



IIT School of Applied Technology

ILLINOIS INSTITUTE OF TECHNOLOGY

information technology & management

527 Data Analytics

April 12,14 2016

Week 14 Presentation

Week 14 Topic: Agenda

- ◆ Housekeeping
- ◆ Discussion Topic Highlights
- ◆ Proxy statement analysis using Yahoo
 - QA Corpus
 - QA Term Document Matrix
 - QA Key Event Indicators (KEIs)

Week 14 Topic: Housekeeping

- ◆ Update the spreadsheet with the company information and presentation date.
- ◆ Online students, make plans to be available for those 10-15 minutes of presentation.
- ◆ In-class students, make sure to be present for all presentation dates.
- ◆ Self-enroll yourself to your group for submissions and grading. If you do not enroll yourself, you will not be able to receive grades.

Week 14 Topic:

Discussion Topic Highlights

R for table parsing – alternative to readHTMLTable -- Using xpath:

```
proxy.table <- htmltab(doc = url, which = "//*[text()][contains(.,'Principal')][text()][contains(.,'Position')]/ancestor::table")
```

```
> URL_YAHOO_2015 <- "https://www.sec.gov/****"
> Comp_Summ_YAHOO_2015_13 <- URL_YAHOO_2015 %>%
+ read_html() %>%
+ html_nodes(xpath='/html/body/document/type/sequence/filename/description/text/table
[286]') %>%
+ html_table(fill = TRUE)
> Comp_Summ_YAHOO_2015_13 <- Comp_Summ_YAHOO_2015_13[[1]]
> Comp_Summ_YAHOO_2015_13 <- Comp_Summ_YAHOO_2015_13[3:5, c("X1", "X4", "X8",
"X12", "X16", "X20", "X24", "X28", "X32")]
> Comp_Summ_YAHOO_2015_13 <- rename(Comp_Summ_YAHOO_2015_13, c("X1"="Name
and Principal Position", "X4"="Year", "X8"="Salary", "X12"="Bonus", "X16"="Stock Awards",
"X20"="Option Awards", "X24"="Non-Equity Incentive Plan Compensation", "X28"="All
Other Compensations", "X32"="Total Compensation"))
> write.csv(Comp_Summ_YAHOO_2015_13, file = "Yahoo.csv")
```

Week 14 Topic:

Discussion Topic Highlights (cont.)

2) Cleansing table:

```
i<-2
while(i != (length(proxy.table)+1))
{ proxy.table[,i] <- replace(as.character(proxy.table[,i]), grep("Â", substr(proxy.table[,i],
1,nchar("Â")),fixed=TRUE) , NA)
  proxy.table[,i] <- replace(as.character(proxy.table[,i]), grep("â€", substr(proxy.table[,i],
1,nchar("â€")),fixed=TRUE) , NA)
  proxy.table[,i][proxy.table[,i] == "$"]<-NA
  i<-i+1
}
```

Week 14 Topic:

readLines Yahoo filings for 2010~2014

```
> yahoo_proxy2014 <- readLines("C:/Users/sshin/Desktop/Yahoo/yahoo_proxy2014.html")  
> yahoo_proxy2013 <- readLines("C:/Users/sshin/Desktop/Yahoo/yahoo_proxy2013.html")  
> yahoo_proxy2012 <- readLines("C:/Users/sshin/Desktop/Yahoo/yahoo_proxy2012.html")  
> yahoo_proxy2011 <- readLines("C:/Users/sshin/Desktop/Yahoo/yahoo_proxy2011.html")  
> yahoo_proxy2010 <- readLines("C:/Users/sshin/Desktop/Yahoo/yahoo_proxy2010.html")
```

Etc.

Better to create a loop per company or all companies...

Week 14 Topic:

Locating QA section Starting Lines

```
> qa_sentence_start <- "QUESTIONS AND ANSWERS ABOUT OUR PROXY MATERIALS"  
> grep(qa_sentence_start, yahoo_proxy2014, ignore.case = TRUE) [1] 237 688 1041  
*Take the third line number as the two are in the Table of Contents and Proposal sections
```

```
> qa_sentence_start <- "QUESTIONS AND ANSWERS ABOUT THE PROXY MATERIALS"  
> grep(qa_sentence_start, yahoo_proxy2013, ignore.case = TRUE) [1] 213 490  
*Take the latter line number as the first is from Table of Contents
```

```
> qa_sentence_start <- "QUESTIONS AND ANSWERS ABOUT THE PROXY MATERIALS"  
> grep(qa_sentence_start, yahoo_proxy2012, ignore.case = TRUE) [1] 218 375  
> grep(qa_sentence_start, yahoo_proxy2011, ignore.case = TRUE) [1] 206 401  
> grep(qa_sentence_start, yahoo_proxy2010, ignore.case = TRUE) [1] 207  
*Take the latter line number as the first is from Table of Contents
```

Etc.

Week 14 Topic:

Locating QA section Ending Lines

```
> qa_sentence_end <- "Your electronic delivery enrollment will be effective until you cancel it"
```

```
> grep(qa_sentence_end, yahoo_proxy2010, ignore.case = TRUE) [1] 813
```

```
> grep(qa_sentence_end, yahoo_proxy2011, ignore.case = TRUE) [1] 877
```

```
> grep(qa_sentence_end, yahoo_proxy2012, ignore.case = TRUE) [1] 904
```

```
> grep(qa_sentence_end, yahoo_proxy2013, ignore.case = TRUE) [1] 970
```

```
> qa_sentence_end <- "If you have questions about electronic delivery"
```

```
> grep(qa_sentence_end, yahoo_proxy2014, ignore.case = TRUE) [1] 1649
```


Week 14 Topic:

Save QA section to file

```
> qa_section2014 <- yahoo_proxy2014[688:1649]
> cat(qa_section2014,
file="C:/Users/sshin/Desktop/Yahoo/qa_section/yahoo_proxyqa_section2014.txt", sep="n",
append = TRUE)
> qa_section2013 <- yahoo_proxy2013[490:970]
> cat(qa_section2013,
file="C:/Users/sshin/Desktop/Yahoo/qa_section/yahoo_proxyqa_section2013.txt", sep="n",
append = TRUE)
> qa_section2012 <- yahoo_proxy2012[375:904]
> cat(qa_section2012,
file="C:/Users/sshin/Desktop/Yahoo/qa_section/yahoo_proxyqa_section2012.txt", sep="n",
append = TRUE)
> qa_section2011 <- yahoo_proxy2011[401:877]
> cat(qa_section2011,
file="C:/Users/sshin/Desktop/Yahoo/qa_section/yahoo_proxyqa_section2011.txt", sep="n",
append = TRUE)
> qa_section2010 <- yahoo_proxy2010[207:813]
> cat(qa_section2010,
file="C:/Users/sshin/Desktop/Yahoo/qa_section/yahoo_proxyqa_section2010.txt", sep="n",
append = TRUE)
```

Week 14 Topic:

Creating a QA section Corpus

```
> cname <- "C:/Users/sshin/Desktop/Yahoo/qa_section/"
```

```
> library(tm)
```

Loading required package: NLP

Warning messages: 1: package 'tm' was built under R version 3.2.3 2: package 'NLP' was built under R version 3.2.3

```
> qa_sections <- Corpus(DirSource(cname))
```

Week 14 Topic:

Inspecting the QA Corpus

```
> inspect(qa_sections[1])
<<VCorpus>> Metadata: corpus specific: 0, document level (indexed): 0 Content: documents:
1 [[1]] <<PlainTextDocument>> Metadata: 7 Content: chars: 70329

> dir(cname)
[1] "yahoo_proxyqa_section2010.txt" "yahoo_proxyqa_section2011.txt" [3]
"yahoo_proxyqa_section2012.txt" "yahoo_proxyqa_section2013.txt" [5]
"yahoo_proxyqa_section2014.txt"

> class(qa_sections) [1] "VCorpus" "Corpus"
> class(qa_sections[[1]]) [1] "PlainTextDocument" "TextDocument"

> summary(qa_sections)
Length Class Mode
yahoo_proxyqa_section2010.txt 2 PlainTextDocument list
yahoo_proxyqa_section2011.txt 2 PlainTextDocument list
yahoo_proxyqa_section2012.txt 2 PlainTextDocument list
yahoo_proxyqa_section2013.txt 2 PlainTextDocument list
yahoo_proxyqa_section2014.txt 2 PlainTextDocument list
```

Week 14 Topic:

Inspecting the QA Corpus Content - Raw

**Define a function to view the Corpus files:*

```
> install.packages("magrittr")
```

```
> library(magrittr)
```

```
> viewDocs <- function(d,n) {d %>% extract2(n) %>% as.character() %>% writeLines()}
```

```
> viewDocs(qa_sections, 1)
```

```
<TD VALIGN="top"> <P STYLE="margin-left:1.00em; text-indent:-1.00em"><FONT
STYLE="font-family:Times New Roman" SIZE="2"><A HREF="#toc25740_1">QUESTIONS
AND ANSWERS ABOUT THE PROXY MATERIALS AND OUR 2010 ANNUAL MEETING
OFnSHAREHOLDERS</A></FONT></P></TD>n<TD VALIGN="bottom"><FONT
SIZE="1">&nbsp;&nbsp;&nbsp;</FONT></TD>n<TD VALIGN="bottom" ALIGN="right"><FONT
STYLE="font-family:Times New Roman" SIZE="2">1</FONT></TD></TR>n<TR>n<TD
VALIGN="top"> <P STYLE="margin-left:1.00em; text-indent:-1.00em"><FONT
STYLE="font-family:Times New Roman" SIZE="2"><A
HREF="#toc25740_2">PROPOSAL&nbsp;&nbsp;&nbsp;NO.&nbsp;&nbsp;&nbsp;1 ELECTION OF
DIRECTORS</A></FONT></P></TD>n<TD VALIGN="bottom"><FONT
SIZE="1">&nbsp;&nbsp;&nbsp;</FONT></TD>n<TD VALIGN="bottom" ALIGN="right"><FONT
STYLE="font-family:Times New Roman" SIZE="2">8</FONT></TD></TR>n<TR>n<TD
VALIGN="top"> <P STYLE="margin-left:3.00em; text-indent:-1.00em"><FONT
```

```
....
```


Week 14 Topic:

List of stop words

```
> length(stopwords("english")) [1] 174
> stopwords("en")
[1] "i" "me" "my" "myself" "we" "our"
[7] "ours" "ourselves" "you" "your" "yours" "yourself"
[13] "yourselves" "he" "him" "his" "himself" "she"
[19] "her" "hers" "herself" "it" "its" "itself"
[25] "they" "them" "their" "theirs" "themselves" "what"
[31] "which" "who" "whom" "this" "that" "these"
[37] "those" "am" "is" "are" "was" "were"
[43] "be" "been" "being" "have" "has" "had"
[49] "having" "do" "does" "did" "doing" "would"
[55] "should" "could" "ought" "i'm" "you're" "he's"
[61] "she's" "it's" "we're" "they're" "i've" "you've"
[67] "we've" "they've" "i'd" "you'd" "he'd" "she'd"
[73] "we'd" "they'd" "i'll" "you'll" "he'll" "she'll"
[79] "we'll" "they'll" "isn't" "aren't" "wasn't" "weren't"
[85] "hasn't" "haven't" "hadn't" "doesn't" "don't" "didn't"
[91] "won't" "wouldn't" "shan't" "shouldn't" "can't" "cannot"
```

Week 14 Topic:

List of stop words (cont.)

[97] "couldn't" "mustn't" "let's" "that's" "who's" "what's"
[103] "here's" "there's" "when's" "where's" "why's" "how's"
[109] "a" "an" "the" "and" "but" "if"
[115] "or" "because" "as" "until" "while" "of"
[121] "at" "by" "for" "with" "about" "against"
[127] "between" "into" "through" "during" "before" "after"
[133] "above" "below" "to" "from" "up" "down"
[139] "in" "out" "on" "off" "over" "under"
[145] "again" "further" "then" "once" "here" "there"
[151] "when" "where" "why" "how" "all" "any"
[157] "both" "each" "few" "more" "most" "other"
[163] "some" "such" "no" "nor" "not" "only"
[169] "own" "same" "so" "than" "too" "very"

Week 14 Topic:

Transforming the QA Corpus - B

Replacing certin expressions with spaces:

```
> toSpace <- content_transformer(function(x, pattern) gsub(pattern, " ",x))
> qa_sections <- tm_map(qa_sections, toSpace, "/|<|>|"|=|@|\\|:|;|-|\\")
```

```
> viewDocs(qa_sections, 1)
```

```
td valign "top" p style "margin left:.em text indent: .em" font style "font family:times new
roman" size "" href "#toc_" questions answers proxy materials annual meeting
ofnshareholders font p td n td valign "bottom" font size "" &nbsp; &nbsp; font td n td valign
"bottom" align "right" font style "font family:times new roman" size "" font td tr n tr n td
valign "top" p style "margin left:.em text indent: .em" font style "font family:times new
roman" size "" href "#toc_" proposal&nbsp; .&nbsp; election directors font p td n td valign
"bottom" font size "" &nbsp; &nbsp; font td n td valign "bottom" align "right" font style "font
family:times new roman" size "" font td tr n tr n td valign "top" p style "margin left:.em text
indent: .em" font style "font family:times new roman" size "" href "#toc_" voting standard font
p td n td valign "bottom" font size "" &nbsp; &nbsp; font td n td valign "bottom" align "right"
font style "font family:times new roman" size "" font td tr n tr n td valign "top" p style
"margin left:.em text indent: .em" font style "font family:times new roman" size "" href ...
```


Week 14 Topic:

Transforming the QA Corpus – C

Remove Punctuations:

```
> qa_sections <- tm_map(qa_sections, removePunctuation)
```

*Punctuation characters: ! " # \$ % & ' () * + , - . / : ; < = > ? @ [\] ^ _ ` { | } ~.

Strip white spaces:

```
> qa_sections <- tm_map(qa_sections, stripWhitespace)
```

```
> viewDocs(qa_sections, 1)
```

```
td valign top p style margin leftem text indent em font style font familytimes new roman size
a href toc questions and answers about the proxy materials and our annual meeting
ofnshareholders a font p td n td valign bottom font size nbsp nbsp font td n td valign bottom
align right font style font familytimes new roman size font td tr n tr n td valign top p style
margin leftem text indent em font style font familytimes new roman size a href toc
proposalnbsp nonbsp election of directors a font p td n td valign bottom font size nbsp nbsp
font td n td valign bottom align right font style font familytimes new roman size font td tr n
tr n td valign top p style margin leftem text indent em font style font familytimes new roman
size a href toc voting standard a font p td n td valign bottom font size nbsp nbsp font td n
td...
```

Week 14 Topic:

Transforming the QA Corpus - D

Remove known often words:

```
> qa_sections <- tm_map(qa_sections, removeWords, c("b", "q", "a", "i", "e", "font",
"style", "n", "trim", "size", "font", "can", "also", "e", "mail", "via", "td", "align", "border",
"familytimes", "roman", "p", "tr", "nbsp", "with", "table", "cellspacing", "valign", "cellpadding",
"width", "top", "left", "sizepx", "telephone", "if", "may", "help", "us", "will", "please", "unless",
"visit", "thnbsp", "toppx", "bottompx", "nnn", "address", "nonbsp", "new", "bottom", "em"))
```

** We remove more words after terms are identified in the matrix later...*

```
> viewDocs(qa_sections, 1)
```

```
margin leftem text indent href toc questions and answers about the proxy materials and our
annual meeting ofnshareholders right margin leftem text indent href toc proposalnbsp
election of directors right margin leftem text indent href toc voting standard right margin
leftem text indent href toc nominees right margin leftem text indent href toc corporate
governance right margin leftem text indent href toc director compensation right margin
leftem text indent href toc proposalnbsp approval of amendments to the directors stock plan
nn right margin leftem text indent href toc summary description directors plan right margin
leftem text indent href toc aggregate past grants under directors plan right margin leftem
text indent
```

Week 14 Topic:

Transforming the QA Corpus - E

Specific Transformation:

```
> toString <- content_transformer(function(x, from, to) gsub(from, to, x))  
> qa_sections <- tm_map(qa_sections, toString, "broker bank", "bb")
```

** We will do more specific transformations after terms are identified in the matrix later...*

Week 14 Topic:

Transforming the QA Corpus - F

Stemming:

```
> install.packages("SnowballC")
```

```
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.2/SnowballC_0.5.1.zip' Content
type 'application/zip' length 3076512 bytes (2.9 MB) downloaded 2.9 MB package 'SnowballC'
successfully unpacked and MD5 sums checked The downloaded binary packages are in
C:\Users\sshin\AppData\Local\Temp\RtmpQBcYVv\downloaded_packages
```

```
> qa_sections <- tm_map(qa_sections, stemDocument)
```

**Stemming uses an algorithm that removes common word endings for English words, such as “es”, “ed” and “s”.*

```
> viewDocs(qa_sections, 1)
```

```
margin leftem text indent href toc question and answer about the proxi materi and our
annual meet ofnsharehold right margin leftem text indent href toc proposalnbsp elect of
director right margin leftem text indent href toc vote standard right margin leftem text
indent href toc nomine right margin leftem text indent href toc corpor govern right margin
leftem text indent href toc director compens right margin leftem text indent href toc
proposalnbsp approv of amend to the director stock plan nn right margin leftem text indent
href toc summari descript director plan ...
```

Week 14 Topic:

Create a TermDocumentMatrix

```
> qa_sections <- tm_map(qa_sections, PlainTextDocument)
```

**We do this because the latest version of tm has a “changes in tm 0.6.0 seems to have broken it. The problem is that the functions tolower and trim won't necessarily return TextDocuments (it looks like the older version may have automatically done the conversion). They instead return characters and the DocumentTermMatrix isn't sure how to handle a corpus of characters.” from <http://stackoverflow.com/questions/24191728/documenttermmatrix-error-on-corpus-argument>*

```
> qa_dtm <- DocumentTermMatrix(qa_sections)
```

```
> qa_dtm
```

```
<<DocumentTermMatrix (documents: 5, terms: 965)>>
```

```
Non-/sparse entries: 2710/2115
```

```
Sparsity          : 44%
```

```
Maximal term length: 25
```

```
Weighting          : term frequency (tf)
```

Week 14 Topic:

Inspecting the TermDocumentMatrix - A

```
> inspect(qa_dtm[1:5, 100:105])  
<<DocumentTermMatrix (documents: 5, terms: 6)>>  
Non-/sparse entries: 20/10  
Sparsity          : 33%  
Maximal term length: 9  
Weighting         : term frequency (tf)
```

	Terms					
Docs	basi	bear	bearingna	becaus	becom	begin
character(0)	0	2	1	0	1	2
character(0)	0	2	1	0	1	2
character(0)	0	2	1	1	1	2
character(0)	2	4	0	0	2	3
character(0)	0	2	0	0	1	2

Week 14 Topic:

Inspecting the TermDocumentMatrix - B

```
> freq <- colSums((as.matrix(qa_dtm)))
```

```
> length(freq)
```

```
[1] 965
```

```
> ord <- order(freq)
```

```
> freq[head(ord)]
```

<i>abovenaddress</i>	<i>absentninstru</i>	<i>acceler</i>	<i>additionnsharehold</i>	<i>affirmativenvot</i>	<i>agggreg</i>
1	1	1	1	1	1

```
> freq[tail(ord)]
```

<i>annual collapsecollaps</i>	<i>sizept</i>	<i>meet</i>	<i>vote</i>	<i>margin</i>
428	440	478	493	595
				1549

```
> head(table(freq), 15)
```

```
freq
```

1	2	3	4	5	6	7	8	9	10	11	12	13	14	16
326	108	54	54	55	65	21	11	17	18	13	15	13	6	6

```
> tail(table(freq), 15)
```

```
freq
```

149	172	240	246	271	273	315	329	413	428	440	478	493	595	1549
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Week 14 Topic:

Conversion to Matrix and Save to CSV

```
> m<-as.matrix(qa_dtm)
```

```
> dim(m)
```

```
[1] 5 965
```

```
> write.csv(m,file= "C:/Users/sshin/Desktop/Yahoo/qa_section/qa_dtm.csv")
```


Week 14 Topic: Next Week Topics

- ◆ Removing sparse terms
- ◆ Identifying Frequent Items and Associations
- ◆ Correlation Plots
- ◆ Plotting Word Frequencies
- ◆ Word Clouds
- ◆ Quantitative Analysis of Text:
 - Counts
 - Frequencies
 - Heatmaps
 - Distances
- ◆ Defining KEIs

Week 14 Topic:

Removing Sparse Terms

```
> dtms <- removeSparseTerms(qa_dtm, 0.1)
> dim(dtms) [1] 5 336
> inspect(dtms)
<<DocumentTermMatrix (documents: 5, terms: 336)>>
```

Non-/sparse entries: 1680/0

Sparsity : 0%

Maximal term length: 19

Weighting : term frequency (tf)

	Terms									
Docs	abl	abstent	access	account	act	action	addit	adjourn	admiss	admit
character(0)	2	3	1	8 6	2	3	4	3	2	
character(0)	2	7	1	6 6	2	3	3	3	2	
character(0)	2	8	1	7 5	1	5	1	3	2	
character(0)	3	12	2	12 12	4	9	2	6	4	
character(0)	2	7	2	8 6	2	4	1	3	3	

Week 14 Topic:

Removing Sparse Terms (cont.)

Docs	adopt	advanc	affirm	against	agent	allow	altern	although	always	amend
character(0)	2	2	2	2	2	1	1	1	7	9
character(0)	2	2	2	4	2	1	1	1	6	2
character(0)	1	2	1	2	2	1	1	1	6	10
character(0)	4	4	2	6	3	2	2	2	11	4
character(0)	1	3	1	7	1	4	1	1	10	15
Terms										

Docs	amount	and	anniversari	announc	annual	answer	appli	applic	appoint
character(0)	1	5	1	2	76	1	2	1	5
character(0)	1	2	1	2	70	1	2	2	4
character(0)	1	2	1	2	66	1	2	2	6
character(0)	2	4	2	3	132	2	2	3	10
character(0)	1	7	1	2	84	9	1	2	7

Week 14 Topic: Removing Sparse Terms (cont.)

	Terms										
Docs	approve	are	assist	attend	attent	audit	author	avail	avenu	ballot	bank
character(0)	6	1	2	11	1	2	3	23	4	1	9
character(0)	3	1	2	12	1	1	3	22	2	1	8
character(0)	10	1	2	13	1	1	2	6	2	2	12
character(0)	12	1	4	26	2	2	6	41	5	4	23
character(0)	13	1	2	13	1	1	11	24	4	2	9

	Terms									
Docs	bear	becom	begin	benefici	board	both	break	broker	brokerag	bulki
character(0)	2	1	2	16	12	1	7	16	2	1
character(0)	2	1	2	15	13	1	6	15	2	1
character(0)	2	1	2	17	12	1	6	23	2	1
character(0)	4	2	3	35	21	2	11	41	4	1
character(0)	2	1	2	19	15	1	10	18	2	1

Week 14 Topic:

Removing Sparse Terms (cont.)

```
> freq <- colSums(as.matrix(dtms))
```

```
> freq
```

abl	abstent	access	account
11	37	7	41
act	action	addit	adjourn
35	11	24	11
admiss	admit	adopt	advanc
18	13	10	13
affirm	against	agent	allow
8	21	10	9
altern	although	alway	amend
6	6	40	40
amount	and	anniversari	announc
6	20	6	11
annual	answer	appli	applic
428	14	9	10
appoint	approv	are	assist
32	44	5	12

Week 14 Topic:

Removing Sparse Terms (cont.)

```
> freq <- colSums(as.matrix(dtms))
```

```
> freq
```

abl	abstent	access	account
11	37	7	41
act	action	addit	adjourn
35	11	24	11
admiss	admit	adopt	advanc
18	13	10	13
affirm	against	agent	allow
8	21	10	9
altern	although	alway	amend
6	6	40	40
amount	and	anniversari	announc
6	20	6	11
annual	answer	appli	applic
428	14	9	10
appoint	approv	are	assist
32	44	5	12

Week 14 Topic:

Week 13 Assignment – due today!

PART 1: Document/Table Parsing (Automation) & Trading Activity Analysis

- ◆ This is automation of Week 11 DEF 14A (Proxy) Filing – Executive Compensation Trend analysis. In essence, you are to automate Week 11 assignment using code. Using XML's readHTMLTable, RCurl, or other available functions in R.
- ◆ Make sure to collect all available DEF 14A filings available for the companies. Make sure to chart the salary, option awards, and/or stock awards over the available years for the company.
- ◆ Highlight any stock trading activity changes before and after filing.

Week 14 Topic:

Week 14 Assignment – due 21st

PART 2: Corpus/Term Document Matrix & Identifying KEIs

- ◆ Create a corpus of filings per company. The Corpus is to house the Q&A sections for each filing. Striped of punctuation, etc.
- ◆ Generate a Term Document Matrix of the Q&A content.
- ◆ Highlight any patterns or trends found in analyzing the results of the Q&A Term Document Matrix statistics.

Week 14 Topic:

Final Project Presentation – Due 26th

The final project should contain the following:

- ◆ Table of Content – well formatted (1 slide)
- ◆ Company/Industry Selection/Background/Overview (1 slide)

PART 1:

Per company (1-5 slides):

- ◆ Make sure to collect all available DEF 14A filings available for the companies. Make sure to chart the salary, option awards, and/or stock awards over the available years for the company.
- ◆ Highlight any stock trading activity changes/activities before and after filing.

For all companies/industry (1-3 slides):

- ◆ Make sure to chart the salary, option awards, and/or stock awards over the available years for the company.
- ◆ Highlight any stock trading activity changes before and after filing.

Week 14 Topic:

Final Project Presentation - Due 26th

PART 2:

Per company (1-5 slides):

- ◆ Highlight any patterns or trends found in analyzing the results of the Q&A Term Document Matrix statistics.

For all companies/industries (1-3 slides):

- ◆ Create a corpus of filings per company. The Corpus is to house the Q&A sections for each filing. Striped of punctuation, etc.
- ◆ Generate a Term Document Matrix of the Q&A content
- ◆ Highlight any patterns or trends found in analyzing the results of the Q&A Term Document Matrix statistics.
- ◆ Highlight any KEIs identified

Week 14 Topic:

Text Mining Definitions

◆ Corpus

A collection of documents is called a *corpus*.

◆ Tokens, Separators, and Terms

A document consists of a set of tokens. A *token* is a contiguous string of characters that does not contain a separator. A *separator* is a special character such as a blank or mark of punctuation. A *term* is a token or a sequence of tokens (such as *White House*) with a specific meaning in a given language.

Week 14 Topic:

Text Extraction by Increasing Complexity

1. Token extraction
2. Term extraction (token + language \Rightarrow term)
3. Concept extraction (nouns, noun phrases)
4. Entity extraction (associates nouns with entities – for example, Person: Mr. White, Location: White House)
5. Atomic fact extraction (associates nouns with verbs, that is, subject \Rightarrow action – for example, terrorist \Rightarrow bombed)
6. Complex fact extraction (natural language understanding)

Week 14 Topic:

Contents of a Document

- ◆ A document consists of the following elements:
 - letters
 - words
 - sentences
 - paragraphs
 - punctuation
 - possible structural items (chapters, sections)
- ◆ The elements of a document can be counted and compared across documents.

Week 14 Topic:

Zipf's Law

Let t_1, t_2, \dots, t_n be the terms in a document collection arranged in order from most frequent to least frequent.

Let f_1, f_2, \dots, f_n be the corresponding frequencies of the terms. The frequency f_k for term t_k is proportional to $1/k$.

- ◆ Zipf's law and its variants help quantify the importance of terms in a document collection. (Konchady 2006)
- ◆ "The product of the frequency of words (f) and their rank (r) is approximately constant."
- ◆ In practice, Zipf's Law is derived as a Power Law, with free parameters that can be estimated based on the document collection.
- ◆ The general formula is shown here: $f_k = C / (\omega + k)^\theta$
- ◆ where C is a constant such that, for given ω and θ , $\sum_{k=1}^n f_k = T$, the total number of words in the document collection. The parameters ω and θ are estimated for a given document collection.
- ◆ Application of Zipf's Law permits identification of important terms for purposes such as describing concepts or topics. You can see the results of Zipf's Law in text mining applications (for example, in the list of terms used to define a topic). Along with methods such as Hidden Markov Models (HMM), the implementation is often hidden from the user. Only the results of the methodology are visible.

Week 14 Topic:

Zipf's Law Relevance to Text Mining

- ◆ Often, a few, very frequent terms are not good discriminators.
 - *stop* words, for example, the, and, an, or, of
 - often words that are described in linguistics as *closed-class* words, which is a grammatical class that does not get new members
- ◆ Typically, there is the following in a document collection:
 - a high number of infrequent terms
 - an average number of average frequency terms
 - a low number of high frequency terms
- ◆ Terms that are neither high nor low frequency are the most informative.

Week 14 Topic:

7 Text Mining Practices

http://datamininglab.com/images/pdfs/PracticalTextMining_Excerpt.pdf:

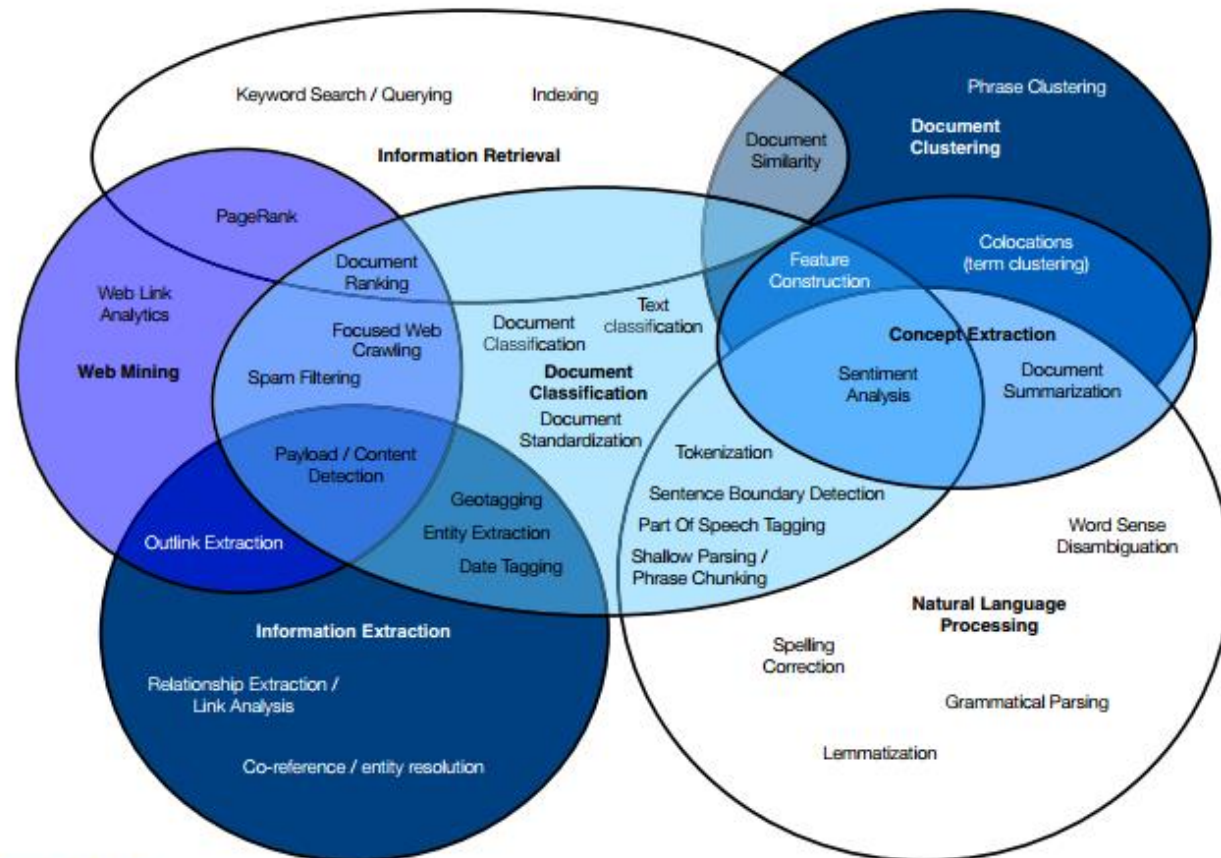


FIGURE 2.3

Visualizing the seven text mining practice areas (ovals) and how specific text mining tasks (labels within ovals) exist at their intersections.

Week 14 Topic:

Key Event Indicator (KEI)

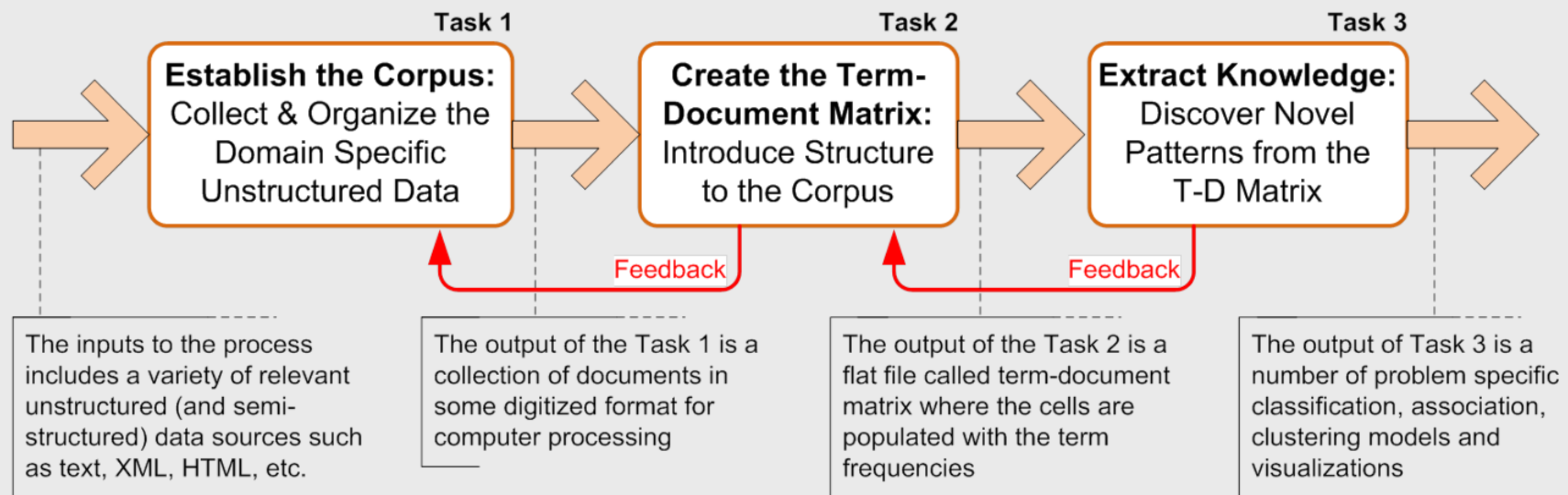
For the final project, we define KEIs as indicators that can be used as alerts or notifications of unusual events that may prompt an investigation for action. For example:

- ◆ Unusual amount of Questions and Answers discussions. Defined by over a certain percentage from the average character count of the section.
- ◆ Unique and/or low frequency terms in the Questions and Answer section with inferred factual context.
- ◆ Entity references to executive position changes that may (with some inference) affect company stock prices.
- ◆ Entity references to company structure changes that may (with some inference) affect company stock prices.
- ◆ Etc.

Week 14 Topic:

Three Step Text Mining Process

www.washburn.edu/faculty/boncella/PPT/CH05.ppt:



The three-step text mining process

Week 14 Topic:

Install tm

```
> library(tm)
```

Loading required package: NLP Warning messages: 1: package 'tm' was built under R version 3.2.3 2: package 'NLP' was built under R version 3.2.3

```
> getSources()
```

*[1] "DataframeSource" "DirSource" "URISource" "VectorSource" "XMLSource" [6]
"ZipSource"*

```
> getReaders()
```

*[1] "readDOC" "readPDF" "readPlain" [4] "readRCV1" "readRCV1asPlain"
"readReut21578XML" [7] "readReut21578XMLasPlain" "readTabular"
"readTagged" [10] "readXML"*

Week 14 Topic:

Set Up – Document Repository

```
> cname <- "C:/Users/sshin/Desktop/Yahoo"
```

```
> cname
```

```
[1] "C:/Users/sshin/Desktop/Yahoo"
```

```
> length(dir(cname))
```

```
[1] 6
```

```
> dir(cname)
```

```
[1] "0001193125-13-187918.txt" "0001193125-14-172132.txt" "0001193125-15-156926.txt" [4] "yahoo_proxy2013.html" "yahoo_proxy2014.html" "yahoo_proxy2015.html"
```

Week 14 Topic: Create a Corpus

```
> proxies <- Corpus(DirSource(cname))
```

```
> proxies
```

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

```
> class(proxies)
```

```
[1] "VCorpus" "Corpus"
```

```
> class(proxies[[1]])
```

```
[1] "PlainTextDocument" "TextDocument"
```

```
> summary(proxies)
```

```
Length Class Mode 0001193125-13-187918.txt 2 PlainTextDocument list 0001193125-14-172132.txt 2 PlainTextDocument list 0001193125-15-156926.txt 2 PlainTextDocument list yahoo_proxy2013.html 2 PlainTextDocument list yahoo_proxy2014.html 2 PlainTextDocument list yahoo_proxy2015.html 2 PlainTextDocument list
```

Week 14 Topic:

Installing stringi

Install:

> install.packages("stringi")

Open library:

>library(stringi)

Getting help:

>library(help=stringi)

stri_extract_all	Extract Occurrences of a Pattern
stri_extract_all_boundaries	Extract Text Between Text Boundaries
stri_flatten	Flatten a String
stri_info	Query Default Settings for 'stringi'
stri_install_check	Installation-Related Utilities [DEPRECATED]
stri_isempty	Determine if a String is of Length Zero
stri_join	Concatenate Character Vectors
stri_length	Count the Number of Code Points
stri_list2matrix	Convert a List to a Character Matrix
stri_locale_info	Query Given Locale
stri_locale_list	List Available Locales
stri_locale_set	Set or Get Default Locale in 'stringi'
stri_locate_all	Locate Occurrences of a Pattern
stri_locate_all_boundaries	Locate Specific Text Boundaries
stri_match_all	Extract Regex Pattern Matches, Together with Capture Groups
stri_numbytes	Count the Number of Bytes
stri_opts_brkiter	Generate a List with BreakIterator Settings
stri_opts_collator	Generate a List with Collator Settings
stri_opts_fixed	Generate a List with Fixed Pattern Search Engine's Settings
stri_opts_regex	Generate a List with Regex Matcher Settings
stri_order	Ordering Permutation and Sorting
stri_pad_both	Pad (Center/Left/Right Align) a String
stri_rand_lipsum	A Lorem Ipsum Generator
stri_rand_shuffle	Randomly Shuffle Code Points in Each String
stri_rand_strings	Generate Random Strings
stri_read_lines	[DRAFT API] Read Text Lines from a Text File
stri_read_raw	[DRAFT API] Read Whole Text File as Raw
stri_replace_all	Replace Occurrences of a Pattern
stri_replace_na	Replace Missing Values in a Character Vector
stri_reverse	Reverse Each String
stri_split	Split a String By Pattern Matches
stri_split_boundaries	Split a String at Specific Text Boundaries
stri_split_lines	Split a String Into Text Lines

Week 14 Topic:

Locating text boundaries

http://finzi.psych.upenn.edu/library/stringi/html/stri_locate_boundaries.html:

Description

These functions locate specific text boundaries (like character, word, line, or sentence boundaries). `stri_locate_all_*` locate all the matches. On the other hand, `stri_locate_first_*` and `stri_locate_last_*` give the first or the last matches, respectively.

Usage

- ◆ `stri_locate_all_boundaries(str, omit_no_match = FALSE, ..., opts_brkiter = NULL)`
- ◆ `stri_locate_last_boundaries(str, ..., opts_brkiter = NULL)`
- ◆ `stri_locate_first_boundaries(str, ..., opts_brkiter = NULL)`
- ◆ `stri_locate_all_words(str, omit_no_match = FALSE, locale = NULL)`
- ◆ `stri_locate_last_words(str, locale = NULL)`
- ◆ `stri_locate_first_words(str, locale = NULL)`

Week 13 Topic:

Yahoo DEF 14A filings

- ◆ Yahoo has 18 filings from 1997 to 2015.

EDGAR Search Results

https://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&CIK=0001011006&type=def+14a&dateb=&owner=exclude&count=40

U.S. Securities and Exchange Commission

EDGAR Search Results

SEC Home » Search the Next-Generation EDGAR System » Company Search » Current Page

YAHOO INC CIK#: 0001011006 (see all company filings)

SIC: 7373 - SERVICES-COMPUTER INTEGRATED SYSTEMS DESIGN
State location: CA | State of Inc.: DE | Fiscal Year End: 1231
(Assistant Director Office: 3)
Note: Ownership filings are available at this link. Reports containing extracted ownership data are temporarily unavailable.

Business Address
YAHOO! INC.
701 FIRST AVENUE
SUNNYVALE CA 94089
4083493300

Mailing Address
701 FIRST AVENUE
SUNNYVALE CA 94089

Filter Results: Filing Type: def 14a Prior to: (YYYYMMDD) Ownership? ☐ include ☒ exclude ☐ only Limit Results Per Page 40 Entries Search Show All

Items 1 - 18 RSS Feed

Filings	Format	Description	Filing Date	File/Film Number
DEF 14A	Documents	Other definitive proxy statements Acc-no: 0001193125-15-156926 (34 Act) Size: 2 MB	2015-04-29	000-28018 15813744
DEF 14A	Documents	Other definitive proxy statements Acc-no: 0001193125-14-172132 (34 Act) Size: 2 MB	2014-04-30	000-28018 14798942
DEF 14A	Documents	Other definitive proxy statements Acc-no: 0001193125-13-187918 (34 Act) Size: 1 MB	2013-04-30	000-28018 13799046
DEF 14A	Documents	Other definitive proxy statements Acc-no: 0001193125-12-258870 (34 Act) Size: 1 MB	2012-06-04	000-28018 12886969
DEF 14A	Documents	Other definitive proxy statements Acc-no: 0001193125-11-118858 (34 Act) Size: 1 MB	2011-04-29	000-28018 11795903
DEF 14A	Documents	Other definitive proxy statements Acc-no: 0001193125-10-099552 (34 Act) Size: 1 MB	2010-04-29	000-28018 10782547
DEF 14A	Documents	Other definitive proxy statements Acc-no: 0001193125-09-092231 (34 Act) Size: 1 MB	2009-04-29	000-28018 09780142
DEF 14A	Documents	Other definitive proxy statements Acc-no: 0000891618-07-000262 (34 Act) Size: 1 MB	2007-04-30	000-28018 07802049
DEF 14A	Documents	Other definitive proxy statements Acc-no: 0001011006-06-005177 (34 Act) Size: 274 KB	2006-04-14	000-28018 06780900

Week 13 Topic:

Yahoo DEF 14A - html version

<https://www.sec.gov/Archives/edgar/data/1011006/000119312515156926/d868077ddef14a.htm>:

DEFINITIVE PROXY STATEMENT

DEF 14A 1 d868077ddef14a.htm DEFINITIVE PROXY STATEMENT

[Table of Contents](#)

**UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
WASHINGTON, D.C. 20549**

SCHEDULE 14A

Proxy Statement Pursuant to Section 14(a) of the
Securities Exchange Act of 1934
(Amendment No.)

Filed by the Registrant ☒ Filed by a Party other than the Registrant ☐

Check the appropriate box:

☐ Preliminary Proxy Statement

☐ Confidential, for Use of the Commission Only (as permitted by Rule 14a-6(e)(2))

☒ Definitive Proxy Statement

☐ Definitive Additional Materials

☐ Soliciting Material Pursuant to §240.14a-11(c) or §240.14a-12

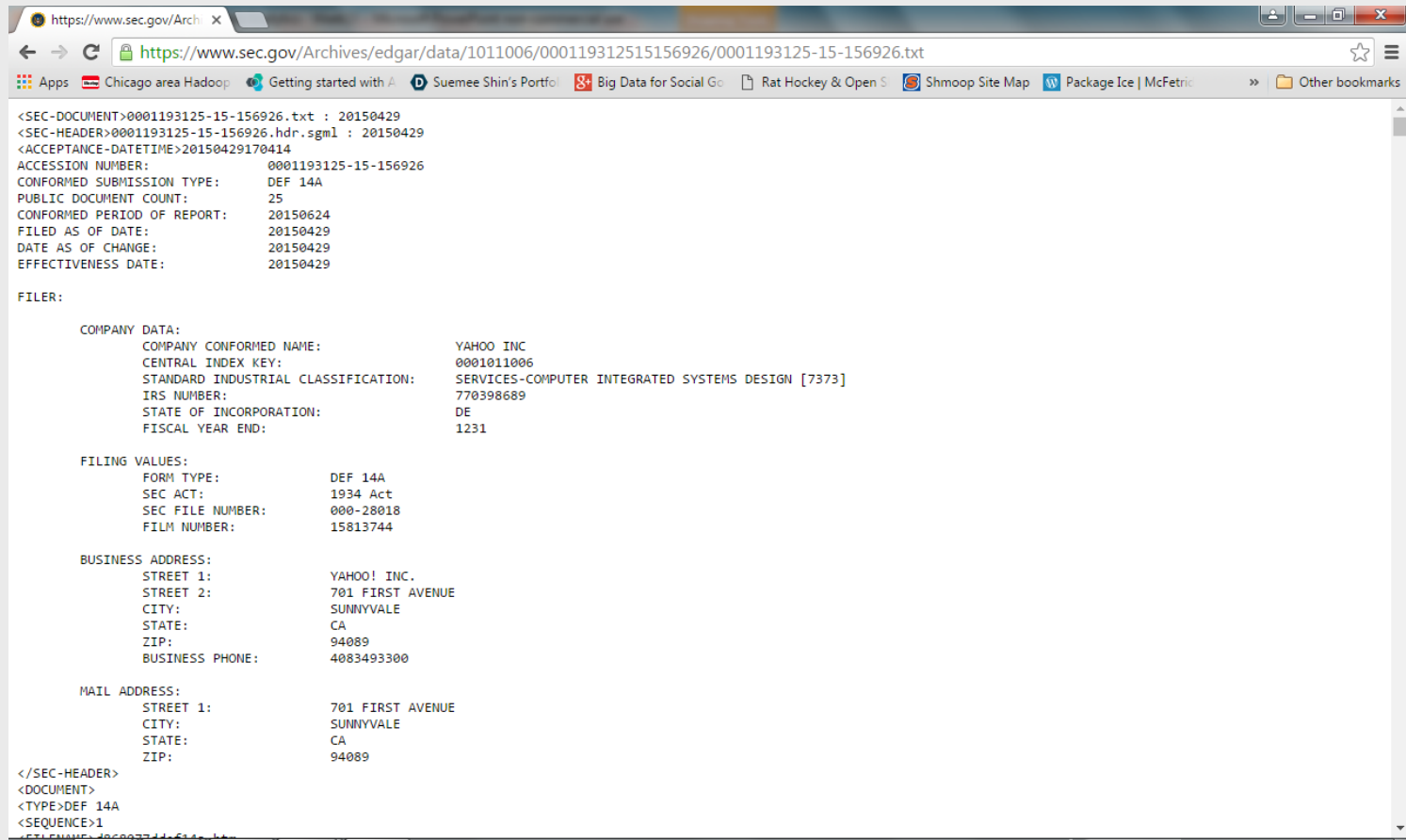
Yahoo! Inc.
(Name of Registrant as Specified In Its Charter)

(Name of Person(s) Filing Proxy Statement, if other than the Registrant)

Week 13 Topic:

Yahoo DEF 14A - txt version

<https://www.sec.gov/Archives/edgar/data/1011006/000119312515156926/0001193125-15-156926.txt>:



The screenshot shows a web browser window with the URL <https://www.sec.gov/Archives/edgar/data/1011006/000119312515156926/0001193125-15-156926.txt>. The browser's address bar and tabs are visible at the top. The main content area displays the text of the SEC filing, which is a DEF 14A form for Yahoo Inc. The text is formatted with line breaks and indentation to represent the original document's structure. The filing is dated 20150429 and is for the period ending 20150624. The filer is Yahoo Inc., with its company data and business address listed. The filing values section indicates the form type is DEF 14A, the SEC act is 1934 Act, the SEC file number is 000-28018, and the film number is 15813744. The business address is 701 First Avenue, Sunnyvale, CA 94089. The mail address is also listed as 701 First Avenue, Sunnyvale, CA 94089. The text ends with the sequence number 1.

```
<SEC-DOCUMENT>0001193125-15-156926.txt : 20150429
<SEC-HEADER>0001193125-15-156926.hdr.sgml : 20150429
<ACCEPTANCE-DATETIME>20150429170414
ACCESSION NUMBER:      0001193125-15-156926
CONFORMED SUBMISSION TYPE: DEF 14A
PUBLIC DOCUMENT COUNT: 25
CONFORMED PERIOD OF REPORT: 20150624
FILED AS OF DATE:      20150429
DATE AS OF CHANGE:      20150429
EFFECTIVENESS DATE:    20150429

FILER:

  COMPANY DATA:
    COMPANY CONFORMED NAME:      YAHOO INC
    CENTRAL INDEX KEY:            0001011006
    STANDARD INDUSTRIAL CLASSIFICATION: SERVICES-COMPUTER INTEGRATED SYSTEMS DESIGN [7373]
    IRS NUMBER:                  770398689
    STATE OF INCORPORATION:      DE
    FISCAL YEAR END:             1231

  FILING VALUES:
    FORM TYPE:                   DEF 14A
    SEC ACT:                    1934 Act
    SEC FILE NUMBER:             000-28018
    FILM NUMBER:                 15813744

  BUSINESS ADDRESS:
    STREET 1:                   YAHOO! INC.
    STREET 2:                   701 FIRST AVENUE
    CITY:                       SUNNYVALE
    STATE:                      CA
    ZIP:                        94089
    BUSINESS PHONE:             4083493300

  MAIL ADDRESS:
    STREET 1:                   701 FIRST AVENUE
    CITY:                       SUNNYVALE
    STATE:                      CA
    ZIP:                        94089

</SEC-HEADER>
<DOCUMENT>
<TYPE>DEF 14A
<SEQUENCE>1
FILED 20150429 17:04:14
```

Week 13 Topic:

Yahoo's Executive Compensation Table

Target Table in HTML View:

Summary Compensation Table—2012–2014

The following table presents 2012–2014 summary compensation information for our Named Executive Officers. As required by SEC rules, stock awards (RSUs) and option awards are shown as compensation for the year in which they were granted (even if they have multi-year vesting schedules), and are valued based on their grant date fair values for accounting purposes. Accordingly, the table includes stock and option awards granted in the years shown even if they were scheduled to vest in later years, and even if they were subsequently forfeited (such as upon the executive's termination). Therefore, the stock and option columns do not report whether the officer realized a financial benefit from the awards (such as by vesting in stock or exercising options).

Name and Principal Position	Year	Salary \$(1)	Bonus \$(1)	Stock Awards \$(2)(3)(4)	Option Awards \$(2)(3)	Non-Equity Incentive Plan Compensation \$(5)	All Other Compensation \$(6)	Total \$(2)
Marissa A. Mayer Chief Executive Officer	2014	1,000,000	0	11,752,355(7)	28,194,288(7)	1,108,800	28,065	42,083,508
	2013	1,000,000	2,250	8,312,316	13,847,283	1,700,000	73,863	24,935,712
	2012	454,862	0	35,000,002	0	1,120,000	40,540	36,615,404
Ken Goldman Chief Financial Officer	2014	600,000	0	2,813,080	9,327,427	300,000	4,549	13,045,056
	2013	600,000	0	2,597,612	2,290,527	500,000	4,615	5,992,754
	2012	116,667	100,000	7,262,357	0	0	29	7,479,053
David Filo Co-Founder and Chief Yahoo	2014	1	0	0	0	0	0	1
	2013	1	0	0	0	0	0	1
	2012	1	0	0	0	0	0	1
Ronald S. Bell General Counsel	2014	600,000	0	3,282,107	0	300,000	4,549	4,186,656
	2013	600,000	0	3,896,386	0	450,000	4,615	4,951,001
	2012	442,763	206,800	558,300	0	443,200	4,424	1,655,487
Henrique de Castro(8) Former Chief Operating Officer	2014	27,083	0	0	0	0	1,177,157	1,204,240
	2013	600,000	0	0	10,307,359	0	37,001	10,944,360
	2012	84,092	1,100,000	37,999,991	0	0	29	39,184,112

(1) Salary and bonus columns include amounts earned in, or awarded for performance during, the specified year (even if paid out early in the following year).

(2) As required by SEC rules, the stock and option award columns present the aggregate grant date fair value of equity awards granted during the years shown as computed for accounting purposes in accordance with FASB ASC 718. As a result, the stock and option columns (as well as the total column) include awards that have not yet vested, awards that were granted but later forfeited (such as upon the executive's termination), and performance-based awards that failed to vest; therefore, these columns are not intended as presentations of pay actually realized by the executive. For information on the assumptions used in the grant date fair value computations, refer to Note 14—"Employee Benefits" in the Notes to Consolidated Financial Statements in our 2014 Form 10-K.

Week 13 Topic:

Yahoo's Executive Compensation Table

Target Table in Source Code View:

```

0002 <div style="width:97%; margin-top:1.5%; margin-left:1.5%; margin-right:-1.25%">
0003 <P STYLE="margin-top:0pt; margin-bottom:0pt; font-size:9pt; font-family:arial" ALIGN="right">EXECUTIVE COMPENSATION </P>
0004 <p STYLE="margin-top:0pt;margin-bottom:0pt ; font-size:8pt">&nbsp;</p></div>
0005 <p STYLE="margin-top:0pt;margin-bottom:0pt ; font-size:8pt">&nbsp;</p>
0006 </P> <P STYLE="margin-top:0pt; margin-bottom:0pt; font-size:16pt; font-family:arial"><FONT COLOR="#7300ff"><B><A NAME="toc868077_29"></A>COMPENSATION TABLES </B></FONT>
0007 </P> <P STYLE="font-size:6pt;margin-top:0pt;margin-bottom:0pt">&nbsp;</P>
0008 <P STYLE="line-height:1.0pt;margin-top:0pt;margin-bottom:2pt;border-bottom:1.00pt solid #000000">&nbsp;</P> <P STYLE="margin-top:12pt; margin-bottom:0pt; text-indent:6%;
0009 font-size:10pt; font-family:arial" ALIGN="justify">The tables on the following
0010 pages present compensation information regarding our Chief Executive Officer, Marissa A. Mayer; our Chief Financial Officer, Ken Goldman; our co-founder and Chief Yahoo,
0011 David Filo; and our General Counsel, Ronald S. Bell. As required by SEC rules,
0012 the tables also include our former Chief Operating Officer, Henrique de Castro, whose service ended during 2014. These five individuals are our &#147;Named Executive
0013 Officers.&#148; We did not have any other executive officers in 2014. </P>
0014 <P STYLE="margin-top:12pt; margin-bottom:0pt; text-indent:6%; font-size:10pt; font-family:arial" ALIGN="justify">As required by SEC rules, in these tables performance-
0015 based awards are treated as having been granted in the year in which their
0016 performance goals were established (and if an award has multiple performance periods, the portion relating to each period is treated as a separate grant). </P> <P
0017 STYLE="margin-top:18pt; margin-bottom:0pt; font-size:16pt; font-family:arial"><FONT
0018 COLOR="#7300ff"><B>Summary Compensation Table&#151;2012&#150;2014 </B></FONT></P> <P STYLE="font-size:6pt;margin-top:0pt;margin-bottom:0pt">&nbsp;</P>
0019 <P STYLE="line-height:1.0pt;margin-top:0pt;margin-bottom:2pt;border-bottom:1.00pt solid #000000">&nbsp;</P> <P STYLE="margin-top:12pt; margin-bottom:0pt; text-indent:6%;
0020 font-size:10pt; font-family:arial" ALIGN="justify"><I></I>The following table
0021 presents 2012&#150;2014 summary compensation information for our Named Executive Officers. As required by SEC rules, stock awards (RSUs) and option awards are shown as
0022 compensation for the year in which they were granted (even if they have
0023 multi-year vesting schedules), and are valued based on their grant date fair values for accounting purposes. Accordingly, the table includes stock and option awards
0024 granted in the years shown even if they were scheduled to vest in later years, and
0025 even if they were subsequently forfeited (such as upon the executive&#146;s termination). Therefore, the stock and option columns do <I>not</I> report whether the officer
0026 realized a financial benefit from the awards (such as by vesting in stock or
0027 exercising options). <I> </I></P> <P STYLE="font-size:12pt;margin-top:0pt;margin-bottom:0pt">&nbsp;</P>
0028 <TABLE CELLSPACING="0" CELLPADDING="0" WIDTH="100%" BORDER="0" STYLE="BORDER-COLLAPSE:COLLAPSE; font-family:arial; font-size:8pt" ALIGN="center">
0029
0030
0031 <TR>
0032 <TD WIDTH="36%"></TD>
0033 <TD VALIGN="bottom" WIDTH="1%"></TD>

```

Week 13 Topic:

Extract tables from HTML

Using readHTMLTable:

```
> install.packages("XML")
```

```
> library(XML)
```

```
> proxy.yahoo.2015.HTML.tables <-  
readHTMLTable("C:/Users/Desktop/yahoo_proxy2015.html")
```

```
> proxy.yahoo.2015.HTML.page[6003:6020] ← to view Summary Compensation  
Table section in HTML
```

Week 13 Topic:

Select the Executive Compensation Table

> `compensation.table <- proxy.yahoo.2015.HTML.tables[[306]]` ← to view
Summary Compensation Table in the array. Table # = 306 for 2015 Yahoo DEF 14A filing

Filter	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16
1																
2	Name and Principal Position	Â	Year	Â	Â	Salary\$(1)	Â	Â	Bonus\$(1)	Â	Â	StockAwards\$(2)(3)(4)	Â	Â	OptionAwards\$(2)(3)	Â
3	Marissa A. Mayer	Â	Â	2014	Â	Â	Â	1,000,000	Â	Â	Â	0	Â	Â	Â	11,752,355
4	Chief Executive Officer	Â	Â	2013	Â	Â	Â	1,000,000	Â	Â	Â	2,250	Â	Â	Â	8,312,316
5		Â	Â	2012	Â	Â	Â	454,862	Â	Â	Â	0	Â	Â	Â	35,000,002
6	Ken Goldman	Â	Â	2014	Â	Â	Â	600,000	Â	Â	Â	0	Â	Â	Â	2,813,080
7	Chief Financial Officer	Â	Â	2013	Â	Â	Â	600,000	Â	Â	Â	0	Â	Â	Â	2,597,612
8		Â	Â	2012	Â	Â	Â	116,667	Â	Â	Â	100,000	Â	Â	Â	7,262,357
9	David Filo	Â	Â	2014	Â	Â	Â	1	Â	Â	Â	0	Â	Â	Â	0
10	Co-Founder and Chief Yahoo	Â	Â	2013	Â	Â	Â	1	Â	Â	Â	0	Â	Â	Â	0
11		Â	Â	2012	Â	Â	Â	1	Â	Â	Â	0	Â	Â	Â	0
12	Ronald S. Bell	Â	Â	2014	Â	Â	Â	600,000	Â	Â	Â	0	Â	Â	Â	3,282,107

Week 13 Topic:

Sub-select relevant table data, rename

```
> compensation.table.data <- compensation.table[3:5, c("V1",  
"V4", "V8", "V12", "V16")]
```

	V1	V4	V8	V12	V16
3	Marissa A. Mayer	2014	1,000,000	0	11,752,355
4	Chief Executive Officer	2013	1,000,000	2,250	8,312,316
5		2012	454,862	0	35,000,002

```
> compensation.table.data <- rename(compensation.table.data, c("V1"="Name and  
Principal Position", "V4"="Year", "V8"="Salary", "V12"="Stock_Awards",  
"V16"="Option_Awards"))
```

	Name and Principal Position	Year	Salary	Stock_Awards	Option_Awards
3	Marissa A. Mayer	2014	1,000,000	0	11,752,355
4	Chief Executive Officer	2013	1,000,000	2,250	8,312,316
5		2012	454,862	0	35,000,002

Week 13 Topic:

Other methods

Using `getURL`:

```
> install.packages("RCurl")  
> library(RCurl)  
> proxy.yahoo.2015.HTML.page <-  
getURL("https://www.sec.gov/Archives/edgar/data/1011006/00011931251515  
6926/d868077ddef14a.htm")  
...
```

Using `readLines`:

```
> proxy.yahoo.2015.HTML.page <-  
readLines("https://www.sec.gov/Archives/edgar/data/1011006/000119312515  
156926/d868077ddef14a.htm")  
> length(proxy.yahoo.2015.HTML.page)  
[1] 10261
```


Week 13 Topic:

Installing plyr

Install:

```
> install.packages("plyr")
```

Open library:

```
> library(plyr)
```

Getting help:

```
> library(help=plyr)
```

<code>m_ply</code>	Call function with arguments in array or data frame, discarding results.
<code>maply</code>	Call function with arguments in array or data frame, returning an array.
<code>mapvalues</code>	Replace specified values with new values, in a vector or factor.
<code>match_df</code>	Extract matching rows of a data frame.
<code>mdply</code>	Call function with arguments in array or data frame, returning a data frame.
<code>mlply</code>	Call function with arguments in array or data frame, returning a list.
<code>mutate</code>	Mutate a data frame by adding new or replacing existing columns.
<code>name_rows</code>	Toggle row names between explicit and implicit.
<code>ozone</code>	Monthly ozone measurements over Central America.
<code>plyr</code>	plyr: the split-apply-combine paradigm for R.
<code>plyr-deprecated</code>	Deprecated Functions in Package plyr
<code>progress_text</code>	Text progress bar.
<code>progress_time</code>	Text progress bar with time.
<code>progress_tk</code>	Graphical progress bar, powered by Tk.
<code>progress_win</code>	Graphical progress bar, powered by Windows.
<code>r_ply</code>	Replicate expression and discard results.
<code>raply</code>	Replicate expression and return results in a array.
<code>rbind.fill</code>	Combine data.frames by row, filling in missing columns.
<code>rbind.fill.matrix</code>	Bind matrices by row, and fill missing columns with NA.
<code>rdply</code>	Replicate expression and return results in a data frame.
<code>rename</code>	Modify names by name, not position.
<code>revalue</code>	Replace specified values with new values, in a factor or character vector.

Week 13 Topic:

Installing RCurl

Install:

```
>install.package("RCurl")
```

Open library:

```
>library(RCurl)
```

Getting help:

```
>library(help=RCurl)
```

Index:

AUTH_ANY	Constants for identifying Authentication Schemes
CFILE	Create a C-level handle for a file
CURLHandle-class	Class "CURLHandle" for synchronous HTTP requests
CurlFeatureBits	Constants for libcurl
HTTP_VERSION_1_0	Symbolic constants for specifying HTTP and SSL versions in libcurl
MultiCURLHandle-class	Class "MultiCURLHandle" for asynchronous, concurrent HTTP requests
base64	Encode/Decode base64 content
basicHeaderGatherer	Functions for processing the response header of a libcurl request
basicTextGatherer	Cumulate text across callbacks (from an HTTP response)
binaryBuffer	Create internal C-level data structure for collecting binary data
chunkToLineReader	Utility that collects data from the HTTP reply into lines and calls user-provided function.
clone	Clone/duplicate an object
coerce,numeric,NetrcEnum-method	Internal functions
complete	Complete an asynchronous HTTP request
curlError	Raise a warning or error about a CURL problem
curlEscape	Handle characters in URL that need to be escaped
curlGlobalInit	Start and stop the Curl library
curlOptions	Constructor and accessors for CURLOPToptions objects
curlPerform	Perform the HTTP query
curlSetOpt	Set values for the CURL options
curlVersion	Information describing the Curl library
dynCurlReader	Dynamically determine content-type of body from HTTP header and set body reader

Week 13 Topic: Installing RCurl (cont.)

Install:

`>install.package("RCurl")`

Open library:

`>library(RCurl)`

Getting help:

`>library(help=RCurl)`

<code>fileUpload</code>	Specify information about a file to upload in an HTTP request
<code>findHTTPHeaderEncoding</code>	Find the encoding of the HTTP response from the HTTP header
<code>ftpUpload</code>	Upload content via FTP
<code>getBinaryURL</code>	Download binary content
<code>getBitIndicators</code>	Operate on bit fields
<code>getCurlErrorClassNames</code>	Retrieve names of all curl error classes
<code>getCurlHandle</code>	Create libcurl handles
<code>getCurlInfo</code>	Access information about a CURL request
<code>getFormParams</code>	Extract parameters from a form query string
<code>getURIAsynchronous</code>	Download multiple URIs concurrently, with inter-leaved downloads
<code>getURL</code>	Download a URI
<code>guessMIMEType</code>	Infer the MIME type from a file name
<code>httpPUT</code>	Simple high-level functions for HTTP PUT and DELETE
<code>merge.list</code>	Method for merging two lists by name
<code>mimeTypeExtensions</code>	Mapping from extension to MIME type
<code>postForm</code>	Submit an HTML form
<code>reset</code>	Generic function for resetting an object
<code>scp</code>	Retrieve contents of a file from a remote host via SCP (Secure Copy)
<code>url.exists</code>	Check if URL exists
<code> ,BitwiseValue,BitwiseValue-method</code>	Classes and coercion methods for enumerations in libcurl

Week 13 Topic:

Installing stringr

Install:

```
>install.packages("stringr")
```

Open library:

```
>library(stringr)
```

Getting help:

```
>library(help=stringr)
```

Index:

case	Convert case of a string.
invert_match	Switch location of matches to location of non-matches.
modifiers	Control matching behaviour with modifier functions.
str_c	Join multiple strings into a single string.
str_conv	Specify the encoding of a string.
str_count	Count the number of matches in a string.
str_detect	Detect the presence or absence of a pattern in a string.
str_dup	Duplicate and concatenate strings within a character vector.
str_extract	Extract matching patterns from a string.
str_length	The length of a string.
str_locate	Locate the position of patterns in a string.
str_match	Extract matched groups from a string.
str_order	Order or sort a character vector.
str_pad	Pad a string.
str_replace	Replace matched patterns in a string.
str_replace_na	Turn NA into "NA"
str_split	Split up a string into pieces.
str_sub	Extract and replace substrings from a character vector.
str_subset	Keep strings matching a pattern.
str_trim	Trim whitespace from start and end of string.
str_wrap	Wrap strings into nicely formatted paragraphs.
stringr	Fast and friendly string manipulation.
word	Extract words from a sentence.

Further information is available in the following vignettes in directory
'C:/Users/sshin/Documents/R/R-3.2.2/library/stringr/doc':

Week 13 Topic:

Installing XML

Install:

>install.package("XML")

Open library:

>library(XML)

Getting help:

>library(help=XML)

Index:

Doctype	Constructor for DTD reference
Doctype-class	Class to describe a reference to an XML DTD
ExternalReference-class	Classes for working with XML Schema
SAXState-class	A virtual base class defining methods for SAX parsing
XMLAttributes-class	Class '"XMLAttributes"'
XMLCodeFile-class	Simple classes for identifying an XML document containing R code
XMLInternalDocument-class	Class to represent reference to C-level data structure for an XML document
XMLNode-class	Classes to describe an XML node object.
[.XMLNode	Convenience accessors for the children of XMLNode objects.
[<-.XMLNode	Assign sub-nodes to an XML node
addChildren	Add child nodes to an XML node
addNode	Add a node to a tree
append.xmlNode	Add children to an XML node
asXMLNode	Converts non-XML node objects to XMLTextNode objects
asXMLTreeNode	Convert a regular XML node to one for use in a "flat" tree
catalogLoad	Manipulate XML catalog contents
catalogResolve	Look up an element via the XML catalog mechanism
coerce,XMLHashTreeNode,XMLHashTree-method	Transform between XML representations
compareXMLDocs	Indicate differences between two XML documents
docName	Accessors for name of XML document
dtdElement	Gets the definition of an element or entity from a DTD.

Week 13 Topic:

Installing XML (cont.)

Install:

>install.package("XML")

Open library:

>library(XML)

Getting help:

>library(help=XML)

<code>getXMLErrors</code>	Get XML/HTML document parse errors
<code>isXMLString</code>	Facilities for working with XML strings
<code>length.XMLNode</code>	Determine the number of children in an XMLNode object.
<code>libxmlVersion</code>	Query the version and available features of the libxml library.
<code>makeClassTemplate</code>	Create S4 class definition based on XML node(s)
<code>names.XMLNode</code>	Get the names of an XML nodes children.
<code>newXMLDoc</code>	Create internal XML node or document object
<code>newXMLNamespace</code>	Add a namespace definition to an XML node
<code>parsedDTD</code>	Read a Document Type Definition (DTD)
<code>parseURI</code>	Parse a URI string into its elements
<code>parseXMLAndAdd</code>	Parse XML content and add it to a node
<code>print.XMLAttributeDef</code>	Methods for displaying XML objects
<code>processXInclude</code>	Perform the XInclude substitutions
<code>readHTMLList</code>	Read data in an HTML list or all lists in a document
<code>readHTMLTable</code>	Read data from one or more HTML tables
<code>readKeyValueDB</code>	Read an XML property-list style document
<code>readSolrDoc</code>	Read the data from a Solr document
<code>removeXMLNamespaces</code>	Remove namespace definitions from a XML node or document
<code>saveXML</code>	Output internal XML Tree
<code>setXMLNamespace</code>	Set the name space on a node
<code>startElement.SAX</code>	Generic Methods for SAX callbacks
<code>supportsExpat</code>	Determines which native XML parsers are being used.
<code>toHTML</code>	Create an HTML representation of the given R object, using internal C-level nodes
<code>toString.XMLNode</code>	Creates string representation of XML node
<code>xmlApply</code>	Applies a function to each of the children of an XMLNode
<code>xmlAttributeType</code>	The type of an XML attribute for element from the DTD

Week 13 Topic:

Installing tm

Install:

`>install.packages("tm")`

Open library:

`>library(tm)`

Getting help:

`>library(help=tm)`

Index:

Corpus	Corpora
DataframeSource	Data Frame Source
DirSource	Directory Source
Docs	Access Document IDs and Terms
MC_tokenizer	Tokenizers
PCorpus	Permanent Corpora
PlainTextDocument	Plain Text Documents
Reader	Readers
Source	Sources
TermDocumentMatrix	Term-Document Matrix
TextDocument	Text Documents
URISource	Uniform Resource Identifier Source
VCorpus	Volatile Corpora
VectorSource	Vector Source
WeightFunction	Weighting Function
XMLSource	XML Source
XMLTextDocument	XML Text Documents
ZipSource	ZIP File Source
Zipf_plot	Explore Corpus Term Frequency Characteristics
acq	50 Exemplary News Articles from the Reuters-21578 Data Set of Topic acq
c.VCorpus	Combine Corpora, Documents, Term-Document Matrices, and Term Frequency Vectors
content_transformer	Content Transformers
crude	20 Exemplary News Articles from the Reuters-21578 Data Set of Topic crude
findAssocs	Find Associations in a Term-Document Matrix
findFreqTerms	Find Frequent Terms
getTokenizers	Tokenizers
getTransformations	Transformations
inspect	Inspect Objects
meta	Metadata Management
plot.TermDocumentMatrix	

Week 13 Topic:

Installing tm (cont.)

Install:

```
>install.package("tm")
```

Open library:

```
>library(tm)
```

Getting help:

```
>library(help=tm)
```

plot.TermDocumentMatrix

	Visualize a Term-Document Matrix
readDOC	Read In a MS Word Document
readPDF	Read In a PDF Document
readPlain	Read In a Text Document
readRCV1	Read In a Reuters Corpus Volume 1 Document
readReut21578XML	Read In a Reuters-21578 XML Document
readTabular	Read In a Text Document
readTagged	Read In a POS-Tagged Word Text Document
readXML	Read In an XML Document
read_dtm_Blei_et_al	Read Document-Term Matrices
removeNumbers	Remove Numbers from a Text Document
removePunctuation	Remove Punctuation Marks from a Text Document
removeSparseTerms	Remove Sparse Terms from a Term-Document Matrix
removeWords	Remove Words from a Text Document
stemCompletion	Complete Stems
stemDocument	Stem Words
stopwords	Stopwords
stripWhitespace	Strip Whitespace from a Text Document
termFreq	Term Frequency Vector
tm_filter	Filter and Index Functions on Corpora
tm_map	Transformations on Corpora
tm_reduce	Combine Transformations
tm_term_score	Compute Score for Matching Terms
weightBin	Weight Binary
weightSMART	SMART Weightings
weightTf	Weight by Term Frequency
weightTfIdf	Weight by Term Frequency - Inverse Document Frequency
writeCorpus	Write a Corpus to Disk

Further information is available in the following vignettes in directory 'C:/Users/sshin/Documents/R/R-3.2.2/library/tm/doc':

extensions: Extensions (source, pdf)

tm: Introduction to the tm Package (source, pdf)