Project #3

Nisha Iyer, Rachel Jordan, Sam Dooley

April 29, 2016

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

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

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

```r
data("acq")

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

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

```r
#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
```

```r
# 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])#the original; 1 with 1287 chars, 2nd with 784 chars
```

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

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

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

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

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

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

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

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

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

```
## $year
##          decide     considering          ending         however
##            0.87            0.75            0.64            0.62
##       shearsons      speculated        spinning          street
##            0.62            0.62            0.62            0.62
##            wall         actions      affiliates           allow
##            0.62            0.61            0.61            0.61
##       ambitious    appreciation      approached       attempted
##            0.61            0.61            0.61            0.61
##           azuma      broadening          caused      circulated
##            0.61            0.61            0.61            0.61
##         clarify        compared     competition       concerned
##            0.61            0.61            0.61            0.61
##         confirm        consider       contacted        decision
##            0.61            0.61            0.61            0.61
##     discussions         divided           drove       employees
##            0.61            0.61            0.61            0.61
##         employs             end        enhanced        external
##            0.61            0.61            0.61            0.61
##             fed         focused          follow        forecast
##            0.61            0.61            0.61            0.61
##           giant          global        improved        industry
##            0.61            0.61            0.61            0.61
##        integral        japanese         jeffrey           kokan
##            0.61            0.61            0.61            0.61
##          little          losses         matters          nippon
##            0.61            0.61            0.61            0.61
##           nkktt     positioning        puzzling        rebuffed
##            0.61            0.61            0.61            0.61
##       recession        rejected         rumored        shearons
##            0.61            0.61            0.61            0.61
##            show         similar         sources       spokesmen
##            0.61            0.61            0.61            0.61
##           stand           steel          steels       strategic
##            0.61            0.61            0.61            0.61
##        strategy      struggling        studying          sunday
##            0.61            0.61            0.61            0.61
##          toshin           tosst      ultimately  unsuccessfully
##            0.61            0.61            0.61            0.61
##           walls       wednesday           weeks         workers
##            0.61            0.61            0.61            0.61
##             yen            yens            firm        services
##            0.61            0.61            0.60            0.60
##     speculation       brokerage        shearson         express
##            0.59            0.58            0.58            0.57
##            last           added        aftertax        american
##            0.57            0.55            0.55            0.55
##        brothers        chairmen     contributed         created
##            0.55            0.55            0.55            0.55
##       divisions          expand             got          highly
##            0.55            0.55            0.55            0.55
##        internal            lane           larry            late
##            0.55            0.55            0.55            0.55
##          lehman            move       positions prudentialbache
##            0.55            0.55            0.55            0.55
##        remained          rumors         selling           sense
##            0.55            0.55            0.55            0.55
##          silent        unlikely          vacant         whether
##            0.55            0.55            0.55            0.55
##          growth            part        analysts         billion
##            0.54            0.54            0.53            0.53
##          beyond           bring   international           major
##            0.52            0.52            0.52            0.52
##         options  reorganization      considered         current
##            0.52            0.52            0.51            0.50
##            need       spokesman           march       statement
##            0.50            0.49            0.48            0.48
##         spinoff         capital             may       financial
##            0.46            0.45            0.45            0.44
##         comment           fully           plans           range
##            0.43            0.43            0.43            0.43
##            said     eckenfelder           place          access
##            0.43            0.42            0.42            0.40
##           alone           close   consideration           given
##            0.40            0.40            0.40            0.40
##            help         improve            loss            meet
##            0.40            0.40            0.40            0.40
##       operating            post         reached     whollyowned
##            0.40            0.40            0.40            0.40
##       worldwide         reflect            also       estimated
##            0.40            0.35            0.34            0.34
##          friday          market      profitable           total
##            0.33            0.32            0.32            0.32
##             can            days           firms           makes
##            0.30            0.30            0.30            0.30
##         related             net        position       president
##            0.30            0.29            0.29            0.29
##          public          higher
##            0.29            0.28
```

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

```
##       term freq
## NA    <NA>   NA
## NA.1  <NA>   NA
## NA.2  <NA>   NA
## NA.3  <NA>   NA
## NA.4  <NA>   NA
## NA.5  <NA>   NA
## NA.6  <NA>   NA
## NA.7  <NA>   NA
## NA.8  <NA>   NA
## NA.9  <NA>   NA
## NA.10 <NA>   NA
## NA.11 <NA>   NA
## NA.12 <NA>   NA
## NA.13 <NA>   NA
## NA.14 <NA>   NA
## NA.15 <NA>   NA
## NA.16 <NA>   NA
## NA.17 <NA>   NA
## NA.18 <NA>   NA
## NA.19 <NA>   NA
## NA.20 <NA>   NA
## NA.21 <NA>   NA
## NA.22 <NA>   NA
## NA.23 <NA>   NA
## NA.24 <NA>   NA
## NA.25 <NA>   NA
## NA.26 <NA>   NA
## NA.27 <NA>   NA
## NA.28 <NA>   NA
## NA.29 <NA>   NA
## NA.30 <NA>   NA
## NA.31 <NA>   NA
## NA.32 <NA>   NA
## NA.33 <NA>   NA
## NA.34 <NA>   NA
## NA.35 <NA>   NA
## NA.36 <NA>   NA
## NA.37 <NA>   NA
## NA.38 <NA>   NA
## NA.39 <NA>   NA
## NA.40 <NA>   NA
## NA.41 <NA>   NA
## NA.42 <NA>   NA
## NA.43 <NA>   NA
## NA.44 <NA>   NA
## NA.45 <NA>   NA
## NA.46 <NA>   NA
## NA.47 <NA>   NA
## NA.48 <NA>   NA
## NA.49 <NA>   NA
```

What happens? Does it yield anything understandable about the documents

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

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

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

```
## Corpus consisting of 50 documents.
##
##   Text Types Tokens Sentences                    author       datetimestamp
##     10   120    233        26                      <NA> 1987-02-26 15:18:06
##     12    89    146        17                      <NA> 1987-02-26 15:19:15
##     44    62     86        13                      <NA> 1987-02-26 15:49:56
##     45   232    431        51 By Cal Mankowski, Reuters 1987-02-26 15:51:17
##     68    42     59         7                      <NA> 1987-02-26 16:08:33
##     96    56     75         8                      <NA> 1987-02-26 16:32:37
##    110   292    666        79     By Patti Domm, Reuter 1987-02-26 16:43:13
##    125    73    112        12                      <NA> 1987-02-26 16:59:25
##    128    34     46         7                      <NA> 1987-02-26 17:01:28
##    134    37     40         6                      <NA> 1987-02-26 17:08:27
##    135    76    110        15                      <NA> 1987-02-26 17:09:47
##    153    77    108        13                      <NA> 1987-02-26 17:36:22
##    157    92    166        19                      <NA> 1987-02-26 17:38:47
##    162    32     39         6                      <NA> 1987-02-26 17:43:59
##    185    35     40         6                      <NA> 1987-02-26 18:12:35
##    186    29     33         4                      <NA> 1987-02-26 18:12:51
##    199    55    101        12                      <NA> 1987-02-26 18:27:56
##    260    91    174        19                      <NA> 1987-03-01 22:20:43
##    302   211    468        45                      <NA> 1987-03-02 04:45:57
##    304    97    201        24                      <NA> 1987-03-02 04:52:58
##    315    66     93        10                      <NA> 1987-03-02 05:48:46
##    331   188    364        39                      <NA> 1987-03-02 06:54:19
##    334    74    114        12                      <NA> 1987-03-02 06:58:00
##    361    71    108        14                      <NA> 1987-03-02 08:16:59
##    362   261    611        69    By Patti Domm, Reuters 1987-03-02 08:17:56
##    366    95    148        19                      <NA> 1987-03-02 08:22:40
##    369    82    121        16                      <NA> 1987-03-02 08:25:56
##    371    72    120        14                      <NA> 1987-03-02 08:26:35
##    372   250    577        67             By Patti Domm 1987-03-02 08:29:05
##    376    31     35         5                      <NA> 1987-03-02 08:41:41
##    379    49     63         8                      <NA> 1987-03-02 08:43:25
##    387    61     93        12                      <NA> 1987-03-02 09:02:51
##    389    85    128        15                      <NA> 1987-03-02 09:03:18
##    393   124    270        30                      <NA> 1987-03-02 09:16:08
##    401    94    140        15                      <NA> 1987-03-02 09:28:21
##    408   108    187        23                      <NA> 1987-03-02 09:33:32
##    424    77    122        12                      <NA> 1987-03-02 09:49:48
##    436    68     82        10                      <NA> 1987-03-02 10:06:32
##    441    76    148        16                      <NA> 1987-03-02 10:20:41
##    442    50     69        10                      <NA> 1987-03-02 10:29:07
##    447    43     64         8                      <NA> 1987-03-02 10:36:04
##    448   143    301        32                      <NA> 1987-03-02 10:36:13
##    467    42     53         7                      <NA> 1987-03-02 10:50:34
##    473   103    199        23                      <NA> 1987-03-02 10:59:16
##    474    59     94        11                      <NA> 1987-03-02 10:59:28
##    478    76    104        13                      <NA> 1987-03-02 11:09:06
##   496   228    555        56                      <NA> 1987-03-02 11:23:31
##   497    58     82        13                      <NA> 1987-03-02 11:23:45
##   498    47     57         9                      <NA> 1987-03-02 11:24:06
##   504   118    197        25                      <NA> 1987-03-02 11:29:26
##  description                                           heading  id
##                COMPUTER TERMINAL SYSTEMS <CPML> COMPLETES SALE  10
##                 OHIO MATTRESS <OMT> MAY HAVE LOWER 1ST QTR NET  12
##                  MCLEAN'S <MII> U.S. LINES SETS ASSET TRANSFER  44
##                 CHEMLAWN <CHEM> RISES ON HOPES FOR HIGHER BIDS  45
##                 <COFAB INC> BUYS GULFEX FOR UNDISCLOSED AMOUNT  68
##                       INVESTMENT FIRMS CUT CYCLOPS <CYL> STAKE  96
##               AMERICAN EXPRESS <AXP> SEEN IN POSSIBLE SPINNOFF 110
##                HONG KONG FIRM UPS WRATHER<WCO> STAKE TO 11 PCT 125
##                            LIEBERT CORP <LIEB> APPROVES MERGER 128
##                   GULF APPLIED TECHNOLOGIES <GATS> SELLS UNITS 134
##                   INVESTMENT GROUP RAISES ROBESON <RBSN> STAKE 135
##               DREXEL OFFICIAL HAS STAKE IN EPSILON DATA <EPSI> 153
##               <NOVA> WINS GOVERNMENT OKAY FOR HUSKY <HYO> DEAL 157
##                    SUFFIELD FINANCIAL <SSBK> GETS FED APPROVAL 162
##                                VERSATILE TO SELL UNIT TO VICON 185
##                            VIDEOTRON BUYS INTO EXHIBIT COMPANY 186
##                        CIRCUIT SYSTEMS <CSYI> BUYS BOARD MAKER 199
##               NIPPON KOKAN STEEL AFFILIATES CONSIDERING MERGER 260
##                  WALLENBERGS FIGHT BID FOR SWEDISH MATCH STAKE 302
##                     SHV SAYS IT MAKING TENDER OFFER FOR IC GAS 304
##                SALE TILNEY BUYS STAKE IN U.S. INSURANCE BROKER 315
##                    EXCO BUYS U.S. GOVERNMENT SECURITIES BROKER 331
##              COLOROLL AGREES TO BUY U.S. WALLCOVERINGS COMPANY 334
##               SCIENTIFIC MICRO SYSTEMS <SMSI> ACUIRES SUPERMAC 361
##                AMERICAN EXPRESS <AXP> VIEWING SHEARSON OPTIONS 362
##                      ROPAK <ROPK> HAS 34 PCT OF BUCKHORN <BKN> 366
##                           PENRIL <PNL> SEEKS TO SELL TWO UNITS 369
##                          <DALE BURDETT INC> FACES DAMAGE CLAIM 371
##                    PUROLATOR <PCC> IN BUYOUT WITH HUTTON <EFH> 372
##                 FINANCIAL SANTA BARBARA <FSB> TO MAKE PURCHASE 376
##                                   MARRIOTT <MHS> TO SELL HOTEL 379
##                       LAROCHE STARTS BID FOR NECO <NPT> SHARES 387
##              SENIOR ENGINEERING MAKES 12.5 MLN DLR US PURCHASE 389
##                       VIACOM <VIA> RECEIVES TWO REVISED OFFERS 393
##                MILLER TABAK HAS 91.8 PCT OF PENN TRAFFIC <PNF> 401
##                     PITTSTON <PCO> AGREES TO ACQUIRE WTC <WAF> 408
##               DIAGNOSTIC <DRS> MAKES A BID FOR ROSPATCH <RPCH> 424
##                         THE JAPAN FUND <JPN> GETS BUYOUT OFFER 436
##                BANK OF NEW YORK <BK> TO HAVE GAIN ON UNIT SALE 441
##               CORNING <GLW>, HAZLETON <HLC> SET EXCAHNGE RATIO 442
##                BALLY <BLY> COMPLETES PURCHASE OF GOLDEN NUGGET 447
##                CONSOLIDATED TVX TO BUY BRAZIL GOLD MINE STAKES 448
##                   AMERICAN NURSERY <ANSY> BUYS FLORIDA NURSERY 467
##                  MULTI-STEP TO SELL LADDER UNIT, CANCEL SHARES 473
##                 ESSELTE BUSINESS <ESB> UNIT BUYS ANTONSON UNIT 474
##                       FOUR SEASONS BUYING MARRIOTT <MHS> HOTEL 478
##                  REDSTONE DETAILS SWEETENED VIACOM <VIA> OFFER 496
##                    MONTEDISON CONCLUDES TALKS WITH ANTIBIOTICOS 497
##                          UTILICORP <UCU> COMPLETES ACQUISITION 498
##                   CARBIDE <UK> LOOKS TO ACQUISITIONS FOR GROWTH 504
##  language            origin topics lewissplit     cgisplit oldid
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5553
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5555
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5587
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5588
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5611
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5639
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5653
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5668
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5671
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5677
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5678
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5696
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5700
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5705
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5728
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5729
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  5742
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET  8345
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12485
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12487
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12498
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12514
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12517
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12543
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12544
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12548
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12551
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12553
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12554
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12558
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12561
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12570
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12572
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12576
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12584
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12591
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12607
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12619
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12624
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12625
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12630
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12631
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12650
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12656
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12657
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12661
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12679
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12680
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12681
##        en Reuters-21578 XML    YES      TRAIN TRAINING-SET 12687
##                 places people orgs exchanges
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                 canada   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                 canada   <NA> <NA>      <NA>
##                 canada   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                  japan   <NA> <NA>      <NA>
##                 sweden   <NA> <NA>      <NA>
##                     uk   <NA> <NA>      <NA>
##         c("usa", "uk")   <NA> <NA>      <NA>
##         c("uk", "usa")   <NA> <NA>      <NA>
##         c("usa", "uk")   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##         c("uk", "usa")   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##  c("canada", "brazil")   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                   <NA>   <NA> <NA>      <NA>
##     c("usa", "sweden")   <NA> <NA>      <NA>
##                 canada   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##    c("italy", "spain")   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##                    usa   <NA> <NA>      <NA>
##
## Source:  Converted from tm VCorpus 'acq'
## Created: Sat Apr 30 11:32:53 2016
## Notes:
```

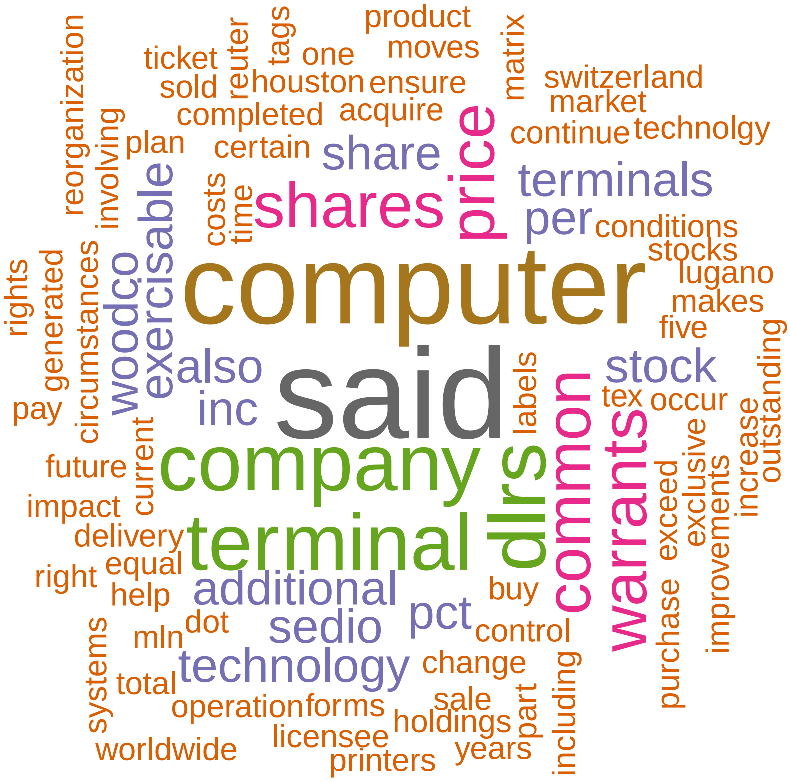
```r
#10 longest documents in the corpus
sort_top10 <- summary_acq %>% arrange(desc(Tokens))
top_10_docs <- subset(sort_top10, select=c(id, heading))[1:10,]
top10 <- top_10_docs[,1]

top10
```

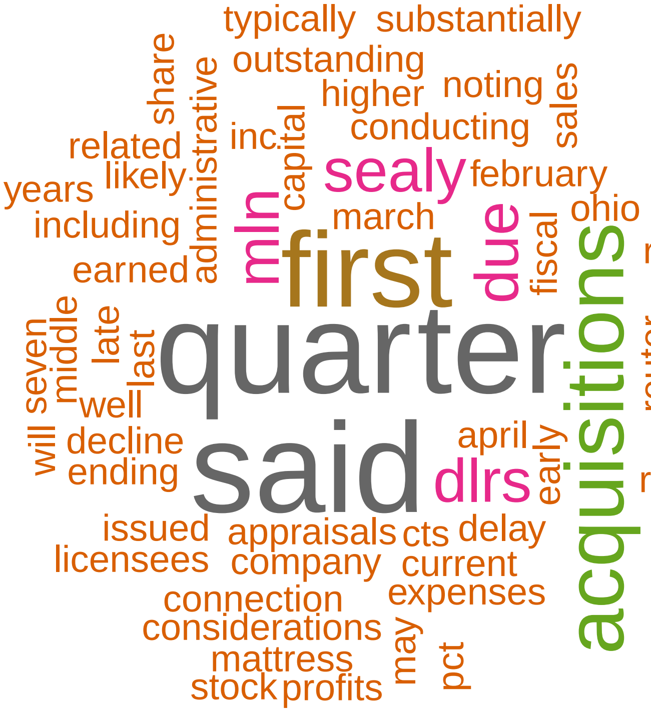
```
##  [1] "110" "362" "372" "496" "302" "45"  "331" "448" "393" "10"
```

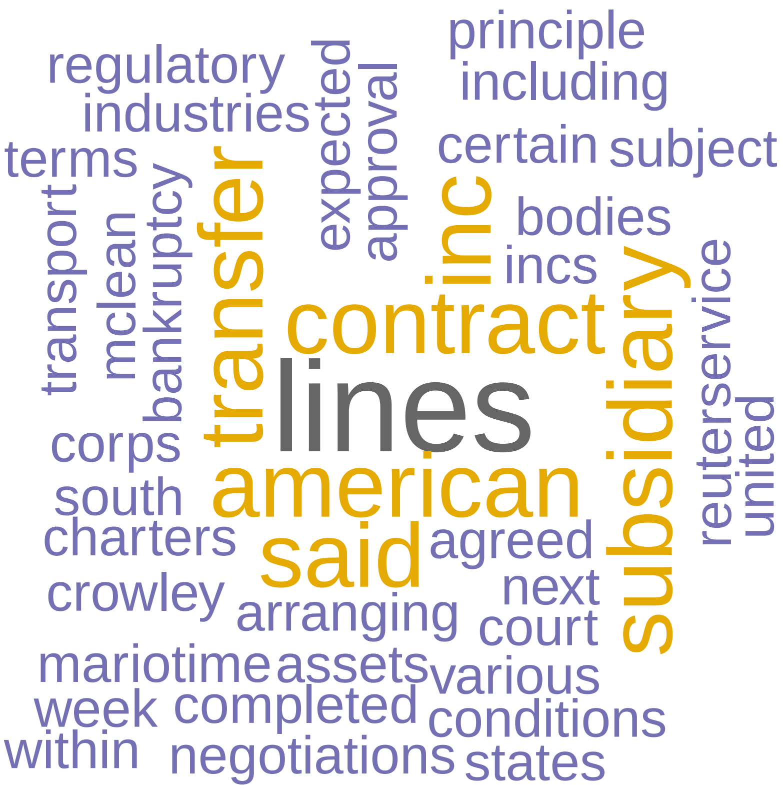
```r
topdocs <- mycorpus[mycorpus$documents$id %in% top10]
topdocs
```

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

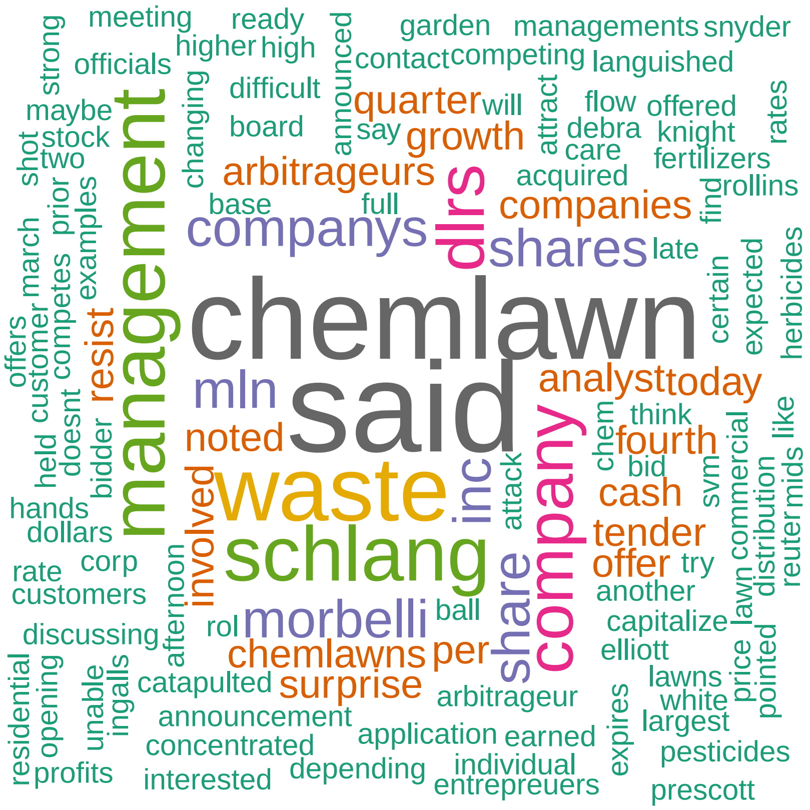
```r
#top 10 dendogram, 1 for each of the top 10 documents
top10.dendogram <- function(tdm2,doc)
{
  acq.mat <- as.matrix(tdm2)
  acq.mat <- as.data.frame(acq.mat)
  acq.mat <- acq.mat[,top10]
  acq.mat <- as.matrix(acq.mat)
  distMatrix <- dist(scale(acq.mat[,doc]))
  fit <- hclust(distMatrix, method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


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



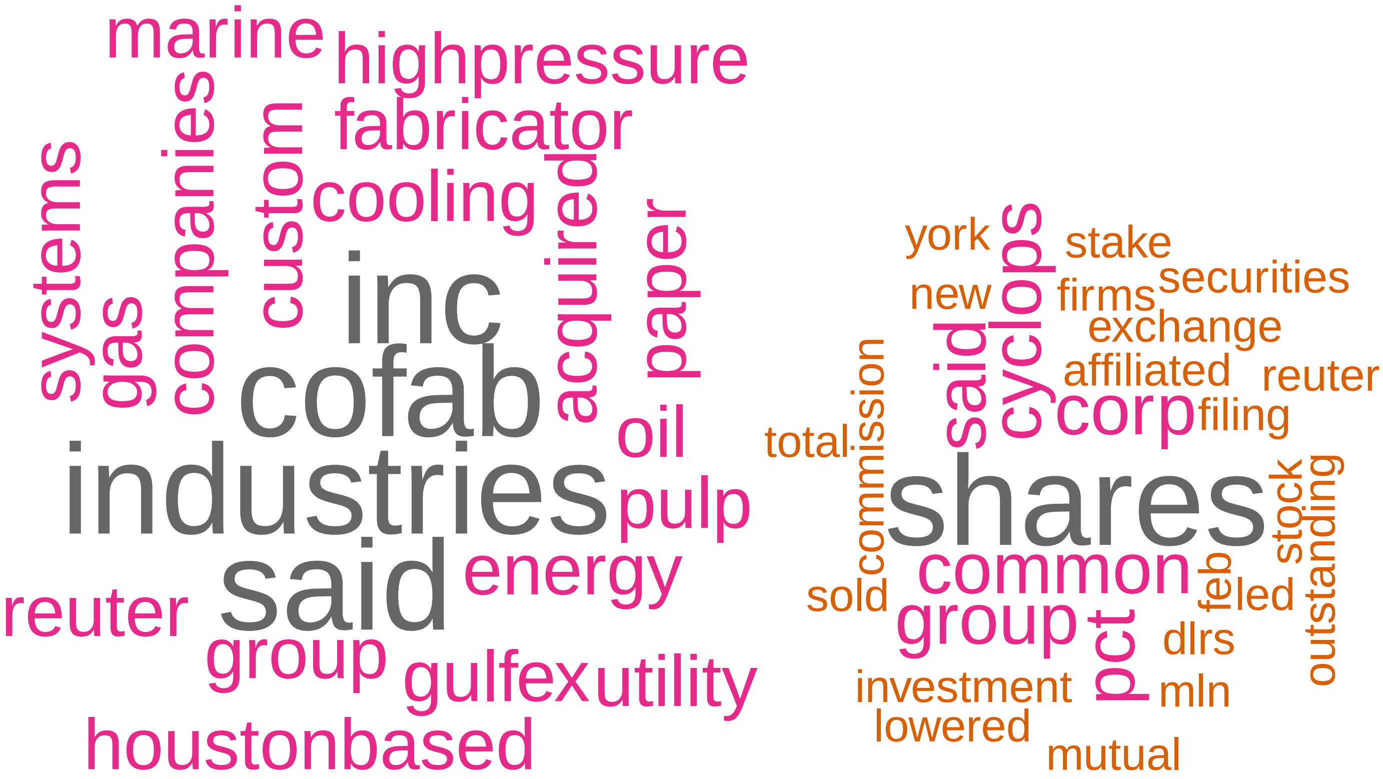




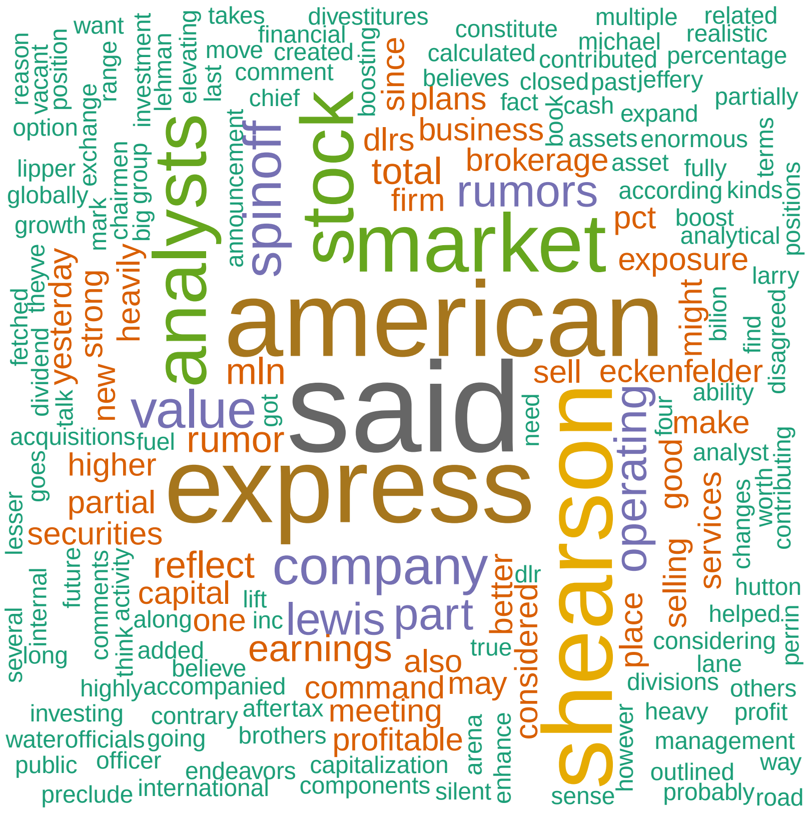
```
## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : overthecounter could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : regularly could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : removal could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : rosemarie could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : scheduled could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : servicemaster could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : sevice could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : slowed could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : spokeswoman could not be fit on page. It will not be plotted.
## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : statement could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : street could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : strohmaier could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : system could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : thousands could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : times could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : totally could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : trading could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : turben could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : valuable could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : wall could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : want could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : wants could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : wastes could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : whether could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : wnx could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : year could not be fit on page. It will not be plotted.
```



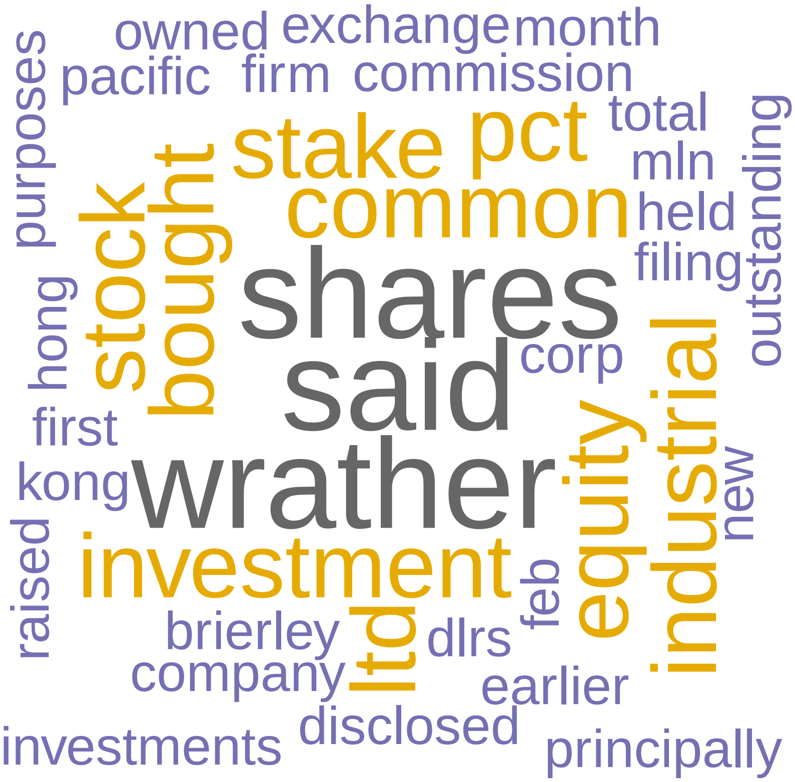
```
## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : petrochemical could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : lubricating could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : manufacture could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : process could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : specialized could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : vessels could not be fit on page. It will not be plotted.
```

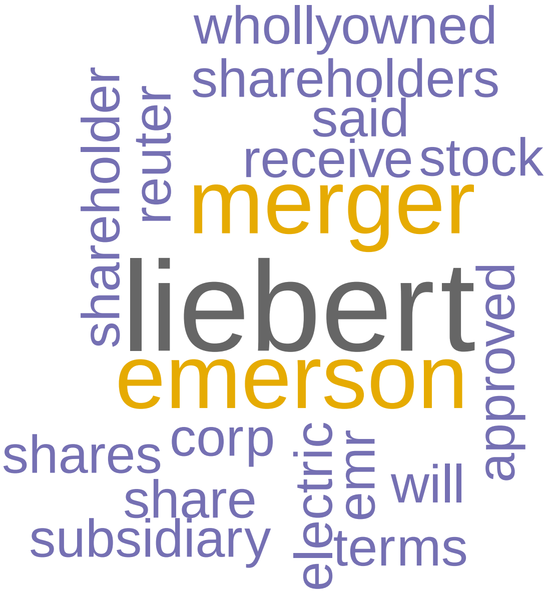


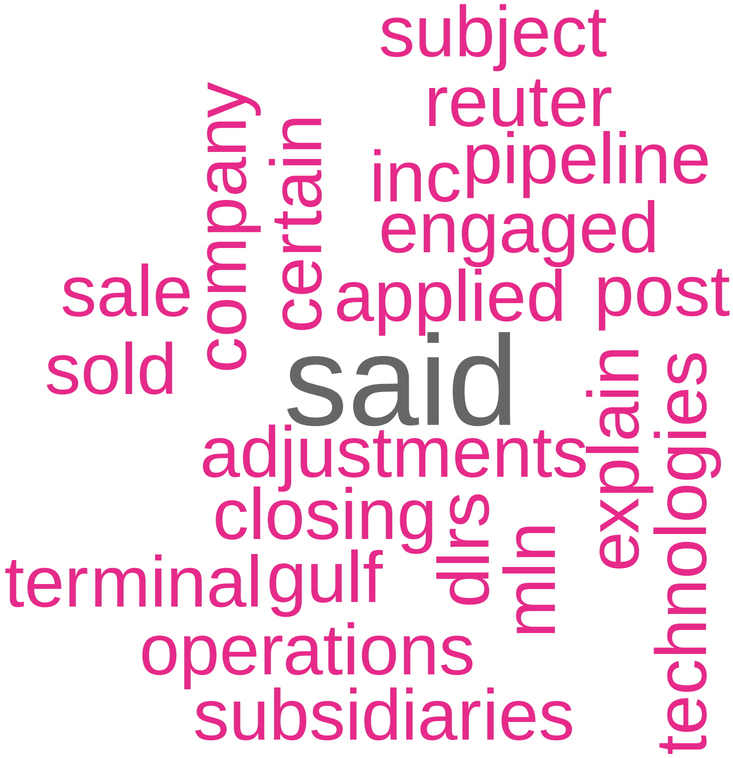
```
## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : performance could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : premium could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : president could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : prudentialbache could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : questioned could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : reduced could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : remained could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : reuter could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : shearsons could not be fit on page. It will not be plotted.
## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : speculated could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : spinning could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : split could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : suggests could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : thereby could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : theyre could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : three could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : tuesday could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : undervalued could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : unhappy could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : unlikely could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : use could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : variation could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : volume could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : well could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : year could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : york could not be fit on page. It will not be plotted.
```



```
## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : publicly could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : reuter could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : securities could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = rownames(v), freq = v, min.freq = 1, max.words
## = 200, : zealand could not be fit on page. It will not be plotted.
```







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

The project helped us learn a lot about text analytics and key principals of analyzing unstructured text.We identified three key areas this project helped you learn about data science

includes (1) the general approach to breaking down texts in R using Corpuses and tokens; (2) The exploratory analysis and derived insights that can be accomplish on a text documents through word counts, frequencies, associations, and character lengths; (3) we were able to learn how to apply data mining techniques to text analytics for deeper insights such as clustering (hierarchical and kmeans).