Project #3

Nisha Iyer, Rachel Jordan, Sam Dooley

April 30, 2016

We will perform analysis on a corpus of 50 documents from the acq dataset.

```
data("acq")

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

Explore using functions from Lecture 7

We can reference information about the document with any of the following commands.

```
#this tell us what information (metadata) about our documents.
# For example, how many chars are within each doc.
alldocs <- 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
```

```
# get the first document
text1 <- ACQ[[1]]

# get the id from the second document
id.2 <- ACQ[[1]]$meta$id
id.2 <- meta(ACQ[[1]], "id") #this is another way to reference
```

The command `meta` will return understandable information about the documents. It will tell you who wrote the article, when it was written, the heading of the article, its language, its origin, etc. This can be useful when searching for particular documents or languages.

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)
```

The DocumentTermMatrix lists as its rows the documents in the corpus, and as it columns the words of the corpus. entries of this matrix are numbered values that indicate how many times given document (row) contains a given a word (column). This can be seen here:

```
inspect(ACQdoc[1:6,1:7])
```

```
## <<DocumentTermMatrix (documents: 6, terms: 7)>>
## Non-/sparse entries: 2/40
## Sparsity           : 95%
## Maximal term length: 11
## Weighting          : term frequency (tf)
##
##     Terms
## Docs -laval .125 .3322 "...that "(american) "any "bridge"
##   10      0    1     0        0           0    0        0
##   12      0    0     0        0           0    0        0
##   44      0    0     0        0           0    0        0
##   45      0    0     0        0           0    1        0
##   68      0    0     0        0           0    0        0
##   96      0    0     0        0           0    0        0
```

termFreq tells us more about an individual doc/text such as term freq within the document. We can also then rank the terms from most frequent to least.

```
test1tf <- as.data.frame(termFreq(text1))
#rank words most to least
rank_words <- as.data.frame(test1tf[order(test1tf, decreasing = T),])
head(rank_words)
```

```
##          test1tf[order(test1tf, decreasing = T), ]
## the                                             15
## said                                             7
## and                                              6
## computer                                         6
## its                                              5
## for                                              4
```

The tm_map and content_transformer transforms the data such as converting the terms to lower case. Converting text to lower case is helpful for matching words that can have different capitalization schemes. For instance, a word might appear at the beginning of the sentence, but it is important to be able to count that word as the same as if it were not capitalized.

```
# to lower case
ACQlow <- tm_map(ACQ, content_transformer(tolower))
```

We also remove characters that are Ebglish letters or spaces. This removes punctuation from the text that can cause issues later on. We note that this is not the ideal method for removing punctuation as hyphenated words like cross-sectional would be distorted to something that isn't a word. For the purposes here, this technique is okay.

```
#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))
```

We also run into a problem if we wanted to analyze frequency of words. The problem is that some words are just obviously more frequent: the, a, of, etc. Thus, we create a class of words, called *stopwords* - which is a part of the tm and quanteda packages - which we wish to remove from the corpus.

```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)
```

This creates an interesting point of analysis: *How much information or text do we lose when we remove stopwords?*

```r
#here we can look at the first two text docs and see how the word count (char) differs
inspect(ACQ[1])
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 1
##
## $`reut-00001.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 1287
```

```r
inspect(ACQstop[1])
```

```
## <<VCorpus>>
## Metadata:  corpus specific: 0, document level (indexed): 0
## Content:  documents: 1
##
## $`reut-00001.xml`
## <<PlainTextDocument>>
## Metadata:  15
## Content:  chars: 1030
```

We see that the first document of ACQ drops from 1287 characters to 1030 characters. This means that this document had abou at 20% reduction in the number of characters. We see that this is a pretty stable reduction across this corpus.

Now that we have removed the document's punctuation and stopwords, we put that corpus back into a `DocumentTermMatrix`.

```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)
```

We also use the function `findFreqTerms` to look through the `DocumentTermMatrix` to find those words that were used a certain number of times or were used in a range of times.

```r
#find terms with a frequency between 15 and 18
freq.terms <- findFreqTerms(ACQdm2, lowfreq=15, highfreq = 18)
freq.terms
```

```
## [1] "acquire"  "bank"     "business" "one"      "rmj"      "value"
## [7] "viacom"
```

We also have a function that will find words in your corpus - really your `DocumentTermMatrix` - and determine which of those words are Associates to another word above a given correlation score. We note that this is a correlation based of of how words are used in the `DocumentTermMatrix`, not similarity of the string like a Levenshtein distance or something.

```r
#the Assocs function finds associations with terms, such as states or year
findAssocs(ACQdm2, "states", 0.6)
```

```
## $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
##          0.63          0.62          0.62          0.60
```

We thus conclude that the different functions allow 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)

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

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

We see from the above that the document IDs in the from the corpus' metadata are listed below. The order is in decreasing order by number of words.

```r
top10
```

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

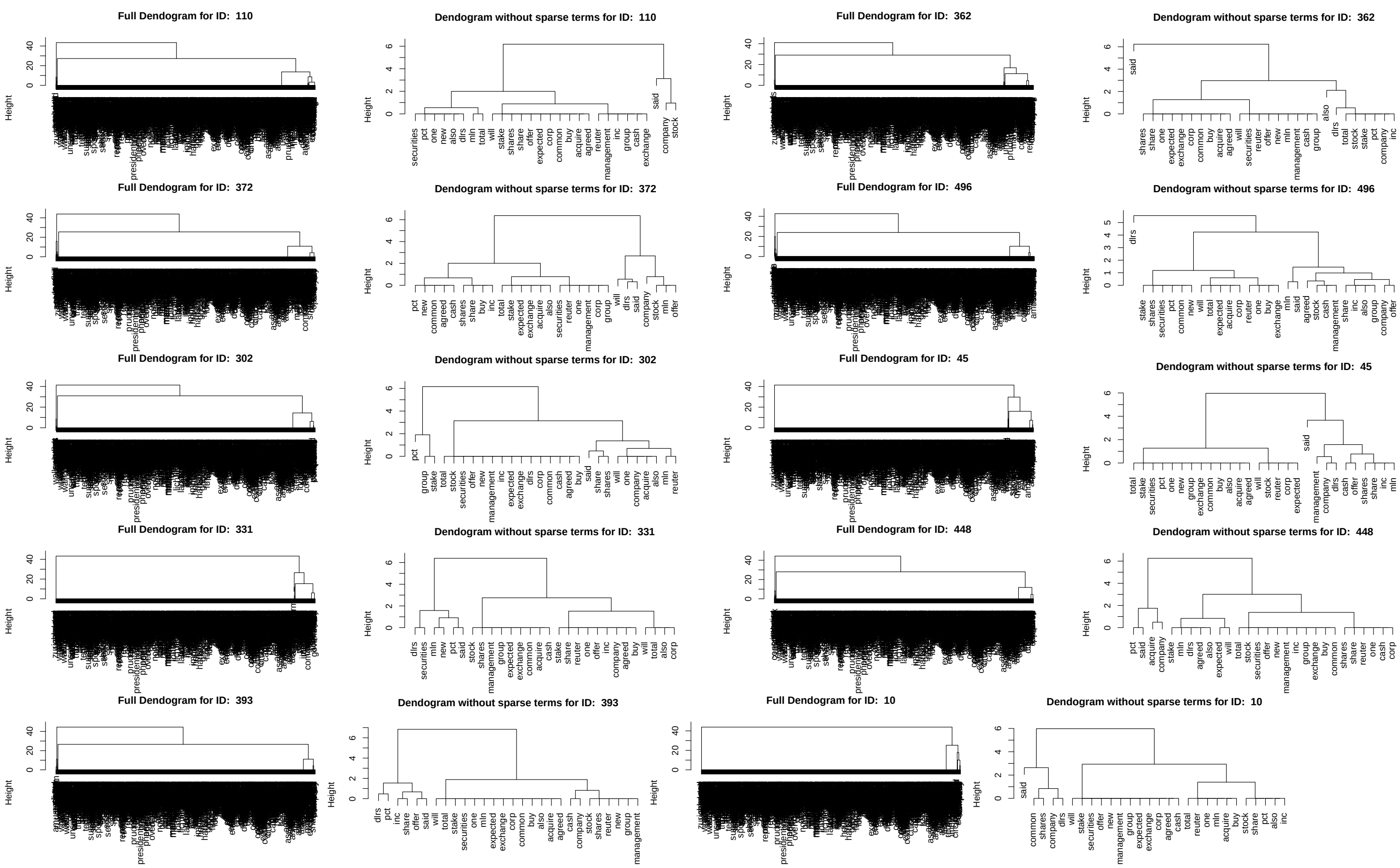
For each document work through the examples given in Lecture 7 to display the dendrogram and the WordCloud

Both the dendograms and the word clouds analyzes the original corpus without punctuation or stopwords. We decided to remove punctuation and stopwords for the visualization because we are not interested in the interaction of common English words. Rather we prefer to ignore the punctuation and stopwords.

For the dendogram, we provide two rednerings. The first dendogram uses all the terms from the corpus without punctuation and stopwords. This reveals very little information as it has all 1,502 words displayed in a dendogram. The dendogram becomes very messy and does not reveal anything interesting about the document. So, we include a dendogram which removes sparse terms at a sparse level of 0.8. This reduces the `DocumentTermMatrix` to only 28 terms. This then makes the dendogram much easier to interpret.

These are dendograms for each of the 10 chosen documents with their ID listed in the title of the figure.

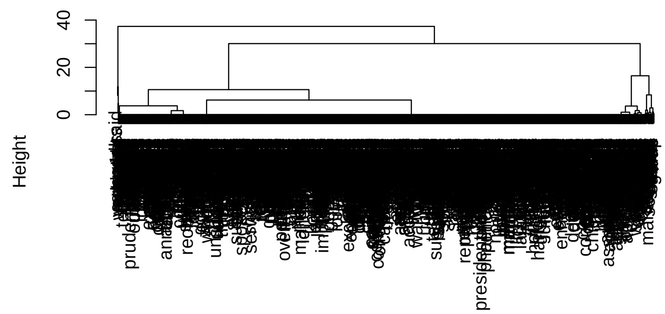
```r
#top 10 dendogram, 1 for each of the top 10 documents
top10.dendogram <- function(doc)
{
  # full dendogram
  acq.mat <- as.matrix( ACQdm2[doc,] )
  distMatrix <- dist(scale(acq.mat[doc,]))
  fit <- hclust(distMatrix, method = "ward.D2")
  plot(fit,main = paste("Full Dendogram for ID: ", doc), xlab ="", sub = "")

  # dendogram with sparse terms removed
  acq.sp <- removeSparseTerms(ACQdm2, sparse = .8)
  acq.mat <- as.matrix( acq.sp[doc,] )
  distMatrix <- dist(scale(acq.mat[doc,]))
  fit <- hclust(distMatrix, method = "ward.D2")
  plot(fit,main = paste("Dendogram without sparse terms for ID: ", doc), xlab = "", sub = "")
}
for (i in 1:10){
  top10.dendogram(top10[i])
}
```



This is a dendogram rendering of the top ten documents in the corpus, without punctuation or stopwords:

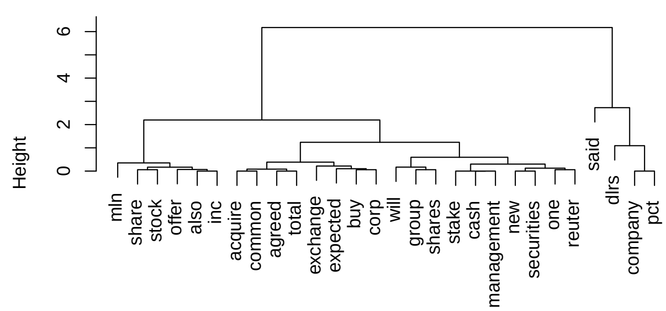
```r
# full dendogram
acq.mat <- as.matrix(ACQdm2)
distMatrix <- dist(scale( colSums(acq.mat[top10,]) ))
fit <- hclust(distMatrix, method = "ward.D2")
plot(fit,main = paste("Full Dendogram for Top 10 Documents"), xlab ="", sub = "")
```

Full Dendogram for Top 10 Documents



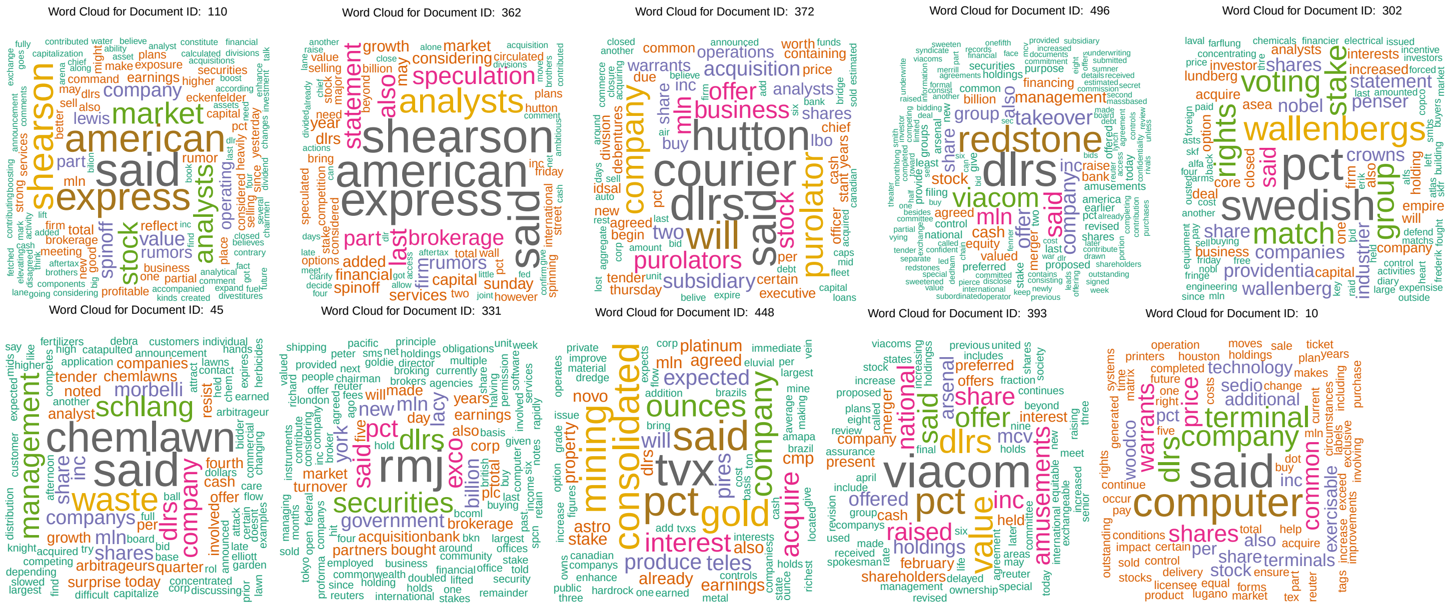

```r
# dendogram with sparse terms removed
acq.sp <- removeSparseTerms(ACQdm2, sparse = .8)
acq.mat <- as.matrix(acq.sp)
distMatrix <- dist(scale( colSums(acq.mat[top10,]) ))
fit <- hclust(distMatrix, method = "ward.D2")
plot(fit,main = paste("Dendogram for Top 10 without Sparse Terms"), xlab ="", sub = "")
```

Dendogram for Top 10 without Sparse Terms



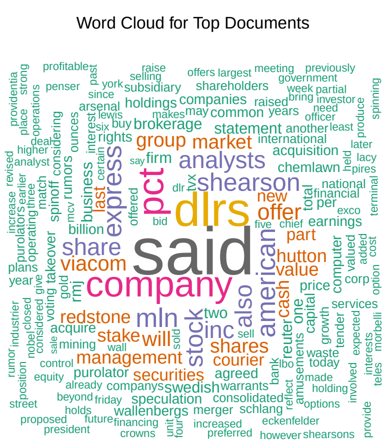
These are word clouds for the individual top ten documents.

```r
#word cloud for top 10
wordcloud.func <- function(ACQstop, doc)
{
  dtm <- TermDocumentMatrix(ACQstop)
  v <- as.matrix(dtm[,doc])
  set.seed(1234)

  layout(matrix(c(1, 2), nrow=2), heights=c(0.5, 4.5))
  par(mar=rep(0, 4))
  plot.new()
  text(x=0.5, y=0.5, paste("Word Cloud for Document ID: ",doc) )
  wordcloud(words = rownames(v), freq = v, 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,top10[i])
}
```



This is a wordcloud for the top ten documents combined.

```
dtm <- TermDocumentMatrix(ACQstop)
v <- t( as.matrix(dtm[,top10]) )
v <- colSums(v)
set.seed(1234)

layout(matrix(c(1, 2), nrow=2), heights=c(0.5, 4.5))
par(mar=rep(0, 4))
plot.new()
text(x=0.5, y=0.5, "Word Cloud for Top Documents" )
wordcloud(words = names(v), freq = v, min.freq = 1,
          max.words=200, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```



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

For this task, we will use the maximum entropy parsers available in the `openNLP`. The results are printed in the table after the following code. We calculate the longest sentence by characters and by consituents (words, etc.). If these two sentences are different, we note that, otherwise, we just list the longest sentence.

```
# find the longest word and longest sentence in each document from the 10 largest documents
for (i in 1:10) {
  ind <- top10[i]
```

```r
  s <- as.String(acq[[ind]]$content)

  ##longest sentence by characters
  sent_token_annotator <- Maxent_Sent_Token_Annotator()
  a1 <- sent_token_annotator(s)
  l <- a1$end - a1$start # table of sentence lengths
  ls.i <- which.max(l) #index of longest sentence by characters
  ls <- as.String( s[a1][ls.i] ) #longest sentence by characters

  ##longest sentence by constituents
  word_token_annotator <- Maxent_Word_Token_Annotator()
  a2 <- word_token_annotator(s, a1)
  a2 <- a2[a2$type=="sentence"]
  l.w <- as.matrix( lapply(a2$features, function(x) length(x$constituents)) ) #sent length
  l.w.ind <- which.max( l.w )
  ls.w <- as.String( s[a1][l.w.ind] )

  ##longest word
  word_token_annotator <- Maxent_Word_Token_Annotator()
  a2 <- word_token_annotator(s, a1)
  a2 <- a2[a2$type=="word"]
  lw.i <- which.max(a2$end - a2$start) #index of longest sentence
  lw <- s[a2][lw.i] #longest sentence

  # print everything so that it's pretty
  print( as.String( paste("Document ID: ", ind) ) )
  print( as.String( paste( "\tLongest Word:\t", lw) ) )
  if (ls == ls.w) {
    print( as.String( paste( "\tLongest Sentence:\t", ls) ) )
  } else {
    print( as.String( paste( "\tLongest Sentence by nchar:\t", ls) ) )
    print( as.String( paste( "\tLongest Sentence by words:\t", ls.w) ) )
    print( as.String(""))
  }
  print( as.String(""))

}
```

```
## Document ID:  110
##  Longest Word:    Prudential-Bache
##  Longest Sentence:    American Express Co remained silent on
## market rumors it would spinoff all or part of its Shearson
## Lehman Brothers Inc, but some analysts said the company may be
## considering such a move because it is unhappy with the market
## value of its stock.
##
## Document ID:  362
##  Longest Word:    Prudential-Bache
##  Longest Sentence by nchar:   In a joint statement, American Express and Shearson said
## the actions under consideration are an integral part of
## American Express' worldwide financial services strategy and
## that the two companies have been having both internal and
## external discussions on the matters.
##  Longest Sentence by words:   American Express Co, rumored to be
## considering a spinoff of part of Shearson Lehman Brothers Inc,
## said it is studying a range of options for its brokerage unit
## that could improve Shearon's access to capital and help it meet
## broadening international competition.
##
```

```
##
## Document ID:  372
##  Longest Word:    Jersey-based
##  Longest Sentence:    If all the shares of Purolator are tendered, shareholders
## would receive for each share 29 dlrs cash, six dlrs in
## debentures, and a warrant to buy shares in a subsidiary of PC
## Acquisition containing the U.S. courier operations.
##
## Document ID:  496
##  Longest Word:    confidentiality
##  Longest Sentence:    The Redstone group, which has a 19.5 pct stake in Viacom,
## and the management group, which has a 5.4 pct stake, have both
## agreed not to buy more shares of the company until a merger is
## completed, unless the purchases are part of a tender offer for
## at least half of the outstanding stock.
##
## Document ID:  302
##  Longest Word:    concentrating
##  Longest Sentence:    But analysts say the Wallenbergs' position in the
## electrical engineering firm ASEA AB <ASEA ST> is also too small
## at 12.6 pct of the voting rights and there has been growing
## speculation that the group will be forced to sell off fringe
## interests to protect its core activities.
##
## Document ID:  45
##  Longest Word:    over-the-counter-
##  Longest Sentence by nchar:   Both Schlang and Morbelli noted that high growth rates had
## catapulted ChemLawn's share price into the mid-30's in 1983 but
## the stock languished as the rate of growth slowed.
##  Longest Sentence by words:   "I think they will resist it," said Elliott Schlang,
## analyst at Prescott, Ball and Turben Inc. "Any company that
## doesn't like a surprise attack would."
##
##
## Document ID:  331
##  Longest Word:    International
##  Longest Sentence:    <Exco International Plc>, a subsidiary of
## British and Commonwealth Shipping Co Plc <BCOM.L>, said it had
## agreed in principle to buy an 80 pct stake in <RMJ Holdings
## Corp> for about 79 mln dlrs.
##
## Document ID:  448
##  Longest Word:    <Consolidated
##  Longest Sentence:    <Consolidated TVX Mining Corp> said it
## agreed to issue 7.8 mln treasury shares to acquire interests in
## three gold mining companies in Brazil and an option to increase
## the company's interest in a platinum property.
##
## Document ID:  393
##  Longest Word:    International
##  Longest Sentence:    Viacom said MCV Holdings, a group which includes the
## company's senior management and the Equitable Life Assurance
## Society of the United States, raised the value of its offer by
## increasing the value of the preferred being offered to 8.50
## dlrs from 8.00 dlrs a share and raising the ownership in the
## new company to be held by present Viacom shareholders to 45 pct
## from 25 pct.
##
## Document ID:  10
##  Longest Word:    reorganization
```

```
##  Longest Sentence by nchar:   Computer Terminal said Sedio also has the right to buy
## additional shares and increase its total holdings up to 40 pct
## of the Computer Terminal's outstanding common stock under
## certain circumstances involving change of control at the
## company.
##  Longest Sentence by words:   Computer Terminal Systems Inc said
## it has completed the sale of 200,000 shares of its common
## stock, and warrants to acquire an additional one mln shares, to
## <Sedio N.V.> of Lugano, Switzerland for 50,000 dlrs.
```

Print a table of the length of each sentence in each of the 10 documents.

We print a table of sentence length for each document, by nchar and by number of words. Since the sentences are long, we only print the first 45 characters.

```r
d.full <- data.frame()
for (i in 1:10) {
  ind <- top10[i]
  s <- as.String(acq[[ind]]$content)

  ##longest sentence by characters
  sent_token_annotator <- Maxent_Sent_Token_Annotator()
  a1 <- sent_token_annotator(s)
  l <- a1$end - a1$start # table of sentence lengths

  ##longest sentence by constituents
  word_token_annotator <- Maxent_Word_Token_Annotator()
  a2 <- word_token_annotator(s, a1)
  a2 <- a2[a2$type=="sentence"]
  l.w <- as.matrix( lapply(a2$features, function(x) length(x$constituents)) ) #sent length


  d<-data.frame(id=ind,lenbychar=l,lenbyword=l.w,sent=sapply( s[a1], function(x) substr(x,0,45)))
  d.full <- rbind(d.full,d)
}
rownames(d.full) <- 1:dim(d.full)[1]
print(d.full)
```

```
##      id lenbychar lenbyword                                          sent
## 1   110       241        45 American Express Co remained silent on\nmarket
## 2   110       188        34  American Express stock got a lift from the ru
## 3   110       111        20  The rumor also was accompanied by talk the fi
## 4   110        90        18  American Express closed on the New York Stock
## 5   110        70        13  American Express would not comment on the rum
## 6   110       147        25  Analysts said comments by the company at an a
## 7   110       154        26  At the meeting, company officials said Americ
## 8   110       142        27  Yesterday, Shearson said it was elevating its
## 9   110        74        13 It also created four new\npositions for chairm
## 10  110       122        22  Analysts speculated a partial spinoff would m
## 11  110       181        34  Some analysts, however, disagreed that any sp
## 12  110       118        24  "I think it is highly unlikely that American 
## 13  110        88        17  He questioned what would be a better investme
## 14  110       126        27  Several analysts said American Express is not
## 15  110       169        33  But others believe the company could very wel
## 16  110       134        20  Larry Eckenfelder of Prudential-Bache Securit
## 17  110        91        19  "Shearson being as profitable as it is would 
## 18  110        49        12 Shearson's book value is in\nthe 1.4 mln dlr r
## 19  110       130        24 Shearson in the market place would\nprobably b
## 20  110        87        15  Some analysts said American Express could use
```

```
## 21  110        60          11  "They have enormous internal growth plans tha
## 22  110       131          25  You want your stock to reflect realistic valu
## 23  110        34           6           Hutton Group analyst Michael Lewis.
## 24  110       133          27  "They've outlined the fact that they're inves
## 25  110        80          16 "...That does not preclude acquisitions and\nd
## 26  110       196          34  Lewis said if American Express reduced its ex
## 27  110        70          15  "It could find its true water mark with a les
## 28  110       166          30  The value of the other components could comma
## 29  110       107          20  Lewis said Shearson contributed 316 mln in af
## 30  110         5           1                                          Reuter
## 31  362       260          46 American Express Co, rumored to be\nconsiderin
## 32  362       266          43  In a joint statement, American Express and Sh
## 33  362       164          28  American Express said no decision has been re
## 34  362       206          37  Last week, rumors circulated on Wall Street t
## 35  362       124          21 Analysts said the\nspeculation also focused on
## 36  362       156          27  There was some speculation that American Expr
## 37  362       149          27  American Express said in the statement on Sun
## 38  362       170          34 The company also remained\nsilent last Thursda
## 39  362       103          18  It said it issued the statement on Sunday bec
## 40  362       194          36  Analysts have been divided on whether it make
## 41  362       178          31  Some analysts said American Express may consi
## 42  362        83          16  Shearson contributed 316 mln dlrs of American
## 43  362       144          25  American Express' ambitious plans for interna
## 44  362        92          17 Analysts speculated that all of\nShearson woul
## 45  362        55          12  To some however, the need for added capital i
## 46  362       137          26  "(American) Express is in a position where th
## 47  362        84          13  Analysts said rumors were fed by the reorgani
## 48  362        88          16 Chief operating officer Jeffrey\nLane got the 
## 49  362       174          29  The reorganization also created four new posi
## 50  362       102          18  Analysts, contacted on Sunday said the statem
## 51  362       237          41 It does\nconfirm, however, that the financial 
## 52  362        52          11  Late last year, Shearson's takeover offer to 
## 53  362       172          29  Hutton Group Inc was rejected by Hutton, and 
## 54  362         5           1                                          Reuter
## 55  372       143          25 New Jersey-based overnight messenger\nPurolato
## 56  372        72          13 Hutton LBO Inc\nand certain managers of Purola
## 57  372        65          13  Analysts have said that Purolator has been fo
## 58  372       132          22 Purolator announced earlier it was mulling a\n
## 59  372        44           9  Hutton LBO, a wholly owned subsidiary of E.F.
## 60  372        55          12 Hutton Group\nInc, will be majority owner of t
## 61  372       155          32  Hutton said the acquiring company, PC Acquisi
## 62  372       163          29 The rest of the shares\nwill be purchased for 
## 63  372       225          42  If all the shares of Purolator are tendered, 
## 64  372       205          37  Hutton said in the merger shareholders would 
## 65  372        78          16 Hutton said the company has valued\nthe warran
## 66  372        55          11  Purolator's stock price closed at 35.125 dlrs
## 67  372       117          26  While some analysts estimated the company was
## 68  372        47           9  This follows sales of two other Purolator uni
## 69  372       132          25 It agreed\nrecently to sell its Canadian Couri
## 70  372        90          16  Purolator retains its Stant division, which m
## 71  372        66          13 A Hutton spokesman said the\nfirm is reviewing
## 72  372       169          33  Purolator's courier business has been lagging
## 73  372         3           1                                            E.F.
## 74  372        73          14  Hutton will provide 279 mln dlrs of its funds
## 75  372       127          24 This so-called "bridge" financing\nwill be rep
## 76  372        88          16 Hutton LBO is committed to\nkeeping the courie
## 77  372       142          27  "Purolator lost 120 mln dlrs over the last tw
## 78  372        73          17 We belive it will be a very\nserious competito
## 79  372       117          21  William Taggart, chief executive officer of U
## 80  372       183          34  The tender offer will be conditioned on a min
```

```
## 81  372       173         28  The offer will begin Thursday, subject to cle
## 82  372         5          1                                         Reuter
## 83  496       201         39 Investor Sumner Redstone, who leads\none of th
## 84  496        99         17  In a filing with the Securities and Exchange 
## 85  496       139         28 National Amusements\nInc, a theater chain oper
## 86  496       100         20  Redstone also raised the face value of the pr
## 87  496       227         40  The Redstone offer, which is being made throu
## 88  496       242         42  Viacom said earlier today it received revised
## 89  496       135         26  The company did not disclose the details of t
## 90  496       285         60  The Redstone group, which has a 19.5 pct stak
## 91  496       149         24  The two rivals also signed confidentiality ag
## 92  496       166         33  In his SEC filing, Redstone, who estimated hi
## 93  496       237         44  Besides the financing it would raise through 
## 94  496       225         40  Merrill Lynch, Pierce Fenner and Smith Inc ha
## 95  496        93         17  Redstone said his group would contribute more
## 96  496       155         30  The Redstone equity contribution to the takeo
## 97  496       173         31  The new offer, the second sweetened deal Reds
## 98  496       229         46  Last week, the management group submitted wha
## 99  496        61         12 Redstone's\nprevious offer had been valued at 
## 100 496         5          1                                         Reuter
## 101 302       203         38 Sweden's Wallenberg group fought back\na bid b
## 102 302       224         39  A statement issued by the Wallenberg holding 
## 103 302       175         35  Thre Wallenbergs paid Nobel Industrier <NOBL 
## 104 302        78         16  Swedish Match's B shares open to foreign buye
## 105 302       115         23 The A shares -- with increased voting\nrights 
## 106 302       236         45  The statement said the deal increased Investo
## 107 302       109         21  The Wallenbergs' stake in Swedish Match had p
## 108 302       238         43  The Swedish Match deal will cost the Wallenbe
## 109 302       221         42  The Wallenbergs originally sold Nobel Industr
## 110 302       172         32  Since then, the Wallenbergs were ousted as th
## 111 302       169         31  Lundberg, a Zurich-based Swedish property tyc
## 112 302       165         29  During 1986, the Wallenbergs have been concen
## 113 302       275         50  But analysts say the Wallenbergs' position in
## 114 302         5          1                                         REUTER
## 115  45       143         26 ChemLawn Corp <CHEM> could attract a\nhigher b
## 116  45       151         27  Shares of ChemLawn shot up 11-5/8 to 29-3/8 i
## 117  45       145         31  "This company could go for 10 times cash flow
## 118  45        85         16 Waste Management's tender offer,\nannounced be
## 119  45        76         16  "This is totally by surprise," said Debra Str
## 120  45       113         18 The company's board held a regularly\nschedule
## 121  45        79         17  She said a statement was expected but it was 
## 122  45       105         19  She was unable to say if there had been any p
## 123  45       150         34  "I think they will resist it," said Elliott S
## 124  45        96         18  Arbitrageurs pointed out it is difficult to r
## 125  45       106         21 Schlang said ChemLawn\ncould try to find a whi
## 126  45       161         25  Analyst Rosemarie Morbelli of Ingalls and Sny
## 127  45       142         25  ChemLawn, with about two mln customers, is th
## 128  45        49          9 Waste Management is involved in removal of\nwa
## 129  45       158         25  Schlang said ChemLawn's customer base could b
## 130  45       172         32  Both Schlang and Morbelli noted that high gro
## 131  45        73         13  Schlang said the company's profits are concen
## 132  45        98         21  In 1986 ChemLawn earned 1.19 dlrs per share f
## 133  45       112         17  Morbelli noted ChemLawn competes with thousan
## 134  45         5          1                                         Reuter
## 135 331       192         38 <Exco International Plc>, a subsidiary of\nBri
## 136 331       186         36  Exco Chairman Richard Lacy told Reuters the a
## 137 331       119         25  Bank of New York and the partners will retain
## 138 331        99         18  RMJ is the holding company of RMJ Securities,
## 139 331       106         18  It is also involved in broking notes, obligat
## 140 331       190         36  Lacy said Exco had been considering buying a 
```

```
## 141 331        40         10      RMJ was then valued at about 50 mln dlrs.
## 142 331       143         29  B and C managing director Peter Goldie said R
## 143 331       120         24  The company's earnings had not been hit by th
## 144 331       181         34  Lacy said that RMJ employed some 300 people,
## 145 331       163         32  RMJ Securities had offices in New York, where
## 146 331       145         28  It was also given permission last week to ope
## 147 331       112         23  The acquisition would contribute between five
## 148 331         5          1                                          REUTER
## 149 448       212         37 <Consolidated TVX Mining Corp> said it\nagreed
## 150 448       207         35  The company said the transactions will bring
## 151 448        41          8    The company did not give\nspecific figures.
## 152 448       169         33  Consolidated TVX said it will acquire 29 pct
## 153 448       169         37  The company also agreed to acquire a 19 pct s
## 154 448       150         31  In addition, Consolidated TVX said it will ac
## 155 448       151         32  CMP earned 11 mln Canadian dlrs in 1986 and e
## 156 448       169         34  Novo Astro operates Brazil's richest gold min
## 157 448       114         20 Mining of\neluvial surface material produced 2
## 158 448       170         33  It also said Teles Pires Mining controls righ
## 159 448         5          1                                          Reuter
## 160 393       117         18 Viacom International Inc said it\nreceived rev
## 161 393        86         16  The company said the special committee plans
## 162 393       152         27  Viacom said National Amusements' Arsenal Hold
## 163 393        52         10 National Amusements holds\n19.6 pct of Viacom'
## 164 393       239         46  The cash value of the offer was raised to 42.
## 165 393       160         36  The interest rate to be used to increase the
## 166 393       134         23  A Viacom spokesman said the Arsenal Holdings'
## 167 393       373         70  Viacom said MCV Holdings, a group which inclu
## 168 393       107         21  MCV called its previous offer, made February
## 169 393         5          1                                          Reuter
## 170  10       208         40 Computer Terminal Systems Inc said\nit has com
## 171  10       103         20  The company said the warrants are exercisable
## 172  10       240         40  Computer Terminal said Sedio also has the rig
## 173  10       183         38  The company said if the conditions occur the
## 174  10       178         33  Computer Terminal also said it sold the techn
## 175  10        98         19  But, it said it would continue to be the excl
## 176  10       134         23  The company said the moves were part of its r
## 177  10        96         16  Computer Terminal makes computer generated la
## 178  10         5          1                                          Reuter
```

For each word print its part of speech

Again, we use the `openNLP` package and the part of speech tagger to determine these parts of speech. Below is a table of every word in the corpus. There will be duplicates in this list as words can have different parts of speech. For example, the word `total` appears in the text as a `JJ` and a `NN`.

```r
d <- data.frame()
pos_tag_annotator = Maxent_POS_Tag_Annotator()
for (i in 1:10) {

  ind <- top10[i]
  s <- as.String(acq[[ind]]$content)
  s <- tolower(s)

  a1 <- sent_token_annotator(s)

  # For each sentence of each document, remove the punctuation.
  s.cl <- sapply( s[a1], removePunctuation )

  for( j in 1:length(s.cl)) {
```

```r
    sub.s <- as.String(s.cl[j])

    a2 <- word_token_annotator(sub.s, a1)
    a3 <- pos_tag_annotator( sub.s, a = a2)

    words <- as.vector( sub.s[a3] )
    pos <- as.matrix( lapply(a3$features, function(x) x$POS) )

    df <- data.frame(words=words, pos=pos)
    d<-rbind(d,df)
  }
}
nums <-sapply( d$words, function(x) suppressWarnings(!is.na(as.numeric(as.character(x)))))
d <- d[!nums,]
d <- unique(d)
print( d[order(d$words),] )
```

```
##              words  pos
## 31               a   DT
## 12             all   DT
## 1         american   JJ
## 628       american   IN
## 23        analysts  NNS
## 28              be   VB
## 33         because   IN
## 19        brothers  NNS
## 21             but   CC
## 3               co   NN
## 26         company   NN
## 29     considering  VBG
## 2          express   JJ
## 74         express   NN
## 924        express  VBP
## 20             inc    ,
## 597            inc   NN
## 1702           inc   IN
## 3537           inc  VBN
## 35              is  VBZ
## 9               it  PRP
## 16             its PRP$
## 18          lehman   NN
## 7           market   NN
## 27             may   MD
## 32            move   NN
## 15              of   IN
## 6               on   IN
## 13              or   CC
## 14            part   NN
## 4         remained  VBD
## 8           rumors  NNS
## 24            said  VBD
## 1518          said  VBN
## 17        shearson   NN
## 5           silent   JJ
## 22            some   DT
## 11         spinoff   NN
## 43           stock   NN
## 30            such   JJ
## 25             the   DT
```

```
## 36          unhappy   JJ
## 40            value   NN
## 37             with   IN
## 10            would   MD
## 53               as   IN
## 363              as   RB
## 68         boosting  VBG
## 56       calculated  VBD
## 62          command   VB
## 50             from   IN
## 64             good   JJ
## 47              got  VBD
## 49             lift   NN
## 58        partially   RB
## 59           public   JJ
## 745          public   NN
## 52            rumor   NN
## 67          thereby   RB
## 70            total   JJ
## 813           total   NN
## 79      accompanied  VBN
## 77             also   RB
## 90              and   CC
## 91            boost   VB
## 80               by   IN
## 93         dividend   NN
## 83        financial   JJ
## 85             firm   NN
## 84         services  NNS
## 87            split   VB
## 81             talk   NN
## 78              was  VBD
## 103              at   IN
## 96           closed  VBD
## 102        exchange   NN
## 108           heavy   JJ
## 99              new   JJ
## 105              up   IN
## 859              up   RP
## 2586             up   RB
## 109          volume   NN
## 100            york   NN
## 121        activity   NN
## 114         comment   VB
## 113             not   RB
## 129              an   DT
## 140    announcement   NN
## 144         changes  NNS
## 124        comments  NNS
## 138             did  VBD
## 134            fuel   VB
## 133          helped  VBD
## 143      management   NN
## 131         meeting  VBG
## 147         meeting   NN
## 132         tuesday   JJ
## 141       yesterday   NN
## 165       according  VBG
## 157            does  VBZ
## 159           fully   RB
```

```
## 149        officials  NNS
## 162      performance   NN
## 160          reflect   VB
## 166               to   TO
## 155      undervalued  VBN
## 182            added  VBN
## 936            added   JJ
## 188             been  VBN
## 175            chief   NN
## 1582           chief   JJ
## 173        elevating  VBG
## 187              had  VBD
## 178          jeffery   NN
## 179             lane   NN
## 177          officer   NN
## 176        operating  VBG
## 200        operating   NN
## 183         position   NN
## 185        president   NN
## 189           vacant   JJ
## 186            which  WDT
## 197         chairmen  NNS
## 192          created  VBD
## 201        divisions  NNS
## 196              for   IN
## 193             four   CD
## 195        positions  NNS
## 211         contrary   NN
## 208             make   VB
## 209             most  JJS
## 1509            most  RBS
## 213              one   CD
## 205          partial   JJ
## 210            sense   NN
## 203       speculated  VBD
## 214        variation   NN
## 245            about   IN
## 227              any   DT
## 240           center   NN
## 244     contributing  VBG
## 225        disagreed  VBD
## 249         earnings  NNS
## 224          however   RB
## 250             last   JJ
## 247              pct   NN
## 2942             pct   JJ
## 239           profit   NN
## 234            since   IN
## 238           strong   JJ
## 226             that   IN
## 480             that  WDT
## 1762            that   DT
## 251             year   NN
## 271       analytical   CC
## 278           better  JJR
## 512           better   RB
## 262            going  VBG
## 272               he  PRP
## 256           highly   RB
## 252                i  PRP
```

```
## 2384              i   NN
## 279      investment   NN
## 270          lipper   NN
## 268            long   RB
## 267          perrin   RB
## 283      profitable   JJ
## 273      questioned  VBD
## 284      securities  NNS
## 264            sell   VB
## 280            than   IN
## 253           think  VBP
## 257        unlikely   JJ
## 282            very   RB
## 274            what   WP
## 310           asset   NN
## 296            cash   NN
## 293              in   IN
## 298           might   MD
## 294            need   NN
## 985            need  VBP
## 301            only   JJ
## 302          reason   NN
## 286         several   JJ
## 313         believe  VBP
## 320      considered  VBN
## 316           could   MD
## 322          option   NN
## 312          others  NNS
## 325             out   IN
## 333         selling  VBG
## 735         selling   NN
## 324        spinning  VBG
## 332        suggests  VBZ
## 318            well   RB
## 349        believes  VBZ
## 343     eckenfelder   NN
## 353            have   VB
## 424            have  VBP
## 342           larry   JJ
## 360            past   NN
## 1473           past   JJ
## 345 prudentialbache   NN
## 362           being  VBG
## 372             big   JJ
## 370         fetched  VBN
## 377           place   NN
## 373         premium   NN
## 379            book   NN
## 386             dlr   NN
## 385             mln   JJ
## 568             mln   NN
## 387           range   NN
## 378       shearsons  NNS
## 400          bilion   CD
## 406  capitalization   NN
## 401            dlrs  NNS
## 394        probably   RB
## 403           terms  NNS
## 397           three   CD
## 396           worth   JJ
```

```
## 416          capital   NN
## 421           expand   VB
## 422         globally   RB
## 419            plans  VBZ
## 687            plans  NNS
## 415              use   VB
## 425         enormous   JJ
## 427           growth   NN
## 426         internal   JJ
## 430            takes  VBZ
## 423             they  PRP
## 443          ability   NN
## 450             down   IN
## 454               ef   NN
## 449        endeavors  NNS
## 441          enhance   VB
## 447            kinds  NNS
## 438        realistic   JJ
## 452             road   NN
## 439       valuations  NNS
## 433             want  VBP
## 2761            want   VB
## 432              you  PRP
## 434             your PRP$
## 457          analyst   NN
## 456            group   NN
## 455           hutton   NN
## 459            lewis    .
## 492            lewis   NN
## 458          michael   NN
## 477            arena   NN
## 463             fact   NN
## 470           future   NN
## 3912          future   JJ
## 472             goes  VBZ
## 467          heavily   RB
## 476    international   JJ
## 474             into   IN
## 466        investing  VBG
## 461         outlined  VBD
## 465           theyre   DT
## 460           theyve   DT
## 484     acquisitions  NNS
## 487            along   IN
## 486     divestitures  NNS
## 483         preclude   VB
## 489              way   NN
## 515           assets  NNS
## 502        brokerage   NN
## 503         business   NN
## 499         exposure   NN
## 494               if   IN
## 514            other   JJ
## 497          reduced  VBD
## 520          related  VBN
## 519           travel   NN
## 525             find   VB
## 532           lesser  JJR
## 529             mark   NN
## 527             true   JJ
```

```
## 528           water   NN
## 541      components  NNS
## 549      constitute  VBP
## 545          higher  JJR
## 546        multiple   JJ
## 552      percentage   NN
## 570        aftertax   NN
## 566     contributed  VBD
## 581          reuter   NN
## 614          access   NN
## 621      broadening   JJ
## 623     competition   NN
## 618            help   VB
## 612         improve   VB
## 620            meet   VB
## 605         options  NNS
## 585         rumored  VBN
## 613        shearons  NNS
## 601        studying  VBG
## 609            unit   NN
## 634         actions  NNS
## 637             are  VBP
## 656            both   DT
## 652       companies  NNS
## 636   consideration   NN
## 660     discussions  NNS
## 659        external   JJ
## 655          having  VBG
## 639        integral   JJ
## 626           joint   JJ
## 663            matt   NN
## 627       statement   NN
## 647        strategy   NN
## 651             two   CD
## 635           under   IN
## 644       worldwide   JJ
## 688         already   RB
## 683          decide   VB
## 668        decision   NN
## 685          follow   VB
## 669             has  VBZ
## 667              no   DT
## 671         reached  VBN
## 674       strategic   JJ
## 682      ultimately   RB
## 694      circulated  VBD
## 702           giant   NN
## 724        japanese   JJ
## 714     speculation   NN
## 721           stake   NN
## 697          street   NN
## 712           there   EX
## 696            wall   NN
## 692            week   NN
## 731         focused  VBD
## 763            plan   NN
## 793          beyond   IN
## 792              go   VB
## 789       spokesman   NN
## 777          sunday   NN
```

```
## 780              will   MD
## 821             bring   VB
## 843          circulat   NN
## 826             close   RB
## 819              days  NNS
## 807             drove  VBD
## 844         employees  NNS
## 804            friday   NN
## 832            issued  VBD
## 2185           issued  VBN
## 839           similar   JJ
## 802          thursday   NN
## 1627         thursday   RB
## 848           divided  VBN
## 858              give   VB
## 1766             give  VBP
## 867          improved  VBD
## 852             makes  VBZ
## 850           whether   IN
## 864       whollyowned   JJ
## 894         concerned  VBN
## 885          consider   VB
## 887               off   IN
## 2549              off   RP
## 898             price   NN
## 918           billion   CD
## 920               net   NN
## 3106              net   JJ
## 925         ambitious   JJ
## 933          enhanced  VBN
## 971          puzzling   JJ
## 980               can   MD
## 981             raise   VB
## 978             where  WRB
## 997               fed  VBN
## 1000  reorganization   NN
## 1004        wednesday    .
## 996              were  VBD
## 1008          jeffrey   NN
## 1015             post   NN
## 1013       previously   RB
## 1036            allow   VB
## 1042            alone   JJ
## 1041            stand   NN
## 1054          clarify   VB
## 1045        contacted  VBD
## 1052           little   RB
## 1056            weeks  NNS
## 1077      acquisition   NN
## 1070        attempted  VBD
## 1061          confirm   VB
## 1092           global   JJ
## 1081          looking  VBG
## 1076            major   JJ
## 1084              own   JJ
## 1089      positioning   NN
## 1069  unsuccessfully   RB
## 1085            walls  NNS
## 1095             late   JJ
## 1100            offer   NN
```

```
## 2912           offer  VBP
## 1099        takeover   NN
## 1126         another   DT
## 1125      approached  VBD
## 1122        rebuffed  VBN
## 1108        rejected  VBN
## 1123            when  WRB
## 1145        acquired  VBN
## 1142          agreed  VBN
## 1405          agreed  VBD
## 1138            corp    ,
## 2559            corp   JJ
## 1137         courier   NN
## 1154          formed  VBN
## 1133     jerseybased   JJ
## 1135       messenger   NN
## 1134       overnight   JJ
## 1136       purolator   NN
## 1161         certain   JJ
## 1158             lbo   NN
## 1201             lbo  VBD
## 1162        managers  NNS
## 1164      purolators  NNS
## 1165              us  PRP
## 1176            sale   NN
## 1179            time   NN
## 1181       announced  VBD
## 2637       announced  VBN
## 1188             bid   NN
## 1182         earlier  JJR
## 1185         mulling  VBG
## 1192       predicted  VBD
## 1191         wrongly   RB
## 1204           owned  VBN
## 1205      subsidiary   NN
## 1203          wholly   RB
## 1213        majority   NN
## 1214           owner   NN
## 1221       acquiring  VBG
## 1244           begin   VB
## 1245         hursday    .
## 1227          paying  VBG
## 1223              pc    .
## 1265              pc   NN
## 1330              pc   JJ
## 1231             per   IN
## 3308             per   FW
## 1232           share   NN
## 1241          tender   NN
## 2739          tender   JJ
## 1259             buy   VB
## 1267      containing  VBG
## 1253       purchased  VBN
## 1247            rest   NN
## 1250          shares  NNS
## 1257        warrants  NNS
## 1271      operations  NNS
## 1299            ares  NNS
## 1292      debentures  NNS
## 1284            each   DT
```

```
## 1282         receive  VB
## 1298               s PRP
## 1280    shareholders NNS
## 1289             six  CD
## 1279        tendered VBN
## 1308               u  NN
## 1295         warrant  NN
## 1322       aggregate  JJ
## 1323          amount  NN
## 1340          common  JJ
## 1327             due  JJ
## 1318             get  VB
## 1325      guaranteed VBN
## 1347            iary  NN
## 1315          merger  NN
## 1346           subsi NNS
## 1334               t  NN
## 1353          valued VBN
## 1383             30s NNS
## 1375       estimated VBD
## 1385           least JJS
## 1382             mid  JJ
## 1372           while  IN
## 1397         follows VBZ
## 1398           sales NNS
## 1396            this  DT
## 1403           units NNS
## 1424            auto  NN
## 1410        canadian  NN
## 1425         filters NNS
## 1414            onex  VB
## 1406        recently  RB
## 1422            sold VBN
## 3900            sold VBD
## 1435            caps NNS
## 1434         closure  NN
## 1431        division  NN
## 1439             gas  NN
## 1437       radiators NNS
## 1428         retains VBZ
## 1430           stant  JJ
## 1452           stant  NN
## 1440           tanks NNS
## 1448       reviewing VBG
## 1458         lagging VBG
## 1477             add  VB
## 1478             air  NN
## 1479        delivery  NN
## 1483           fleet  NN
## 1482          ground  NN
## 1467            high  JJ
## 1470            paid VBD
## 2223            paid VBN
## 1463          rivals NNS
## 1475           years NNS
## 1495        complete  VB
## 1493           funds NNS
## 1487         provide  VB
## 1497     transaction  NN
## 1515            bank  NN
```

```
## 1500          bridge   NN
## 1508            debt   NN
## 1501       financing   NN
## 1513            form   NN
## 1505           later   RB
## 1510          likely   JJ
## 1516           loans  NNS
## 1507        longterm   JJ
## 1504        replaced  VBN
## 1499        socalled   JJ
## 1522       committed  VBN
## 1531           idsal   NN
## 1524         keeping  VBG
## 1530          warren   NN
## 1543         largely   RB
## 1534            lost  VBD
## 1538            over   IN
## 1556          around   RB
## 1555         turning  VBG
## 1550              we  PRP
## 1558          belive  VBP
## 1565      competitor   NN
## 1564         serious   JJ
## 1574       executive   NN
## 1583       executive   JJ
## 1572         taggart   NN
## 1571         william   NN
## 1594     conditioned  VBN
## 1622      conditions  NNS
## 1616              er   NN
## 1612      expiration   NN
## 1597         minimum   NN
## 1600          thirds  NNS
## 1609       withdrawn  VBN
## 1645           after   IN
## 1630      clearances  NNS
## 1637        commerce   NN
## 1638      commission   NN
## 1641          expire   VB
## 1648        extended  VBN
## 1636      interstate   JJ
## 1646     ommencement   NN
## 1633           staff   NN
## 1628         subject   JJ
## 1647          unless   IN
## 1662         control   NN
## 1697        controls  VBZ
## 3495        controls  NNS
## 1698          dedham   NN
## 1688          filing   NN
## 1659          groups  NNS
## 1670             his PRP$
## 1650        investor   NN
## 1654           leads  VBZ
## 1699       massbased  VBN
## 1667         offered  VBD
## 2572         offered  VBN
## 1652        redstone   NN
## 1651          sumner   NN
## 1669         sweeten   VB
```

```
## 1664          viacom   IN
## 1715          viacom   NN
## 1660           vying  VBG
## 1653             who   WP
## 1701      amusements  NNS
## 1705           chain   NN
## 1700        national   JJ
## 1706        operator   NN
## 1712         portion   NN
## 1704         theater   NN
## 1729            face   NN
## 1737        offering  VBG
## 1733       preferred   JJ
## 2123       preferred  VBN
## 1727          raised  VBD
## 3560          raised  VBN
## 1752         arsenal   NN
## 1774         arsenal   JJ
## 1753        holdings  NNS
## 1750            made  VBN
## 3755            made  VBD
## 1769        onefifth   IN
## 1763         purpose   NN
## 1759             set  VBN
## 1751         through   IN
## 1817       agreement   NN
## 1788            bids  NNS
## 1803       competing  VBG
## 1815          formal   JJ
## 1797             led  VBN
## 1792             mcv   NN
## 1785        received  VBD
## 1786         revised  VBN
## 1783           today   NN
## 1843           today   RB
## 1838           board   NN
## 1835       committee   NN
## 1826         details  NNS
## 1824        disclose   VB
## 1830          offers  NNS
## 1840          review   VB
## 1834         special   JJ
## 1841            them  PRP
## 1880       completed  VBN
## 1893            half   NN
## 1871            more  JJR
## 1896     outstanding   JJ
## 1883       purchases  NNS
## 1876           until   IN
## 1904      agreements  NNS
## 1903 confidentiality   NN
## 1917     information   NN
## 1915            keep   VB
## 1913        provided  VBN
## 3157        provided  VBD
## 1912         records  NNS
## 1918          secret   NN
## 1902          signed  VBD
## 1910         viacoms  NNS
## 1939         america   NN
```

```
## 1929      completing  VBG
## 1941       confident   JJ
## 1927            cost   NN
## 2353            cost   VB
## 1921             sec   JJ
## 1948         besides   IN
## 1973         limited   JJ
## 1967        separate   JJ
## 1957       syndicate   NN
## 2000      commitment   NN
## 1992          fenner   NN
## 1997       increased  VBN
## 1990           lynch   NN
## 1989         merrill   JJ
## 1991          pierce   NN
## 1994           smith   NN
## 2006    subordinated   JJ
## 2021      underwrite   VB
## 1999    underwriting   NN
## 2030      contribute   VB
## 2037          equity   NN
## 2038          toward   IN
## 2049         consist   VB
## 2044    contribution   NN
## 2081         bidding   NN
## 2086        contains  VBZ
## 2074            deal   NN
## 2092       documents  NNS
## 2088           drawn  VBN
## 2080       monthlong   JJ
## 2087           newly   RB
## 2077        proposed  VBN
## 2072          second   JJ
## 2073       sweetened   JJ
## 2082             war   NN
## 2103          called  VBD
## 2116      consisting  VBG
## 2127           eight   CD
## 2100       submitted  VBD
## 2138        previous   JJ
## 2137       redstones  NNS
## 2152            back   RB
## 2178            core   NN
## 2182          empire   NN
## 2160            erik  VBD
## 2159       financier   NN
## 2151          fought  VBD
## 2165           large   JJ
## 2157     londonbased   JJ
## 2169           match   NN
## 2161          penser   NN
## 2163          secure   VB
## 2170            smbs  NNS
## 2171              st   JJ
## 2227              st   NN
## 2481              st   RB
## 2148         swedens  NNS
## 2158         swedish   JJ
## 2180           their PRP$
## 2149      wallenberg   JJ
```

```
## 2188      wallenberg   NN
## 2191              ab   IN
## 2209              ab   NN
## 2442              ab   JJ
## 2211         acquire   VB
## 2194    forvaltnings  NNS
## 2204            held  VBN
## 2659            held  VBD
## 2189         holding  VBG
## 2207      industrier   NN
## 2394      industrier   IN
## 2218               n   NN
## 2206           nobel   NN
## 2196     providentia   NN
## 2217             rig   NN
## 2208          sweden   JJ
## 2200           taken  VBN
## 2216          voting   NN
## 2226            nobl   JJ
## 2221            thre   JJ
## 2222     wallenbergs  NNS
## 2254               b   NN
## 2259          buyers  NNS
## 2263          crowns  NNS
## 2258         foreign   JJ
## 2253          matchs   NN
## 2256            open   JJ
## 3202            open   VB
## 2284            free   JJ
## 2279      restricted  VBN
## 2272          rights  NNS
## 2292       investors  NNS
## 2311            left  VBN
## 2325           pital   NN
## 2324             sha   NN
## 2337        amounted  VBN
## 2381          defend   VB
## 2369       expensise   JJ
## 2383        farflung   NN
## 2363          making  VBG
## 2370           moves  NNS
## 2387         outside   JJ
## 2388       predators  NNS
## 2385             sts  NNS
## 2374      undertaken  VBN
## 2396            arms  NNS
## 2420            asts  NNS
## 2418           atlas   IN
## 2407          buying  VBG
## 2398       chemicals  NNS
## 2419           copco   NN
## 2415             key   JJ
## 2424      koppabergs  NNS
## 2391      originally   RB
## 2405             pay   VB
## 2423           stora   NN
## 2409            volv   NN
## 2408           volvo   NN
## 2446        frederik   JJ
## 2451       incentive   NN
```

```
## 2434         largest  JJS
## 2447        lundberg   NN
## 2455        lundberg  VBG
## 2431          ousted  VBN
## 2441         skanska   DT
## 2443            skbs  NNS
## 2437             skf   NN
## 2438            skfr   NN
## 2427            then   RB
## 2448         wrested  VBD
## 2477            alfa   NN
## 2480            alfs   RB
## 2474           diary   NN
## 2475       equipment   NN
## 2478           laval   NN
## 2462         managed  VBD
## 2459        property   NN
## 2460          tycoon   NN
## 2457     zurichbased   JJ
## 2490        building  VBG
## 2488   concentrating  VBG
## 2482          during   IN
## 2504           heart   NN
## 2499         prevent   VB
## 2501            raid   NN
## 2556      activities  NNS
## 2520            asea   NN
## 2546              ed  VBN
## 2517      electrical   JJ
## 2518     engineering   NN
## 2550          fringe   NN
## 2539         growing  VBG
## 2551       interests  NNS
## 2553         protect   VB
## 2511             say  VBP
## 2692             say   VB
## 2527           small   JJ
## 2526             too   RB
## 2580    arbitrageurs  NNS
## 2562         attract   VB
## 2560            chem   NN
## 2558        chemlawn   NN
## 2574           waste   NN
## 2577             wnx   IN
## 2606       afternoon   NN
## 2602        changing  VBG
## 2598        companys  NNS
## 2603           hands  NNS
## 2591  overthecounter   JJ
## 2585            shot  VBD
## 2592         trading   NN
## 2631     arbitrageur   NN
## 2629          bidder   NN
## 2622       depending  VBG
## 2621         dollars  NNS
## 2615            flow   NN
## 2619           maybe   RB
## 2613           times  NNS
## 2638          before   IN
## 2654           cheml   NN
```

```
## 2651          debra  NN
## 2642        expires VBZ
## 2634    managements NNS
## 2643          march  VB
## 2640        opening  NN
## 2655     pokeswoman  NN
## 2652     strohmaier VBD
## 2649       surprise  NN
## 2647        totally  RB
## 2667     discussing VBG
## 2661      regularly  RB
## 2662      scheduled  JJ
## 2677       expected VBN
## 2687          ready  JJ
## 2672            she PRP
## 2700        between  IN
## 2699        contact  NN
## 2698          prior  JJ
## 2690         unable  JJ
## 2718           ball  NN
## 2713        elliott  NN
## 2717       prescott  NN
## 2710         resist  VB
## 2714        schlang  NN
## 2720         turben  NN
## 2729         attack  NN
## 2725         doesnt  NN
## 2726           like  IN
## 2736      difficult  JJ
## 2732        pointed VBD
## 2757         knight  NN
## 2752            try  VB
## 2756          white  JJ
## 2785       examples NNS
## 2772        ingalls NNS
## 2778             lp VBP
## 2770       morbelli NNS
## 2791         rested VBN
## 2783            rol  NN
## 2781        rollins NNS
## 2769      rosemarie VBD
## 2776  servicemaster  NN
## 2774         snyder  NN
## 2779            svm  JJ
## 2797      customers NNS
## 2805    application  NN
## 2807    fertilizers NNS
## 2810     herbicides NNS
## 2803       involved VBN
## 2812          lawns NNS
## 2808     pesticides NNS
## 2818        removal  NN
## 2820         wastes NNS
## 2825           base  NN
## 2835     capitalize  VB
## 2823      chemlawns NNS
## 2841     commercial  JJ
## 2824       customer  NN
## 2842         distri  NN
## 2839    residential  JJ
```

```
## 2844          system   NN
## 2828        valuable   JJ
## 2833           wants  VBZ
## 2869             ate  VBD
## 2855       catapulted  VBN
## 2867      languished  VBD
## 2861          mid30s  NNS
## 2849           noted  VBD
## 2853           rates  NNS
## 2872          slowed  VBD
## 2879    concentrated  VBN
## 2898              dl   NN
## 2887          earned  VBD
## 2882          fourth   JJ
## 2894            full   JJ
## 2877         profits  NNS
## 2883         quarter   NN
## 2916            care   NN
## 2905        competes  VBZ
## 2910    entreprеuers  NNS
## 2915          garden   NN
## 2909      individual   JJ
## 2913            lawn   NN
## 2917          sevice   NN
## 2907       thousands  NNS
## 2931           bcoml   NN
## 2925         british   JJ
## 2927    commonwealth   NN
## 2919            exco   NN
## 2921             plc   NN
## 2937       principle   NN
## 2945             rmj  NNP
## 2980             rmj   JJ
## 3010             rmj   NN
## 2928        shipping   NN
## 2971             bkn   NN
## 2954        chairman   NN
## 2973       currently   RB
## 2983            hold  VBP
## 2974           holds  VBZ
## 2956            lacy   NN
## 2981        partners  NNS
## 2985       remainder   NN
## 2958         reuters  NNS
## 2955         richard   JJ
## 2957            told  VBD
## 3004          bought  VBN
## 3007            next   JJ
## 2994          retain   VB
## 3001          stakes  NNS
## 3000           these   DT
## 3025         brokers  NNS
## 3023      government   NN
## 3031         broking  VBG
## 3036     instruments  NNS
## 3032           notes  NNS
## 3033     obligations  NNS
## 3037       sponsored  VBN
## 3041        agencies  NNS
## 3040         federal   JJ
```

```
## 3053          broker   NN
## 3072         pacific   NN
## 3071        security   NN
## 3074            spcn   NN
## 3088               c   NN
## 3090        director   NN
## 3092          goldie   NN
## 3107          income   NN
## 3089        managing   NN
## 3091           peter   NN
## 3101            same   JJ
## 3105      suggesting  VBG
## 3129             ago   RB
## 3134         doubled  VBN
## 3125            fees  NNS
## 3122         halving   NN
## 3119             hit  VBN
## 3128          months  NNS
## 3131         volumes  NNS
## 3164       community   NN
## 3158        computer   NN
## 3139        employed  VBD
## 3142          people  NNS
## 3154             sms  NNS
## 3159        software   NN
## 3168         offices  NNS
## 3175        turnover   NN
## 3185             day   NN
## 3188          london   RB
## 3241           basis   NN
## 3233              cs  NNS
## 3227               d  VBD
## 3226              fi   NN
## 3216            five   CD
## 3197           given  VBN
## 3212          lifted  VBN
## 3204          office   NN
## 3198      permission   NN
## 3240        proforma   FW
## 3213         rapidly   RB
## 3206           tokyo   NN
## 3265          brazil   JJ
## 3243    consolidated   JJ
## 3261            gold   JJ
## 3330            gold   NN
## 3270        increase   VB
## 3273        interest   NN
## 3251           issue   VB
## 3245          mining   NN
## 3494          mining  VBG
## 3276        platinum   NN
## 3254        treasury   NN
## 3244             tvx   JJ
## 3291             tvx   TO
## 3319             tvx   NN
## 3343             tvx   IN
## 3285       immediate   JJ
## 3295           metal   NN
## 3296       potential   NN
## 3294        precious   JJ
```

```
## 3286     production   NN
## 3282   transactions  NNS
## 3317        figures  NNS
## 3316       specific   JJ
## 3327            cmp   NN
## 3347    shareholder   NN
## 3346         single   JJ
## 3381       addition   NN
## 3360          astro   IN
## 3441          astro   NN
## 3388              e   NN
## 3373     increasing  VBG
## 3359           novo   NN
## 3376      ownership   NN
## 3405           owns  VBZ
## 3371          pires  NNS
## 3362        private   JJ
## 3390          right   NN
## 3370          teles  NNS
## 3375           tvxs   NN
## 3419        expects  VBZ
## 3423         ounces  NNS
## 3421        produce   VB
## 3436          ounce   NN
## 3449          amapa   DT
## 3453        average   JJ
## 3443        brazils  NNS
## 3454          grade   NN
## 3464       hardrock   NN
## 3447        located  VBN
## 3446           mine   NN
## 3442       operates  VBZ
## 3465         quartz   NN
## 3444        richest  JJS
## 3450          state   NN
## 3461            ton   NN
## 3466           vein   NN
## 3472        eluvial   JJ
## 3474       material   NN
## 3475       produced  VBD
## 3473        surface   NN
## 3509         dredge   NN
## 3500      kilometer   NN
## 3506          river   NN
## 3501        section   NN
## 3574             al   JJ
## 3578          areas  NNS
## 3618   exchangeable   JJ
## 3611              f   NN
## 3606       february   JJ
## 3613       fraction   NN
## 3610             va   FW
## 3648          april   JJ
## 3646        delayed  VBN
## 3655              m   NN
## 3653           nine   CD
## 3633           rate   NN
## 3636           used  VBN
## 3684     areholders  NNS
## 3672      continues  VBZ
```

```
## 3670       holdingss  NN
## 3674         include  VB
## 3682         present  JJ
## 3683            viac  NN
## 3701       assurance  NN
## 3699       equitable  JJ
## 3692        includes VBZ
## 3700            life  NN
## 3719              ng VBG
## 3718        preferre  NN
## 3730         raising VBG
## 3695          senior  JJ
## 3702          societ  NN
## 3705          states NNS
## 3704          united VBN
## 3759           final  JJ
## 3761        revision  NN
## 3790      additional  JJ
## 3798          lugano  NN
## 3796              nv  NN
## 3795           sedio  VB
## 3825           sedio  NN
## 3799     switzerland  NN
## 3770         systems NNS
## 3769        terminal  NN
## 3809     exercisable  JJ
## 3815        purchase  NN
## 3854              ch  NN
## 3852   circumstances NNS
## 3853       involving VBG
## 3846       terminals NNS
## 3875           equal  JJ
## 3890          exceed  VB
## 3866           occur VBP
## 3882          stocks NNS
## 3906             dot  NN
## 3918         houston  NN
## 3908          impact  NN
## 3913    improvements NNS
## 3910       including VBG
## 3907          matrix  NN
## 3902       technolgy  NN
## 3909      technology  NN
## 3919             tex  NN
## 3915          woodco  VB
## 3939          woodco  NN
## 3928        continue  VB
## 3932       exclusive  JJ
## 3934        licensee  NN
## 3957           costs NNS
## 3955         current  JJ
## 3959          ensure  VB
## 3956       operation  NN
## 3960         product  NN
## 3968           forms NNS
## 3966       generated VBD
## 3967          labels NNS
## 3972        printers NNS
## 3969            tags NNS
## 3971          ticket  NN
```

Analyze word frequency using functions from package zipfR

Discussion

The project helped us learn a lot about text analytics and key principals of analyzing unstructured text. Key themes of this project that helped us 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; and (3) learning how to apply data mining techniques to text analytics for deeper insights such as clustering (hierarchical and means).

There were a few key considerations/issues we realized through this project about text analytics within data science. For example, when breaking down text for mining you might go through the exercises of removing punctuation. When removing punctuation you run the risk of losing hyphenated words or variations of words used such as those with apostrophes. Additionally, a common problem is misspellings and variations of spelling of terms or words. For example, when trying to identify key terms and themes through text analytics/text mining you might dilute popular trends based on not summarizing the different variations of spelling of a term into one. For example, if we were analyzing top terms, "profit", "profitable", and "profits" needs to be considered as one term in order to full capture true trends of words. If the variations aren't considered then the total frequencies (therefore top trends and categories) might not get captured.

Lastly, the use of text analytics really depends on what we're trying to accomplish. Word clouds are interesting and good tools for data exploration but may not be helpful nor a tool for one to make actionable decisions. The application and use case of association of terms as well as dendograms are interesting because if someone was interested to categorize or summarize key concepts on a website or through a content service, it may provide actionable insight on where to summarize or collapse specific sub-pages or categories of content and sources/themes.

In summary, we learn a lot about text analytics as it relates to data science. We learned the general approach to breaking down texts in R, how to explore text through different analysis approaches, and how to apply data mining techniques to text analytics for deeper insights such as clustering (hierarchical and kmeans).