Project

Nisha Iyer, Rachel Jordan, Sam Dooley

April 29, 2016

For the complete set of documents, try the functions in lecture 7:

You can start by following the slides in Lecture 7.You should do at least the following:For the complete set of documents, try the functions in lecture 7. What happens? Does it yield anything understandable about the documents. [answered below]

```
data("acq")
head(acq)
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 6
```

```
#compilation of 50 news articles with additional meta information form the
#Reuters-21578 data se of topic acq. 13 documents
ACQ <- acq
ACQ
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 50
```

```
#this tell us what information (metadata) about our documents.  For example, how many chars are within
inspect(ACQ) #all 50 docs
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 50
##
## $`reut-00001.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 1287
##
## $`reut-00002.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 784
##
## $`reut-00003.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 502
##
## $`reut-00004.xml`
```

```
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 2308
##
## $`reut-00005.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 337
##
## $`reut-00006.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 381
##
## $`reut-00007.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 3635
##
## $`reut-00008.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 593
##
## $`reut-00009.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 248
##
## $`reut-00010.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 234
##
## $`reut-00011.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 620
##
## $`reut-00012.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 596
##
## $`reut-00013.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 850
##
## $`reut-00014.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 235
##
```

```
## $`reut-00015.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 229
##
## $`reut-00016.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 168
##
## $`reut-00017.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 528
##
## $`reut-00018.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 871
##
## $`reut-00020.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 2457
##
## $`reut-00021.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 1009
##
## $`reut-00022.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 429
##
## $`reut-00023.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 1873
##
## $`reut-00024.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 587
##
## $`reut-00025.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 602
##
## $`reut-00026.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 3516
```

```
##
## $`reut-00027.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 863
##
## $`reut-00028.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 652
##
## $`reut-00029.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 697
##
## $`reut-00030.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 3109
##
## $`reut-00031.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 213
##
## $`reut-00032.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 336
##
## $`reut-00034.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 448
##
## $`reut-00035.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 637
##
## $`reut-00036.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 1465
##
## $`reut-00039.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 773
##
## $`reut-00040.xml`
## <<PlainTextDocument>>
## Metadata:  15
```

```
## Content:  chars: 1043
##
## $`reut-00042.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 618
##
## $`reut-00043.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 460
##
## $`reut-00045.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 805
##
## $`reut-00046.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 429
##
## $`reut-00047.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 362
##
## $`reut-00048.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 1607
##
## $`reut-00049.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 295
##
## $`reut-00050.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 1022
##
## $`reut-00051.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 510
##
## $`reut-00052.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 547
##
## $`reut-00053.xml`
## <<PlainTextDocument>>
```

```
## Metadata:  15
## Content:  chars: 3013
##
## $`reut-00054.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 502
##
## $`reut-00055.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 343
##
## $`reut-00056.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 1068
```

```
inspect(ACQ[1:2]) #just the first 2
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 2
##
## $`reut-00001.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 1287
##
## $`reut-00002.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 784
```

```
meta(ACQ[[2]], "id") #this is another way to reference
```

```
## [1] "12"
```

```
#extract one document
text1 <- ACQ[[1]]
text1
```

```
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 1287
```

```
#This function tells us more information about the texts (all 50)
#For example, the maximal term length, non/sparse entries
ACQdoc <-DocumentTermMatrix(ACQ)
ACQdoc
```

```
## <<DocumentTermMatrix (documents: 50, terms: 2103)>>
## Non-/sparse entries: 4135/101015
## Sparsity           : 96%
## Maximal term length: 21
## Weighting          : term frequency (tf)
```

```
nrow(ACQdoc) #50 rows
```

```
## [1] 50
```

```
ncol(ACQdoc) #2103 cols
```

```
## [1] 2103
```

```
inspect(ACQdoc[1:6,1:10])
```

```
## <<DocumentTermMatrix (documents: 6, terms: 10)>>
## Non-/sparse entries: 2/58
## Sparsity           : 97%
## Maximal term length: 11
## Weighting          : term frequency (tf)
##
##     Terms
## Docs -laval .125 .3322 "...that "(american) "any "bridge" "final" "it
##   10      0    1     0        0           0    0        0       0   0
##   12      0    0     0        0           0    0        0       0   0
##   44      0    0     0        0           0    0        0       0   0
##   45      0    0     0        0           0    1        0       0   0
##   68      0    0     0        0           0    0        0       0   0
##   96      0    0     0        0           0    0        0       0   0
##     Terms
## Docs "purolator
##   10          0
##   12          0
##   44          0
##   45          0
##   68          0
##   96          0
```

```
#termFreq tells us more about an individual doc/text such as term freq within the doc
test1tf <- as.data.frame(termFreq(text1))
#rank words most to least
rank_of_words <- cbind(as.data.frame(rownames(test1tf)),test1tf %>% arrange(desc(termFreq(text1))))

#the tm_map and content_transformer transforms the data
#such as converting the terms to lower case
ACQlow <- tm_map(ACQ, content_transformer(tolower))
ACQlow
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 50
```

```r
#the next function removes anything other than English letters or spaces
removeNumPunct <- function(x) gsub("[^[:alpha:][:space:]]*", "", x)
ACQcl <- tm_map(ACQlow,content_transformer(removeNumPunct))
ACQcl
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 50
```

```r
#after converting the text to lower case, and removing punctionation
#we are going to remove stopwords (filler words such as a, an, the, etc.)
stopwords <- c(stopwords('english'))
ACQstop <- tm_map(ACQcl, removeWords, stopwords)

#here we can look at the first two text docs and see how the word count (char) differs
inspect(ACQ[1:2])#the original; 1 with 1287 chars, 2nd with 784 chars
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 2
##
## $`reut-00001.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 1287
##
## $`reut-00002.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 784
```

```r
inspect(ACQstop[1:2]) #the amount of words is much less; first with 1030 chars, second with 620 chars
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 2
##
## $`reut-00001.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 1030
##
## $`reut-00002.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 620
```

```r
#now we are putting the terms without punctuation and stopwords into a matrix
ACQdm2 <- DocumentTermMatrix(ACQstop, control= list(wordLenghts = c(1,Inf)))
ACQdm2
```

```
## <<DocumentTermMatrix (documents: 50, terms: 1502)>>
## Non-/sparse entries: 2998/72102
## Sparsity           : 96%
## Maximal term length: 20
## Weighting          : term frequency (tf)
```

```
#find terms with a frequency of 5 or more
freq.terms <- findFreqTerms(ACQdm2, lowfreq=5)
freq.terms
```

```
##   [1] "acquire"       "acquired"      "acquisition"   "acquisitions"
##   [5] "added"         "agreed"        "agreement"     "already"
##   [9] "also"          "american"      "amusements"    "analysts"
##  [13] "another"       "approval"      "around"        "arsenal"
##  [17] "assets"        "bank"          "barbara"       "bid"
##  [21] "billion"       "board"         "bought"        "brokerage"
##  [25] "burdett"       "business"      "buy"           "capital"
##  [29] "cash"          "certain"       "chemlawn"      "chief"
##  [33] "circuit"       "closed"        "commission"    "common"
##  [37] "companies"     "company"       "companys"      "completed"
##  [41] "completion"    "computer"      "considered"    "considering"
##  [45] "consolidated"  "control"       "corp"          "courier"
##  [49] "current"       "deal"          "debt"          "division"
##  [53] "dlr"           "dlrs"          "due"           "earlier"
##  [57] "earnings"      "equity"        "esselte"       "exchange"
##  [61] "expected"      "express"       "february"      "filing"
##  [65] "financial"     "financing"     "firm"          "first"
##  [69] "five"          "four"          "friday"        "gas"
##  [73] "give"          "gold"          "government"    "group"
##  [77] "growth"        "held"          "holding"       "holdings"
##  [81] "hotel"         "husky"         "hutton"        "inc"
##  [85] "increase"      "industries"    "interest"      "international"
##  [89] "investment"    "issued"        "last"          "ltd"
##  [93] "made"          "management"    "march"         "market"
##  [97] "match"         "may"           "meeting"       "merger"
## [101] "mining"        "mln"           "multistep"     "national"
## [105] "need"          "net"           "new"           "now"
## [109] "offer"         "offered"       "officer"       "one"
## [113] "operating"     "operations"    "option"        "ordinary"
## [117] "ounces"        "outstanding"   "owned"         "owns"
## [121] "part"          "pct"           "penn"          "per"
## [125] "pittston"      "plan"          "plans"         "plc"
## [129] "position"      "preferred"     "president"     "pretax"
## [133] "previously"    "price"         "profit"        "profitable"
## [137] "profits"       "public"        "purchase"      "purolator"
## [141] "purolators"    "quarter"       "raised"        "received"
## [145] "redstone"      "reuter"        "rights"        "rmj"
## [149] "rumors"        "said"          "sale"          "santa"
## [153] "schlang"       "securities"    "sell"          "services"
## [157] "share"         "shareholders"  "shares"        "shearson"
## [161] "six"           "sold"          "speculation"   "spinoff"
## [165] "spokesman"     "stake"         "statement"     "steel"
## [169] "stock"         "subject"       "subsidiary"    "swedish"
## [173] "systems"       "takeover"      "technology"    "tender"
```

```
## [177] "terminal"      "terms"         "three"         "today"
## [181] "total"         "traffic"       "transaction"   "tvx"
## [185] "two"           "undisclosed"   "union"         "unit"
## [189] "value"         "valued"        "viacom"        "voting"
## [193] "wallenbergs"   "warrants"      "waste"         "will"
## [197] "worth"         "wtc"           "year"          "years"
## [201] "york"
```

```
#there are 201 terms with a frequency of 5 or more

#the Assocs function finds associations with terms, such as states or year
findAssocs(ACQdm2, "states", 0.25)
```

```
## $states
##        areas    arranging    assurance   bankruptcy       bodies
##         0.70         0.70         0.70         0.70         0.70
##     charters    continues     contract        court      crowley
##         0.70         0.70         0.70         0.70         0.70
##      delayed    equitable exchangeable        final     fraction
##         0.70         0.70         0.70         0.70         0.70
##    holdingss      include     includes         life        lines
##         0.70         0.70         0.70         0.70         0.70
##    mariotime       mclean      present      raising     revision
##         0.70         0.70         0.70         0.70         0.70
##      society    transport         used       united          mcv
##         0.70         0.70         0.70         0.69         0.66
##       raised   amusements     transfer     national      arsenal
##         0.63         0.62         0.62         0.60         0.57
##       offers      offered     holdings        value       called
##         0.56         0.52         0.49         0.49         0.48
##    committee        corps        eight   increasing         incs
##         0.48         0.48         0.48         0.48         0.48
##         meet         nine    ownership         rate   regulatory
##         0.48         0.48         0.48         0.48         0.48
##      service      various       viacom      viacoms       within
##         0.48         0.48         0.48         0.48         0.48
##         held     february        april    including negotiations
##         0.47         0.42         0.38         0.38         0.38
##    preferred     previous    principle       review          inc
##         0.38         0.38         0.38         0.38         0.34
##   conditions        holds        later         next      special
##         0.32         0.32         0.32         0.32         0.32
##         week       beyond      revised        south shareholders
##         0.32         0.29         0.29         0.29         0.28
##       assets    spokesman    increased     proposed       senior
##         0.27         0.27         0.25         0.25         0.25
```

```
findAssocs(ACQdm2, "year", 0.25)
```

```
## $year
##          decide     considering          ending         however
##            0.87            0.75            0.64            0.62
##       shearsons      speculated        spinning          street
```

```
##           0.62           0.62           0.62           0.62
##           wall        actions     affiliates          allow
##           0.62           0.61           0.61           0.61
##      ambitious    appreciation     approached      attempted
##           0.61           0.61           0.61           0.61
##          azuma     broadening         caused     circulated
##           0.61           0.61           0.61           0.61
##        clarify       compared    competition      concerned
##           0.61           0.61           0.61           0.61
##        confirm       consider      contacted       decision
##           0.61           0.61           0.61           0.61
##    discussions        divided          drove      employees
##           0.61           0.61           0.61           0.61
##        employs            end       enhanced       external
##           0.61           0.61           0.61           0.61
##            fed        focused         follow       forecast
##           0.61           0.61           0.61           0.61
##          giant         global       improved       industry
##           0.61           0.61           0.61           0.61
##       integral       japanese        jeffrey          kokan
##           0.61           0.61           0.61           0.61
##         little         losses        matters         nippon
##           0.61           0.61           0.61           0.61
##          nkktt    positioning       puzzling       rebuffed
##           0.61           0.61           0.61           0.61
##      recession       rejected        rumored       shearons
##           0.61           0.61           0.61           0.61
##           show        similar        sources      spokesmen
##           0.61           0.61           0.61           0.61
##          stand          steel         steels      strategic
##           0.61           0.61           0.61           0.61
##       strategy     struggling       studying         sunday
##           0.61           0.61           0.61           0.61
##         toshin          tosst     ultimately unsuccessfully
##           0.61           0.61           0.61           0.61
##          walls      wednesday          weeks        workers
##           0.61           0.61           0.61           0.61
##            yen           yens           firm       services
##           0.61           0.61           0.60           0.60
##    speculation      brokerage       shearson        express
##           0.59           0.58           0.58           0.57
##           last          added       aftertax       american
##           0.57           0.55           0.55           0.55
##       brothers       chairmen    contributed        created
##           0.55           0.55           0.55           0.55
##      divisions         expand            got         highly
##           0.55           0.55           0.55           0.55
##       internal           lane          larry           late
##           0.55           0.55           0.55           0.55
##         lehman           move      positions prudentialbache
##           0.55           0.55           0.55           0.55
##       remained        rumors        selling          sense
##           0.55           0.55           0.55           0.55
##         silent       unlikely         vacant        whether
```

```
##            0.55            0.55            0.55            0.55
##          growth            part        analysts         billion
##            0.54            0.54            0.53            0.53
##          beyond           bring   international           major
##            0.52            0.52            0.52            0.52
##         options  reorganization      considered         current
##            0.52            0.52            0.51            0.50
##            need       spokesman           march       statement
##            0.50            0.49            0.48            0.48
##         spinoff         capital             may       financial
##            0.46            0.45            0.45            0.44
##         comment           fully           plans           range
##            0.43            0.43            0.43            0.43
##            said     eckenfelder           place          access
##            0.43            0.42            0.42            0.40
##           alone           close   consideration           given
##            0.40            0.40            0.40            0.40
##            help         improve            loss            meet
##            0.40            0.40            0.40            0.40
##       operating            post         reached     whollyowned
##            0.40            0.40            0.40            0.40
##       worldwide         reflect            also       estimated
##            0.40            0.35            0.34            0.34
##          friday          market      profitable           total
##            0.33            0.32            0.32            0.32
##             can            days           firms           makes
##            0.30            0.30            0.30            0.30
##         related             net        position       president
##            0.30            0.29            0.29            0.29
##          public          higher
##            0.29            0.28
```

```
#Next we're going to put the terms with frequency count of 5 or more into a dataframe
term.freq <- rowSums(as.matrix(ACQdm2))
term.freq <- subset(term.freq, term.freq <= 5)
termdf <- data.frame(term = names(term.freq),freq=term.freq)
term_sort <- termdf %>% arrange(desc(freq))
term_sort[1:50,]
```

```
##       term freq
## NA    <NA>   NA
## NA.1  <NA>   NA
## NA.2  <NA>   NA
## NA.3  <NA>   NA
## NA.4  <NA>   NA
## NA.5  <NA>   NA
## NA.6  <NA>   NA
## NA.7  <NA>   NA
## NA.8  <NA>   NA
## NA.9  <NA>   NA
## NA.10 <NA>   NA
## NA.11 <NA>   NA
## NA.12 <NA>   NA
## NA.13 <NA>   NA
```

```
## NA.14 <NA>   NA
## NA.15 <NA>   NA
## NA.16 <NA>   NA
## NA.17 <NA>   NA
## NA.18 <NA>   NA
## NA.19 <NA>   NA
## NA.20 <NA>   NA
## NA.21 <NA>   NA
## NA.22 <NA>   NA
## NA.23 <NA>   NA
## NA.24 <NA>   NA
## NA.25 <NA>   NA
## NA.26 <NA>   NA
## NA.27 <NA>   NA
## NA.28 <NA>   NA
## NA.29 <NA>   NA
## NA.30 <NA>   NA
## NA.31 <NA>   NA
## NA.32 <NA>   NA
## NA.33 <NA>   NA
## NA.34 <NA>   NA
## NA.35 <NA>   NA
## NA.36 <NA>   NA
## NA.37 <NA>   NA
## NA.38 <NA>   NA
## NA.39 <NA>   NA
## NA.40 <NA>   NA
## NA.41 <NA>   NA
## NA.42 <NA>   NA
## NA.43 <NA>   NA
## NA.44 <NA>   NA
## NA.45 <NA>   NA
## NA.46 <NA>   NA
## NA.47 <NA>   NA
## NA.48 <NA>   NA
## NA.49 <NA>   NA
```

What happens? Does it yield anything understandable about the documents

Yes, the different functions allows us to break down the different text documents we were able to see how many stopwords and punctuation was included in the total character count of the texts the term frequencies allowed us insight into the top frequented words in the text the functions provided a lot of insight into the general documents, text, and words used in the texts

Find the 10 longest documents (in number of words).

```
#using quanteda for the next few questions
mycorpus <- corpus(acq)
summary_acq <- as.data.frame(summary(mycorpus))
```

```
## Corpus consisting of 50 documents.
##
##  Text Types Tokens Sentences                    author       datetimestamp
```

```
##    10   120    233        26                      <NA> 1987-02-26 15:18:06
##    12    89    146        17                      <NA> 1987-02-26 15:19:15
##    44    62     86        13                      <NA> 1987-02-26 15:49:56
##    45   232    431        51 By Cal Mankowski, Reuters 1987-02-26 15:51:17
##    68    42     59         7                      <NA> 1987-02-26 16:08:33
##    96    56     75         8                      <NA> 1987-02-26 16:32:37
##   110   292    666        79     By Patti Domm, Reuter 1987-02-26 16:43:13
##   125    73    112        12                      <NA> 1987-02-26 16:59:25
##   128    34     46         7                      <NA> 1987-02-26 17:01:28
##   134    37     40         6                      <NA> 1987-02-26 17:08:27
##   135    76    110        15                      <NA> 1987-02-26 17:09:47
##   153    77    108        13                      <NA> 1987-02-26 17:36:22
##   157    92    166        19                      <NA> 1987-02-26 17:38:47
##   162    32     39         6                      <NA> 1987-02-26 17:43:59
##   185    35     40         6                      <NA> 1987-02-26 18:12:35
##   186    29     33         4                      <NA> 1987-02-26 18:12:51
##   199    55    101        12                      <NA> 1987-02-26 18:27:56
##   260    91    174        19                      <NA> 1987-03-01 22:20:43
##   302   211    468        45                      <NA> 1987-03-02 04:45:57
##   304    97    201        24                      <NA> 1987-03-02 04:52:58
##   315    66     93        10                      <NA> 1987-03-02 05:48:46
##   331   188    364        39                      <NA> 1987-03-02 06:54:19
##   334    74    114        12                      <NA> 1987-03-02 06:58:00
##   361    71    108        14                      <NA> 1987-03-02 08:16:59
##   362   261    611        69    By Patti Domm, Reuters 1987-03-02 08:17:56
##   366    95    148        19                      <NA> 1987-03-02 08:22:40
##   369    82    121        16                      <NA> 1987-03-02 08:25:56
##   371    72    120        14                      <NA> 1987-03-02 08:26:35
##   372   250    577        67             By Patti Domm 1987-03-02 08:29:05
##   376    31     35         5                      <NA> 1987-03-02 08:41:41
##   379    49     63         8                      <NA> 1987-03-02 08:43:25
##   387    61     93        12                      <NA> 1987-03-02 09:02:51
##   389    85    128        15                      <NA> 1987-03-02 09:03:18
##   393   124    270        30                      <NA> 1987-03-02 09:16:08
##   401    94    140        15                      <NA> 1987-03-02 09:28:21
##   408   108    187        23                      <NA> 1987-03-02 09:33:32
##   424    77    122        12                      <NA> 1987-03-02 09:49:48
##   436    68     82        10                      <NA> 1987-03-02 10:06:32
##   441    76    148        16                      <NA> 1987-03-02 10:20:41
##   442    50     69        10                      <NA> 1987-03-02 10:29:07
##   447    43     64         8                      <NA> 1987-03-02 10:36:04
##   448   143    301        32                      <NA> 1987-03-02 10:36:13
##   467    42     53         7                      <NA> 1987-03-02 10:50:34
##   473   103    199        23                      <NA> 1987-03-02 10:59:16
##   474    59     94        11                      <NA> 1987-03-02 10:59:28
##   478    76    104        13                      <NA> 1987-03-02 11:09:06
##   496   228    555        56                      <NA> 1987-03-02 11:23:31
##   497    58     82        13                      <NA> 1987-03-02 11:23:45
##   498    47     57         9                      <NA> 1987-03-02 11:24:06
##   504   118    197        25                      <NA> 1987-03-02 11:29:26
##  description                                          heading  id
##                COMPUTER TERMINAL SYSTEMS <CPML> COMPLETES SALE  10
##                 OHIO MATTRESS <OMT> MAY HAVE LOWER 1ST QTR NET  12
##                  MCLEAN'S <MII> U.S. LINES SETS ASSET TRANSFER  44
```

```
##                 CHEMLAWN <CHEM> RISES ON HOPES FOR HIGHER BIDS  45
##                 <COFAB INC> BUYS GULFEX FOR UNDISCLOSED AMOUNT  68
##                       INVESTMENT FIRMS CUT CYCLOPS <CYL> STAKE  96
##               AMERICAN EXPRESS <AXP> SEEN IN POSSIBLE SPINNOFF 110
##                HONG KONG FIRM UPS WRATHER<WCO> STAKE TO 11 PCT 125
##                           LIEBERT CORP <LIEB> APPROVES MERGER 128
##                  GULF APPLIED TECHNOLOGIES <GATS> SELLS UNITS 134
##                  INVESTMENT GROUP RAISES ROBESON <RBSN> STAKE 135
##               DREXEL OFFICIAL HAS STAKE IN EPSILON DATA <EPSI> 153
##               <NOVA> WINS GOVERNMENT OKAY FOR HUSKY <HYO> DEAL 157
##                    SUFFIELD FINANCIAL <SSBK> GETS FED APPROVAL 162
##                              VERSATILE TO SELL UNIT TO VICON 185
##                           VIDEOTRON BUYS INTO EXHIBIT COMPANY 186
##                     CIRCUIT SYSTEMS <CSYI> BUYS BOARD MAKER 199
##               NIPPON KOKAN STEEL AFFILIATES CONSIDERING MERGER 260
##                  WALLENBERGS FIGHT BID FOR SWEDISH MATCH STAKE 302
##                     SHV SAYS IT MAKING TENDER OFFER FOR IC GAS 304
##                SALE TILNEY BUYS STAKE IN U.S. INSURANCE BROKER 315
##                    EXCO BUYS U.S. GOVERNMENT SECURITIES BROKER 331
##              COLOROLL AGREES TO BUY U.S. WALLCOVERINGS COMPANY 334
##               SCIENTIFIC MICRO SYSTEMS <SMSI> ACUIRES SUPERMAC 361
##                AMERICAN EXPRESS <AXP> VIEWING SHEARSON OPTIONS 362
##                      ROPAK <ROPK> HAS 34 PCT OF BUCKHORN <BKN> 366
##                           PENRIL <PNL> SEEKS TO SELL TWO UNITS 369
##                         <DALE BURDETT INC> FACES DAMAGE CLAIM 371
##                    PUROLATOR <PCC> IN BUYOUT WITH HUTTON <EFH> 372
##                  FINANCIAL SANTA BARBARA <FSB> TO MAKE PURCHASE 376
##                                MARRIOTT <MHS> TO SELL HOTEL 379
##                       LAROCHE STARTS BID FOR NECO <NPT> SHARES 387
##              SENIOR ENGINEERING MAKES 12.5 MLN DLR US PURCHASE 389
##                       VIACOM <VIA> RECEIVES TWO REVISED OFFERS 393
##                MILLER TABAK HAS 91.8 PCT OF PENN TRAFFIC <PNF> 401
##                     PITTSTON <PCO> AGREES TO ACQUIRE WTC <WAF> 408
##               DIAGNOSTIC <DRS> MAKES A BID FOR ROSPATCH <RPCH> 424
##                         THE JAPAN FUND <JPN> GETS BUYOUT OFFER 436
##                BANK OF NEW YORK <BK> TO HAVE GAIN ON UNIT SALE 441
##               CORNING <GLW>, HAZLETON <HLC> SET EXCAHNGE RATIO 442
##                BALLY <BLY> COMPLETES PURCHASE OF GOLDEN NUGGET 447
##                CONSOLIDATED TVX TO BUY BRAZIL GOLD MINE STAKES 448
##                   AMERICAN NURSERY <ANSY> BUYS FLORIDA NURSERY 467
##                  MULTI-STEP TO SELL LADDER UNIT, CANCEL SHARES 473
##                 ESSELTE BUSINESS <ESB> UNIT BUYS ANTONSON UNIT 474
##                       FOUR SEASONS BUYING MARRIOTT <MHS> HOTEL 478
##                  REDSTONE DETAILS SWEETENED VIACOM <VIA> OFFER 496
##                   MONTEDISON CONCLUDES TALKS WITH ANTIBIOTICOS 497
##                          UTILICORP <UCU> COMPLETES ACQUISITION 498
##                  CARBIDE <UK> LOOKS TO ACQUISITIONS FOR GROWTH 504
##  language            origin topics lewissplit     cgisplit oldid
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5553
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5555
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5587
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5588
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5611
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5639
```

```
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5653
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5668
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5671
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5677
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5678
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5696
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5700
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5705
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5728
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5729
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5742
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  8345
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12485
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12487
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12498
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12514
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12517
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12543
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12544
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12548
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12551
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12553
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12554
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12558
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12561
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12570
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12572
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12576
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12584
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12591
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12607
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12619
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12624
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12625
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12630
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12631
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12650
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12656
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12657
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12661
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12679
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12680
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12681
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12687
##                 places people orgs exchanges
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
```

```
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                 canada   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                 canada   <NA> <NA>      <NA>
##                 canada   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                  japan   <NA> <NA>      <NA>
##                 sweden   <NA> <NA>      <NA>
##                     uk   <NA> <NA>      <NA>
##         c("usa", "uk")   <NA> <NA>      <NA>
##         c("uk", "usa")   <NA> <NA>      <NA>
##         c("usa", "uk")   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##         c("uk", "usa")   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##  c("canada", "brazil")   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                   <NA>   <NA> <NA>      <NA>
##     c("usa", "sweden")   <NA> <NA>      <NA>
##                 canada   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##    c("italy", "spain")   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##
## Source:  Converted from tm VCorpus 'acq'
## Created: Sat Apr 30 13:31:08 2016
## Notes:
```

```
#10 longest documents in the corpus
sort_top10 <- summary_acq %>% arrange(desc(Tokens))
top_10_docs <- subset(sort_top10, select=c(id, heading))[1:10,]
top10 <- top_10_docs[,1]

top10
```

```
##  [1] "110" "362" "372" "496" "302" "45"  "331" "448" "393" "10"
```

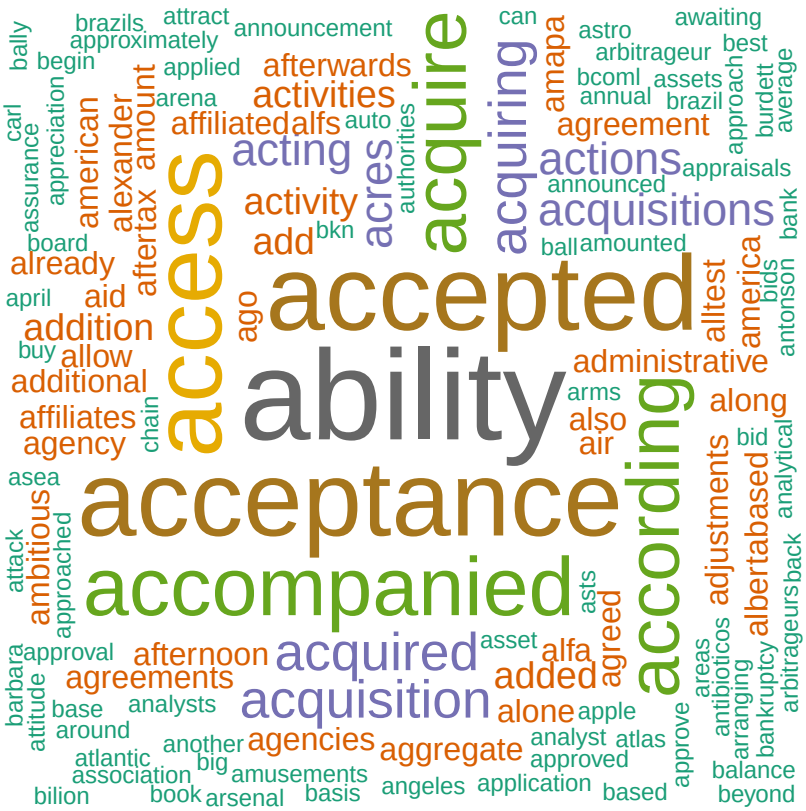
```
topdocs <- mycorpus[mycorpus$documents$id %in% top10]
topdocs
```

```
##
##
##
##
##
## "American Express Co remained silent on\nmarket rumors it would spinoff all or part of its Shearson\r
##
##
##
##
##
##
##
##
##
##
##
##
##
##
```

```
#top 10 dendogram, 1 for each of the top 10 documents
top10.dendogram <- function(tdm2,doc)
{
  acq.mat <- as.matrix(t(tdm2))
  acq.mat <- as.data.frame(acq.mat)
  acq.mat <- acq.mat[,top10]
  acq.mat <- as.matrix(acq.mat)
  distMatrix <- dist(scale(acq.mat[,doc]))
  fit <- hclust(distMatrix, method = "ward.D2")
  print(plot(fit,main = "Dendogram"))
}


#word cloud for top 10
wordcloud.func <- function(ACQstop, doc)
{
  dtm <- TermDocumentMatrix(ACQstop)
  m <- as.data.frame(as.matrix(dtm))
  m <- m[,top10]
  m <- as.matrix(m)
  v <- sort(m[,doc],decreasing=TRUE)
  d <- data.frame(word = row.names(m),freq=v)
  set.seed(1234)
  print(wordcloud(words = d$word, freq = d$freq, min.freq = 1,
            max.words=200, random.order=FALSE, rot.per=0.35,
            colors=brewer.pal(8, "Dark2")))
}
for (i in 1:10){
  wordcloud.func(ACQstop,i)
```

```
}
```


brazils
attract
announcement
can
awaiting
bally
approximately
astro
amapa
begin
applied
afterwards
arbitrageur
best
bcoml
assets
approach
burdett
average
arena
activities
annual
brazil
carl
affiliated
alfs
auto
agreement
assurance
appreciation
american
alexander
amount
acting
acres
authorities
acquire
acquiring
actions
appraisals
activity
announced
acquisitions
bank
board
add
bkn
ball
amounted
already
aftertax
access
accepted
april
aid
ago
alltest
america
bids
antonson
addition
ability
administrative
buy
allow
additional
arms
along
affiliates
chain
also
agency
air
bid
asea
acceptance
adjustments
albertabased
analytical
attack
ambitious
approached
accompanied
according
asts
agreed
back
approval
afternoon
acquired
asset
alfa
areas
barbara
agreements
added
antibioticos
arbitrageurs
attitude
acquisition
alone
apple
arranging
base
analysts
around
agencies
aggregate
analyst
atlas
approve
bankruptcy
another
approved
atlantic
big
balance
association
amusements
application
bilion
book
arsenal
basis
angeles
based
beyond

```
## NULL
```


angeles
alone
afterwards
amounted
asea
approach
american
amount
agencies
additional
agreement
annual
activities
ball
affiliates
bidder
bid
air
alfs
acquisition
affiliated
approve
ambitious
bridge
access
accompanied
alexander
according
accepted
alltest
albertabased
buy
back
attack
asts
acting
bank
adjustments
ability
agreed
alfa
add
added
april
amapa
acceptance
acquiring
announcement
acres
aftertax
acquire
acquired
ago
best
america
activity
antonson
bkn
allow
big
aid
actions
agency
agreements
acquisitions
also
calls
better
afternoon
arms
addition
aggregate
another
apple
believes
analyst
along
areas
already
become
administrative
bids
arena

```
## NULL
```


alltest
auto
also
can
afternoon
analysts
addition
ago
analyst
atlas
bid
according
access
acres
acquisitions
adjustments
arms
alfs
acquisition
aftertax
acquiring
bidder
analytical
alfa
add
acquired
arena
accepted
attack
air
apple
agency
ability
added
american
affiliates
aid
acquire
ambitious
acceptance
antibioticos
accompanied
benficiaries
america
along
affiliated
asts
acting
actions
additional
announced
agencies
activity
afterwards
board
bkn
amounted
activities
asea
allow
alone
atlantic

```
## NULL
```


ability
acceptance
accompanied
accepted
access
acquired
acquisitions
acquisition
acquire
acquiring
acres
according
administrative
affiliated
activity
activities
acting
actions
addition
add
added
additional
adjustments
agreement
aggregate
allow
alone
affiliates
agreed
albertabased
afternoon
afterwards
air
alfa
amapa
aftertax
alexander
ambitious
agreements
agencies
aid
agency
alfs
along
also
already
alltest
america
bought
book
attempted
arena
authorization
approved
billion
boost
boosting
bcoml
basis
applied
announcement
awaiting
auto
announced
believes
brierley
average
bidding
borrowing
amusements
banking
brokers
based
british
belive
approval
areas
asea
bluffton
assets
angeles
biltmores
broker
barbara
another
ago
annual
boards
around
became
antibiotico
american
associationicgsl
approve
believe
biltmore
big
appreciation
authorities
antibiotics
appraisals
azuma
arbitrageurs
amounted
attitude
bankruptcy
approached
assurance
bidder
bridge
benficiaries
arms
bally
back
board
analyst
bank
april
balance
base
approximately
arsenal
bid
attack
asts
application
bring
arbitrageur
begin
astro
bancorp
broadening
amount
analytical
arranging
bilion
bkn
attract
apple
analysts
become
asset
antonson
approach
ball
brazil
brazils
broking
beyond
association
better
atlas
besides
brothers
brokerage
bids
assumption
atlantic
best
bodies

```
## NULL
```


atlas
arena
amusements
agreements
applied
asea
better
angeles
agencies
administrative
ball
appraisals
back
afterwards
additional
assets
auto
asts
barbara
approved
amapa
alone
asset
allow
apple
activities
afternoon
acquiring
approve
activity
acquisition
alfa
alfs
antonson
approach
acquire
accompanied
access
america
also
accepted
american
aid
aftertax
amount
ago
acres
according
april
bid
ability
albertabased
analyst
acting
air
agreed
acceptance
ambitious
astro
along
antibioticos
analytical
azuma
alexander
adjustments
acquired
actions
alltest
acquisitions
added
appreciation
bkn
analysts
big
affiliates
add
affiliated
already
annual
bank
areas
arms
agency
addition
agreement
application
announced
aggregate
amounted
another
approval

```
## NULL
```


asset
billion
angeles
another
analysts
arsenal
approached
bank
announced
amusements
announcement
afternoon
approve
arbitrageurs
approved
affiliated
addition
already
big
appreciation
antonson
ambitious
agencies
added
actions
allow
atlas
approach
additional
bid
acquisition
agreement
aggregate
alfs
acquire
access
accompanied
apple
around
annual
albertabased
alexander
agreed
accepted
add
alone
amount
aftertax
alltest
asts
ago
according
adjustments
astro
activities
ability
amapa
attack
american
acceptance
analytical
back
antibioticos
affiliates
acquired
air
aid
acres
acquiring
activity
analyst
along
areas
april
acquisitions
alfa
amounted
agency
assurance
afterwards
acting
arms
america
also
applied
appraisals
administrative
application
auto
azuma
approximately
agreements
ball
approval
asea
bids
arena

```
## NULL
```

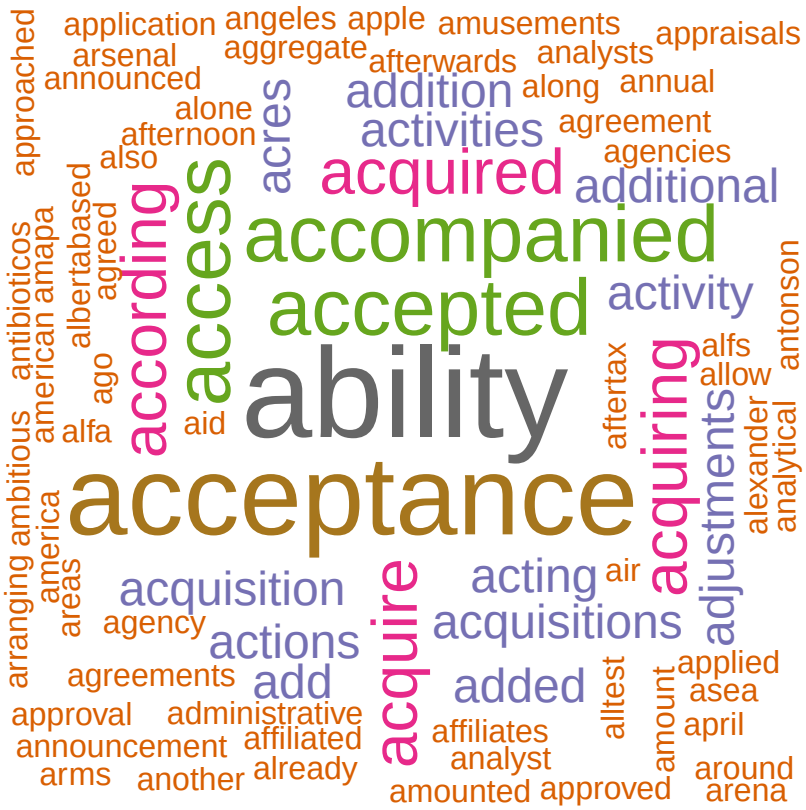
```
## NULL
```


around
along
allow
arena
ago
alfa
albertabased
according
access
acres
announced
added
affiliated
angeles
activities
agreement
aid
alone
areas
acquisition
acquiring
acquired
aggregate
activity
accepted
aftertax
ability
agencies
agency
air
add
alexander
asts
acquire
acceptance
adjustments
american
alltest
agreed
accompanied
amapa
also
acquisitions
actions
alfs
already
addition
acting
additional
amount
america
analytical
analyst
agreements
administrative
affiliates

NULL


announced
agreements
aftertax
acres
afternoon
application
appreciation
annual
april
along
administrative
arena
activities
arms
amounted
added
acquired
already
agency
angeles
ambitious
according
accompanied
apple
antibioticos
america
acquiring
analyst
accepted
activity
american
agreed
also
alone
alexander
ability
air
ago
albertabased
adjustments
amapa
antonson
add
areas
acceptance
access
allow
aid
alltest
acting
acquisition
acquire
amount
acquisitions
analysts
actions
addition
afterwards
applied
analytical
additional
alfs
alfa
agreement
approval
affiliates
affiliated
announcement
around
agencies
aggregate
amusements
another

```
## NULL
```




```
## NULL
```

Prior to removing punctuation find the longest word and longest sentence in each of 10 docs my corpus is before removing punctuation

```
#####FIND LONGEST WORD in 10 docs
max_length <- c()
word <- c()
id <- c()
for (i in 1:10){
  words <- tokenize_words(topdocs[[i]][[1]])
  word[i] <- words[nchar(words) == max(nchar(words))]
  max_length[i] <- max(nchar(words))
  id[i] <- names(topdocs[i])
}
```

```
## Warning in word[i] <- words[nchar(words) == max(nchar(words))]: number of
## items to replace is not a multiple of replacement length

## Warning in word[i] <- words[nchar(words) == max(nchar(words))]: number of
## items to replace is not a multiple of replacement length

## Warning in word[i] <- words[nchar(words) == max(nchar(words))]: number of
## items to replace is not a multiple of replacement length
```

```
final.longest_word <- data.frame(max_length = max_length,word=word,id = id)

#####FIND LONGEST SENTENCE in 10 docs
topdocs
```

```
##
##
##
##
##
## "American Express Co remained silent on\nmarket rumors it would spinoff all or part of its Shearson\r
##
##
##
##
##
##
##
##
##
##
##
##
##
##
```

```
topdocs[[2]]
```

```
## [1] "ChemLawn Corp <CHEM> could attract a\nhigher bid than the 27 dlrs per share offered by Waste\nMa
```

```
names(topdocs[1])
```

```
## [1] "10"
```

```
#split into sentences
get_sentence_df_func <- function(x){
  sentence_df <- data.frame(sentence = character(0),
                            document = character(0))
  for (i in 1:10){
    temp <- data.frame(sentence=tokenize_sentences(x[[i]][[1]]),id=names(x[i]))
    sentence_df <- rbind(sentence_df,temp)
  }
  return(sentence_df)
}

#ALL sentences in the top 10 documents
text_sent <- get_sentence_df_func(topdocs)

#word count for each sentence
text_sent$sentence <- as.character(text_sent$sentence)
count <- c()
sapply(strsplit(text_sent$sentence[23], " "), length)
```

```
## [1] 2
```

```
for (i in 1:nrow(text_sent)){
  count[i] <- sapply(strsplit(text_sent$sentence[i], " "), length)
}

#length of each sentence in each document
count_sentences <- cbind(count,text_sent)

#top 10 lengths
longest_10 <- count_sentences %>% group_by(id) %>%
  arrange(desc(count)) %>% top_n(1,count) %>% distinct(id)

#remove punctuation for each sentence
#remove punctuation from topdocs
str(count_sentences$sentence)
```

```
##  chr [1:495] "computer terminal systems inc said" ...
```

```
#loop to remove punctuation from each sentence
nopunct <- c()
for (i in 1:nrow(count_sentences)){
  nopunct[i] <-  ( gsub("[[:punct:]]", "", count_sentences$sentence[i]) )
}
#bind final output together together
final_nopunct_df <- cbind(nopunct,count_sentences)
```

The project helped us learn a lot about text analytics and key principals of analyzing unstructured text.We identified three key areas this project helped you learn about data science includes (1) the general approach to breaking down texts in R using Corpuses and tokens; (2) The exploratory analysis and derived insights that can be accomplish on a text documents through word counts, frequencies, associations, and character lengths; (3) we were able to learn how to apply data mining techniques to text analytics for deeper insights such as clustering (hierarchical and kmeans).