The Reuters Corpus

Lindsay Bartol

2024-08-17

```
knitr::opts_chunk$set(echo = TRUE)

library(tm)

## Loading required package: NLP

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2

## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ ggplot2::annotate() masks NLP::annotate()
## ✖ dplyr::filter()     masks stats::filter()
## ✖ dplyr::lag()        masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors

library(slam)
library(proxy)

##
## Attaching package: 'proxy'
##
## The following objects are masked from 'package:stats':
##
##     as.dist, dist
##
## The following object is masked from 'package:base':
##
##     as.matrix

#Here is the reader function
readerPlain = function(fname){
  readPlain(elem=list(content=readLines(fname)),
            id=fname, language='en') }
```

```r
#defining the path to my dataset
data_path <-"C:\\Users\\linds\\OneDrive\\Documents\\MSBA-lindsayslaptop\\STA
S380\\ReutersC50\\C50train"
author_dirs <- list.dirs(data_path, full.names =TRUE, recursive =FALSE)

# Initialize an empty list to store documents
all_documents <-list()
all_mynames <-list()

# Iterate over each author's directory
for(author_dir in author_dirs){
  file_list <- Sys.glob(file.path(author_dir,'*.txt'))
  author_documents <- lapply(file_list, readerPlain)# Clean up the file names
  mynames <- file_list %>%{ strsplit(.,'/', fixed=TRUE)}%>%{ lapply(., tail,
n=2)}%>%{ lapply(., paste0, collapse ='')}%>%
    unlist

# Store the documents and their cleaned names in the lists
  all_documents <-c(all_documents, author_documents)
  all_mynames <-c(all_mynames, mynames)}# Combine all documents into one list
  names(all_documents)<- all_mynames

# Create a text corpus from the combined documents
documents_raw <- Corpus(VectorSource(all_documents))

my_documents <- documents_raw
my_documents <- tm_map(my_documents, content_transformer(tolower))# make
everything lowercase

## Warning in tm_map.SimpleCorpus(my_documents,
content_transformer(tolower)):
## transformation drops documents

my_documents <- tm_map(my_documents, content_transformer(removeNumbers))#
remove numbers

## Warning in tm_map.SimpleCorpus(my_documents,
## content_transformer(removeNumbers)): transformation drops documents

my_documents <- tm_map(my_documents, content_transformer(removePunctuation))#
remove punctuation

## Warning in tm_map.SimpleCorpus(my_documents,
## content_transformer(removePunctuation)): transformation drops documents

my_documents <- tm_map(my_documents, content_transformer(stripWhitespace))#
remove excess white-space

## Warning in tm_map.SimpleCorpus(my_documents,
## content_transformer(stripWhitespace)): transformation drops documents
```

```
my_documents <- tm_map(my_documents, content_transformer(removeWords),
stopwords("en"))# remove stopwords

## Warning in tm_map.SimpleCorpus(my_documents,
content_transformer(removeWords),
## : transformation drops documents

custom_stopwords <-
c(stopwords("en"),"cuserslindsonedrivedocumentsmsbalindsayslaptopsta","dateti
mestamp","meta","gmt","isdst","language","mday","mon","month","wday","yday","
year","zone","datetimestamp","listauthor","listcontent","listsec","sreuterscc
trainlynneodonnellnewsmltxt","sreuterscctrainpeterhumphreynewsmltxt")
my_documents <- tm_map(my_documents, removeWords, custom_stopwords)

## Warning in tm_map.SimpleCorpus(my_documents, removeWords,
custom_stopwords):
## transformation drops documents

custom_stopwords <-
c(stopwords("en"),"said","character","gmtoff","heading","origin","hour")
my_documents <- tm_map(my_documents, removeWords, custom_stopwords)

## Warning in tm_map.SimpleCorpus(my_documents, removeWords,
custom_stopwords):
## transformation drops documents

# Create a Document-Term Matrix for all documents
DTM_all <- DocumentTermMatrix(my_documents)# Display basic summary statistics
of the DTM
DTM_all  # This will show you the number of documents and terms

## <<DocumentTermMatrix (documents: 2500, terms: 32548)>>
## Non-/sparse entries: 494732/80875268
## Sparsity           : 99%
## Maximal term length: 41
## Weighting          : term frequency (tf)

# Inspect the first 10 documents and the first 20 terms
inspect(DTM_all[1:10,1:20])

## <<DocumentTermMatrix (documents: 10, terms: 20)>>
## Non-/sparse entries: 42/158
## Sparsity           : 79%
## Maximal term length: 11
## Weighting          : term frequency (tf)
## Sample             :
##     Terms
## Docs access accounts agencies also announced bogus business called charged
##   1       1        1        1    1         1     2        2      1       1
##   10      4        0        0    1         0     0        1      0       0
##   2       0        0        0    2         1     0        1      0       0
##   3       2        0        0    0         0     0        0      0       0
```

```
##   4       0        0        0    0         1     0        1      0       0
##   5       0        0        0    0         1     0        1      0       0
##   6       0        0        0    0         0     0        0      0       0
##   7       0        0        1    1         0     0        0      1       0
##   8       0        0        0    0         0     0        1      0       1
##   9       0        0        1    0         0     0        1      0       1
##     Terms
## Docs commission
##   1           2
##   10          0
##   2           0
##   3           0
##   4           0
##   5           0
##   6           0
##   7           5
##   8           2
##   9           2
```

```r
# Find words that appear in at least 750 documents
frequent_terms <- findFreqTerms(DTM_all,750)
print(frequent_terms)
```

```
##   [1] "also"          "business"      "computer"      "description"
##   [5] "group"         "internet"      "investors"     "local"
##   [9] "major"         "may"           "million"       "min"
##  [13] "new"           "one"           "services"      "shares"
##  [17] "state"         "still"         "trade"         "tuesday"
##  [21] "wednesday"     "world"         "can"           "corp"
##  [25] "just"          "many"          "now"           "people"
##  [29] "plans"         "president"     "service"       "trading"
##  [33] "will"          "executive"     "companies"     "end"
##  [37] "government"    "international" "market"        "months"
##  [41] "next"          "three"         "two"           "week"
##  [45] "analyst"       "banks"         "company"       "financial"
##  [49] "last"          "much"          "officials"     "sales"
##  [53] "states"        "take"          "already"       "another"
##  [57] "big"           "billion"       "expected"      "foreign"
##  [61] "investment"    "markets"       "say"           "told"
##  [65] "united"        "added"         "chief"         "made"
##  [69] "since"         "stock"         "years"         "around"
##  [73] "chairman"      "friday"        "half"          "inc"
##  [77] "time"          "first"         "growth"        "monday"
##  [81] "news"          "strong"        "well"          "bank"
##  [85] "deal"          "thursday"      "going"         "industry"
##  [89] "make"          "share"         "reuters"       "think"
##  [93] "analysts"      "percent"       "price"         "prices"
##  [97] "profit"        "second"        "earnings"      "british"
## [101] "beijing"       "profits"       "quarter"       "pounds"
```

```
## [105] "hong"          "kong"          "china"         "chinas"
## [109] "chinese"
```

```
#a lot of china - these article proabbly talk a lot about global affairs
#as expected, a lot of business/finance terms - market, growth, prices, etc.

# Find words that are associated with"china" with a correlation of at least
0.25
associations <- findAssocs(DTM_all,"china",0.25)
print(associations)
```

```
## $china
##     chinese     beijing      chinas        hong        kong      cchina
##        0.61        0.57        0.55        0.43        0.43        0.40
##    official      taiwan    beijings        ties       visit       kongs
##        0.35        0.35        0.33        0.32        0.31        0.30
##    economic     taiwans      states   relations  diplomatic     tenghui
##        0.29        0.29        0.28        0.28        0.28        0.28
##     imports      colony       zemin    colonial       trade  washington
##        0.27        0.27        0.27        0.27        0.26        0.26
## cooperation sovereignty   officials agriculture      taipei
##        0.26        0.26        0.25        0.25        0.25
```

```
# Remove terms that appear in fewer than 5% of documents
DTM_all_reduced <- removeSparseTerms(DTM_all,0.95)# Display the reduced DTM
DTM_all_reduced
```

```
## <<DocumentTermMatrix (documents: 2500, terms: 783)>>
## Non-/sparse entries: 237783/1719717
## Sparsity           : 88%
## Maximal term length: 18
## Weighting          : term frequency (tf)
```

```
inspect(DTM_all_reduced[1:50,1:50])
```

```
## <<DocumentTermMatrix (documents: 50, terms: 50)>>
## Non-/sparse entries: 640/1860
## Sparsity           : 74%
## Maximal term length: 11
## Weighting          : term frequency (tf)
## Sample             :
##     Terms
## Docs access also computer description federal internet law min new one
##   1       1    1        1           1       2        9   1   1   1   1
##   11      1    1        3           1       1        4   1   1   3   0
##   23      0    2        0           1       0       12   0   1   7   2
##   26     11    1        0           1       1        7   1   1   2   0
##   27      6    0        0           1       1        8   1   1   1   0
##   32      4    2        1           1       1       18   0   1   3   1
##   33      4    2        1           1       1       18   0   1   3   2
##   39      0    1        0           1       0       14   1   1   5   1
```

```
##   48      0    0        0           1       7        0   3   1   0   1
##   7       0    1        1           1       7        4   3   1   2   0
```

```
#wow, this reduced it from like 32k word to around 800. That's a huge
reduction. Might want to come back later and mess with this.

# Compute TF-IDF weights
tfidf_all <- weightTfIdf(DTM_all_reduced)# Inspect the TF-IDF matrix for the
first document
inspect(tfidf_all[1,])
```

```
## <<DocumentTermMatrix (documents: 1, terms: 783)>>
## Non-/sparse entries: 63/720
## Sparsity           : 92%
## Maximal term length: 18
## Weighting          : term frequency - inverse document frequency
(normalized) (tf-idf)
## Sample             :
##     Terms
## Docs commission   consumer consumers    federal    initial  internet
investors
##    1 0.07725414 0.07685393 0.1761798 0.08326403 0.04946848 0.3697142
0.08835436
##     Terms
## Docs        may    quality   reports
##    1 0.06065897 0.04835383 0.1383413
```

Okay, so there was some interesting stuff in that last section. Looking at the most frequent terms, a lot of them were expected - things like market, price, analyst, financial. Something I found interesting, though, was that China made it up there pretty high. Potentially, there are a lot of articles on golobal affairs, specifically China.

Reducing the terms led to a huge reduction- from over 30k to around 800. That's crazy.

After that, I wanted to look at words associated with "China." A lot of it was fairly expected - ties, visit, economic, taiwan, relations, diplomatic, imports, etc. Terms relating to foreign relations.

After applying tf-idf weights - there are 64 terms that have non-zero scores, versus 725 that have zero values

```
#I want to see the top words for every tenth document to get an idea for
topics
tfidf_dense <- as.matrix(tfidf_all)
top_terms <-list()
for(i in seq(1, nrow(tfidf_dense), by =10)){
  doc_tfidf <- tfidf_dense[i,]# This should now be a numeric vector
  max_index <- which.max(doc_tfidf)
  top_term <- colnames(tfidf_dense)[max_index]
  top_terms[[i]]<-list(term = top_term, score = doc_tfidf[max_index])
```

```
  cat("Document", i,"top term:", top_term,"with score:",
doc_tfidf[max_index],"\n")}# Combine the results into a data frame

## Document 1 top term: internet with score: 0.3697142
## Document 11 top term: internet with score: 0.1024007
## Document 21 top term: data with score: 0.2441374
## Document 31 top term: credit with score: 0.4444044
## Document 41 top term: insurance with score: 0.1972414
## Document 51 top term: czech with score: 0.1668588
## Document 61 top term: czech with score: 0.2190021
## Document 71 top term: czech with score: 0.1514089
## Document 81 top term: profit with score: 0.1004773
## Document 91 top term: house with score: 0.1912287
## Document 101 top term: venture with score: 0.07496784
## Document 111 top term: banking with score: 0.1039153
## Document 121 top term: plan with score: 0.09424564
## Document 131 top term: britain with score: 0.08934242
## Document 141 top term: restructuring with score: 0.2454108
## Document 151 top term: direct with score: 0.2085693
## Document 161 top term: long with score: 0.1365868
## Document 171 top term: china with score: 0.2180068
## Document 181 top term: court with score: 0.2532035
## Document 191 top term: china with score: 0.1335826
## Document 201 top term: profits with score: 0.1102244
## Document 211 top term: merger with score: 0.1311287
## Document 221 top term: home with score: 0.1796587
## Document 231 top term: found with score: 0.2216261
## Document 241 top term: internet with score: 0.2783438
## Document 251 top term: banks with score: 0.1591198
## Document 261 top term: officer with score: 0.1001381
## Document 271 top term: sales with score: 0.1749377
## Document 281 top term: banking with score: 0.1463723
## Document 291 top term: increase with score: 0.08291448
## Document 301 top term: bank with score: 0.2005014
## Document 311 top term: toronto with score: 0.164085
## Document 321 top term: rates with score: 0.1665031
## Document 331 top term: bank with score: 0.1638714
## Document 341 top term: services with score: 0.07573129
## Document 351 top term: quarter with score: 0.102819
## Document 361 top term: production with score: 0.0703254
## Document 371 top term: workers with score: 0.1905157
## Document 381 top term: local with score: 0.1708087
## Document 391 top term: workers with score: 0.1123222
## Document 401 top term: tough with score: 0.1049451
## Document 411 top term: followed with score: 0.07899622
## Document 421 top term: deals with score: 0.08822796
## Document 431 top term: launch with score: 0.2333202
## Document 441 top term: million with score: 0.1760889
## Document 451 top term: computer with score: 0.2176896
## Document 461 top term: internet with score: 0.1288009
```

```
## Document 471 top term: house with score: 0.05611171
## Document 481 top term: network with score: 0.08772933
## Document 491 top term: stocks with score: 0.1962106
## Document 501 top term: banks with score: 0.1509488
## Document 511 top term: profit with score: 0.09890731
## Document 521 top term: ministry with score: 0.2815114
## Document 531 top term: bank with score: 0.1495762
## Document 541 top term: tax with score: 0.08367796
## Document 551 top term: markets with score: 0.1406514
## Document 561 top term: trading with score: 0.2533742
## Document 571 top term: shares with score: 0.09247585
## Document 581 top term: shares with score: 0.1625333
## Document 591 top term: investors with score: 0.1407125
## Document 601 top term: tonnes with score: 0.1728991
## Document 611 top term: takeover with score: 0.1112435
## Document 621 top term: health with score: 0.153944
## Document 631 top term: process with score: 0.09790207
## Document 641 top term: government with score: 0.09081213
## Document 651 top term: officials with score: 0.08117845
## Document 661 top term: party with score: 0.3561956
## Document 671 top term: china with score: 0.1687834
## Document 681 top term: trade with score: 0.1411983
## Document 691 top term: daily with score: 0.1164982
## Document 701 top term: average with score: 0.1610396
## Document 711 top term: results with score: 0.08493795
## Document 721 top term: party with score: 0.1015079
## Document 731 top term: data with score: 0.2591925
## Document 741 top term: never with score: 0.134346
## Document 751 top term: pacific with score: 0.1223067
## Document 761 top term: demand with score: 0.1807878
## Document 771 top term: kong with score: 0.1772813
## Document 781 top term: building with score: 0.2976979
## Document 791 top term: base with score: 0.1511613
## Document 801 top term: shareholders with score: 0.09242266
## Document 811 top term: french with score: 0.1193792
## Document 821 top term: bank with score: 0.1022557
## Document 831 top term: funds with score: 0.3901021
## Document 841 top term: debt with score: 0.2625989
## Document 851 top term: czech with score: 0.1784462
## Document 861 top term: czech with score: 0.1168011
## Document 871 top term: house with score: 0.1637232
## Document 881 top term: czech with score: 0.1892611
## Document 891 top term: czech with score: 0.2197871
## Document 901 top term: human with score: 0.2265699
## Document 911 top term: data with score: 0.08049456
## Document 921 top term: sales with score: 0.09614221
## Document 931 top term: amp with score: 0.09599562
## Document 941 top term: pence with score: 0.09059652
## Document 951 top term: pence with score: 0.2713242
## Document 961 top term: amp with score: 0.1042506
```

```
## Document 971 top term: pence with score: 0.1416207
## Document 981 top term: east with score: 0.2333979
## Document 991 top term: bid with score: 0.1383567
## Document 1001 top term: thursday with score: 0.07637111
## Document 1011 top term: countries with score: 0.07191821
## Document 1021 top term: central with score: 0.08259748
## Document 1031 top term: countrys with score: 0.1189233
## Document 1041 top term: washington with score: 0.1230933
## Document 1051 top term: pounds with score: 0.09913488
## Document 1061 top term: newspaper with score: 0.07380687
## Document 1071 top term: expansion with score: 0.08540361
## Document 1081 top term: pounds with score: 0.2304393
## Document 1091 top term: television with score: 0.09906052
## Document 1101 top term: consumer with score: 0.1231572
## Document 1111 top term: united with score: 0.3233104
## Document 1121 top term: economy with score: 0.1020388
## Document 1131 top term: insurance with score: 0.1834557
## Document 1141 top term: cents with score: 0.1193507
## Document 1151 top term: sector with score: 0.1312539
## Document 1161 top term: news with score: 0.204672
## Document 1171 top term: australian with score: 0.1188172
## Document 1181 top term: australian with score: 0.2040295
## Document 1191 top term: sector with score: 0.1433929
## Document 1201 top term: similar with score: 0.0952055
## Document 1211 top term: merger with score: 0.09101173
## Document 1221 top term: pounds with score: 0.07908259
## Document 1231 top term: japan with score: 0.1762211
## Document 1241 top term: software with score: 0.2045702
## Document 1251 top term: rose with score: 0.1269848
## Document 1261 top term: revenue with score: 0.1590589
## Document 1271 top term: software with score: 0.2237026
## Document 1281 top term: software with score: 0.2316655
## Document 1291 top term: software with score: 0.1793923
## Document 1301 top term: points with score: 0.2095896
## Document 1311 top term: stocks with score: 0.1163732
## Document 1321 top term: index with score: 0.122072
## Document 1331 top term: points with score: 0.1564542
## Document 1341 top term: stocks with score: 0.1363565
## Document 1351 top term: tonnes with score: 0.220551
## Document 1361 top term: official with score: 0.2243132
## Document 1371 top term: chinese with score: 0.1180379
## Document 1381 top term: tonnes with score: 0.2179801
## Document 1391 top term: chinese with score: 0.09888332
## Document 1401 top term: tonnes with score: 0.255617
## Document 1411 top term: list with score: 0.135173
## Document 1421 top term: tonnes with score: 0.193252
## Document 1431 top term: export with score: 0.1687721
## Document 1441 top term: export with score: 0.1194719
## Document 1451 top term: french with score: 0.2070864
## Document 1461 top term: french with score: 0.2552763
```

```
## Document 1471 top term: profit with score: 0.08674927
## Document 1481 top term: french with score: 0.1460034
## Document 1491 top term: asset with score: 0.0859595
## Document 1501 top term: national with score: 0.1546675
## Document 1511 top term: demand with score: 0.07800491
## Document 1521 top term: quarter with score: 0.1686874
## Document 1531 top term: national with score: 0.1224177
## Document 1541 top term: court with score: 0.2455485
## Document 1551 top term: shareholder with score: 0.07571466
## Document 1561 top term: launch with score: 0.1142793
## Document 1571 top term: court with score: 0.3118569
## Document 1581 top term: deal with score: 0.08141975
## Document 1591 top term: took with score: 0.07494167
## Document 1601 top term: system with score: 0.09146292
## Document 1611 top term: rates with score: 0.1571569
## Document 1621 top term: manager with score: 0.1349808
## Document 1631 top term: production with score: 0.07472074
## Document 1641 top term: quality with score: 0.2214548
## Document 1651 top term: news with score: 0.1689574
## Document 1661 top term: products with score: 0.09073269
## Document 1671 top term: sales with score: 0.251801
## Document 1681 top term: give with score: 0.2214093
## Document 1691 top term: television with score: 0.2852412
## Document 1701 top term: chinese with score: 0.156692
## Document 1711 top term: china with score: 0.1853799
## Document 1721 top term: court with score: 0.1491325
## Document 1731 top term: chinese with score: 0.1726268
## Document 1741 top term: region with score: 0.1971713
## Document 1751 top term: communications with score: 0.09551507
## Document 1761 top term: post with score: 0.08242989
## Document 1771 top term: john with score: 0.06563219
## Document 1781 top term: local with score: 0.06803
## Document 1791 top term: local with score: 0.07702247
## Document 1801 top term: weve with score: 0.09429596
## Document 1811 top term: board with score: 0.1239485
## Document 1821 top term: quarter with score: 0.1364662
## Document 1831 top term: communications with score: 0.117778
## Document 1841 top term: speculation with score: 0.09407248
## Document 1851 top term: hong with score: 0.2934596
## Document 1861 top term: committee with score: 0.1781773
## Document 1871 top term: hong with score: 0.2186562
## Document 1881 top term: members with score: 0.103048
## Document 1891 top term: hong with score: 0.2080497
## Document 1901 top term: debt with score: 0.146977
## Document 1911 top term: air with score: 0.3292143
## Document 1921 top term: france with score: 0.171648
## Document 1931 top term: merger with score: 0.09656768
## Document 1941 top term: sale with score: 0.06893006
## Document 1951 top term: american with score: 0.138848
## Document 1961 top term: british with score: 0.175872
```

```
## Document 1971 top term: southern with score: 0.315525
## Document 1981 top term: southern with score: 0.3353851
## Document 1991 top term: inc with score: 0.08158496
## Document 2001 top term: local with score: 0.1347678
## Document 2011 top term: fund with score: 0.1173625
## Document 2021 top term: calls with score: 0.2341235
## Document 2031 top term: system with score: 0.2169313
## Document 2041 top term: rules with score: 0.3596212
## Document 2051 top term: loss with score: 0.191055
## Document 2061 top term: network with score: 0.1243563
## Document 2071 top term: data with score: 0.09847739
## Document 2081 top term: software with score: 0.1581851
## Document 2091 top term: technology with score: 0.1127821
## Document 2101 top term: southern with score: 0.2344924
## Document 2111 top term: currency with score: 0.2132672
## Document 2121 top term: committee with score: 0.1583459
## Document 2131 top term: hong with score: 0.1806951
## Document 2141 top term: hong with score: 0.09566209
## Document 2151 top term: kong with score: 0.1027718
## Document 2161 top term: trade with score: 0.1402632
## Document 2171 top term: software with score: 0.09298643
## Document 2181 top term: industries with score: 0.2630258
## Document 2191 top term: remains with score: 0.141081
## Document 2201 top term: exchange with score: 0.1513521
## Document 2211 top term: worldwide with score: 0.07261834
## Document 2221 top term: life with score: 0.153842
## Document 2231 top term: pounds with score: 0.1079674
## Document 2241 top term: pounds with score: 0.1469892
## Document 2251 top term: hong with score: 0.2252821
## Document 2261 top term: chinas with score: 0.1201805
## Document 2271 top term: hong with score: 0.2445497
## Document 2281 top term: kong with score: 0.1207022
## Document 2291 top term: hong with score: 0.2505947
## Document 2301 top term: internet with score: 0.2957714
## Document 2311 top term: quarter with score: 0.1506417
## Document 2321 top term: internet with score: 0.1620562
## Document 2331 top term: service with score: 0.1708775
## Document 2341 top term: internet with score: 0.1672343
## Document 2351 top term: growth with score: 0.07565705
## Document 2361 top term: dividend with score: 0.1071965
## Document 2371 top term: profits with score: 0.06789184
## Document 2381 top term: profits with score: 0.1232773
## Document 2391 top term: sales with score: 0.1160337
## Document 2401 top term: vice with score: 0.0957016
## Document 2411 top term: workers with score: 0.2030086
## Document 2421 top term: workers with score: 0.219566
## Document 2431 top term: effect with score: 0.1544476
## Document 2441 top term: campaign with score: 0.1502855
## Document 2451 top term: venture with score: 0.3176415
## Document 2461 top term: banks with score: 0.1187689
```

```
## Document 2471 top term: trade with score: 0.172193
## Document 2481 top term: economy with score: 0.1006216
## Document 2491 top term: investors with score: 0.1298885
```

Themes I'm seeing relate to location. I feel like that will be a big easy for grouping documents. There are words like china, czech, french, toronto, etc. Other topics include the business topic- technology, housing, data, restructuring, rates, tax, bank. Other terms I'm seeing that are harder to group include television, health, human, and local.

Now, let's get into topic modeling!

```
library(topicmodels)# Set the number of topics
k <- 5

lda_model <- LDA(DTM_all_reduced, k = k, control =list(seed =1234))
terms(lda_model,10)
```

```
##       Topic 1    Topic 2   Topic 3       Topic 4       Topic 5
##  [1,] "million"  "hong"    "will"        "will"        "percent"
##  [2,] "last"     "china"   "percent"     "new"         "million"
##  [3,] "new"      "kong"    "market"      "billion"     "billion"
##  [4,] "market"   "last"    "million"     "description" "will"
##  [5,] "also"     "two"     "also"        "analysts"    "company"
##  [6,] "will"     "company" "one"         "industry"    "profit"
##  [7,] "share"    "chinese" "company"     "expected"    "government"
##  [8,] "may"      "min"     "description" "business"    "bank"
##  [9,] "years"    "bank"    "new"         "last"        "min"
## [10,] "computer" "british" "can"         "companies"   "quarter"
```

```
#these terms are all super similar, so I want to see the terms with the
highest tf idf scores for each group.

doc_topic_distr <- posterior(lda_model)$topics
doc_topics <- apply(doc_topic_distr,1, which.max)

topic_top_tfidf_terms <-list()

for(topic in 1:k){
  topic_docs <- which(doc_topics == topic)
  tfidf_subset <- tfidf_all[topic_docs,]
  mean_tfidf <- colMeans(as.matrix(tfidf_subset))
  top_terms <- sort(mean_tfidf, decreasing =TRUE)[1:20]
  topic_top_tfidf_terms[[paste("Topic", topic)]]<- top_terms
}

for(topic in 1:k){
  cat("Top TF-IDF terms for", paste("Topic", topic),":\n")
  print(topic_top_tfidf_terms[[paste("Topic", topic)]])
  cat("\n")}
```

```
## Top TF-IDF terms for Topic 1 :
##    computer    software    internet  technology     service        corp
## 0.020520829 0.018533877 0.017006268 0.015476249 0.012451436 0.012212420
##         inc     workers       plant         bid       stock         new
## 0.011360274 0.010830093 0.010704804 0.010488844 0.009965429 0.009730128
##     company       offer   customers    southern   companies   president
## 0.009655515 0.009501892 0.009308560 0.009255788 0.008724711 0.008327835
##      people     million
## 0.008300343 0.008272316
##
## Top TF-IDF terms for Topic 2 :
##       hong      china       kong    chinese    beijing     chinas
kongs
## 0.05262690 0.04359429 0.04303971 0.03245328 0.02643835 0.01743554
0.01678351
##    foreign      trade      party   official  officials       bank
states
## 0.01475791 0.01314787 0.01254706 0.01174128 0.01165138 0.01142942
0.01133996
##     rights      human     united government        law      court
## 0.01077804 0.01068109 0.01046364 0.01045480 0.01042780 0.01023507
##
## Top TF-IDF terms for Topic 3 :
##      pounds         air       pence     million      tonnes      france
## 0.025066414 0.021570335 0.020471470 0.018739194 0.015731580 0.015699155
##    internet     percent     company      market       group     profits
## 0.012584842 0.012518385 0.012432555 0.011859082 0.011474568 0.011138071
##       sales      french   companies     british    business   insurance
## 0.011024977 0.010952340 0.009707389 0.009704512 0.009404505 0.009217234
##        will         plc
## 0.009199925 0.009125677
##
## Top TF-IDF terms for Topic 4 :
##        workers          banks        quarter       internet          local
##    0.018562232    0.017589741    0.014184565    0.013605853    0.013164651
##      companies       industry        billion          union communications
##    0.012450534    0.012144339    0.012089486    0.011162047    0.011137007
##        company          court            new           bank         system
##    0.011001631    0.010986688    0.010978383    0.010845330    0.010663666
##            amp           deal       analysts          plant           corp
##    0.010423314    0.010385389    0.010318647    0.009869230    0.009839169
##
## Top TF-IDF terms for Topic 5 :
##    quarter    percent       bank     profit      sales    million
billion
## 0.02375807 0.01941232 0.01786969 0.01776564 0.01739186 0.01676070
0.01636436
##   earnings      cents      share     shares      stock      index
analysts
## 0.01595233 0.01395427 0.01392559 0.01355987 0.01297639 0.01258281
```

```
0.01202753
##      czech        ltd       rose     points     prices  investors
## 0.01195750 0.01188389 0.01146860 0.01145969 0.01132450 0.01090022

#after running this with 15 topics, multiple seem like they could be grouped
- there are a number  of topics that could probably be grouped -
computer/tech stuff and global affairs. I am going to reduce to 10. I think
that should be plenty

#ten still feels like too much- I really think we could narrow it down to
about 5 main topics. This will push out more specific topics, but I want to
know overarching ideas.

#Topic 1 - Technology
#Topic 2 - China and International Relations
#Topic 3 - European Markets
#Topic 4 - Labor & Company Management
#Topic 5 - Financial Performance & Market Analysis



#now I want to rename topics
topic_names <-c("Topic 1"="Technology","Topic 2"="China and International
Relations","Topic 3"="European Markets","Topic 4"="Labor & Company
Management","Topic 5"="Financial Performance & Market Analysis")

for(i in 1:k){
  cat("Top TF-IDF terms for", topic_names[paste("Topic", i)],":\n")
  print(topic_top_tfidf_terms[[paste("Topic", i)]])
  cat("\n")}

## Top TF-IDF terms for Technology :
##    computer    software    internet  technology     service        corp
## 0.020520829 0.018533877 0.017006268 0.015476249 0.012451436 0.012212420
##         inc     workers       plant         bid       stock         new
## 0.011360274 0.010830093 0.010704804 0.010488844 0.009965429 0.009730128
##     company       offer   customers    southern   companies   president
## 0.009655515 0.009501892 0.009308560 0.009255788 0.008724711 0.008327835
##      people     million
## 0.008300343 0.008272316
##
## Top TF-IDF terms for China and International Relations :
##       hong      china       kong    chinese    beijing     chinas
kongs
## 0.05262690 0.04359429 0.04303971 0.03245328 0.02643835 0.01743554
0.01678351
##    foreign      trade      party   official  officials       bank
states
## 0.01475791 0.01314787 0.01254706 0.01174128 0.01165138 0.01142942
0.01133996
##     rights      human     united government        law      court
```

```
## 0.01077804 0.01068109 0.01046364 0.01045480 0.01042780 0.01023507
##
## Top TF-IDF terms for European Markets :
##      pounds         air       pence     million      tonnes      france
## 0.025066414 0.021570335 0.020471470 0.018739194 0.015731580 0.015699155
##    internet     percent     company      market       group     profits
## 0.012584842 0.012518385 0.012432555 0.011859082 0.011474568 0.011138071
##       sales      french   companies     british    business   insurance
## 0.011024977 0.010952340 0.009707389 0.009704512 0.009404505 0.009217234
##        will         plc
## 0.009199925 0.009125677
##
## Top TF-IDF terms for Labor & Company Management :
##        workers          banks        quarter       internet          local
##    0.018562232    0.017589741    0.014184565    0.013605853    0.013164651
##      companies       industry        billion          union communications
##    0.012450534    0.012144339    0.012089486    0.011162047    0.011137007
##        company          court            new           bank         system
##    0.011001631    0.010986688    0.010978383    0.010845330    0.010663666
##            amp           deal       analysts          plant           corp
##    0.010423314    0.010385389    0.010318647    0.009869230    0.009839169
##
## Top TF-IDF terms for Financial Performance & Market Analysis :
##    quarter    percent       bank     profit      sales    million
billion
## 0.02375807 0.01941232 0.01786969 0.01776564 0.01739186 0.01676070
0.01636436
##   earnings      cents      share     shares      stock      index
analysts
## 0.01595233 0.01395427 0.01392559 0.01355987 0.01297639 0.01258281
0.01202753
##      czech        ltd       rose     points     prices  investors
## 0.01195750 0.01188389 0.01146860 0.01145969 0.01132450 0.01090022

#let's look at how many documents focus on each topic
doc_topic_distr <- posterior(lda_model)$topics

# Assign each document to the topic with the highest probability
doc_topics <- apply(doc_topic_distr,1, which.max)# Count the number of
documents assigned to each topic
topic_counts <- table(doc_topics)# Print the results
print(topic_counts)

## doc_topics
##   1   2   3   4   5
## 423 649 512 376 540

#let's look at overall distributions
doc_topic_distr <- posterior(lda_model)$topics
```

```
# Sum the topic probabilities across all documents
topic_distribution_sums <- colSums(doc_topic_distr)# Print the summed
distributions
print("Summed Topic Distributions (should sum to 2500):")

## [1] "Summed Topic Distributions (should sum to 2500):"

print(topic_distribution_sums)

##        1        2        3        4        5
## 499.7739 500.7115 499.8173 499.6627 500.0346

#the distributions are very even - this probably points to a lot of overlap
```