2] " Such a noob maneuver"
## [3] " #sendambien #andchexmix"
```

2. Read the text into R and manipulate it in order to create a data frame with the following summaries for each speech:

You now have the fundamental R tools to complete this exercise, but you will may still have to explore new techniques and packages. You will work with the full text of the State of the Union speeches from 1790 until 2012. The speeches are all in the file stateoftheunion1790-2012.txt on the GitHub site.

Loading the state of the union data

```
SoTU <- readLines("https://raw.githubusercontent.com/jdwilson4/Intro-Data-Science/maste
r/Data/stateoftheunion1790-2012.txt")
```

```
## Warning in readLines("https://raw.githubusercontent.com/jdwilson4/Intro-Data-
## Science/master/Data/stateoftheunion1790-2012.txt"): incomplete final line found
## on 'https://raw.githubusercontent.com/jdwilson4/Intro-Data-Science/master/Data/
## stateoftheunion1790-2012.txt'
```

```
SoTU.selected <- SoTU[11:236]
```

(a) the President's name who gave the speech

The following list shows the name of the President who gave each speech:

```
presidents.names <- str_extract_all(SoTU.selected, "^(.+?),")
presidents.names.no.comma <- str_replace_all(presidents.names, ",", "")
presidents.names.no.comma
```

```
##   [1] "  George Washington"     "  George Washington"
##   [3] "  George Washington"     "  George Washington"
##   [5] "  George Washington"     "  George Washington"
##   [7] "  George Washington"     "  George Washington"
##   [9] "  John Adams"            "  John Adams"
##  [11] "  John Adams"            "  John Adams"
##  [13] "  Thomas Jefferson"      "  Thomas Jefferson"
##  [15] "  Thomas Jefferson"      "  Thomas Jefferson"
##  [17] "  Thomas Jefferson"      "  Thomas Jefferson"
##  [19] "  Thomas Jefferson"      "  Thomas Jefferson"
##  [21] "  James Madison"         "  James Madison"
##  [23] "  James Madison"         "  James Madison"
##  [25] "  James Madison"         "  James Madison"
##  [27] "  James Madison"         "  James Madison"
##  [29] "  James Monroe"          "  James Monroe"
##  [31] "  James Monroe"          "  James Monroe"
##  [33] "  James Monroe"          "  James Monroe"
##  [35] "  James Monroe"          "  James Monroe"
##  [37] "  John Quincy Adams"     "  John Quincy Adams"
##  [39] "  John Quincy Adams"     "  John Quincy Adams"
##  [41] "  Andrew Jackson"        "  Andrew Jackson"
##  [43] "  Andrew Jackson"        "  Andrew Jackson"
##  [45] "  Andrew Jackson"        "  Andrew Jackson"
##  [47] "  Andrew Jackson"        "  Andrew Jackson"
##  [49] "  Martin van Buren"      "  Martin van Buren"
##  [51] "  Martin van Buren"      "  Martin van Buren"
##  [53] "  John Tyler"            "  John Tyler"
##  [55] "  John Tyler"            "  John Tyler"
##  [57] "  James Polk"            "  James Polk"
##  [59] "  James Polk"            "  James Polk"
##  [61] "  Zachary Taylor"        "  Millard Fillmore"
##  [63] "  Millard Fillmore"      "  Millard Fillmore"
##  [65] "  Franklin Pierce"       "  Franklin Pierce"
##  [67] "  Franklin Pierce"       "  Franklin Pierce"
##  [69] "  James Buchanan"        "  James Buchanan"
##  [71] "  James Buchanan"        "  James Buchanan"
##  [73] "  Abraham Lincoln"       "  Abraham Lincoln"
##  [75] "  Abraham Lincoln"       "  Abraham Lincoln"
##  [77] "  Andrew Johnson"        "  Andrew Johnson"
##  [79] "  Andrew Johnson"        "  Andrew Johnson"
##  [81] "  Ulysses S. Grant"      "  Ulysses S. Grant"
##  [83] "  Ulysses S. Grant"      "  Ulysses S. Grant"
##  [85] "  Ulysses S. Grant"      "  Ulysses S. Grant"
##  [87] "  Ulysses S. Grant"      "  Ulysses S. Grant"
##  [89] "  Rutherford B. Hayes"   "  Rutherford B. Hayes"
##  [91] "  Rutherford B. Hayes"   "  Rutherford B. Hayes"
##  [93] "  Chester A. Arthur"     "  Chester A. Arthur"
##  [95] "  Chester A. Arthur"     "  Chester A. Arthur"
##  [97] "  Grover Cleveland"      "  Grover Cleveland"
##  [99] "  Grover Cleveland"      "  Grover Cleveland"
## [101] "  Benjamin Harrison"     "  Benjamin Harrison"
## [103] "  Benjamin Harrison"     "  Benjamin Harrison"
## [105] "  Grover Cleveland"      "  Grover Cleveland"
## [107] "  Grover Cleveland"      "  Grover Cleveland"
## [109] "  William McKinley"      "  William McKinley"
## [111] "  William McKinley"      "  William McKinley"
## [113] "  Theodore Roosevelt"    "  Theodore Roosevelt"
## [115] "  Theodore Roosevelt"    "  Theodore Roosevelt"
## [117] "  Theodore Roosevelt"    "  Theodore Roosevelt"
## [119] "  Theodore Roosevelt"    "  Theodore Roosevelt"
## [121] "  William H. Taft"       "  William H. Taft"
## [123] "  William H. Taft"       "  William H. Taft"
## [125] "  Woodrow Wilson"        "  Woodrow Wilson"
## [127] "  Woodrow Wilson"        "  Woodrow Wilson"
## [129] "  Woodrow Wilson"        "  Woodrow Wilson"
## [131] "  Woodrow Wilson"        "  Woodrow Wilson"
## [133] "  Warren Harding"        "  Warren Harding"
## [135] "  Calvin Coolidge"       "  Calvin Coolidge"
## [137] "  Calvin Coolidge"       "  Calvin Coolidge"
## [139] "  Calvin Coolidge"       "  Calvin Coolidge"
## [141] "  Herbert Hoover"        "  Herbert Hoover"
## [143] "  Herbert Hoover"        "  Herbert Hoover"
## [145] "  Franklin D. Roosevelt" "  Franklin D. Roosevelt"
## [147] "  Franklin D. Roosevelt" "  Franklin D. Roosevelt"
## [149] "  Franklin D. Roosevelt" "  Franklin D. Roosevelt"
## [151] "  Franklin D. Roosevelt" "  Franklin D. Roosevelt"
## [153] "  Franklin D. Roosevelt" "  Franklin D. Roosevelt"
## [155] "  Franklin D. Roosevelt" "  Franklin D. Roosevelt"
## [157] "  Harry S. Truman"       "  Harry S. Truman"
## [159] "  Harry S. Truman"       "  Harry S. Truman"
## [161] "  Harry S. Truman"       "  Harry S. Truman"
## [163] "  Harry S. Truman"       "  Harry S. Truman"
## [165] "  Dwight D. Eisenhower"  "  Dwight D. Eisenhower"
## [167] "  Dwight D. Eisenhower"  "  Dwight D. Eisenhower"
## [169] "  Dwight D. Eisenhower"  "  Dwight D. Eisenhower"
## [171] "  Dwight D. Eisenhower"  "  Dwight D. Eisenhower"
## [173] "  Dwight D. Eisenhower"  "  John F. Kennedy"
## [175] "  John F. Kennedy"       "  John F. Kennedy"
## [177] "  Lyndon B. Johnson"     "  Lyndon B. Johnson"
## [179] "  Lyndon B. Johnson"     "  Lyndon B. Johnson"
## [181] "  Lyndon B. Johnson"     "  Lyndon B. Johnson"
## [183] "  Richard Nixon"         "  Richard Nixon"
## [185] "  Richard Nixon"         "  Richard Nixon"
## [187] "  Richard Nixon"         "  Gerald R. Ford"
## [189] "  Gerald R. Ford"        "  Gerald R. Ford"
## [191] "  Jimmy Carter"          "  Jimmy Carter"
## [193] "  Jimmy Carter"          "  Jimmy Carter"
## [195] "  Ronald Reagan"         "  Ronald Reagan"
## [197] "  Ronald Reagan"         "  Ronald Reagan"
## [199] "  Ronald Reagan"         "  Ronald Reagan"
## [201] "  Ronald Reagan"         "  George H.W. Bush"
## [203] "  George H.W. Bush"      "  George H.W. Bush"
## [205] "  George H.W. Bush"      "  William J. Clinton"
## [207] "  William J. Clinton"    "  William J. Clinton"
## [209] "  William J. Clinton"     "  William J. Clinton"
## [211] "  William J. Clinton"     "  William J. Clinton"
## [213] "  William J. Clinton"     "  George W. Bush"
## [215] "  George W. Bush"         "  George W. Bush"
## [217] "  George W. Bush"         "  George W. Bush"
## [219] "  George W. Bush"         "  George W. Bush"
## [221] "  George W. Bush"         "  George W. Bush"
## [223] "  Barack Obama"           "  Barack Obama"
## [225] "  Barack Obama"           "  Barack Obama"
```

(b) the year the speech was given

The following list shows the year each speech was given:

```
years <- unlist(str_extract_all(SoTU.selected, "\\b[^,]+$"))
years
```

```
##   [1] "1790" "1790" "1791" "1792" "1793" "1794" "1795" "1796" "1797" "1798"
##  [11] "1799" "1800" "1801" "1802" "1803" "1804" "1805" "1806" "1807" "1808"
##  [21] "1809" "1810" "1811" "1812" "1813" "1814" "1815" "1816" "1817" "1818"
##  [31] "1819" "1820" "1821" "1822" "1823" "1824" "1825" "1826" "1827" "1828"
##  [41] "1829" "1830" "1831" "1832" "1833" "1834" "1835" "1836" "1837" "1838"
##  [51] "1839" "1840" "1841" "1842" "1843" "1844" "1845" "1846" "1847" "1848"
##  [61] "1849" "1850" "1851" "1852" "1853" "1854" "1855" "1856" "1857" "1858"
##  [71] "1859" "1860" "1861" "1862" "1863" "1864" "1865" "1866" "1867" "1868"
##  [81] "1869" "1870" "1871" "1872" "1873" "1874" "1875" "1876" "1877" "1878"
##  [91] "1879" "1880" "1881" "1882" "1883" "1884" "1885" "1886" "1887" "1888"
## [101] "1889" "1890" "1891" "1892" "1893" "1894" "1895" "1896" "1897" "1898"
## [111] "1899" "1900" "1901" "1902" "1903" "1904" "1905" "1906" "1907" "1908"
## [121] "1909" "1910" "1911" "1912" "1913" "1914" "1915" "1916" "1917" "1918"
## [131] "1919" "1920" "1921" "1922" "1923" "1924" "1925" "1926" "1927" "1928"
## [141] "1929" "1930" "1931" "1932" "1934" "1935" "1936" "1937" "1938" "1939"
## [151] "1940" "1941" "1942" "1943" "1944" "1945" "1946" "1947" "1948" "1949"
## [161] "1950" "1951" "1952" "1953" "1953" "1954" "1955" "1956" "1957" "1958"
## [171] "1959" "1960" "1961" "1961" "1962" "1963" "1964" "1965" "1966" "1967"
## [181] "1968" "1969" "1970" "1971" "1972" "1973" "1974" "1975" "1976" "1977"
## [191] "1978" "1979" "1980" "1981" "1982" "1983" "1984" "1985" "1986" "1987"
## [201] "1988" "1989" "1990" "1991" "1992" "1993" "1994" "1995" "1996" "1997"
## [211] "1998" "1999" "2000" "2001" "2001" "2002" "2003" "2004" "2005" "2006"
## [221] "2007" "2008" "2009" "2010" "2011" "2012"
```

(c) the month the speech was given

The following list shows the month each speech was given:

```
raw.monthANDday <- str_extract_all(SoTU.selected, "((?:[^,]+,){2})([^,]+)")
monthANDday <- str_extract_all(raw.monthANDday, "\\b[^,]+$")
monthANDday.split <- unlist(str_split(monthANDday, " "))
monthANDday.split
```

```
##   [1] "January"   "8"         "December"  "8"         "October"   "25"
##   [7] "November"  "6"         "December"  "3"         "November"  "19"
##  [13] "December"  "8"         "December"  "7"         "November"  "22"
##  [19] "December"  "8"         "December"  "3"         "November"  "11"
##  [25] "December"  "8"         "December"  "15"        "October"   "17"
##  [31] "November"  "8"         "December"  "3"         "December"  "2"
##  [37] "October"   "27"        "November"  "8"         "November"  "29"
##  [43] "December"  "5"         "November"  "5"         "November"  "4"
##  [49] "December"  "7"         "September" "20"        "December"  "5"
##  [55] "December"  "3"         "December"  "12"        "November"  "16"
##  [61] "December"  "7"         "November"  "14"        "December"  "3"
##  [67] "December"  "3"         "December"  "2"         "December"  "7"
##  [73] "December"  "6"         "December"  "5"         "December"  "4"
##  [79] "December"  "2"         "December"  "8"         "December"  "6"
##  [85] "December"  "6"         "December"  "4"         "December"  "3"
##  [91] "December"  "1"         "December"  "7"         "December"  "5"
##  [97] "December"  "5"         "December"  "3"         "December"  "2"
## [103] "December"  "5"         "December"  "7"         "December"  "6"
## [109] "December"  "6"         "December"  "3"         "December"  "2"
## [115] "December"  "8"         "December"  "7"         "December"  "5"
## [121] "December"  "4"         "December"  "2"         "December"  "2"
## [127] "December"  "6"         "December"  "5"         "December"  "4"
## [133] "December"  "31"        "December"  "2"         "December"  "8"
## [139] "December"  "6"         "December"  "19"        "December"  "3"
## [145] "December"  "3"         "December"  "1"         "December"  "8"
## [151] "December"  "6"         "December"  "4"         "December"  "3"
## [157] "December"  "3"         "December"  "9"         "December"  "6"
## [163] "December"  "5"         "December"  "4"         "December"  "2"
## [169] "December"  "1"         "December"  "7"         "December"  "7"
## [175] "December"  "5"         "December"  "3"         "December"  "2"
## [181] "December"  "1"         "December"  "6"         "December"  "6"
## [187] "December"  "4"         "December"  "4"         "December"  "1"
## [193] "December"  "8"         "December"  "6"         "December"  "6"
## [199] "December"  "3"         "December"  "3"         "December"  "1"
## [205] "December"  "9"         "December"  "6"         "December"  "3"
## [211] "December"  "2"         "December"  "7"         "December"  "4"
## [217] "December"  "6"         "December"  "5"         "December"  "5"
## [223] "December"  "3"         "December"  "3"         "December"  "2"
## [229] "December"  "7"         "December"  "6"         "December"  "5"
## [235] "December"  "3"         "December"  "3"         "December"  "8"
## [241] "December"  "7"         "December"  "6"         "December"  "5"
## [247] "December"  "3"         "December"  "2"         "December"  "8"
## [253] "December"  "7"         "December"  "5"         "December"  "4"
## [259] "December"  "2"         "December"  "2"         "December"  "7"
## [265] "December"  "6"         "December"  "8"         "December"  "6"
## [271] "December"  "3"         "December"  "8"         "December"  "7"
## [277] "December"  "6"         "December"  "4"         "December"  "3"
## [283] "December"  "2"         "December"  "8"         "December"  "6"
## [289] "January"   "3"         "January"   "4"         "January"   "3"
## [295] "January"   "6"         "January"   "3"         "January"   "4"
## [301] "January"   "3"         "January"   "6"         "January"   "6"
## [307] "January"   "7"         "January"   "11"        "January"   "6"
## [313] "January"   "21"        "January"   "6"         "January"   "7"
## [319] "January"   "5"         "January"   "4"         "January"   "8"
## [325] "January"   "9"         "January"   "7"         "February"  "2"
## [331] "January"   "7"         "January"   "6"         "January"   "5"
## [337] "January"   "10"        "January"   "9"         "January"   "9"
## [343] "January"   "7"         "January"   "12"        "January"   "30"
## [349] "January"   "11"        "January"   "14"        "January"   "8"
## [355] "January"   "4"         "January"   "12"        "January"   "10"
## [361] "January"   "17"        "January"   "14"        "January"   "22"
## [367] "January"   "22"        "January"   "20"        "February"  "2"
## [373] "January"   "30"        "January"   "15"        "January"   "19"
## [379] "January"   "12"        "January"   "19"        "January"   "25"
## [385] "January"   "21"        "January"   "16"        "January"   "26"
## [391] "January"   "25"        "January"   "25"        "February"  "6"
## [397] "February"  "4"         "January"   "27"        "January"   "25"
## [403] "February"  "9"         "January"   "31"        "January"   "29"
## [409] "January"   "28"        "February"  "17"        "January"   "25"
## [415] "January"   "24"        "January"   "23"        "February"  "4"
## [421] "January"   "27"        "January"   "19"        "January"   "27"
## [427] "February"  "27"        "September" "20"        "January"   "29"
## [433] "January"   "28"        "January"   "20"        "February"  "2"
## [439] "January"   "31"        "January"   "23"        "January"   "28"
## [445] "February"  "24"        "January"   "27"        "January"   "25"
## [451] "January"   "23"
```

```
unlisted.monthANDday <- unlist(monthANDday.split)
Months <- unlisted.monthANDday[seq(from = 1, to = length(unlisted.monthANDday), by = 2)]
Months
```

```
##   [1] "January"   "December"  "October"   "November"  "December"  "November"
##   [7] "December"  "December"  "November"  "December"  "December"  "November"
##  [13] "December"  "December"  "October"   "November"  "December"  "December"
##  [19] "October"   "November"  "November"  "December"  "November"  "November"
##  [25] "December"  "September" "December"  "December"  "December"  "November"
##  [31] "December"  "November"  "December"  "December"  "December"  "December"
##  [37] "December"  "December"  "December"  "December"  "December"  "December"
##  [43] "December"  "December"  "December"  "December"  "December"  "December"
##  [49] "December"  "December"  "December"  "December"  "December"  "December"
##  [55] "December"  "December"  "December"  "December"  "December"  "December"
##  [61] "December"  "December"  "December"  "December"  "December"  "December"
##  [67] "December"  "December"  "December"  "December"  "December"  "December"
##  [73] "December"  "December"  "December"  "December"  "December"  "December"
##  [79] "December"  "December"  "December"  "December"  "December"  "December"
##  [85] "December"  "December"  "December"  "December"  "December"  "December"
##  [91] "December"  "December"  "December"  "December"  "December"  "December"
##  [97] "December"  "December"  "December"  "December"  "December"  "December"
## [103] "December"  "December"  "December"  "December"  "December"  "December"
## [109] "December"  "December"  "December"  "December"  "December"  "December"
## [115] "December"  "December"  "December"  "December"  "December"  "December"
## [121] "December"  "December"  "December"  "December"  "December"  "December"
## [127] "December"  "December"  "December"  "December"  "December"  "December"
## [133] "December"  "December"  "December"  "December"  "December"  "December"
## [139] "December"  "December"  "December"  "December"  "December"  "December"
## [145] "January"   "January"   "January"   "January"   "January"   "January"
## [151] "January"   "January"   "January"   "January"   "January"   "January"
## [157] "January"   "January"   "January"   "January"   "January"   "January"
## [163] "January"   "January"   "February"  "January"   "January"   "January"
## [169] "January"   "January"   "January"   "January"   "January"   "January"
## [175] "January"   "January"   "January"   "January"   "January"   "January"
## [181] "January"   "January"   "January"   "January"   "January"   "February"
## [187] "January"   "January"   "January"   "January"   "January"   "January"
## [193] "January"   "January"   "January"   "January"   "January"   "February"
## [199] "February"  "January"   "January"   "February"  "January"   "January"
## [205] "January"   "February"  "January"   "January"   "January"   "February"
## [211] "January"   "January"   "January"   "February"  "September" "January"
## [217] "January"   "January"   "February"  "January"   "January"   "January"
## [223] "February"  "January"   "January"   "January"
```

(d) day of the week the speech was given

The following list shows the day each speech was given:

```
Days <- unlisted.monthANDday[seq(from = 2, to = length(unlisted.monthANDday), by = 2)]
Days
```

```
##   [1] "8"  "8"  "25" "6"  "3"  "19" "8"  "7"  "22" "8"  "3"  "11" "8"  "15" "17"
##  [16] "8"  "3"  "2"  "27" "8"  "29" "5"  "5"  "4"  "7"  "20" "5"  "3"  "12" "16"
##  [31] "7"  "14" "3"  "3"  "2"  "7"  "6"  "5"  "4"  "2"  "8"  "6"  "6"  "4"  "3" 
##  [46] "1"  "7"  "5"  "5"  "3"  "2"  "5"  "7"  "6"  "6"  "3"  "2"  "8"  "7"  "5" 
##  [61] "4"  "2"  "2"  "6"  "5"  "4"  "31" "2"  "8"  "6"  "19" "3"  "3"  "1"  "8" 
##  [76] "6"  "4"  "3"  "3"  "9"  "6"  "5"  "4"  "2"  "1"  "7"  "7"  "5"  "3"  "2" 
##  [91] "1"  "6"  "6"  "4"  "4"  "1"  "8"  "6"  "6"  "3"  "3"  "1"  "9"  "6"  "3" 
## [106] "2"  "7"  "4"  "6"  "5"  "5"  "3"  "3"  "2"  "7"  "6"  "5"  "3"  "3"  "8" 
## [121] "7"  "6"  "5"  "3"  "2"  "8"  "7"  "5"  "4"  "2"  "2"  "7"  "6"  "8"  "6" 
## [136] "3"  "8"  "7"  "6"  "4"  "3"  "2"  "8"  "6"  "3"  "4"  "3"  "6"  "3"  "4" 
## [151] "3"  "6"  "6"  "7"  "11" "6"  "21" "6"  "7"  "5"  "4"  "8"  "9"  "7"  "2" 
## [166] "7"  "6"  "5"  "10" "9"  "9"  "7"  "12" "30" "11" "14" "8"  "4"  "12" "10"
## [181] "17" "14" "22" "22" "20" "2"  "30" "15" "19" "12" "19" "25" "21" "16" "26"
## [196] "25" "25" "6"  "4"  "27" "25" "9"  "31" "29" "28" "17" "25" "24" "23" "4" 
## [211] "27" "19" "27" "27" "20" "29" "28" "20" "2"  "31" "23" "28" "24" "27" "25"
## [226] "23"
```

(e) the number of sentences in the speech

```
x <- grep("^\\*{3}", SoTU)
list.speeches <- list()
for(i in 1:length(x)){
  if(i == 1){
    list.speeches[[i]] <- paste(SoTU[1:x[1]], collapse = " ")
  }
  else{
  list.speeches[[i]] <- paste(SoTU[x[i-1]:x[i]], collapse = " ")}
}
```

The following list shows the number of sentences in each speech:

```
num.sentences <- c()
for(speech in list.speeches){
  num <- length(gregexpr('[[:alnum:] ][.!?]', speech)[[1]])
  num.sentences <- append(num.sentences, num)
}

# Exclude first item in list.speeches , it is not a speech
num.sentences.final <- num.sentences[2:227]
num.sentences.final
```

```
##   [1]   24   39   59   62   54   76   52   81   59   58   36   40   90   63   50
##  [16]   52   80   78   64   75   40   64   47   89   71   50   62   72  130  118
##  [31]  134   88  168  129  204  257  214  178  168  203  308  394  168  200  186
##  [46]  318  240  294  285  266  344  184  212  212  196  271  465  497  455  625
##  [61]  220  242  380  289  233  268  275  255  422  552  438  506  225  359  226
##  [76]  237  275  204  385  302  253  298  217  123  303  280  363  205  238  272
##  [91]  349  207  158  138  196  315  587  446  129  260  458  385  492  467  432
## [106]  553  368  476  387  655  473  656  684  338  585  540  805  675  893  639
## [121]  392  312  752  763  109  174  227   53  138  163  162   73  210  226  365
## [136]  278  484  447  452  380  409  163  251   61   80  143  174  110  176  174
## [151]  124  151  179  204  172  339 1370  297  287  188  240  246  298  432  354
## [166]  289  339  380  188  257  266  258  279  216  278  247  143  257  251  371
## [181]  266  216  203  198  174   60  211  230  279  219  251  161  167 1570  269
## [196]  274  306  237  189  219  241  299  243  250  301  297  377  461  367  358
## [211]  371  406  421  290  195  217  310  282  243  280  295  314  286  417  417
## [226]  426
```

f) the number of words in the speech

The following list shows the number of words in each speech:

```
num.words <- c()
for(speech in list.speeches){
  words <- sapply(gregexpr("\\S+", speech), length)
  num.words <- append(num.words, words)
}
# Exclude first item in list.speeches , it is not a speech
num.words.final <- num.words[2:227]
num.words.final
```

```
##   [1]  1095  1415  2315  2110  1977  2927  1998  2879  2069  2230  1517  1384
##  [13]  3236  2213  2291  2113  2955  2880  2406  2700  1843  2458  2287  3254
##  [25]  3270  2126  3166  3379  4439  4389  4723  3453  5835  4746  6394  8430
##  [37]  9016  7761  7000  7326 10535 15084  7201  7887  7913 13458 10821 12367
##  [49] 11451 11491 13432  8992  8252  8413  8038  9321 16124 18233 16423 21302
##  [61]  7631  8330 13255  9934  9599 10146 11622 10485 13663 16359 12348 14037
##  [73]  6988  8404  6123  6001  9230  7139 11993  9829  7715  8749  6471  4002
##  [85] 10035  9207 12216  6806  8030  7894 11645  6710  3835  3080  3797  8960
##  [97] 19758 15147  5301  9032 13013 11534 16304 13690 12295 15904 14683 15455
## [109] 12122 20220 15145 19146 19607  9768 14888 17418 25045 23587 27394 19395
## [121] 13902  6776 23717 25163  3564  4545  7698  2129  3925  5477  4767  2717
## [133]  5616  5756  6713  6976 10855 10320  8789  8072 11005  4558  5698   992
## [145]  2236  3528  3823  2743  4690  3749  3183  3301  3483  4546  3768  8137
## [157] 27735  6046  5104  3410  5139  4001  5347  9612  6954  5987  7247  8258
## [169]  4142  4924  4882  5641  6201  5178  6449  5339  3200  4407  5268  7124
## [181]  4868  4120  4467  4487  3979  1667  5166  4116  4964  4653  4592  3255
## [193]  3408 33576  5207  5584  4985  4262  3519  3815  4867  4833  3783  3778
## [205]  4731  7019  7413  9156  6296  6755  7296  7485  7400  4409  2979  3837
## [217]  5385  5184  5062  5334  5611  5765  5953  7118  6952  7004
```