CPP 527: Code-Through Assignment

Joanna Garcia Arellano

2/29/2020

Sentiment Analysis

This code-through will take the text from a single article to conduct a sentiment analysis.

The article used for this code through can be located through the link below:

https://www.washingtonpost.com/world/2020/02/29/coronavirus-live-updates/ (https://www.washingtonpost.com/world/2020/02/29/coronavirus-live-updates/)

Install packages required

For this simple sentiment analysis, we will be using the text mining packages “tm” and "stringr

```
#install text mining package
library(tm)
```

```
## Loading required package: NLP
```

```
#install packages
library(stringr)
```

Step 1:

Locate the article/ text you to analyze. You will need to save the plain text of this article/ text into a simple .txt file in your working directory.

You can confirm the file path for your working directory using the ‘get working directory’ function ’getwd().

```
#save the text only of your selected article into an .txt file
#save file into your working directory
#you can check file path for working directory by using getwd()
 getwd()
```

```
## [1] "/Users/joannagarciaarellano/Documents/R"
```

Step 2:

Once your plain text document has been saved, you can utilize the readLines() function to read the text from your selected document into R.

This function will read the text from the document line by line, we will change this in the next step.

```
#use function readLines() to read text from your selected document
#ensure that your document is saved as plan text- no rich text!
#this function will read the text from the document line by line

readLines("article1.txt")
```

```
##  [1] "Live updates: First U.S. death confirmed; travel restrictions announced affecting Iran, Italy and South
Korea in response to coronavirus"
##  [2] "Standing with President Trump on Feb. 29, Vice President Pence announced new travel restrictions affect
ing Iran, Italy and South Korea. (Reuters)"
##  [3] ""
##  [4] "By "
##  [5] "Gerry Shih, "
##  [6] "Marisa Iati, "
##  [7] "Derek Hawkins, "
##  [8] "Katie Mettler and "
##  [9] "Miriam Berger "
## [10] "Feb. 29, 2020 at 10:28 p.m. EST"
## [11] ""
## [12] "The Trump administration Saturday announced additional travel restrictions affecting Iran, Italy and So
uth Korea in response to the coronavirus outbreak following the first death from the virus in the United States.
"
## [13] "Vice President Pence said the existing travel ban on Iran would extend to foreign nationals who had bee
n in that country the past 14 days. The State Department also is increasing its warning advising Americans not t
o travel to parts of Italy and South Korea affected by the virus."
## [14] "Right before the White House’s news conference, health officials in Washington state confirmed that a p
erson diagnosed with coronavirus in King County has died."
## [15] ""
## [16] "President Trump described the patient as a “wonderful woman” and a “medically high-risk patient” in her
late 50s, at the news conference. The Centers for Disease Control and Prevention later said it had mistakenly de
scribed the patient’s gender in a briefing to Trump and Vice President Pence, and local health officials clarifi
ed the deceased was a man with underlying conditions."
## [17] ""
## [18] "More coronavirus infections were reported from South Korea to France to Qatar on Saturday as health off
icials in Washington state, Oregon and California reported another worrying development: new cases among people
who have not traveled recently to countries hit hard by the outbreak or come into contact with anyone known to h
ave the disease, which public health officials refer to as community transmission."
## [19] "Washington state on Saturday announced three new cases of the virus — including the person who died — w
ith circumstances that suggest person-to-person spread in the community. The other patients were a health-care w
orker at a long-term nursing facility and a female resident in her 70s from the same center."
## [20] ""
## [21] "California has reported three cases of community transmission, two of which are in Santa Clara County a
nd one of which is in Solano County. Illinois reported a third case Saturday but did not say how the person may
have gotten the virus."
## [22] "Here are the latest developments:"
## [23] "\t•\tThe five new cases announced Saturday bring the number of infections in the United States to 24, e
xcluding repatriations, according to the CDC. Forty-seven other people who have been repatriated to the United S
tates from Wuhan, China, and from the Diamond Princess cruise ship also have the virus."
## [24] "\t•\tThe new cases in Washington state included the first possible outbreak in a long-term nursing faci
lity. Health officials have said older people and adults in poor health face the highest risk from the virus."
## [25] "\t•\tThe Food and Drug Administration expanded coronavirus testing by speeding up hospitals’ abilities
to test, though some worried the changes fell short in reducing logistical burdens."
## [26] "\t•\tItaly became the third country, in addition to China and South Korea, to confirm more than 1,000 c
ases of the virus. Of the 1,049 patients, 401 are hospitalized, and the rest are self-isolated at home."
## [27] "Mapping the spread of the coronavirus | What you need to know about the virus | How to prepare for coro
navirus in the U.S. | Post Reports: Your questions about coronavirus, answered"
```

Step 3:

Since the readLines function treats each line as a separate element in a character vector, we collapse the text into one single character vector element.

Note that you can always review the structure of your file using the structure function ‘str()’.

```
#readLines treats each line as separate element in character vector
#you can review structure of the file using structure function str(readLines("article1.txt))
#we need to treat text as one element, can do so using paste function and collapse argument

paste(readLines("article1.txt"), collapse = " ")
```

```
## [1] "Live updates: First U.S. death confirmed; travel restrictions announced affecting Iran, Italy and South
Korea in response to coronavirus Standing with President Trump on Feb. 29, Vice President Pence announced new tr
avel restrictions affecting Iran, Italy and South Korea. (Reuters)  By  Gerry Shih,  Marisa Iati,  Derek Hawkins
,  Katie Mettler and  Miriam Berger  Feb. 29, 2020 at 10:28 p.m. EST  The Trump administration Saturday announce
d additional travel restrictions affecting Iran, Italy and South Korea in response to the coronavirus outbreak f
ollowing the first death from the virus in the United States. Vice President Pence said the existing travel ban
on Iran would extend to foreign nationals who had been in that country the past 14 days. The State Department al
so is increasing its warning advising Americans not to travel to parts of Italy and South Korea affected by the
virus. Right before the White House's news conference, health officials in Washington state confirmed that a per
son diagnosed with coronavirus in King County has died.  President Trump described the patient as a "wonderful w
oman" and a "medically high-risk patient" in her late 50s, at the news conference. The Centers for Disease Contr
ol and Prevention later said it had mistakenly described the patient's gender in a briefing to Trump and Vice Pr
esident Pence, and local health officials clarified the deceased was a man with underlying conditions.  More cor
onavirus infections were reported from South Korea to France to Qatar on Saturday as health officials in Washing
ton state, Oregon and California reported another worrying development: new cases among people who have not trav
eled recently to countries hit hard by the outbreak or come into contact with anyone known to have the disease,
which public health officials refer to as community transmission. Washington state on Saturday announced three n
ew cases of the virus — including the person who died — with circumstances that suggest person-to-person spread
in the community. The other patients were a health-care worker at a long-term nursing facility and a female resi
dent in her 70s from the same center.  California has reported three cases of community transmission, two of whi
ch are in Santa Clara County and one of which is in Solano County. Illinois reported a third case Saturday but d
id not say how the person may have gotten the virus. Here are the latest developments: \t•\tThe five new cases a
nnounced Saturday bring the number of infections in the United States to 24, excluding repatriations, according
to the CDC. Forty-seven other people who have been repatriated to the United States from Wuhan, China, and from
the Diamond Princess cruise ship also have the virus. \t•\tThe new cases in Washington state included the first
possible outbreak in a long-term nursing facility. Health officials have said older people and adults in poor he
alth face the highest risk from the virus. \t•\tThe Food and Drug Administration expanded coronavirus testing by
speeding up hospitals' abilities to test, though some worried the changes fell short in reducing logistical burd
ens. \t•\tItaly became the third country, in addition to China and South Korea, to confirm more than 1,000 cases
of the virus. Of the 1,049 patients, 401 are hospitalized, and the rest are self-isolated at home. Mapping the s
pread of the coronavirus | What you need to know about the virus | How to prepare for coronavirus in the U.S. |
Post Reports: Your questions about coronavirus, answered"
```

```
article1<- paste(readLines("article1.txt"), collapse = " ")
```

Step 4:

We will now work on cleaning up the text.

We will begin by removing any special characters using the 'gsub' function.

```
#begin cleaning up text
#remove punctuation and any special characters that are not needed for analysis using gsub function
#\\W refers to anything that is not a word
#we are replacing any of these characters in our text with a blank space

gsub(pattern="\\W", replace=" ", article1)
```

```
## [1] "Live updates  First U S  death confirmed  travel restrictions announced affecting Iran  Italy and South Korea in response to coronavirus Standing with President Trump on Feb  29  Vice President Pence announced new tr avel restrictions affecting Iran  Italy and South Korea  Reuters  By  Gerry Shih  Marisa Iati  Derek Hawkins Katie Mettler and  Miriam Berger  Feb  29  2020 at 10 28 p m  EST  The Trump administration Saturday announced a dditional travel restrictions affecting Iran  Italy and South Korea in response to the coronavirus outbreak foll owing the first death from the virus in the United States  Vice President Pence said the existing travel ban on Iran would extend to foreign nationals who had been in that country the past 14 days  The State Department also is increasing its warning advising Americans not to travel to parts of Italy and South Korea affected by the vir us  Right before the White House s news conference  health officials in Washington state confirmed that a person diagnosed with coronavirus in King County has died  President Trump described the patient as a  wonderful woman and a  medically high risk patient  in her late 50s  at the news conference  The Centers for Disease Control and Prevention later said it had mistakenly described the patient s gender in a briefing to Trump and Vice President Pence  and local health officials clarified the deceased was a man with underlying conditions  More coronavirus infections were reported from South Korea to France to Qatar on Saturday as health officials in Washington state Oregon and California reported another worrying development  new cases among people who have not traveled recent ly to countries hit hard by the outbreak or come into contact with anyone known to have the disease  which publi c health officials refer to as community transmission  Washington state on Saturday announced three new cases of the virus  including the person who died  with circumstances that suggest person to person spread in the commu nity  The other patients were a health care worker at a long term nursing facility and a female resident in her 70s from the same center  California has reported three cases of community transmission  two of which are in Sa nta Clara County and one of which is in Solano County  Illinois reported a third case Saturday but did not say h ow the person may have gotten the virus  Here are the latest developments  The five new cases announced Satur day bring the number of infections in the United States to 24  excluding repatriations  according to the CDC  Fo rty seven other people who have been repatriated to the United States from Wuhan  China  and from the Diamond Pr incess cruise ship also have the virus  The new cases in Washington state included the first possible outbrea k in a long term nursing facility  Health officials have said older people and adults in poor health face the hi ghest risk from the virus  The Food and Drug Administration expanded coronavirus testing by speeding up hospi tals  abilities to test  though some worried the changes fell short in reducing logistical burdens  Italy bec ame the third country  in addition to China and South Korea  to confirm more than 1 000 cases of the virus  Of t he 1 049 patients  401 are hospitalized  and the rest are self isolated at home  Mapping the spread of the coron avirus  What you need to know about the virus  How to prepare for coronavirus in the U S  Post Reports  Your questions about coronavirus  answered"
```

```
article1_2<- gsub(pattern="\\W", replace=" ", article1)
```

Next, we will remove any digits using 'gsub' function

```
#\\d refers to any digits, 0-9
#again, we will replace any digit characters with a blank space
gsub(pattern = "\\d", replace= " ", article1_2)
```

```
## [1] "Live updates  First U S  death confirmed  travel restrictions announced affecting Iran  Italy and South
Korea in response to coronavirus Standing with President Trump on Feb      Vice President Pence announced new tr
avel restrictions affecting Iran  Italy and South Korea   Reuters   By  Gerry Shih   Marisa Iati   Derek Hawkins
Katie Mettler and  Miriam Berger  Feb            at       p m  EST  The Trump administration Saturday announced a
dditional travel restrictions affecting Iran  Italy and South Korea in response to the coronavirus outbreak foll
owing the first death from the virus in the United States  Vice President Pence said the existing travel ban on
Iran would extend to foreign nationals who had been in that country the past    days  The State Department also
is increasing its warning advising Americans not to travel to parts of Italy and South Korea affected by the vir
us  Right before the White House s news conference  health officials in Washington state confirmed that a person
diagnosed with coronavirus in King County has died   President Trump described the patient as a  wonderful woman
and a  medically high risk patient  in her late   s  at the news conference  The Centers for Disease Control and
Prevention later said it had mistakenly described the patient s gender in a briefing to Trump and Vice President
Pence  and local health officials clarified the deceased was a man with underlying conditions   More coronavirus
infections were reported from South Korea to France to Qatar on Saturday as health officials in Washington state
Oregon and California reported another worrying development  new cases among people who have not traveled recent
ly to countries hit hard by the outbreak or come into contact with anyone known to have the disease  which publi
c health officials refer to as community transmission  Washington state on Saturday announced three new cases of
the virus   including the person who died   with circumstances that suggest person to person spread in the commu
nity  The other patients were a health care worker at a long term nursing facility and a female resident in her
s from the same center   California has reported three cases of community transmission  two of which are in Sant
a Clara County and one of which is in Solano County  Illinois reported a third case Saturday but did not say how
the person may have gotten the virus  Here are the latest developments      The five new cases announced Saturday
bring the number of infections in the United States to     excluding repatriations  according to the CDC  Forty
seven other people who have been repatriated to the United States from Wuhan  China  and from the Diamond Prince
ss cruise ship also have the virus     The new cases in Washington state included the first possible outbreak in
a long term nursing facility  Health officials have said older people and adults in poor health face the highest
risk from the virus     The Food and Drug Administration expanded coronavirus testing by speeding up hospitals
abilities to test  though some worried the changes fell short in reducing logistical burdens     Italy became th
e third country  in addition to China and South Korea  to confirm more than       cases of the virus  Of the
patients      are hospitalized  and the rest are self isolated at home  Mapping the spread of the coronavirus
What you need to know about the virus   How to prepare for coronavirus in the U S    Post Reports  Your question
s about coronavirus  answered"
```

```
article1_3<-gsub(pattern = "\\d", replace= " ", article1_2)
```

Use the ‘tolower’ function to make all letters lower case

```
#use 'tolower' function to make all letters lower case
tolower(article1_3)
```

```
## [1] "live updates  first u s  death confirmed  travel restrictions announced affecting iran  italy and south
korea in response to coronavirus standing with president trump on feb      vice president pence announced new tr
avel restrictions affecting iran  italy and south korea  reuters  by  gerry shih  marisa iati  derek hawkins
katie mettler and  miriam berger  feb           at      p m  est  the trump administration saturday announced a
dditional travel restrictions affecting iran  italy and south korea in response to the coronavirus outbreak foll
owing the first death from the virus in the united states  vice president pence said the existing travel ban on
iran would extend to foreign nationals who had been in that country the past    days  the state department also
is increasing its warning advising americans not to travel to parts of italy and south korea affected by the vir
us  right before the white house s news conference  health officials in washington state confirmed that a person
diagnosed with coronavirus in king county has died  president trump described the patient as a  wonderful woman
and a  medically high risk patient  in her late   s  at the news conference  the centers for disease control and
prevention later said it had mistakenly described the patient s gender in a briefing to trump and vice president
pence  and local health officials clarified the deceased was a man with underlying conditions  more coronavirus
infections were reported from south korea to france to qatar on saturday as health officials in washington state
oregon and california reported another worrying development  new cases among people who have not traveled recent
ly to countries hit hard by the outbreak or come into contact with anyone known to have the disease  which publi
c health officials refer to as community transmission  washington state on saturday announced three new cases of
the virus  including the person who died  with circumstances that suggest person to person spread in the commu
nity  the other patients were a health care worker at a long term nursing facility and a female resident in her
s from the same center  california has reported three cases of community transmission  two of which are in sant
a clara county and one of which is in solano county  illinois reported a third case saturday but did not say how
the person may have gotten the virus  here are the latest developments     the five new cases announced saturday
bring the number of infections in the united states to     excluding repatriations  according to the cdc  forty
seven other people who have been repatriated to the united states from wuhan  china  and from the diamond prince
ss cruise ship also have the virus     the new cases in washington state included the first possible outbreak in
a long term nursing facility  health officials have said older people and adults in poor health face the highest
risk from the virus     the food and drug administration expanded coronavirus testing by speeding up hospitals
abilities to test  though some worried the changes fell short in reducing logistical burdens     italy became th
e third country  in addition to china and south korea  to confirm more than       cases of the virus  of the
patients      are hospitalized  and the rest are self isolated at home  mapping the spread of the coronavirus
what you need to know about the virus  how to prepare for coronavirus in the u s  post reports  your question
s about coronavirus  answered"
```

```
article1_4<- tolower(article1_3)
```

Remove filler words easily by using the ‘removeWords’ function

```
#remove filler words like the, and, a, to, or using the stopwords function
removeWords(article1_4, stopwords())
```

```
## [1] "live updates  first u s  death confirmed  travel restrictions announced affecting iran  italy  south kor
ea  response  coronavirus standing  president trump  feb       vice president pence announced new travel restrict
ions affecting iran  italy  south korea   reuters      gerry shih   marisa iati   derek hawkins   katie mettler
miriam berger  feb                    p m  est   trump administration saturday announced additional travel restric
tions affecting iran  italy  south korea  response   coronavirus outbreak following  first death   virus   unite
d states  vice president pence said  existing travel ban  iran  extend  foreign nationals       country  past
days   state department also  increasing  warning advising americans   travel  parts  italy  south korea affecte
d   virus  right   white house s news conference  health officials  washington state confirmed   person diagnose
d  coronavirus  king county  died   president trump described  patient    wonderful woman      medically high ris
k patient    late   s    news conference   centers  disease control  prevention later said   mistakenly describe
d  patient s gender   briefing  trump  vice president pence   local health officials clarified  deceased   man
underlying conditions    coronavirus infections  reported  south korea  france  qatar  saturday  health official
s  washington state  oregon  california reported another worrying development  new cases among people    travele
d recently  countries hit hard   outbreak  come  contact  anyone known    disease   public health officials refe
r   community transmission  washington state  saturday announced three new cases   virus   including  person  di
ed    circumstances  suggest person  person spread   community    patients   health care worker   long term nurs
ing facility   female resident      s    center   california  reported three cases  community transmission  two
santa clara county  one      solano county  illinois reported  third case saturday    say   person may  gotten  v
irus     latest developments      five new cases announced saturday bring  number  infections   united states
excluding repatriations  according   cdc  forty seven  people    repatriated  united states  wuhan  china     d
iamond princess cruise ship also   virus      new cases  washington state included  first possible outbreak   lo
ng term nursing facility  health officials  said older people  adults  poor health face  highest risk   virus
food  drug administration expanded coronavirus testing  speeding  hospitals  abilities  test  though  worried  c
hanges fell short  reducing logistical burdens     italy became  third country   addition  china  south korea
confirm        cases   virus          patients       hospitalized    rest  self isolated  home  mapping  spread
coronavirus     need  know   virus     prepare  coronavirus   u s    post reports   questions  coronavirus  answ
ered"
```

```
article1_5<- removeWords(article1_4, stopwords())
```

Remove any single letter characters remaining in the text

```
#remove single letter words remaining in text
gsub(pattern ="\\b[A-z]\\b{1}", replace=" ", article1_5)
```

```
## [1] "live updates  first      death confirmed  travel restrictions announced affecting iran  italy  south kor
ea  response  coronavirus standing  president trump  feb       vice president pence announced new travel restrict
ions affecting iran  italy  south korea   reuters      gerry shih   marisa iati   derek hawkins   katie mettler
miriam berger  feb                      est   trump administration saturday announced additional travel restric
tions affecting iran  italy  south korea  response   coronavirus outbreak following  first death   virus   unite
d states  vice president pence said  existing travel ban  iran  extend  foreign nationals       country  past
days   state department also  increasing  warning advising americans   travel  parts  italy  south korea affecte
d   virus  right   white house   news conference  health officials  washington state confirmed   person diagnose
d  coronavirus  king county  died   president trump described  patient    wonderful woman      medically high ris
k patient    late          news conference   centers  disease control  prevention later said   mistakenly describe
d  patient   gender   briefing  trump  vice president pence   local health officials clarified  deceased   man
underlying conditions    coronavirus infections  reported  south korea  france  qatar  saturday  health official
s  washington state  oregon  california reported another worrying development  new cases among people    travele
d recently  countries hit hard   outbreak  come  contact  anyone known    disease   public health officials refe
r   community transmission  washington state  saturday announced three new cases   virus   including  person  di
ed    circumstances  suggest person  person spread   community    patients   health care worker   long term nurs
ing facility   female resident           center   california  reported three cases  community transmission  two
santa clara county  one      solano county  illinois reported  third case saturday    say   person may  gotten  v
irus     latest developments      five new cases announced saturday bring  number  infections   united states
excluding repatriations  according   cdc  forty seven  people    repatriated  united states  wuhan  china     d
iamond princess cruise ship also   virus      new cases  washington state included  first possible outbreak   lo
ng term nursing facility  health officials  said older people  adults  poor health face  highest risk   virus
food  drug administration expanded coronavirus testing  speeding  hospitals  abilities  test  though  worried  c
hanges fell short  reducing logistical burdens     italy became  third country   addition  china  south korea
confirm        cases   virus          patients       hospitalized    rest  self isolated  home  mapping  spread
coronavirus     need  know   virus     prepare  coronavirus         post reports   questions  coronavirus  answ
ered"
```

```
article1_6<-gsub(pattern ="\\b[A-z]\\b{1}", replace=" ", article1_5)
```

Clean up any extra blanks that we have created using the stripWhitespace function

```
#clean up any extra blanks that have been created using "stripWhitespace" function
stripWhitespace(article1_6)
```

```
## [1] "live updates first death confirmed travel restrictions announced affecting iran italy south korea respon
se coronavirus standing president trump feb vice president pence announced new travel restrictions affecting ira
n italy south korea reuters gerry shih marisa iati derek hawkins katie mettler miriam berger feb est trump admin
istration saturday announced additional travel restrictions affecting iran italy south korea response coronaviru
s outbreak following first death virus united states vice president pence said existing travel ban iran extend f
oreign nationals country past days state department also increasing warning advising americans travel parts ital
y south korea affected virus right white house news conference health officials washington state confirmed perso
n diagnosed coronavirus king county died president trump described patient wonderful woman medically high risk p
atient late news conference centers disease control prevention later said mistakenly described patient gender br
iefing trump vice president pence local health officials clarified deceased man underlying conditions coronaviru
s infections reported south korea france qatar saturday health officials washington state oregon california repo
rted another worrying development new cases among people traveled recently countries hit hard outbreak come cont
act anyone known disease public health officials refer community transmission washington state saturday announce
d three new cases virus including person died circumstances suggest person person spread community patients heal
th care worker long term nursing facility female resident center california reported three cases community trans
mission two santa clara county one solano county illinois reported third case saturday say person may gotten vir
us latest developments five new cases announced saturday bring number infections united states excluding repatri
ations according cdc forty seven people repatriated united states wuhan china diamond princess cruise ship also
virus new cases washington state included first possible outbreak long term nursing facility health officials sa
id older people adults poor health face highest risk virus food drug administration expanded coronavirus testing
speeding hospitals abilities test though worried changes fell short reducing logistical burdens italy became thi
rd country addition china south korea confirm cases virus patients hospitalized rest self isolated home mapping
spread coronavirus need know virus prepare coronavirus post reports questions coronavirus answered"
```

```
article1_7<- stripWhitespace(article1_6)
```

Step 5:

Now that our text data has been cleaned up, we will use the string split function to split up all the words in our text.

Then, we must ensure our vectors are character outputs, not lists.

Note: you can use the class function 'class()' to check the type of output vector you are receiving.

```
#use the string split function to split up all the words in our text
#"\\s+" refers to any number of spaces

str_split(article1_7, pattern="\\s+")
```

```
## [[1]]
##   [1] "live"          "updates"       "first"         "death"
##   [5] "confirmed"     "travel"        "restrictions"  "announced"
##   [9] "affecting"     "iran"          "italy"         "south"
##  [13] "korea"         "response"      "coronavirus"   "standing"
##  [17] "president"     "trump"         "feb"           "vice"
##  [21] "president"     "pence"         "announced"     "new"
##  [25] "travel"        "restrictions"  "affecting"     "iran"
##  [29] "italy"         "south"         "korea"         "reuters"
##  [33] "gerry"         "shih"          "marisa"        "iati"
##  [37] "derek"         "hawkins"       "katie"         "mettler"
##  [41] "miriam"        "berger"        "feb"           "est"
##  [45] "trump"         "administration" "saturday"      "announced"
##  [49] "additional"    "travel"        "restrictions"  "affecting"
##  [53] "iran"          "italy"         "south"         "korea"
##  [57] "response"      "coronavirus"   "outbreak"      "following"
##  [61] "first"         "death"         "virus"         "united"
##  [65] "states"        "vice"          "president"     "pence"
##  [69] "said"          "existing"      "travel"        "ban"
```

```
##  [73] "iran"          "extend"        "foreign"       "nationals"
##  [77] "country"       "past"          "days"          "state"
##  [81] "department"    "also"          "increasing"    "warning"
##  [85] "advising"      "americans"     "travel"        "parts"
##  [89] "italy"         "south"         "korea"         "affected"
##  [93] "virus"         "right"         "white"         "house"
##  [97] "news"          "conference"    "health"        "officials"
## [101] "washington"    "state"         "confirmed"     "person"
## [105] "diagnosed"     "coronavirus"   "king"          "county"
## [109] "died"          "president"     "trump"         "described"
## [113] "patient"       "wonderful"     "woman"         "medically"
## [117] "high"          "risk"          "patient"       "late"
## [121] "news"          "conference"    "centers"       "disease"
## [125] "control"       "prevention"    "later"         "said"
## [129] "mistakenly"    "described"     "patient"       "gender"
## [133] "briefing"      "trump"         "vice"          "president"
## [137] "pence"         "local"         "health"        "officials"
## [141] "clarified"     "deceased"      "man"           "underlying"
## [145] "conditions"    "coronavirus"   "infections"    "reported"
## [149] "south"         "korea"         "france"        "qatar"
## [153] "saturday"      "health"        "officials"     "washington"
## [157] "state"         "oregon"        "california"    "reported"
## [161] "another"       "worrying"      "development"   "new"
## [165] "cases"         "among"         "people"        "traveled"
## [169] "recently"      "countries"     "hit"           "hard"
## [173] "outbreak"      "come"          "contact"       "anyone"
## [177] "known"         "disease"       "public"        "health"
## [181] "officials"     "refer"         "community"     "transmission"
## [185] "washington"    "state"         "saturday"      "announced"
## [189] "three"         "new"           "cases"         "virus"
## [193] "including"     "person"        "died"          "circumstances"
## [197] "suggest"       "person"        "person"        "spread"
## [201] "community"     "patients"      "health"        "care"
## [205] "worker"        "long"          "term"          "nursing"
## [209] "facility"      "female"        "resident"      "center"
## [213] "california"    "reported"      "three"         "cases"
## [217] "community"     "transmission"  "two"           "santa"
## [221] "clara"         "county"        "one"           "solano"
## [225] "county"        "illinois"      "reported"      "third"
## [229] "case"          "saturday"      "say"           "person"
## [233] "may"           "gotten"        "virus"         "latest"
## [237] "developments"  "five"          "new"           "cases"
## [241] "announced"     "saturday"      "bring"         "number"
## [245] "infections"    "united"        "states"        "excluding"
## [249] "repatriations" "according"     "cdc"           "forty"
## [253] "seven"         "people"        "repatriated"   "united"
## [257] "states"        "wuhan"         "china"         "diamond"
## [261] "princess"      "cruise"        "ship"          "also"
## [265] "virus"         "new"           "cases"         "washington"
## [269] "state"         "included"      "first"         "possible"
## [273] "outbreak"      "long"          "term"          "nursing"
## [277] "facility"      "health"        "officials"     "said"
## [281] "older"         "people"        "adults"        "poor"
## [285] "health"        "face"          "highest"       "risk"
## [289] "virus"         "food"          "drug"          "administration"
## [293] "expanded"      "coronavirus"   "testing"       "speeding"
## [297] "hospitals"     "abilities"     "test"          "though"
## [301] "worried"       "changes"       "fell"          "short"
## [305] "reducing"      "logistical"    "burdens"       "italy"
## [309] "became"        "third"         "country"       "addition"
## [313] "china"         "south"         "korea"         "confirm"
## [317] "cases"         "virus"         "patients"      "hospitalized"
## [321] "rest"          "self"          "isolated"      "home"
## [325] "mapping"       "spread"        "coronavirus"   "need"
## [329] "know"          "virus"         "prepare"       "coronavirus"
## [333] "post"          "reports"       "questions"     "coronavirus"
## [337] "answered"
```

```
text<- str_split(article1_7, pattern ="\\s+")

#use the unlist function to make sure our outputs are character vectors rather than a list
unlist(text)
```

```
##   [1] "live"          "updates"       "first"         "death"
##   [5] "confirmed"     "travel"        "restrictions"  "announced"
##   [9] "affecting"     "iran"          "italy"         "south"
##  [13] "korea"         "response"      "coronavirus"   "standing"
##  [17] "president"     "trump"         "feb"           "vice"
##  [21] "president"     "pence"         "announced"     "new"
##  [25] "travel"        "restrictions"  "affecting"     "iran"
##  [29] "italy"         "south"         "korea"         "reuters"
##  [33] "gerry"         "shih"          "marisa"        "iati"
##  [37] "derek"         "hawkins"       "katie"         "mettler"
##  [41] "miriam"        "berger"        "feb"           "est"
##  [45] "trump"         "administration" "saturday"     "announced"
##  [49] "additional"    "travel"        "restrictions"  "affecting"
##  [53] "iran"          "italy"         "south"         "korea"
##  [57] "response"      "coronavirus"   "outbreak"      "following"
##  [61] "first"         "death"         "virus"         "united"
##  [65] "states"        "vice"          "president"     "pence"
##  [69] "said"          "existing"      "travel"        "ban"
##  [73] "iran"          "extend"        "foreign"       "nationals"
##  [77] "country"       "past"          "days"          "state"
##  [81] "department"    "also"          "increasing"    "warning"
##  [85] "advising"      "americans"     "travel"        "parts"
##  [89] "italy"         "south"         "korea"         "affected"
##  [93] "virus"         "right"         "white"         "house"
##  [97] "news"          "conference"    "health"        "officials"
## [101] "washington"    "state"         "confirmed"     "person"
## [105] "diagnosed"     "coronavirus"   "king"          "county"
## [109] "died"          "president"     "trump"         "described"
## [113] "patient"       "wonderful"     "woman"         "medically"
## [117] "high"          "risk"          "patient"       "late"
## [121] "news"          "conference"    "centers"       "disease"
## [125] "control"       "prevention"    "later"         "said"
## [129] "mistakenly"    "described"     "patient"       "gender"
## [133] "briefing"      "trump"         "vice"          "president"
## [137] "pence"         "local"         "health"        "officials"
## [141] "clarified"     "deceased"      "man"           "underlying"
## [145] "conditions"    "coronavirus"   "infections"    "reported"
## [149] "south"         "korea"         "france"        "qatar"
## [153] "saturday"      "health"        "officials"     "washington"
## [157] "state"         "oregon"        "california"    "reported"
## [161] "another"       "worrying"      "development"   "new"
## [165] "cases"         "among"         "people"        "traveled"
## [169] "recently"      "countries"     "hit"           "hard"
## [173] "outbreak"      "come"          "contact"       "anyone"
## [177] "known"         "disease"       "public"        "health"
## [181] "officials"     "refer"         "community"     "transmission"
## [185] "washington"    "state"         "saturday"      "announced"
## [189] "three"         "new"           "cases"         "virus"
## [193] "including"     "person"        "died"          "circumstances"
## [197] "suggest"       "person"        "person"        "spread"
## [201] "community"     "patients"      "health"        "care"
## [205] "worker"        "long"          "term"          "nursing"
## [209] "facility"      "female"        "resident"      "center"
## [213] "california"    "reported"      "three"         "cases"
## [217] "community"     "transmission"  "two"           "santa"
## [221] "clara"         "county"        "one"           "solano"
## [225] "county"        "illinois"      "reported"      "third"
## [229] "case"          "saturday"      "say"           "person"
## [233] "may"           "gotten"        "virus"         "latest"
## [237] "developments"  "five"          "new"           "cases"
```

```
## [241] "announced"      "saturday"      "bring"         "number"
## [245] "infections"     "united"        "states"        "excluding"
## [249] "repatriations"  "according"     "cdc"           "forty"
## [253] "seven"          "people"        "repatriated"   "united"
## [257] "states"         "wuhan"         "china"         "diamond"
## [261] "princess"       "cruise"        "ship"          "also"
## [265] "virus"          "new"           "cases"         "washington"
## [269] "state"          "included"      "first"         "possible"
## [273] "outbreak"       "long"          "term"          "nursing"
## [277] "facility"       "health"        "officials"     "said"
## [281] "older"          "people"        "adults"        "poor"
## [285] "health"         "face"          "highest"       "risk"
## [289] "virus"          "food"          "drug"          "administration"
## [293] "expanded"       "coronavirus"   "testing"       "speeding"
## [297] "hospitals"      "abilities"     "test"          "though"
## [301] "worried"        "changes"       "fell"          "short"
## [305] "reducing"       "logistical"    "burdens"       "italy"
## [309] "became"         "third"         "country"       "addition"
## [313] "china"          "south"         "korea"         "confirm"
## [317] "cases"          "virus"         "patients"      "hospitalized"
## [321] "rest"           "self"          "isolated"      "home"
## [325] "mapping"        "spread"        "coronavirus"   "need"
## [329] "know"           "virus"         "prepare"       "coronavirus"
## [333] "post"           "reports"       "questions"     "coronavirus"
## [337] "answered"
```

```
text<- unlist(text)

#output will look almost identical but you can check type of vector using class function

class(text)
```

```
## [1] "character"
```

Step 6:

We can now begin the process of sentiment analysis.

First, we will scan in our text documents containing the positive and negative words that will need to be saved in our working directory as plain .txt docs.

The link for both lists can be located below:

[http://ptrckprry.com/course/ssd/data/positive-words.txt (http://ptrckprry.com/course/ssd/data/positive-words.txt)]
[http://ptrckprry.com/course/ssd/data/negative-words.txt (http://ptrckprry.com/course/ssd/data/negative-words.txt)]

Note: this is not the only way to conduct sentiment analysis. In fact, the tidytext package already contains several sentiment lexicons to access. However, for the purposes of this simplistic code-through, we will make this as simple as possible.

```
#save lexicon of positive and negative words into txt file and save files into working directory in separate doc
uments
#scan in the txt document containing positive words

positivelexicon<- scan('positivelexicon.txt', what='character', comment.char = ";")
#scan in txt document containing negative words

negativelexicon<- scan('negativelexicon.txt', what='character', comment.char = ";")
```

Step 7:

We will use the match function to determine how many words in our text match with the words in the positive and negative lexicons.

```
#use match() function to
match(text, positivelexicon)
```

```
##   [1]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
##  [15]   NA   NA   NA 1841   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
##  [29]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
##  [43]   NA   NA 1841   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
##  [57]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
##  [71]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
##  [85]   NA   NA   NA   NA   NA   NA   NA   NA   NA 1532   NA   NA   NA   NA
##  [99]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA 1841   NA
## [113] 1266 1979   NA   NA   NA   NA 1266   NA   NA   NA   NA   NA   NA   NA
## [127]   NA   NA   NA   NA 1266   NA   NA 1841   NA   NA   NA   NA   NA   NA
## [141]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## [155]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## [169]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## [183]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## [197]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## [211]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## [225]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## [239]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## [253]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## [267]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## [281]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## [295]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## [309]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## [323]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## [337]   NA
```

```
match(text, negativelexicon)
```

```
##   [1]   NA   NA   NA  846   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
##  [15]   NA   NA   NA   NA   NA 4631   NA   NA   NA   NA   NA   NA   NA   NA
##  [29]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
##  [43]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
##  [57]   NA   NA 3132   NA   NA  846 4659   NA   NA 4631   NA   NA   NA   NA
##  [71]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA 4681
##  [85]   NA   NA   NA   NA   NA   NA   NA   NA 4659   NA   NA   NA   NA   NA
##  [99]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA 1068   NA   NA   NA
## [113]   NA   NA   NA   NA   NA 3605   NA   NA   NA   NA   NA   NA   NA   NA
## [127]   NA   NA 2923   NA   NA   NA   NA   NA 4631   NA   NA   NA   NA   NA
## [141]   NA   NA   NA   NA   NA   NA 2387   NA   NA   NA   NA   NA   NA   NA
## [155]   NA   NA   NA   NA   NA   NA   NA 4742   NA   NA   NA   NA   NA   NA
## [169]   NA   NA   NA 1949 3132   NA   NA   NA   NA   NA   NA   NA   NA   NA
## [183]   NA   NA   NA   NA   NA   NA   NA   NA   NA 4659   NA   NA 1068   NA
## [197]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## [211]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
## [225]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA 4659   NA   NA   NA
## [239]   NA   NA   NA   NA   NA   NA 2387   NA   NA   NA   NA   NA   NA   NA
## [253]   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA 4659   NA
## [267]   NA   NA   NA   NA   NA   NA 3132   NA   NA   NA   NA   NA   NA   NA
## [281]   NA   NA   NA 3336   NA   NA   NA 3605 4659   NA   NA   NA   NA   NA
## [295]   NA   NA   NA   NA   NA   NA 4736   NA 1645   NA   NA   NA   NA   NA
## [309]   NA   NA   NA   NA   NA   NA   NA   NA   NA 4659   NA   NA   NA   NA
## [323] 2583   NA   NA   NA   NA   NA   NA 4659   NA   NA   NA   NA   NA   NA
## [337]   NA
```

We will make this easy to read by using the sum 'sum()' and is not na function '!is.na() to determine number of matching words in each category: positive and negative.

```
#is not na will provide a FALSE for any NAs and a TRUE for any non NAs (which are matches for positive/ negative words)
#we will then use the sum function to obtain totals of matches



!is.na(match(text,positivelexicon ))
```

```
##   [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [12] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
##  [23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [34] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [45]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [56] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [67] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [78] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [89] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
## [100] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [111]  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## [122] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
## [133] FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [144] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [155] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [166] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [177] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [188] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [199] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [210] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [221] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [232] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [243] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [254] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [265] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [276] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [287] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [298] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [309] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [320] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [331] FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
sum(!is.na(match(text,positivelexicon )))
```

```
## [1] 9
```

```
#complete same steps as above for the negative lexicon
!is.na(match(text,negativelexicon ))
```

```
##   [1] FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
##  [23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [34] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [45] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [56] FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE
##  [67] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [78] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
##  [89] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## [100] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
## [111] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
## [122] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
## [133] FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [144] FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [155] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
## [166] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE
## [177] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [188] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE
## [199] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [210] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [221] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [232] FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [243] FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [254] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [265]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## [276] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## [287] FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [298] FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
## [309] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
## [320] FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
## [331] FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
sum(!is.na(match(text,negativelexicon )))
```

```
## [1] 30
```

Now we can calculate the sentiment score analysis by finding the difference between the positive word count and the negative word count.

```
#calculate the sentiment score analysis by finding the difference between the positive word count and negative w
ord count

score<- sum(!is.na(match(text,positivelexicon ))) - sum(!is.na(match(text,negativelexicon )))


score
```

```
## [1] -21
```

Conclusion:

A score of -21 indicates an overall negative sentiment in this article. In a real world implementation, we would likely be looking at a collection of text documents. A collection of text documents is referred to as a corpus. You could then calculate average sentiment scores for each of your text documents and even create a histogram to view distribution of scores in the corpus.