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.

```
#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?*

```
#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
```

```
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`.

```
#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.

```
#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"
```

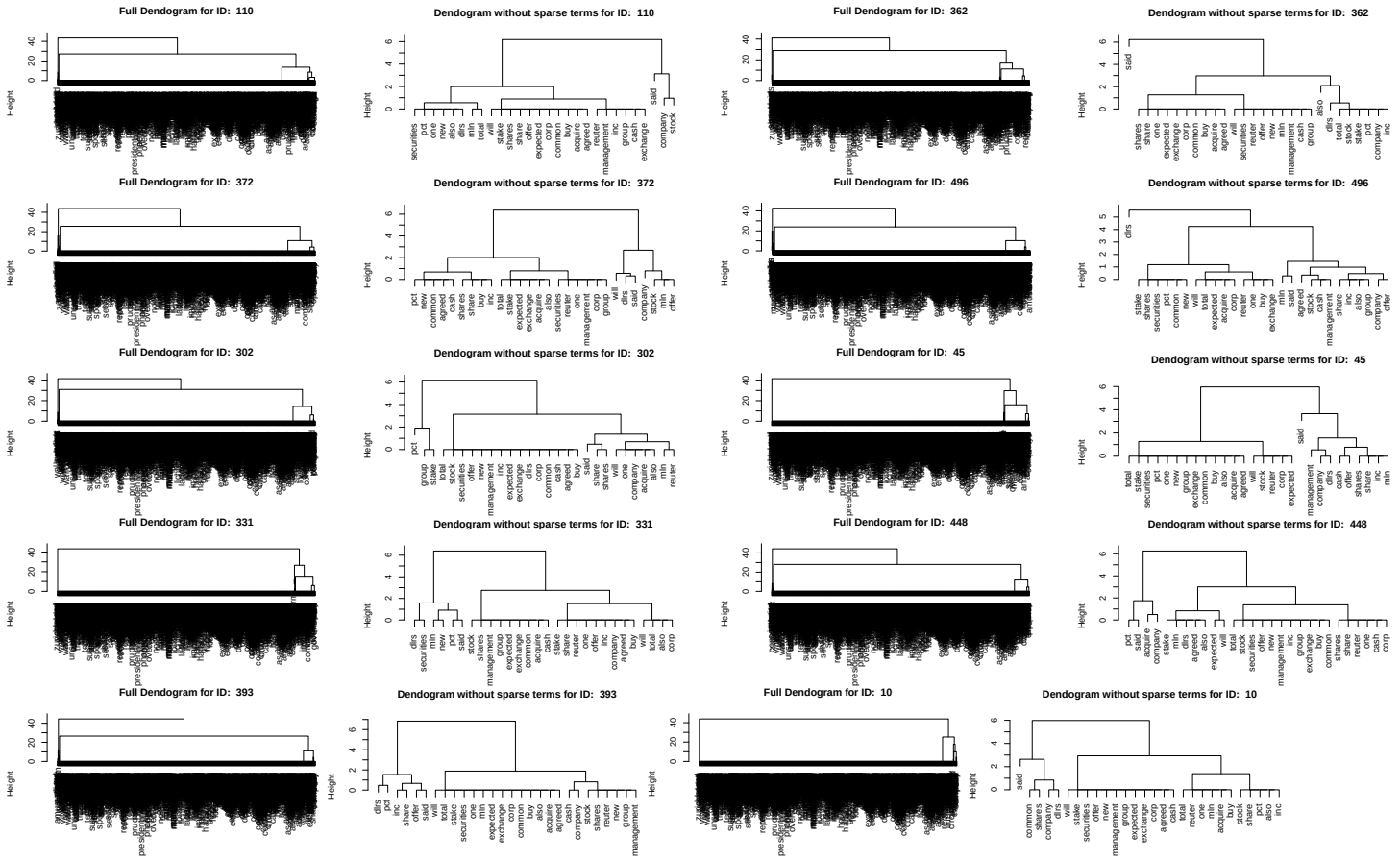
For each document work through the examples given in Lecture 7 to display the dendrogram and the WordCloud

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.

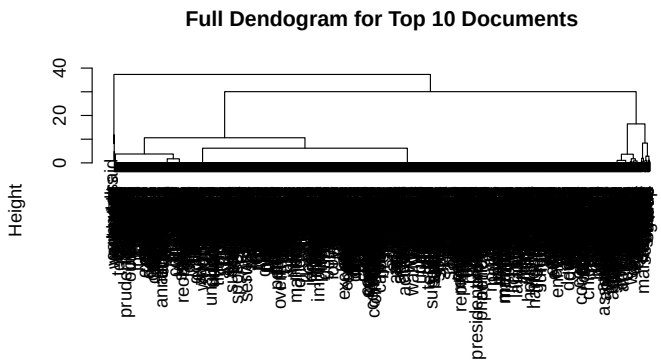
These are dendograms for each of the 10 chosen documents with their ID listed in the title of the figure.

```
#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])
}
```

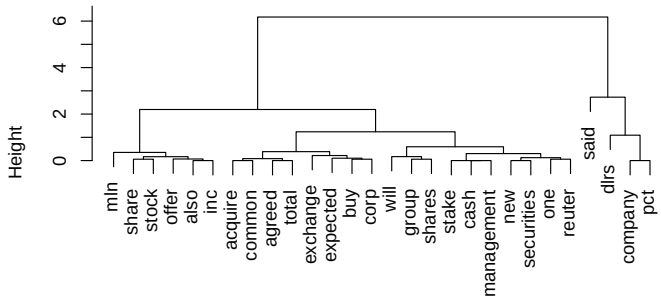


This is a dendogram rendering of the top ten documents in the corpus, without punctuation or stopwords:

```
# 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 = "")
```

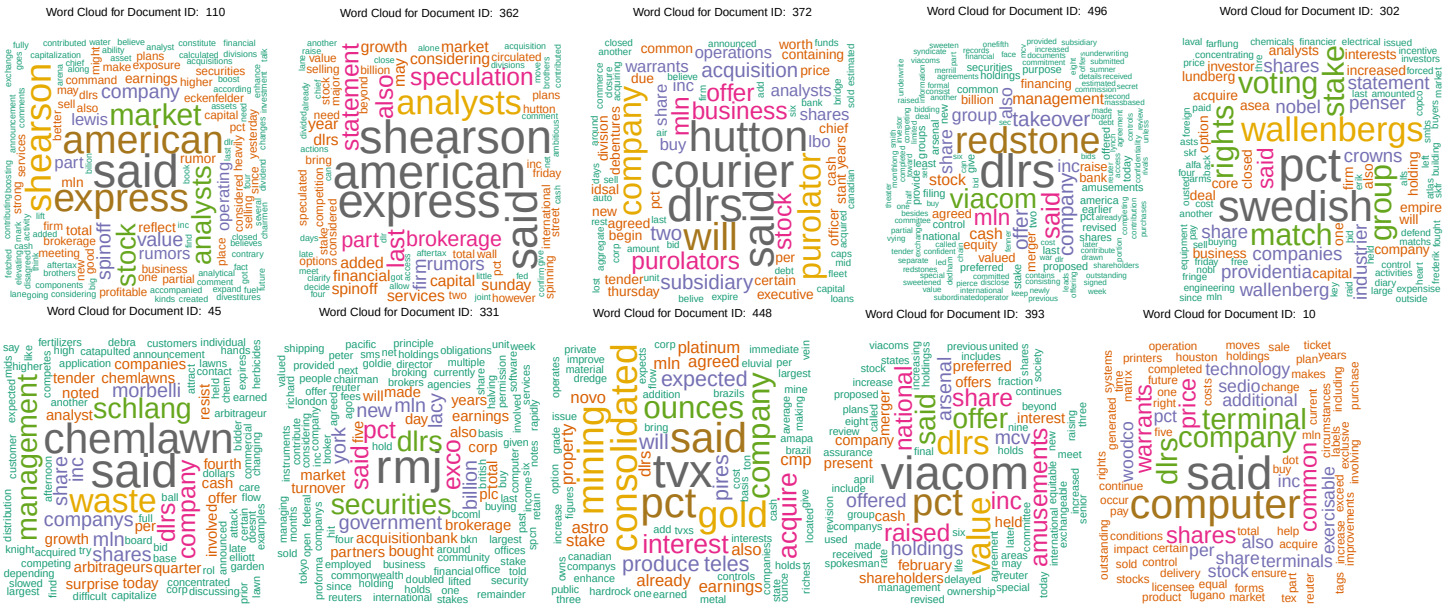



```
# dendogram with sparse terms removed
acq.sp <- removeSparseTerms(ACQdm2, sparse = .8)
acq.mat <- as.matrix(acq.sp)
distMatrix <- dist(scale( colSums(acq.mat[top10,]) ))
fit <- hclust(distMatrix, method = "ward.D2")
plot(fit,main = paste("Dendogram for Top 10 without Sparse Terms"), xlab ="", sub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



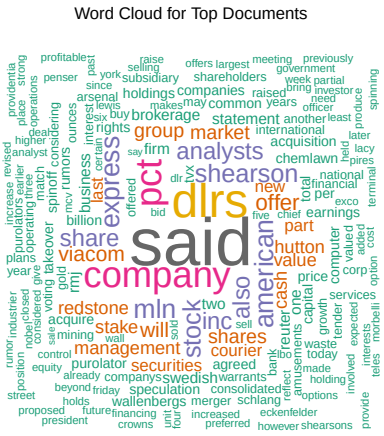
These are word clouds for the individual top ten documents.

```
#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]
  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.

```
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
```

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).