

Introduction to Data Science-Topic 8

- Instructor: Professor Henry Horng-Shing Lu,
Institute of Statistics, National Chiao Tung University, Taiwan
Email: hslu@stat.nctu.edu.tw
- WWW: <http://www.stat.nctu.edu.tw/misg/hslu/course/DataScience.htm>
- Reference:
M. A. Pathak, Beginning Data Science with R, 2014, Springer-Verlag.
- Evaluation: Homework: 50%, Term Project: 50%
- Office hours: By appointment

Course Outline

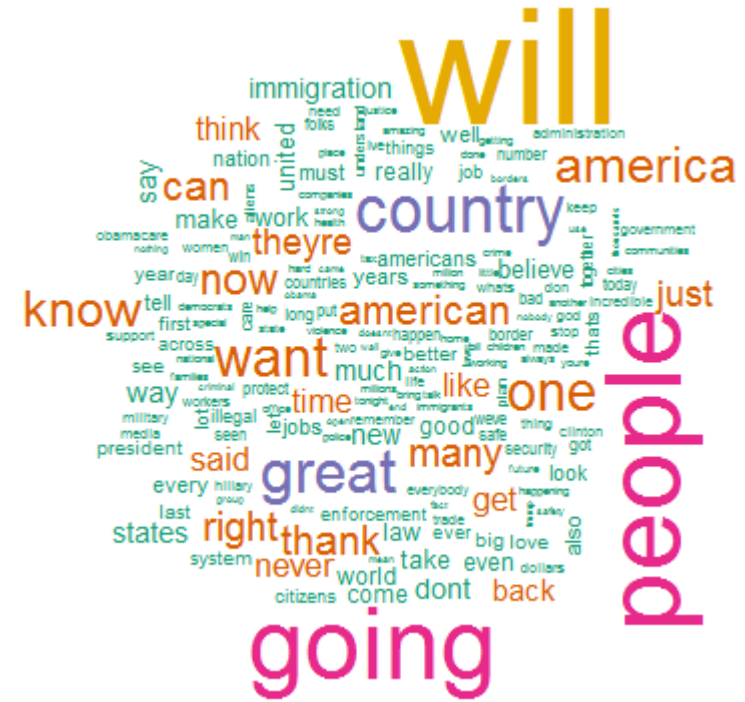

- Introduction of Data Science
- Introduction of R
- More on R
- Process Real Data by R
- Data Visualization
- Exploratory Data Analysis
- Regression
- Classification
- **Text Mining**
- Clustering

Text Mining with R

References:

Ch. 8, M. A. Pathak, Beginning Data Science with R, 2014, Springer-Verlag.





4

Data and packages prepare

- Code:
 - `Needed <- c("tm", "SnowballCC", "RColorBrewer", "ggplot2", "wordcloud", "biclust", "cluster", "igraph", "fpc")`
 - `install.packages(Needed, dependencies = TRUE)`
 - `install.packages("Rcampdf", repos = "http://datacube.wu.ac.at/", type = "source")`
- Data:
 - Trump speech: <https://drive.google.com/drive/folders/0B914dXn0AXvlaWhmVXdPclZrTjA>
- Please put the data you want analysis into one folder and set the work path EX:
 - `Cname<-file.path("C:/Users/USER/Desktop/text")`

Data summary

- **Code:**

- library(tm)
- docs <- VCorpus(DirSource(cname))
- summary(docs)

- **File list :**

	Length	Class	Mode
Trump Black History Month Speech.txt	2	PlainTextDocument	list
Trump CIA Speech.txt	2	PlainTextDocument	list
Trump Congressional Address.txt	2	PlainTextDocument	list
Trump CPAC Speech.txt	2	PlainTextDocument	list
Trump Florida Rally 2-18-17.txt	2	PlainTextDocument	list
Trump Immigration Speech 8-31-16.txt	2	PlainTextDocument	list
Trump Inauguration Speech.txt	2	PlainTextDocument	list
Trump National Prayer Breakfast.txt	2	PlainTextDocument	list
Trump Nomination Speech.txt	2	PlainTextDocument	list
Trump Police Chiefs Speech.txt	2	PlainTextDocument	list
Trump Response to Healthcare Bill Failure.txt	2	PlainTextDocument	list

Data summary

- **Code:**

- inspect(docs[1])
- writeLines(as.character(docs[1]))

- **Firs file word:**

```
list(list(content = c("Well, the election, it came out really well. Next time we<e2>\u0080 triple the number or quadruple it. We want to get it over 51, right? At least 51.", "", "Well this is Black History Month, so this is our little breakfast, our little get-together. Hi Lynn, how are you? Just a few notes. During this month, we honor the tremendous history of African-Americans throughout our country. Throughout the world, if you really think about it, right? And their story is one of unimaginable sacrifice, hard work, and faith in America. I<e2>\u0080e gotten a real glimpse<e2>\u0080uring the campaign, I<e2>\u0080 go around with Ben to a lot of different places I wasn<e2>\u0080 so familiar with. They<e2>\u0080e incredible people. And I want to thank Ben Carson, who<e2>\u0080 gonna be heading up HUD. That<e2>\u0080 a big job. That<e2>\u0080 a job that<e2>\u0080 not only housing, but it<e2>\u0080 mind and spirit. Right, Ben? And you understand, nobody<e2>\u0080 gonna be better than Ben.", "", "Last month, we celebrated the life of Reverend Martin Luther King, Jr., whose incredible example is unique in American history. You read all about Dr. Martin Luther King a week ago when somebody said I took the statue out of my office. It turned out that that was fake news. Fake news. The statue is cherished, it
```

Data preprocessing

- **Remove Punctuation**

- docs <- tm_map(docs,removePunctuation)
- writeLines(as.character(docs[1])) # Check to see if it worked.

- **Output:**

```
list(list(content = c("well the election it came out really well Next time wee2\u00801 triple the number  
or quadruple it we want to get it over 51 right At least 51", "", "well this is Black History Month so thi  
s is our little breakfast our little gettogether Hi Lynn how are you Just a few notes During this month we  
honor the tremendous history of AfricanAmericans throughout our country Throughout the world if you reall  
y think about it right And their story is one of unimaginable sacrifice hard work and faith in America Ie2  
\u0080e gotten a real glimpse2\u0080uring the campaign Ie2\u0080 go around with Ben to a lot of differ  
ent places I wasne2\u0080 so familiar with Theye2\u0080e incredible people And I want to thank Ben Carso  
n whoe2\u0080 gonna be heading up HUD Thate2\u0080 a big job Thate2\u0080 a job thate2\u0080 not only  
housing but ite2\u0080 mind and spirit Right Ben And you understand nobodye2\u0080 gonna be better than  
Ben",
```


Data preprocessing

- **Remove ascii character that did not translate, so it had to be removed**

- `for (j in seq(docs)){`
- `docs[[j]] <- gsub("/", " ", docs[[j]])`
- `docs[[j]] <- gsub("@", " ", docs[[j]])`
- `docs[[j]] <- gsub("\\\\|", " ", docs[[j]])`
- `docs[[j]] <- gsub('e2\\u0080e ', " ", docs[[j]])`
- `docs[[j]] <- gsub('e2\\u0080', " ", docs[[j]])`
- `docs[[j]] <- gsub("e2\\u2028e", " ", docs[[j]])`
- `docs[[j]] <- gsub("e2\\u2028", " ", docs[[j]])`
- `#docs[[j]] <- gsub("\\u0080 ", " ", docs[[j]])# This is an ascii character that did not translate, so it had to be removed.`
- `}`
- `writeLines(as.character(docs[1])) # You can check a document (in this case`
- `# the first) to see if it worked.`

- **Output**

```
list(c("Well the election it came out really well Next time we l triple the number or quadruple it We wan  
t to get it over 51 right At least 51", "", "Well this is Black History Month so this is our little breakf  
ast our little gettogether Hi Lynn how are you Just a few notes During this month we honor the tremendous  
history of AfricanAmericans throughout our country Throughout the world if you really think about it right  
And their story is one of unimaginable sacrifice hard work and faith in America I e gotten a real glimps  
e uring the campaign I go around with Ben to a lot of different places I wasn so familiar with They  
e incredible people And I want to thank Ben Carson who gonna be heading up HUD That a big job That  
a job that not only housing but it mind and spirit Right Ben And you understand nobody gonna be bett  
er than Ben",
```

Data preprocessing

- **Remove number and Converting to lowercase**

- docs <- tm_map(docs, removeNumbers)
- docs <- tm_map(docs, tolower)
- docs <- tm_map(docs, PlainTextDocument)
- DocsCopy <- docs
- writeLines(as.character(docs[1])) # Check to see if it worked.

- **Output**

```
list(list(content = c("well the election it came out really well next time we l triple the number or quad  
ruple it we want to get it over right at least ", "", "well this is black history month so this is our li  
ttle breakfast our little gettogether hi lynn how are you just a few notes during this month we honor the  
tremendous history of africanamericans throughout our country throughout the world if you really think abo  
ut it right and their story is one of unimaginable sacrifice hard work and faith in america i e gotten a  
real glimpse uring the campaign i go around with ben to a lot of different places i wasn so familiar  
with they e incredible people and i want to thank ben carson who gonna be heading up hud that a big j  
ob that a job that not only housing but it mind and spirit right ben and you understand nobody gon  
na be better than ben",  
"", "last month we celebrated the life of reverend martin luther king jr whose incredible example is uniqu  
e in american history you read all about dr martin luther king a week ago when somebody said i took the st  
atue out of my office it turned out that that was fake news fake news the statue is cherished it one of
```

Data preprocessing

- **Remove stop words**

- docs <- tm_map(docs, removeWords, stopwords("english"))
- docs <- tm_map(docs, PlainTextDocument)
- writeLines(as.character(docs[1])) # Check to see if it worked.
- docs <- tm_map(docs, removeWords, c("syllogism", "tautology"))

- **Output**

```
list(list(content = c("well election came really well next time 1 triple number quadruple want g  
et right least ", "", "well black history month little breakfast little gettogether hi lynn  
just notes month honor tremendous history africanamericans throughout country throughout world  
really think right story one unimaginable sacrifice hard work faith america e gotten real glimps  
se uring campaign go around ben lot different places wasn familiar e incredible people w  
ant thank ben carson gonna heading hud big job job housing mind spirit right ben  
understand nobody gonna better ben",  
"", "last month celebrated life reverend martin luther king jr whose incredible example unique americ  
an history read dr martin luther king week ago somebody said took statue office turned fake  
news fake news statue cherished one favorite things nd good ones lincoln jefferson d  
r martin luther king said statue bust martin luther king taken office never even touched thi  
nk disgrace way press unfortunate", "", " proud now museum national mall people can le  
arn reverend king many things frederick douglass example somebody done amazing job recognized  
noticed harriet tubman rosa parks millions black americans made america today big impact",
```

Data preprocessing

- **Combining words that should stay together**

- for (j in seq(docs))
- {
- docs[[j]] <- gsub("fake news", "fake_news", docs[[j]])
- docs[[j]] <- gsub("inner city", "inner-city", docs[[j]])
- docs[[j]] <- gsub("politically correct", "politically_correct", docs[[j]])
- }
- docs <- tm_map(docs, PlainTextDocument)
- writeLines(as.character(docs[1])) # Check to see if it worked.

- **Output**

```
just notes month honor tremendous history africanamericans throughout country throughout world
really think right story one unimaginable sacrifice hard work faith america e gotten real glimps
se uring campaign go around ben lot different places wasn familiar e incredible people w
ant thank ben carson gonna heading hud big job job housing mind spirit right ben
understand nobody gonna better ben",
"", "last month celebrated life reverend martin luther king jr whose incredible example unique americ
an history read dr martin luther king week ago somebody said took statue office turned fake
_news fake_news statue cherished one favorite things nd good ones lincoln jefferson d
r martin luther king said statue bust martin luther king taken office never even touched thi
nk disgrace way press unfortunate", "", " proud now museum national mall people can le
arn reverend king many things frederick douglass example somebody done amazing job recognized
noticed harriet tubman rosa parks millions black americans made america today big impact",
"", " proud honor heritage will honoring folks table almost cases great friends supporte
rs darrell met darrell defending television people side argument didn chance right pa
ris done amazing job hostile cnn community l seven people paris l take paris seven d
on watch cnn don get see much used don like watching fake_news fox treated nice whereve
```

Data preprocessing

- **Combining words that should stay together**

- for (j in seq(docs))
- {
- docs[[j]] <- gsub("fake news", "fake_news", docs[[j]])
- docs[[j]] <- gsub("inner city", "inner-city", docs[[j]])
- docs[[j]] <- gsub("politically correct", "politically_correct", docs[[j]])
- }
- docs <- tm_map(docs, PlainTextDocument)
- writeLines(as.character(docs[1])) # Check to see if it worked.

- **Output**

```
list(list(content = c("well election came really well next time l triple number quadruple want g
et right least ", "", "well black history month little breakfast little gettogether hi lynn
just notes month honor tremendous history africanamericans throughout country throughout world
really think right story one unimaginable sacrifice hard work faith america e gotten real glimp
se uring campaign go around ben lot different places wasn familiar e incredible people w
ant thank ben carson gonna heading hud big job job housing mind spirit right ben
understand nobody gonna better ben",
"", "last month celebrated life reverend martin luther king jr whose incredible example unique americ
an history read dr martin luther king week ago somebody said took statue office turned fake
_news fake_news statue cherished one favorite things nd good ones lincoln jefferson d
r martin luther king said statue bust martin luther king taken office never even touched thi
nk disgrace way press unfortunate", "", " proud now museum national mall people can le
arn reverend king many things frederick douglass example somebody done amazing job recognized
noticed harriet tubman rosa parks millions black americans made america today big impact",
"", " proud honor heritage will honoring folks table almost cases great friends supporte
rs darrell met darrell defending television people side argument didn chance right pa
```

Data preprocessing

- **Removing common word endings (e.g., “ing”, “es”, “s”)**

- docs_st <- tm_map(docs, stemDocument)
- docs_st <- tm_map(docs_st, PlainTextDocument)
- writeLines(as.character(docs_st[1])) # Check to see if it worked.

- **Output**

```
", "", "well black histori month littl breakfast littl gettogeth hi lynn just note month honor tremend his  
tori africanamerican throughout countri throughout world realli think right stori one unimagin sacrific ha  
rd work faith america e gotten real glimps ure campaign go around ben lot differ place wasn familiar  
e incred peopl want thank ben carson gonna head hud big job job hous mind spirit right ben und  
erstand nobodi gonna better ben",  
"", "last month celebr life reverend martin luther king jr whose incred exampl uniqu american histori read  
dr martin luther king week ago somebodi said took statu offic turn fake_new fake_new statu cherish one  
favorit thing nd good one lincoln jefferson dr martin luther king said statu bust martin luther king take  
n offic never even touch think disgrac way press unfortun", "", "proud now museum nation mall peopl can  
learn reverend king mani thing frederick douglass exampl somebodi done amaz job recogn notic harriet tub  
man rosa park million black american made america today big impact",  
"", " proud honor heritag will honor folk tabl almost case great friend support darrel met darrel defen  
d televis peopl side argument didn chanc right pari done amaz job hostil cnn communiti l seven peopl  
pari l take pari seven don watch cnn don get see much use don like watch fake_new fox treat nice wh  
erev fox thank", "", "e gonna need better school need soon need job need better wage lot better wage e g
```


Data preprocessing

- **Add common endings to improve interpretability**

- `docs_stc <- tm_map(docs_st, stemCompletion, dictionary = DocsCopy, lazy=TRUE)`
- `docs_stc <- tm_map(docs_stc, PlainTextDocument)`
- `writeLines(as.character(docs_stc[1]))` # Check to see if it worked.

- **Output**

```
list(list(content = c("well elect came realli well next time 1 tripl number quadrupl want get right least  
", "", "well black histori month littl breakfast littl gettogeth hi lynn just note month honor tremend his  
tori africanamerican throughout countri throughout world realli think right stori one unimagin sacrific ha  
rd work faith america e gotten real glimps ure campaign go around ben lot differ place wasn familiar  
e incred peopl want thank ben carson gonna head hud big job job hous mind spirit right ben und  
erstand nobodi gonna better ben",  
"", "last month celebr life reverend martin luther king jr whose incred exampl uniqu american histori read  
dr martin luther king week ago somebodi said took statu offic turn fake_new fake_new statu cherish one  
favorit thing nd good one lincoln jefferson dr martin luther king said statu bust martin luther king take  
n offic never even touch think disgrac way press unfortun", "", "proud now museum nation mall peopl can  
learn reverend king mani thing frederick douglass exampl somebodi done amaz job recogn notic harriet tub  
man rosa park million black american made america today big impact",  
"", " proud honor heritag will honor folk tabl almost case great friend support darrel met darrel defen
```

Data preprocessing

- **Stripping unnecessary whitespace**

- `docs <- tm_map(docs, stripWhitespace)`
- `writeLines(as.character(docs[1]))` # Check to see if it worked.

- **Output**

```
list(list(content = c("well election came really well next time l triple number quadruple want get right  
least ", "", "well black history month little breakfast little gettogether hi lynn just notes month honor  
tremendous history africanamericans throughout country throughout world really think right story one unima  
ginable sacrifice hard work faith america e gotten real glimpse uring campaign go around ben lot diffe  
rent places wasn familiar e incredible people want thank ben carson gonna heading hud big job job  
housing mind spirit right ben understand nobody gonna better ben",  
"", "last month celebrated life reverend martin luther king jr whose incredible example unique american hi  
story read dr martin luther king week ago somebody said took statue office turned fake_news fake_news stat  
ue cherished one favorite things nd good ones lincoln jefferson dr martin luther king said statue bust  
martin luther king taken office never even touched think disgrace way press unfortunate", "", "proud no  
w museum national mall people can learn reverend king many things frederick douglass example somebody do  
ne amazing job recognized noticed harriet tubman rosa parks millions black americans made america today bi  
a impact"
```


Data preprocessing

- **Finish preprocessing**

- docs <- tm_map(docs, PlainTextDocument)
- writeLines(as.character(docs[1])) # Check to see if it worked.

- **Output**

```
list(list(content = c("well election came really well next time l triple number quadruple want get right  
least ", "", "well black history month little breakfast little gettogether hi lynn just notes month honor  
tremendous history africanamericans throughout country throughout world really think right story one unima  
ginable sacrifice hard work faith america e gotten real glimpse uring campaign go around ben lot diffe  
rent places wasn familiar e incredible people want thank ben carson gonna heading hud big job job  
housing mind spirit right ben understand nobody gonna better ben",  
"", "last month celebrated life reverend martin luther king jr whose incredible example unique american hi  
story read dr martin luther king week ago somebody said took statue office turned fake_news fake_news stat  
ue cherished one favorite things nd good ones lincoln jefferson dr martin luther king said statue bust  
martin luther king taken office never even touched think disgrace way press unfortunate", "", "proud no  
w museum national mall people can learn reverend king many things frederick douglass example somebody do
```

Stage the Data

- **Create a document term matrix**

- dtm <- DocumentTermMatrix(docs)
- dtm
- tdm <- TermDocumentMatrix(docs)
- Tdm

- **Output**

```
> dtm
<<DocumentTermMatrix (documents: 11, terms: 3624)>>
Non-/sparse entries: 8317/31547
Sparsity           : 79%
Maximal term length: 19
Weighting          : term frequency (tf)
> tdm <- TermDocumentMatrix(docs)
> tdm
<<TermDocumentMatrix (terms: 3624, documents: 11)>>
Non-/sparse entries: 8317/31547
Sparsity           : 79%
Maximal term length: 19
Weighting          : term frequency (tf)
```

Explore data

- **Terms frequency**

- `freq <- colSums(as.matrix(dtm))`
- `length(freq)`

- **Output**

```
> length(freq)
[1] 3624
```

Explore data

- **Start by removing sparse terms**

- `dtms <- removeSparseTerms(dtm, 0.2)` # This makes a matrix that is 20% empty space, maximum.
- `Dtms`

- **Output**

```
<<DocumentTermMatrix (documents: 11, terms: 87)>>  
Non-/sparse entries: 848/109  
Sparsity           : 11%  
Maximal term length: 11  
Weighting          : term frequency (tf)
```

Definition

- Term frequency (TF)
 - how many times a term occurs in a document
- Word Frequency (WF)
 - how many times a word occurs in all document
- Frequency-inverse document frequency (TF-IDF)
 - all words in the corpus are not equally important
 - IDF: $\text{Log}(N/d)$,
 - A corpus has N documents, and a term appears in d of them
 - A terms most of the documents will have, the IDF value will be low. Otherwise, will be large

Word Frequency

- **Most and least frequently occurring words**

- `head(table(freq), 20)` # The ", 20" indicates that we only want the first 20 frequencies. Feel free to change that number.
- `tail(table(freq), 20)` # The ", 20" indicates that we only want the last 20 frequencies. Feel free to change that number, as needed.
- `freq <- colSums(as.matrix(dtms))`
- `freq`

- **Output**

```
> freq
also      always    america  american  another    back      bad      believe
  54         24      128      107         22       75      35         60
big       came      can       care      come      country   day      different
  45        20      107        37        55      174      36         16
done enforcement even      ever      every      get      getting   give
  24         43        55        42        49      79      24         25
going      good     great     group    happen    job      just      know
 265         58      163        20        36      39      88      127
last      law      let      life     like     little    long      look
  44         59       45        27        79      24      36         52
lot      love     made     many    much     must      nation    need
  44         45       32       101       68      53      50      32
never     new      now      office  one      people   president  put
  83        69     111       24     139     279      48      35
```

Word Frequency

- **Most and least frequently occurring words (TOP 14)**

- `freq <- sort(colSums(as.matrix(dtm)), decreasing=TRUE)`
- `head(freq, 14)`
- `findFreqTerms(dtm, lowfreq=50)` # Change "50" to whatever is most appropriate for your text data.
- `wf <- data.frame(word=names(freq), freq=freq)`
- `head(wf)`

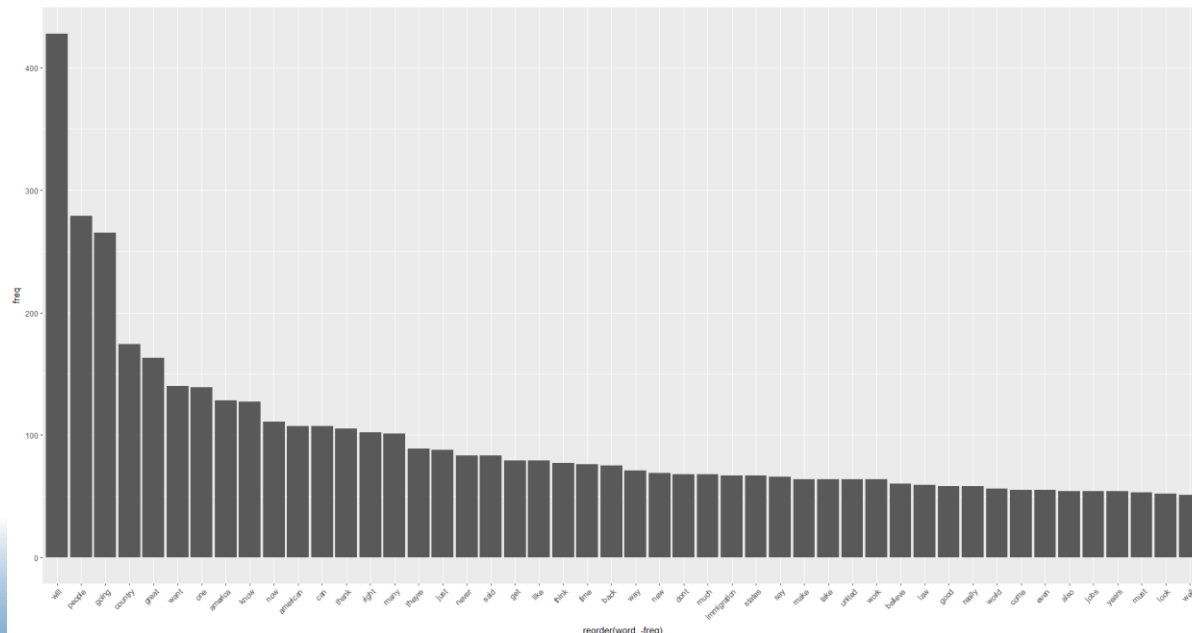
- **Output**

```
will    people    going    country    great    want    one    america    know    now    american
428      279      265      174      163      140      139      128      127      111      107
can      thank    right
107      105      102

> head(wf)
  word freq
will  428
people 279
going 265
country 174
great 163
want 140
```

Plot Word Frequencies

- **Plot words that appear at least 50 times**
 - `library(ggplot2)`
 - `p <- ggplot(subset(wf, freq>50), aes(x = reorder(word, -freq), y = freq)) +`
 - `geom_bar(stat = "identity") +`
 - `theme(axis.text.x=element_text(angle=45, hjust=1))`
 - `p`
- **Output**



Relationships Between Terms

- **Term Correlations**

- `findAssocs(dtm, c("country", "american"), corlimit=0.85)` # specifying a correlation limit of 0.85
- `findAssocs(dtms, "think", corlimit=0.70)` # specifying a correlation limit of 0.95

- **Output**

```
$country
nothing    cities countries      jobs      come    biggest    donors
  0.95      0.94      0.94      0.92      0.91      0.90      0.90
second     begin    border    plan    crimes    globe      meant
  0.90      0.88      0.88      0.88      0.87      0.87      0.87
thousands means workers    also    despite    take
  0.87      0.86      0.86      0.85      0.85      0.85

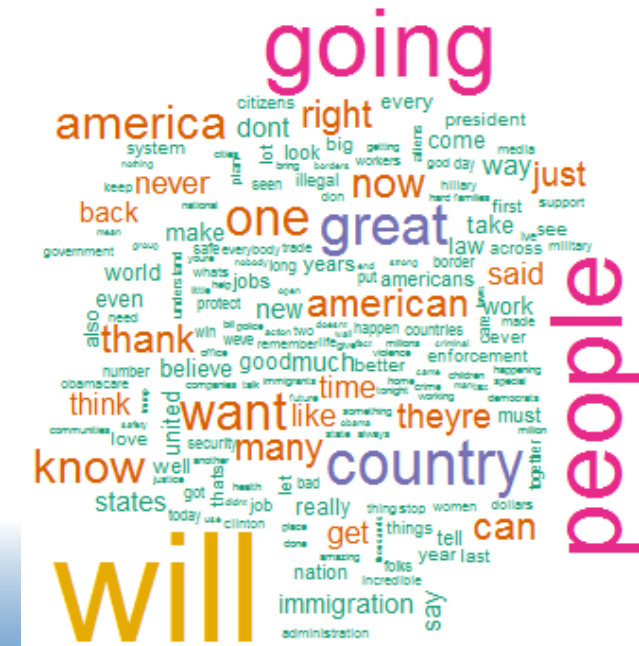
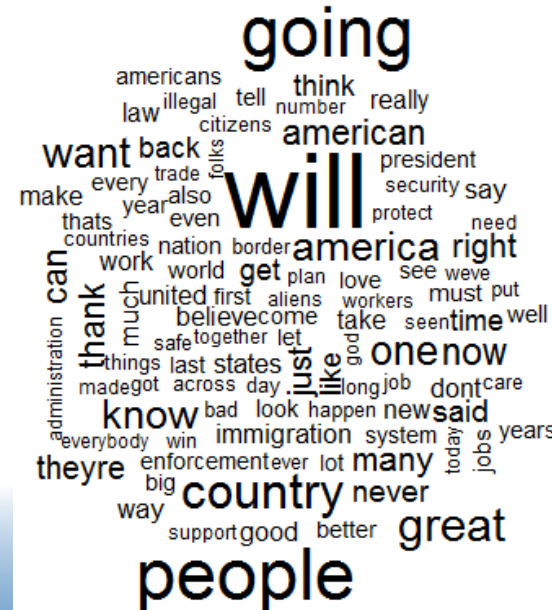
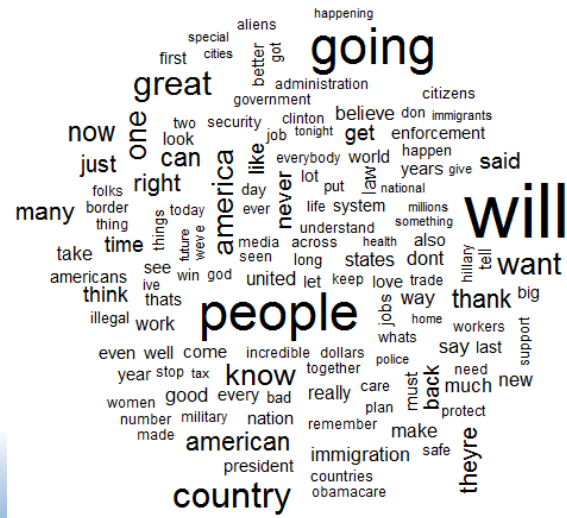
$think
well really  care    lot happen    see    good
  0.88    0.87    0.86    0.76    0.72    0.72    0.71
```

Relationships Between Terms

- **Term Correlations**

- `set.seed(142)`
- `wordcloud(names(freq), freq, min.freq=25)`
- `set.seed(142)`
- `wordcloud(names(freq), freq, max.words=100)`
- `set.seed(142)`
- `wordcloud(names(freq), freq, min.freq=20, scale=c(5, .1), colors=brewer.pal(6, "Dark2"))`

- **Output**



Clustering by Term Similarity

- **Remove uninteresting or infrequent words**

- `dtmss <- removeSparseTerms(dtm, 0.15)` # This makes a matrix that is only 15% empty space, maximum.
- `Dtmss`

- **Output**

```
<<DocumentTermMatrix (documents: 11, terms: 43)>>  
Non-/sparse entries: 452/21  
Sparsity           : 4%  
Maximal term length: 9  
Weighting          : term frequency (tf)
```

Clustering by Term Similarity

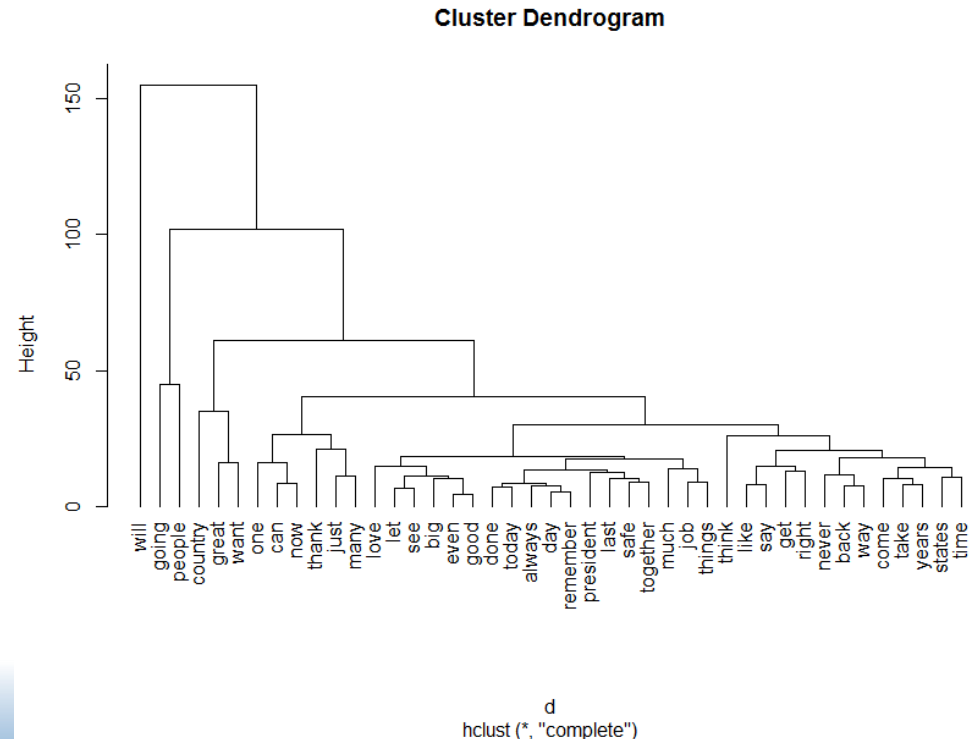
• Hierarchical Clustering

- `library(cluster)`
- `d <- dist(t(dtmss), method="euclidian")`
- `fit <- hclust(d=d, method="complete")` # for a different look try substituting: `method="ward.D"`
- `fit`
- `plot(fit, hang=-1)`

• Output

```
call:
hclust(d = d, method = "complete")

Cluster method      : complete
Distance            : euclidean
Number of objects: 43
```

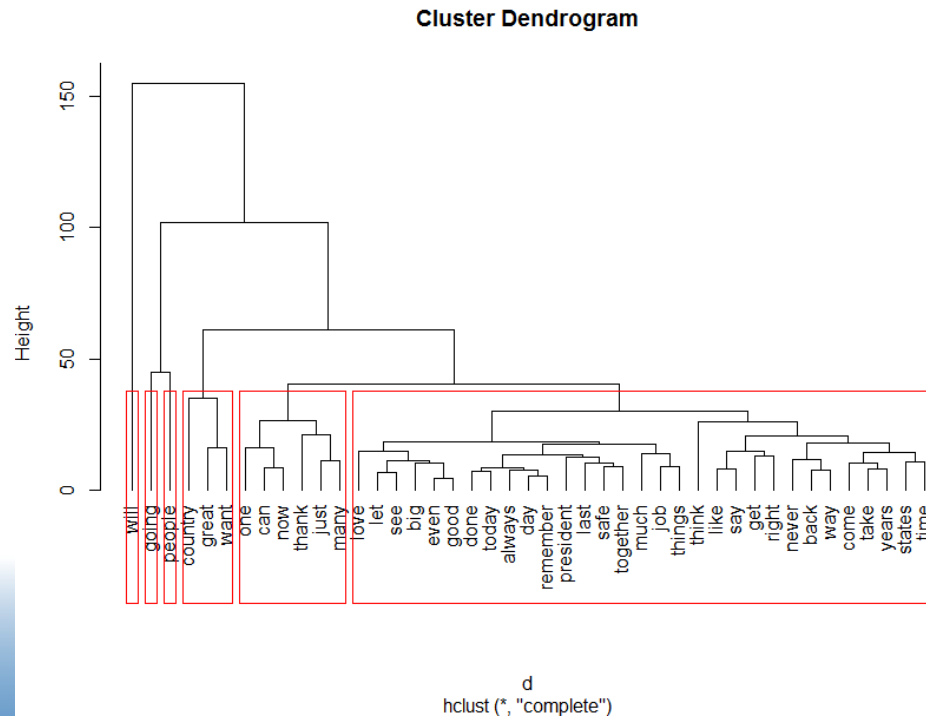


Clustering by Term Similarity

- **Helping to Read a Dendrogram**

- `plot.new()`
- `plot(fit, hang=-1)`
- `groups <- cutree(fit, k=6)` # "k=" defines the number of clusters you are using
- `rect.hclust(fit, k=6, border="red")` # draw dendrogram with red borders around the 6 clusters

- **Output**

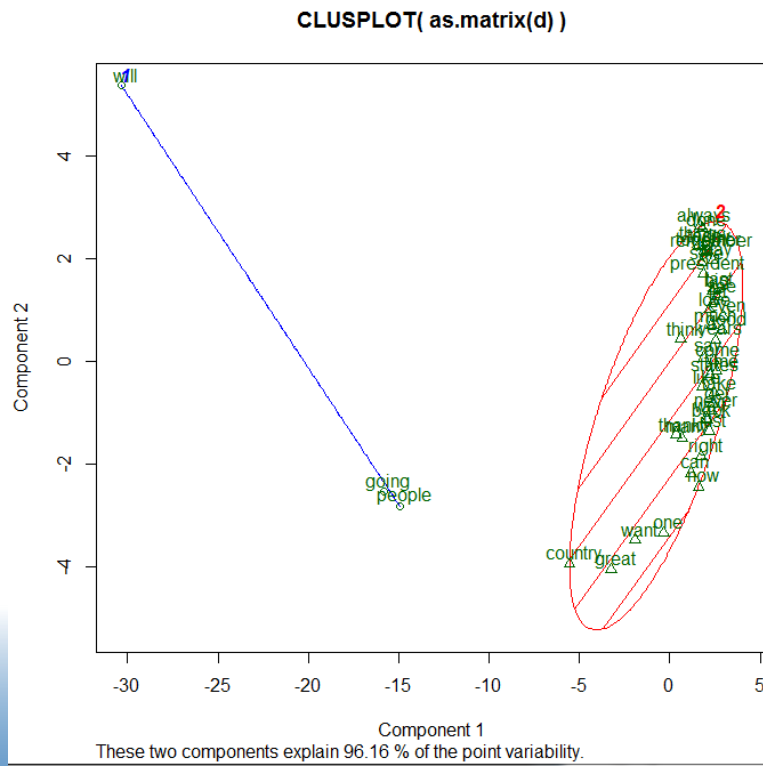


Clustering by Term Similarity

- **K-means clustering**

- library(fpc)
- d <- dist(t(dtmss), method="euclidian")
- kfit <- kmeans(d, 2)
- clusplot(as.matrix(d), kfit\$cluster, color=T, shade=T, labels=2, lines=0)

- **Output**



Classification data

- Movie comments
 - 1000 positive and 1000 negative processed reviews. Introduced in Pang/Lee ACL 2004. Released June 2004.
 - <http://www.cs.cornell.edu/people/pabo/movie-review-data/>

TF-IDF Weighting Function

- Term frequency (TF)
 - how many times a term occurs in a document
- Frequency-inverse document frequency (TF-IDF)
 - all words in the corpus are not equally important
 - IDF: $\text{Log}(N/d)$,
 - A corpus has N documents, and a term appears in d of them
 - A terms most of the documents will have, the IDF value will be low. Otherwise, will be large

TF-IDF Weighting Function

- **Code:**

- `tdm.pos3 <- TermDocumentMatrix(corpus.pos[1:3], control = list(weighting = weightTfIdf, stopwords = T, removePunctuation = T, stemming = T))`
- `inspect(tdm.pos3)`

- **Output:**

Terms	Docs		
	cv000_29590.txt	cv001_18431.txt	cv002_15918.txt
broderick	0	0.02264232	0.00000000
elect	0	0.04528464	0.00000000
fox	0	0.00000000	0.02264232
manipul	0	0.00000000	0.02264232
rushmor	0	0.02717079	0.00000000
ryan	0	0.00000000	0.02264232
school	0	0.02264232	0.00000000
shop	0	0.00000000	0.02264232
student	0	0.02264232	0.00000000
witherspoon	0	0.02264232	0.00000000

Classification

- **Preprocessing:**

- `Source<-DirSource("work_path",recursive=T)`
- `corpus <- Corpus(source)`
- `dtm <- DocumentTermMatrix(corpus,control = list(weighting = weightTfIdf,`
 - `stopwords = T, removePunctuation = T,`
 - `stemming = T))`
- `x.df <- as.data.frame(as.matrix(dtm))`
- `dim(x.df)`
- `x.df$class_label <- c(rep(0,1000),rep(1,1000))`
- `index <- sample(1:1000,300)`

- `train <- x.df[-c(index,index+1000),]`
- `test <- x.df[c(index,index+1000),]`
- `table(train$class_label)`
- `s <- findFreqTerms(dtm,4)#the lower bound value of IDF is 4`
- `s <- c(s,"class_label")`

Classification

- **Logistic Regression:**

- `model.glm <- glm(class_label ~ ., train[,s], family='binomial')`
- `coef(model.glm)`

- **Output:**

```
> coef(model.glm)
(Intercept)      act      action      actor      actual
-8.290933e-02 -1.766547e+01 -9.532001e+00 -2.175362e+01 -2.635650e+01
alien          also      although      american      anim
 3.597223e+00  1.112705e+02  2.490209e+01  3.489623e+01  3.763168e+00
anoth        audienc      back      bad      becom
-2.417609e+01 -2.421298e+01  8.131895e+01 -1.569754e+02 -3.976716e+01
best         better      big      brother      can
 3.345572e+01 -4.905861e+01 -3.733192e+01 -1.236567e+01 -5.080901e+01
cast         charact      come      comedi      day
 8.234391e-03 -6.839820e+00 -3.756932e+01 -8.845282e+00  1.613605e+01
direct       director      doesn      don      effect
 2.508545e+01 -2.254644e+01 -1.640883e+01 -2.382740e+01  4.716120e+00
end          enough      even      everi      famili
 7.671934e+00 -2.801125e+01 -9.795454e+01 -7.182451e+00  2.156777e+01
feel         film      final      find      first
 3.603780e+01 -2.603441e+01  3.604573e+01  2.805977e+01  3.070028e+01
friend       funni      get      girl      give
-2.508410e+00 -2.690193e+01  3.131949e+01 -1.175206e+01  2.247310e+00
good         great      guy      happen      hard
 3.372868e+01  1.287330e+02  3.185141e+01  5.776861e-01 -5.528525e+01
```

Classification

- **Verification:**

- `P<-ifelse(predict(model.glm, newdata = test) < 0,0,1)`
- `table(p , test$class_label)`

- **Output:**

p	0	1
0	229	94
1	71	206

Classification

- **Support Vector Machine:**

- `model.svm<-tune.svm(class_label ~., data = train[,s],kernel = "radial",gamma = gamma_list, cost=cost_list)`
- `model.svm`

- **Output:**

```
Parameter tuning of 'svm':  
- sampling method: 10-fold cross validation  
  
- best parameters:  
      gamma cost  
0.00390625 0.25  
  
- best performance: 0.1885059
```

Classification

- **Support Vector Machine:**

- `model_svm<-svm(class_label ~., data = train[,s],kernel = "radial",gamma = 0.00390625,cost = 0.25)`
- `model_svm`

- **Output:**

```
Call:
svm(formula = class_label ~ ., data = train[, s], kernel = "radial",
     gamma = 0.00390625, cost = 0.25)
```

Parameters:

```
SVM-Type:  eps-regression
SVM-Kernel: radial
cost:      0.25
gamma:     0.00390625
epsilon:   0.1
```

```
Number of Support Vectors: 1332
```

Classification

- **Support Vector Machine:**

- `model_svm<-svm(class_label ~., data = train[,s],kernel = "radial",gamma = 0.00390625,cost = 0.25)`
- `model_svm`

- **Output:**

```
Call:
svm(formula = class_label ~ ., data = train[, s], kernel = "radial",
     gamma = 0.00390625, cost = 0.25)
```

Parameters:

```
SVM-Type:  eps-regression
SVM-Kernel: radial
cost:      0.25
gamma:     0.00390625
epsilon:   0.1
```

```
Number of Support Vectors: 1332
```

Classification

- **Support Vector Machine:**

- `p.svm<-ifelse(predict(model_svm,newdata = test) < 0.5,0,1)`
- `table(p.svm , test$class_label)`

- **Output:**

```
p.svm  0  1
0 237  92
1  63 208
```


Chapter Summary

- **R packages: tm**
 - load the text dataset
 - Remove the noise from the data
- **TF and TF-IDF**
 - Frequency of occurrence of the term in the document
 - The importance of terms based on how infrequently the term occurs in the corpus
- **Frequency analysis**
- **Visualization**
- **Clustering**
- **Classification**

Homework 8 (You can practice by yourself and you do not need to submit.)

- Basic
 - Find a dataset and do text mining with data preprocessing and frequency analysis
- Advance
 - Practice machine learning with your text data

References

1. <http://blog.revolutionanalytics.com/2011/03/how-the-new-york-times-uses-r-for-data-visualization.html>
2. <http://www.stevefenton.co.uk/Content/Pie-Charts-Are-Bad/>.
3. Lewis, M. (2004). *Moneyball: The art of winning an unfair game*. NewYork: W.W. Norton &Company.
4. Teach yourself shiny. <http://shiny.rstudio.com/tutorial/>.
5. Tufte, E. (2001). *The visual display of quantitative information*. Cheshire: Graphics Press.
6. Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. Use R!. NewYork: Springer.
7. Wilkinson, L. (2005). *The grammar of graphics*. NewYork: Springer.
8. <https://www.analyticsvidhya.com/blog/2016/10/creating-interactive-data-visualization-using-shiny-app-in-r-with-examples/>
9. https://www.mailman.columbia.edu/sites/default/files/media/fdawg_ggplot2.html