Health News Tweets

Prem Kumar Kamasani

October 28, 2018

Data for this project is taken from <https://archive.ics.uci.edu/ml/machine-learning-databases/00438/>

Data consists of tweets about health news in Newyork Times.

library(tm)

## Warning: package 'tm' was built under R version 3.5.3

## Loading required package: NLP

## Warning: package 'NLP' was built under R version 3.5.2

setwd("D:/Downloads")  
text<-readLines("Health-Tweets/nytimeshealth.txt")  
text[1:10]

## [1] "548662191340421120|Sat Dec 27 02:10:34 +0000 2014|Risks in Using Social Media to Spot Signs of Mental Distress http://nyti.ms/1rqi9I1"   
## [2] "548579831169163265|Fri Dec 26 20:43:18 +0000 2014|RT @paula\_span: The most effective nationwide diabetes prevention program you've probably never heard of: http://newoldage.blogs.nytimes.com/2014/12/26/diabetes-prevention-that-works/"  
## [3] "548579045269852161|Fri Dec 26 20:40:11 +0000 2014|The New Old Age Blog: Diabetes Prevention That Works http://nyti.ms/1xm7fTi"   
## [4] "548444679529041920|Fri Dec 26 11:46:15 +0000 2014|Well: Comfort Casseroles for Winter Dinners http://nyti.ms/1xTNoO0"   
## [5] "548311901227474944|Fri Dec 26 02:58:39 +0000 2014|High-Level Knowledge Before Veterans Affairs Scandal http://nyti.ms/13yCpvS"   
## [6] "548305625449787392|Fri Dec 26 02:33:42 +0000 2014|Your Money: Affordable Care Actâ\200\231s Tax Effects Now Loom for Filers http://nyti.ms/13yAtUf"   
## [7] "548283182853160960|Fri Dec 26 01:04:32 +0000 2014|Well: Christmas in the Hospital http://nyti.ms/1vtPNcm"   
## [8] "548278414504108033|Fri Dec 26 00:45:35 +0000 2014|The Texas Tribune: For Members of Health Ministries in Texas, Caring Means Sharing the Bills http://nyti.ms/1wLtAen"   
## [9] "548176399299649536|Thu Dec 25 18:00:12 +0000 2014|A q&amp;a with Laura Hillenbrand, author of \"Unbroken,\" who suffers from chronic fatigue syndrome. http://bit.ly/1ziCflj"   
## [10] "547949409288609792|Thu Dec 25 02:58:14 +0000 2014|Race to Deliver Nicotineâ\200\231s Punch, With Less Risk http://nyti.ms/1xNzyg4"

Above Shown is the first 10 lines of the data. Inorder to proceed with the further analysis we need to clean the data.

# Build Corpus  
  
corpus<- iconv(text)  
corpus<-Corpus(VectorSource(corpus))

Removing numbers from the data.

# clean Text  
  
inspect(corpus[1:10])

## <<SimpleCorpus>>  
## Metadata: corpus specific: 1, document level (indexed): 0  
## Content: documents: 10  
##   
## [1] 548662191340421120|Sat Dec 27 02:10:34 +0000 2014|Risks in Using Social Media to Spot Signs of Mental Distress http://nyti.ms/1rqi9I1   
## [2] 548579831169163265|Fri Dec 26 20:43:18 +0000 2014|RT @paula\_span: The most effective nationwide diabetes prevention program you've probably never heard of: http://newoldage.blogs.nytimes.com/2014/12/26/diabetes-prevention-that-works/  
## [3] 548579045269852161|Fri Dec 26 20:40:11 +0000 2014|The New Old Age Blog: Diabetes Prevention That Works http://nyti.ms/1xm7fTi   
## [4] 548444679529041920|Fri Dec 26 11:46:15 +0000 2014|Well: Comfort Casseroles for Winter Dinners http://nyti.ms/1xTNoO0   
## [5] 548311901227474944|Fri Dec 26 02:58:39 +0000 2014|High-Level Knowledge Before Veterans Affairs Scandal http://nyti.ms/13yCpvS   
## [6] 548305625449787392|Fri Dec 26 02:33:42 +0000 2014|Your Money: Affordable Care Actâ\200\231s Tax Effects Now Loom for Filers http://nyti.ms/13yAtUf   
## [7] 548283182853160960|Fri Dec 26 01:04:32 +0000 2014|Well: Christmas in the Hospital http://nyti.ms/1vtPNcm   
## [8] 548278414504108033|Fri Dec 26 00:45:35 +0000 2014|The Texas Tribune: For Members of Health Ministries in Texas, Caring Means Sharing the Bills http://nyti.ms/1wLtAen   
## [9] 548176399299649536|Thu Dec 25 18:00:12 +0000 2014|A q&amp;a with Laura Hillenbrand, author of "Unbroken," who suffers from chronic fatigue syndrome. http://bit.ly/1ziCflj   
## [10] 547949409288609792|Thu Dec 25 02:58:14 +0000 2014|Race to Deliver Nicotineâ\200\231s Punch, With Less Risk http://nyti.ms/1xNzyg4

corpus<- tm\_map(corpus, removeNumbers)

## Warning in tm\_map.SimpleCorpus(corpus, removeNumbers): transformation drops  
## documents

inspect(corpus[1:10])

## <<SimpleCorpus>>  
## Metadata: corpus specific: 1, document level (indexed): 0  
## Content: documents: 10  
##   
## [1] |Sat Dec :: + |Risks in Using Social Media to Spot Signs of Mental Distress http://nyti.ms/rqiI   
## [2] |Fri Dec :: + |RT @paula\_span: The most effective nationwide diabetes prevention program you've probably never heard of: http://newoldage.blogs.nytimes.com////diabetes-prevention-that-works/  
## [3] |Fri Dec :: + |The New Old Age Blog: Diabetes Prevention That Works http://nyti.ms/xmfTi   
## [4] |Fri Dec :: + |Well: Comfort Casseroles for Winter Dinners http://nyti.ms/xTNoO   
## [5] |Fri Dec :: + |High-Level Knowledge Before Veterans Affairs Scandal http://nyti.ms/yCpvS   
## [6] |Fri Dec :: + |Your Money: Affordable Care Actâ\200\231s Tax Effects Now Loom for Filers http://nyti.ms/yAtUf   
## [7] |Fri Dec :: + |Well: Christmas in the Hospital http://nyti.ms/vtPNcm   
## [8] |Fri Dec :: + |The Texas Tribune: For Members of Health Ministries in Texas, Caring Means Sharing the Bills http://nyti.ms/wLtAen   
## [9] |Thu Dec :: + |A q&amp;a with Laura Hillenbrand, author of "Unbroken," who suffers from chronic fatigue syndrome. http://bit.ly/ziCflj   
## [10] |Thu Dec :: + |Race to Deliver Nicotineâ\200\231s Punch, With Less Risk http://nyti.ms/xNzyg

Removing URL’s

removeUrlPat<-function(x) gsub("(ftp|http)(s?)://.\*\\b", "", x)  
corpus<-tm\_map(corpus, removeUrlPat)

## Warning in tm\_map.SimpleCorpus(corpus, removeUrlPat): transformation drops  
## documents

inspect(corpus[1:10])

## <<SimpleCorpus>>  
## Metadata: corpus specific: 1, document level (indexed): 0  
## Content: documents: 10  
##   
## [1] |Sat Dec :: + |Risks in Using Social Media to Spot Signs of Mental Distress   
## [2] |Fri Dec :: + |RT @paula\_span: The most effective nationwide diabetes prevention program you've probably never heard of:   
## [3] |Fri Dec :: + |The New Old Age Blog: Diabetes Prevention That Works   
## [4] |Fri Dec :: + |Well: Comfort Casseroles for Winter Dinners   
## [5] |Fri Dec :: + |High-Level Knowledge Before Veterans Affairs Scandal   
## [6] |Fri Dec :: + |Your Money: Affordable Care Actâ\200\231s Tax Effects Now Loom for Filers   
## [7] |Fri Dec :: + |Well: Christmas in the Hospital   
## [8] |Fri Dec :: + |The Texas Tribune: For Members of Health Ministries in Texas, Caring Means Sharing the Bills   
## [9] |Thu Dec :: + |A q&amp;a with Laura Hillenbrand, author of "Unbroken," who suffers from chronic fatigue syndrome.   
## [10] |Thu Dec :: + |Race to Deliver Nicotineâ\200\231s Punch, With Less Risk

Removing dates

removeDates<-function(x) gsub('\\|.\*\\|',"", x)  
corpus<-tm\_map(corpus,removeDates)

## Warning in tm\_map.SimpleCorpus(corpus, removeDates): transformation drops  
## documents

inspect(corpus[1:10])

## <<SimpleCorpus>>  
## Metadata: corpus specific: 1, document level (indexed): 0  
## Content: documents: 10  
##   
## [1] Risks in Using Social Media to Spot Signs of Mental Distress   
## [2] RT @paula\_span: The most effective nationwide diabetes prevention program you've probably never heard of:   
## [3] The New Old Age Blog: Diabetes Prevention That Works   
## [4] Well: Comfort Casseroles for Winter Dinners   
## [5] High-Level Knowledge Before Veterans Affairs Scandal   
## [6] Your Money: Affordable Care Actâ\200\231s Tax Effects Now Loom for Filers   
## [7] Well: Christmas in the Hospital   
## [8] The Texas Tribune: For Members of Health Ministries in Texas, Caring Means Sharing the Bills   
## [9] A q&amp;a with Laura Hillenbrand, author of "Unbroken," who suffers from chronic fatigue syndrome.   
## [10] Race to Deliver Nicotineâ\200\231s Punch, With Less Risk

Converting whole data to lower case.

corpus<-tm\_map(corpus, tolower)

## Warning in tm\_map.SimpleCorpus(corpus, tolower): transformation drops  
## documents

inspect(corpus[1:10])

## <<SimpleCorpus>>  
## Metadata: corpus specific: 1, document level (indexed): 0  
## Content: documents: 10  
##   
## [1] risks in using social media to spot signs of mental distress   
## [2] rt @paula\_span: the most effective nationwide diabetes prevention program you've probably never heard of:   
## [3] the new old age blog: diabetes prevention that works   
## [4] well: comfort casseroles for winter dinners   
## [5] high-level knowledge before veterans affairs scandal   
## [6] your money: affordable care actâ\200\231s tax effects now loom for filers   
## [7] well: christmas in the hospital   
## [8] the texas tribune: for members of health ministries in texas, caring means sharing the bills   
## [9] a q&amp;a with laura hillenbrand, author of "unbroken," who suffers from chronic fatigue syndrome.   
## [10] race to deliver nicotineâ\200\231s punch, with less risk

remove punctuations

corpus<-tm\_map(corpus, removePunctuation)

## Warning in tm\_map.SimpleCorpus(corpus, removePunctuation): transformation  
## drops documents

inspect(corpus[1:10])

## <<SimpleCorpus>>  
## Metadata: corpus specific: 1, document level (indexed): 0  
## Content: documents: 10  
##   
## [1] risks in using social media to spot signs of mental distress   
## [2] rt paulaspan the most effective nationwide diabetes prevention program youve probably never heard of   
## [3] the new old age blog diabetes prevention that works   
## [4] well comfort casseroles for winter dinners   
## [5] highlevel knowledge before veterans affairs scandal   
## [6] your money affordable care actâ\200\231s tax effects now loom for filers   
## [7] well christmas in the hospital   
## [8] the texas tribune for members of health ministries in texas caring means sharing the bills   
## [9] a qampa with laura hillenbrand author of unbroken who suffers from chronic fatigue syndrome   
## [10] race to deliver nicotineâ\200\231s punch with less risk

Remove stopwords

corpus<-tm\_map(corpus, removeWords, stopwords("english"))

## Warning in tm\_map.SimpleCorpus(corpus, removeWords, stopwords("english")):  
## transformation drops documents

inspect(corpus[1:10])

## <<SimpleCorpus>>  
## Metadata: corpus specific: 1, document level (indexed): 0  
## Content: documents: 10  
##   
## [1] risks using social media spot signs mental distress   
## [2] rt paulaspan effective nationwide diabetes prevention program youve probably never heard   
## [3] new old age blog diabetes prevention works   
## [4] well comfort casseroles winter dinners   
## [5] highlevel knowledge veterans affairs scandal   
## [6] money affordable care actâ\200\231s tax effects now loom filers   
## [7] well christmas hospital   
## [8] texas tribune members health ministries texas caring means sharing bills   
## [9] qampa laura hillenbrand author unbroken suffers chronic fatigue syndrome   
## [10] race deliver nicotineâ\200\231s punch less risk

Remove unnecessary whitespaces

corpus<-tm\_map(corpus, stripWhitespace)

## Warning in tm\_map.SimpleCorpus(corpus, stripWhitespace): transformation  
## drops documents

inspect(corpus[1:10])

## <<SimpleCorpus>>  
## Metadata: corpus specific: 1, document level (indexed): 0  
## Content: documents: 10  
##   
## [1] risks using social media spot signs mental distress   
## [2] rt paulaspan effective nationwide diabetes prevention program youve probably never heard   
## [3] new old age blog diabetes prevention works   
## [4] well comfort casseroles winter dinners   
## [5] highlevel knowledge veterans affairs scandal   
## [6] money affordable care actâ\200\231s tax effects now loom filers   
## [7] well christmas hospital   
## [8] texas tribune members health ministries texas caring means sharing bills   
## [9] qampa laura hillenbrand author unbroken suffers chronic fatigue syndrome   
## [10] race deliver nicotineâ\200\231s punch less risk

After cleaning the data we need to count the word frequency

# Term Document Matrix  
  
tdm<-TermDocumentMatrix(corpus, control = list(minWordLength=c(1,Inf)))  
tdm<-as.matrix(tdm)  
termFrequency<- rowSums(tdm)  
termFrequency[1:10]

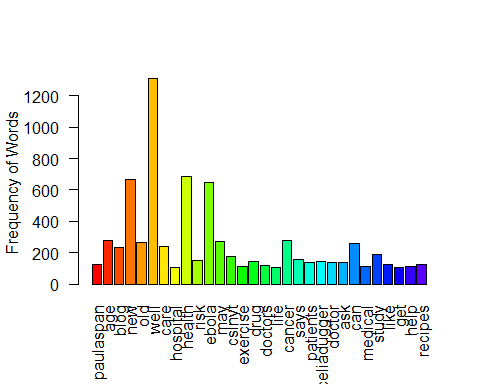
## distress media mental risks signs social spot   
## 3 3 51 47 13 13 3   
## using diabetes effective   
## 18 48 19

Since there are many words, we keep only words with frequency >= 100

termFrequency<-subset(termFrequency, termFrequency>=100)  
termFrequency[1:10]

## paulaspan age blog new old well care   
## 123 277 230 664 263 1309 242   
## hospital health risk   
## 103 685 147

#Bar Plot  
  
barplot(termFrequency,las=2, col=rainbow(40), ylab = 'Frequency of Words')



library(wordcloud)

## Warning: package 'wordcloud' was built under R version 3.5.3

## Loading required package: RColorBrewer

## Warning: package 'RColorBrewer' was built under R version 3.5.2

wordfreq<-sort(rowSums(tdm),decreasing = TRUE )  
#wordfreq<-subset(wordfreq, wordfreq>=50)  
wordcloud(words=names(wordfreq), freq = wordfreq, random.order = FALSE, max.words = 100, colors = rainbow(12))

