# DSI Summer Workshops Series

## June 7, 2018

Peggy Lindner
Center for Advanced Computing & Data Science (CACDS)
Data Science Institute (DSI)
University of Houston
plindner@uh.edu

Please make sure you have a copy of R up and running, as well as a Python 3 installation (ideally from Anacodna).

## Goals for today

Understand basics of text analysis using R

(well enough so that you can Google your problems, find the answer, and implement it.)

### More specifically

1. Up and running with R & IPython
2. Understand a basic exploratory data analysis workflow
3. Basics of R and Topic Modeling

### Why R and not Python

It's good for data exploration!

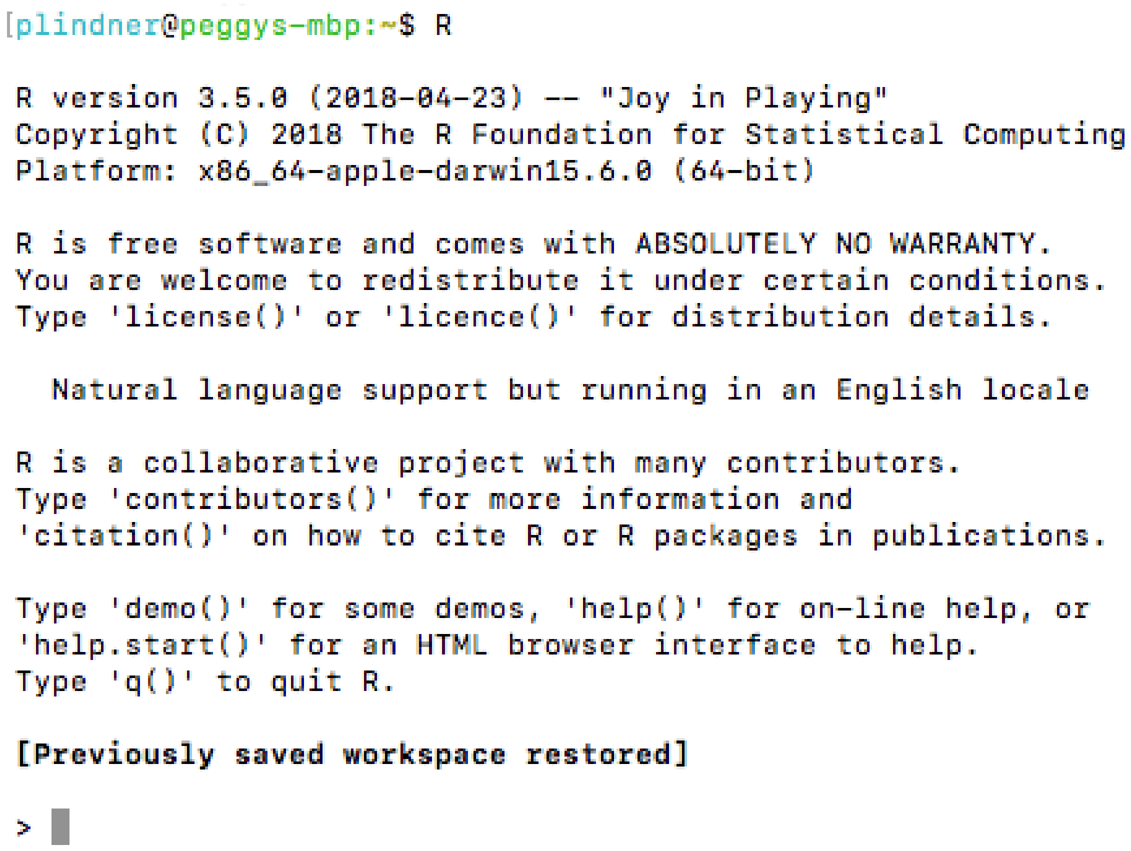
# Part 1: Getting yourself ready

## First: Install software on your computer

- R CRAN (https://www.anaconda.com/download/)
- PythonAnaconda (https://www.anaconda.com/download/)

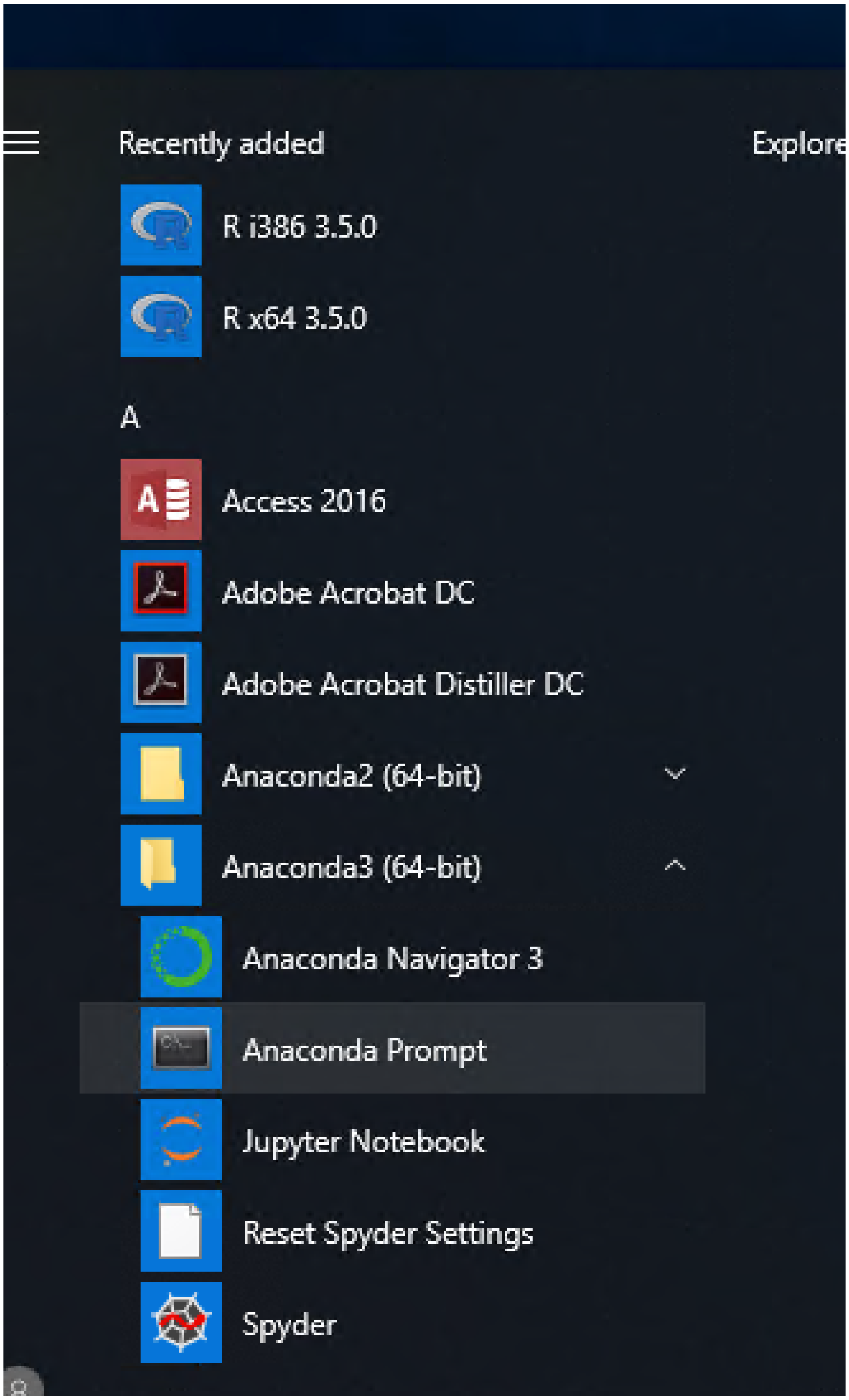
## Second: Prep your R environment

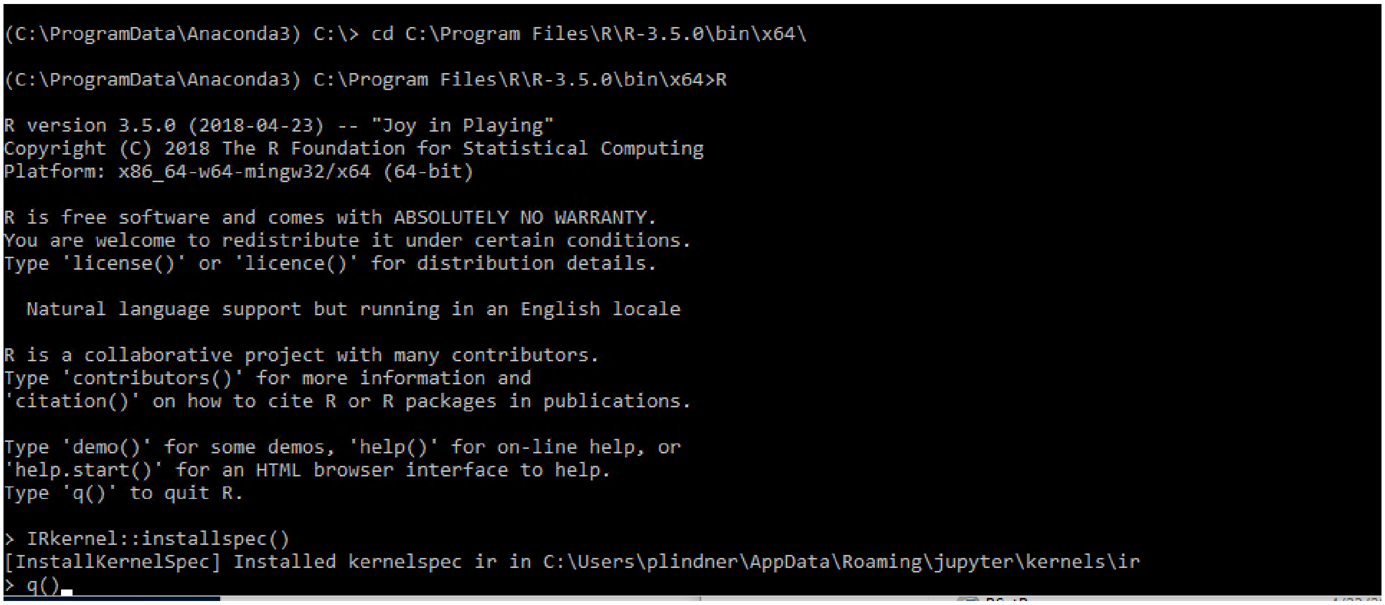
On a Mac open a terminal and start R



```
[plindner@peggys-mbp:~$ R

R version 3.5.0 (2018-04-23) -- "Joy in Playing"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

>
```

On Windows: Open the Anaconda Command line and start R



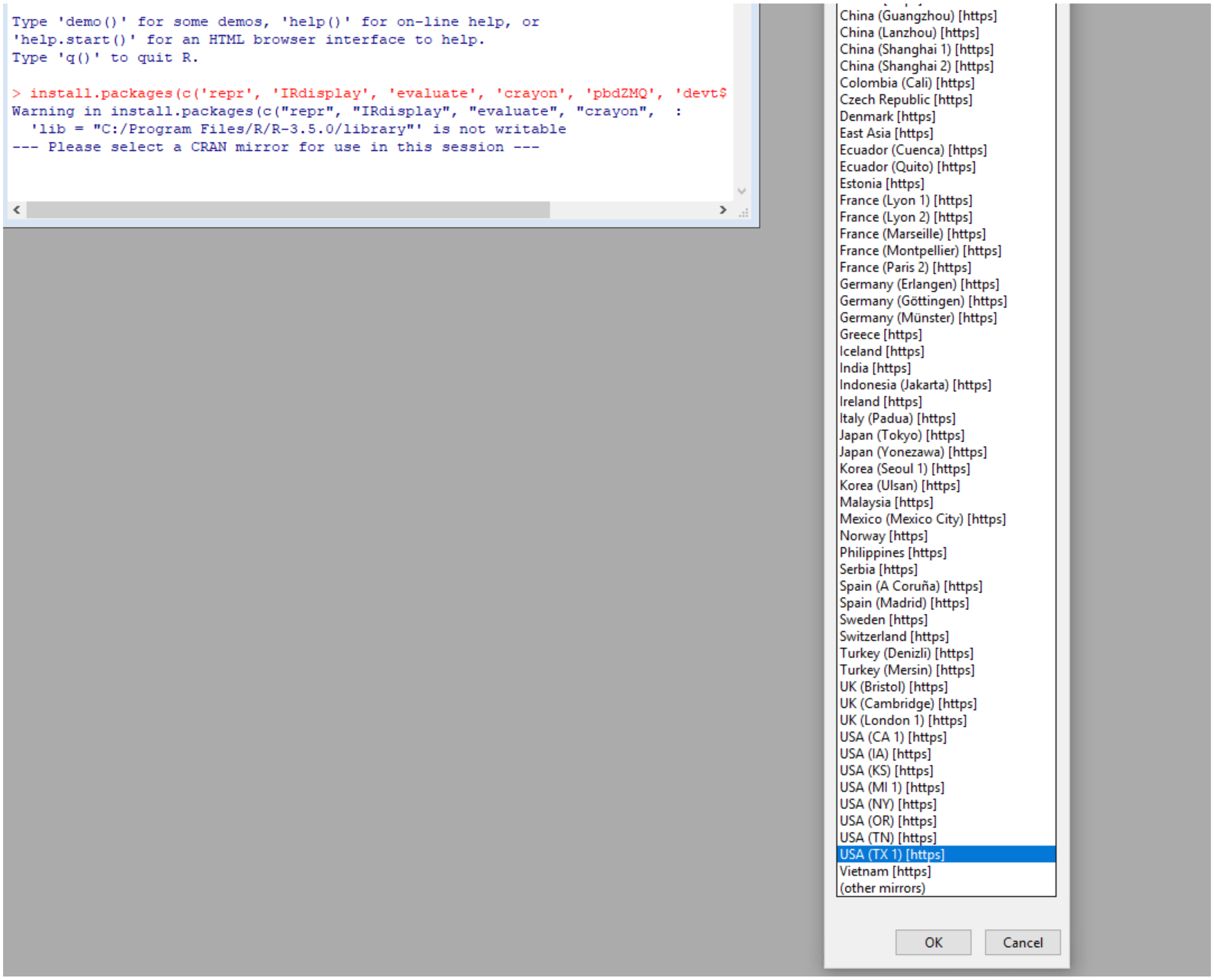


```
(C:\ProgramData\Anaconda3) C:\> cd C:\Program Files\R\R-3.5.0\bin\x64\

(C:\ProgramData\Anaconda3) C:\Program Files\R\R-3.5.0\bin\x64>R

R version 3.5.0 (2018-04-23) -- "Joy in Playing"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> IRkernel::installspec()
[InstallKernelSpec] Installed kernelspec ir in C:\Users\plindner\AppData\Roaming\jupyter\kernels\ir
> q()
```

Now let's install some packages ...

```
> install.packages(c('readr', 'stringr', 'SnowballC', 'w
ordcloud', 'RColorBrewer'))
> install.packages(c('tm', 'ggplot2', 'topicmodels'))
> install.packages(c('repr', 'IRdisplay', 'evaluate', 'c
rayon', 'pbdZMQ', 'devtools', 'uuid', 'digest'))
> devtools::install_github('IRkernel/IRkernel')
```

When you see "Please select a CRAN mirror" , well select one.



```
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> install.packages(c('repr', 'IRdisplay', 'evaluate', 'crayon', 'pbdZMQ', 'devt$
Warning in install.packages(c("repr", "IRdisplay", "evaluate", "crayon",  :
  'lib = "C:/Program Files/R/R-3.5.0/library"' is not writable
--- Please select a CRAN mirror for use in this session ---
```

... one last step - installing the Kernel

```
> IRkernel::installspec()
```

Now we can close the R environment (but leave your terminal and console open)

```
> quit()
```

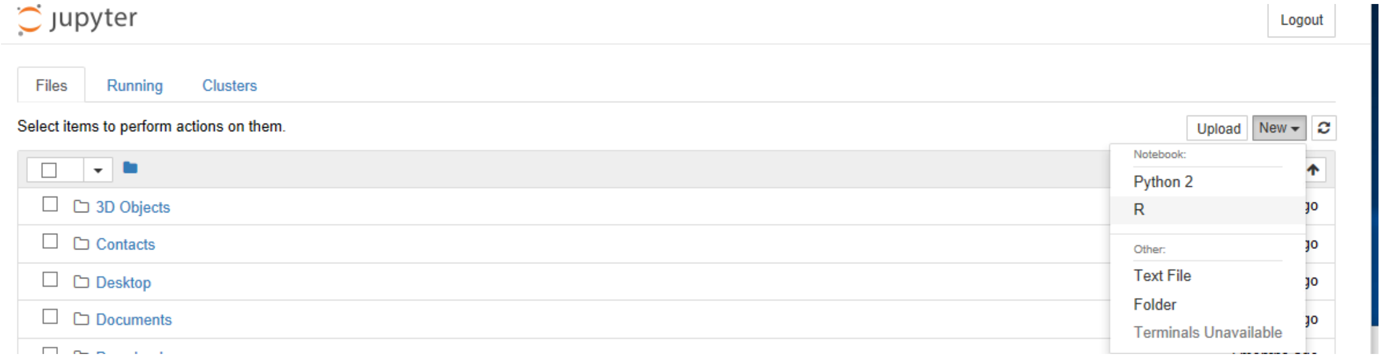
Say "N" (no) when asked to save the workspace.

## Jupyter Notebooks is what we will be going to use

We are now ready to start up our Jupyter Environment from the terminal or the console:

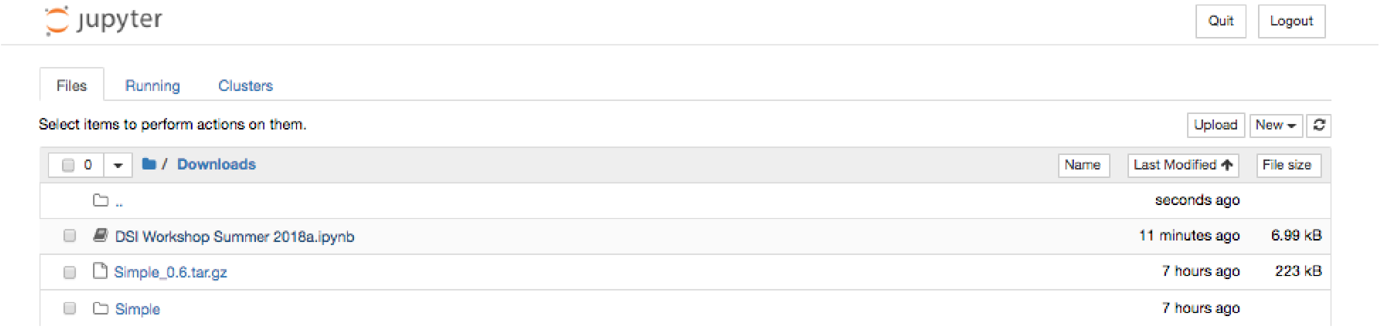
```
$ jupyter notebook --notebook-dir C:/Users/[your username]

or on a Mac

$ jupyter notebook --notebook-dir /Users/[your username]
```

And your browser should open at the address: http://localhost:8888/tree (http://localhost:8888/tree)



Select items to perform actions on them.

## Open the downloaded notebook on your computer



Select items to perform actions on them.

| Name | Last Modified | File size |
|---|---|---|
| .. | seconds ago | |
| DSI Workshop Summer 2018a.ipynb | 11 minutes ago | 6.99 kB |
| Simple_0.6.tar.gz | 7 hours ago | 223 kB |
| Simple | 7 hours ago | |

## Quick intro to Jupyter notebooks

Cells can be Markdown (like this one) or code

### To start off with

Make sure you hit `Shift-Enter` or `Ctrl-Enter` when you are done. You can also use the "Run" button.

```
In [1]:
```

```
2 + 2
```

```
4
```

## Part 2: The Exploratory Analysis Workflow

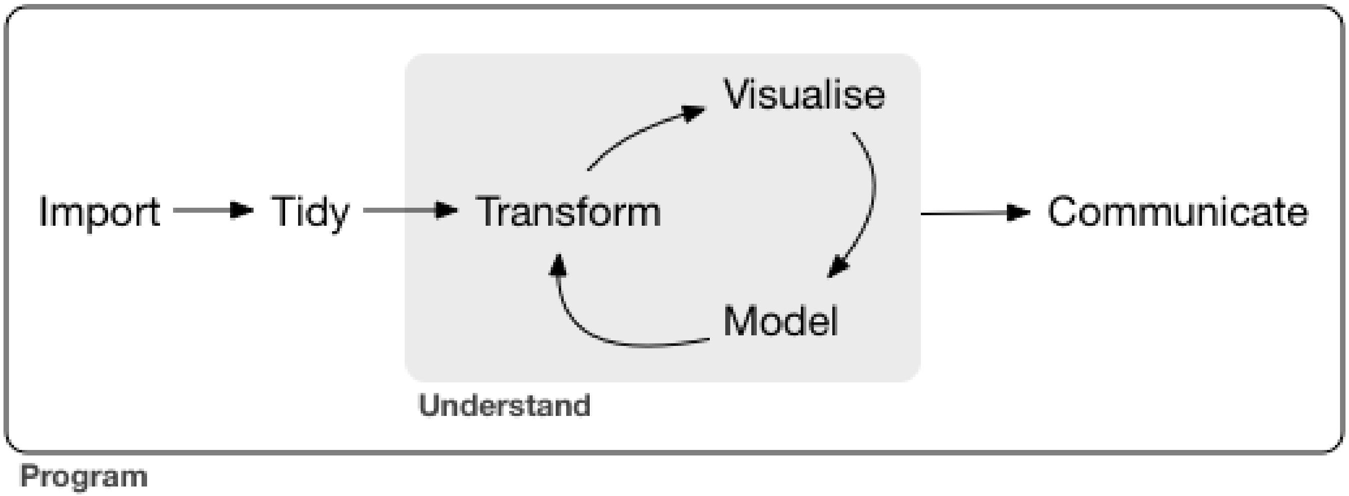
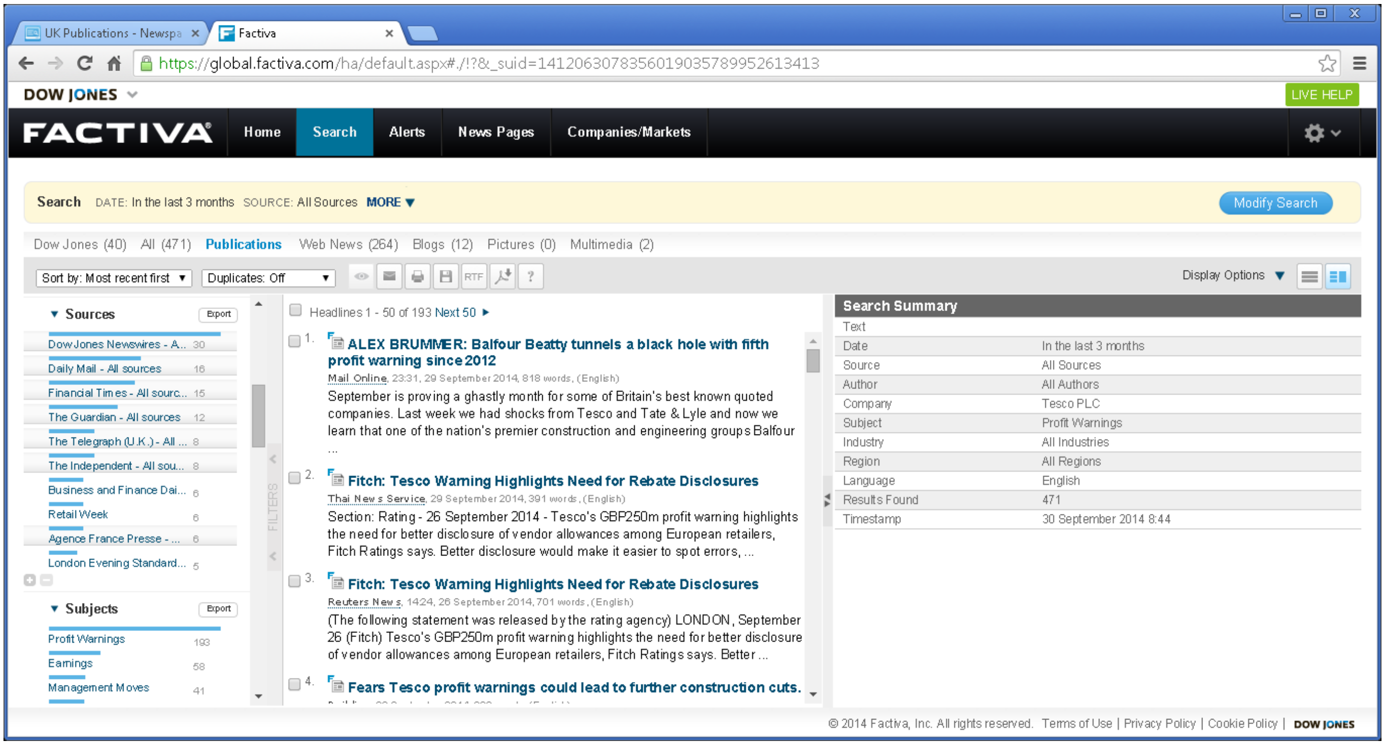


Image source: Hadley Wickham, R for Data Science

## Our Example

Media Analysis of a bunch of articles downloaded from a database called "Factiva"



Search DATE: In the last 3 months SOURCE: All Sources MORE ▼

Sort by: Most recent first ▼ Duplicates: Off ▼

▼ Sources

| | |
|---|---|
| Dow Jones Newswires - A... | 30 |
| Daily Mail - All sources | 16 |
| Financial Times - All sourc... | 15 |
| The Guardian - All sources | 12 |
| The Telegraph (U.K.) - All ... | 8 |
| The Independent - All sou... | 8 |
| Business and Finance Dai... | 6 |
| Retail Week | 6 |
| Agence France Presse - ... | 6 |
| London Evening Standard... | 5 |

▼ Subjects

| | |
|---|---|
| Profit Warnings | 193 |
| Earnings | 58 |
| Management Moves | 41 |

□ Headlines 1 - 50 of 193 Next 50 ▶

□ 1. **ALEX BRUMMER: Balfour Beatty tunnels a black hole with fifth profit warning since 2012**
Mail Online, 23:31, 29 September 2014, 818 words, (English)
September is proving a ghastly month for some of Britain's best known quoted companies. Last week we had shocks from Tesco and Tate & Lyle and now we learn that one of the nation's premier construction and engineering groups Balfour ...

□ 2. **Fitch: Tesco Warning Highlights Need for Rebate Disclosures**
Thai News Service, 29 September 2014, 391 words, (English)
Section: Rating - 26 September 2014 - Tesco's GBP250m profit warning highlights the need for better disclosure of vendor allowances among European retailers, Fitch Ratings says. Better disclosure would make it easier to spot errors, ...

□ 3. **Fitch: Tesco Warning Highlights Need for Rebate Disclosures**
Reuters News, 14:24, 26 September 2014, 701 words, (English)
(The following statement was released by the rating agency) LONDON, September 26 (Fitch) Tesco's GBP250m profit warning highlights the need for better disclosure of vendor allowances among European retailers, Fitch Ratings says. Better ...

□ 4. **Fears Tesco profit warnings could lead to further construction cuts.**

| Search Summary | |
|---|---|
| Text | |
| Date | In the last 3 months |
| Source | All Sources |
| Author | All Authors |
| Company | Tesco PLC |
| Subject | Profit Warnings |
| Industry | All Industries |
| Region | All Regions |
| Language | English |
| Results Found | 471 |
| Timestamp | 30 September 2014 8:44 |

© 2014 Factiva, Inc. All rights reserved. Terms of Use | Privacy Policy | Cookie Policy | DOW JONES

Make sure you download the data source file: https://raw.githubusercontent.com/peggylind/Materials_Summer2018/master/dataJun (https://raw.githubusercontent.com/peggylind/Materials_Summer2018/master/dataJur and store it in a folder called dataJune7th next the Jupyter notebook directory (next to the *.ipynb file).

### Frequently used R Packages in conjunction with text data

- readr (https://cran.r-project.org/web/packages/readr/readr.pdf) Import data

Data Analysis of text based material

- stringr (https://cran.r-project.org/web/packages/stringr/vignettes/stringr.html) Clean up text
- SnowballC (https://cran.r-project.org/web/packages/SnowballC/SnowballC.pdf) Stemming of words
- tm (https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf) Text mining
- Quanteda (https://quanteda.io/) veratile text analysis tool
- topicmodels (https://www.tidytextmining.com/topicmodeling.html) Topic Modeling

Visualization

- ggplot2 (http://ggplot2.tidyverse.org/) Modern R visulaizations
- wordcloud (http://developer.marvel.com) Make some nice word clouds
- RColorBrewer (https://dataset.readthedocs.org/en/latest/) Get color into your visualizations

```
In [2]:
```

```
#load all required libraries
library(readr)
library(stringr)
library(SnowballC)
library(wordcloud)
library(RColorBrewer)
library(tm)
library(ggplot2)
library(topicmodels)
```

```
Loading required package: RColorBrewer
Loading required package: NLP

Attaching package: ‘NLP’

The following object is masked from ‘package:ggplot
2’:

    annotate
```

### Data Import

```
In [3]:
```

```
# put the name of your csv file
inputfile <- "dataJune7th/sample.txt"
# read the data
alldata <- read_file(inputfile)
# look at the dat
# what type is our data?
str(alldata)
```

```
 chr " \n\n\nTurkey rejects US 'double standard' in
Syria ceasefire\n\n474 words\n6 March 2018\nAl Jaze
era English\nA"| __truncated__
```

### Prepare data

In [4]:

```
# data wrangling - split the file in different articles
split.word <- "Document AJAZEN(.*)"

# split up into individual documents
list_alldata_splitted <- str_split(alldata, split.word)
# convert to vector and remove last element (which is a leftover)
alldata_splitted <- unlist(list_alldata_splitted)
alldata_splitted <- alldata_splitted[-length(alldata_splitted)]
str(alldata_splitted)
```

```
 chr [1:71] " \n\n\nTurkey rejects US 'double stand
ard' in Syria ceasefire\n\n474 words\n6 March 2018
\nAl Jazeera English\nA"| __truncated__ ...
```

In [5]:

```
### create corpus
article.corpus <- Corpus(VectorSource((alldata_splitted)))

article.corpus
```

```
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (inde
xed): 0
Content:  documents: 71
```

In [6]:

```
#inspect a particular document
writeLines(as.character(article.corpus[[30]]))
```

# Qatar-Gulf crisis: All the latest updates

630 words
17 February 2018
Al Jazeera English
AJAZEN
English

Here are all the latest updates of the Qatar-Gulf crisis, now in its eighth month.

Seven months ago, an air, sea and land blockade was imposed on Qatar by neighbouring countries. Here are the latest developments as of Saturday, February 17:

## Munich Security Conference

"It has been a futile crisis, manufactured by our neighbours," Qatari Emir Sheikh Tamim bin Hamad Al Thani told a major security conference held in Germany on Friday. "By defusing the impact of the illegal and aggressive measures imposed on our people, Qatar has preserved its sovereignty."

* "Those aggressive actors wish to use smaller states as pawns within their power games and sectarian conflicts. It is vital to the interests of the people of the Middle East to guarantee the sovereignty of states like Qatar."

## Tillerson in Kuwait

* On Tuesday, the US Secretary of State Rex Tillerson said that the restoration of Arab Gulf unity was in the best interest of all parties in the region.

* Tillerson made the assertion at a press conference held in Kuwait, where he is attending a high-leve

l meeting between members of a US-led coalition against the Islamic State of Iraq and the Levant (ISIL, also known as ISIS).

### Asian Championships League

* On Monday, Al Gharafa of Qatar opened its Asian Championships League campaign in Abu Dhabi against Al Jazira of United Arab Emirates.

* UAE requested to play those games in a third country, but the idea was rejected by the Asian Football Confederation, which organises the tournament featuring 32 teams split into eight groups of four.

* "Clubs from Qatar, Saudi Arabia and the United Arab Emirates should be played on a home and away basis in 2018 as per the AFC regulations," the AFC said in a recent statement.

* The soccer federations of the UAE and Saudi Arabia accepted the decision though expressed reservations about how it had been made.

### Anti-Qatar conference

* On Sunday, a report on Buzzfeed revealed that Daniel Kawczynski, a British parliamentarian, was paid 15,000 British pounds ($20,700) to help organise an anti-Qatar conference in London.

* At the time, analysts described the conference as an attempt to gather support for a coup in Qatar and accused Saudi Arabia and the UAE of funding it.

### Russian president's invitation

* On Wednesday, the Emir Tamim bin Hamad Al Thani met the President of the Russian Republic of Ingushetia Yunus-bek Yevkurov in Doha.

* During the meeting, they reviewed the relations between Qatar and Ingushetia and discussed aspects of cooperation in economic, trade and investment, as

well as means of developing them.

* President Yunus-bek Yevkurov handed a written message from Russian President Vladimir Putin, including an invitation to visit Russia.

## Calls to end the blockade

* On Tuesday, local media reported that Dr Ali bin Smaikh al-Marri, Chairman of the National Human Rights Committee (NHRC), called on the European Union countries and some other nations to join international human rights organisations demanding an immediate end to the blockade imposed on Qatar.

## Louvre Map

* On Monday, Qatar's National Human Rights Committee reported that the Louvre museum apologised and opened an official inquiry into the incident where Abu Dhabi's Louvre Museum map omitted the Qatari Peninsula.

## Bilateral agreements

* Anti-terrorism: On Monday, both countries held a session of official talks and signed a memorandum of understanding (MoU) on the "fight against terrorism" and organised crime."

* Sierra Leone: On Monday, both countries signed an agreement encouraging and protecting mutual investments and an agreement on air services between the two governments.

* Sierra Leone was not among the eight countries that downgraded ties with Qatar at the start of the crisis.

June 5, 2017 - February 02, 2018 For all previous developments click here Al Jazeera and news agencies

In [ ]:

```
#Check details (look at bunched up corpus to find anomalies)
inspect(article.corpus)
```

### Data cleaning

In [8]:

```
#create the toSpace content transformer
toSpace <- content_transformer(function(x, pattern) { return (
gsub(pattern, " ", x))})
#to remove potentially problematic symbols
article.corpus <- tm_map(article.corpus, toSpace, "-")
article.corpus <- tm_map(article.corpus, toSpace, ":")
article.corpus <- tm_map(article.corpus, toSpace, "'")
article.corpus <- tm_map(article.corpus, toSpace, "'")
article.corpus <- tm_map(article.corpus, toSpace, " -")

#Good practice to check after each step.
writeLines(as.character(article.corpus[[30]]))
```

# Qatar Gulf crisis  All the latest updates

630 words
17 February 2018
Al Jazeera English
AJAZEN
English

Here are all the latest updates of the Qatar Gulf crisis, now in its eighth month.

Seven months ago, an air, sea and land blockade was imposed on Qatar by neighbouring countries. Here are the latest developments as of Saturday, February 17

## Munich Security Conference

"It has been a futile crisis, manufactured by our neighbours," Qatari Emir Sheikh Tamim bin Hamad Al Thani told a major security conference held in Germany on Friday. "By defusing the impact of the illegal and aggressive measures imposed on our people, Qatar has preserved its sovereignty."

* "Those aggressive actors wish to use smaller states as pawns within their power games and sectarian conflicts. It is vital to the interests of the people of the Middle East to guarantee the sovereignty of states like Qatar."

## Tillerson in Kuwait

* On Tuesday, the US Secretary of State Rex Tillerson said that the restoration of Arab Gulf unity was in the best interest of all parties in the region.

* Tillerson made the assertion at a press conference held in Kuwait, where he is attending a high leve

l meeting between members of a US led coalition against the Islamic State of Iraq and the Levant (ISIL, also known as ISIS).

## Asian Championships League

* On Monday, Al Gharafa of Qatar opened its Asian Championships League campaign in Abu Dhabi against Al Jazira of United Arab Emirates.

* UAE requested to play those games in a third country, but the idea was rejected by the Asian Football Confederation, which organises the tournament featuring 32 teams split into eight groups of four.

* "Clubs from Qatar, Saudi Arabia and the United Arab Emirates should be played on a home and away basis in 2018 as per the AFC regulations," the AFC said in a recent statement.

* The soccer federations of the UAE and Saudi Arabia accepted the decision though expressed reservations about how it had been made.

## Anti Qatar conference

* On Sunday, a report on Buzzfeed revealed that Daniel Kawczynski, a British parliamentarian, was paid 15,000 British pounds ($20,700) to help organise an anti Qatar conference in London.

* At the time, analysts described the conference as an attempt to gather support for a coup in Qatar and accused Saudi Arabia and the UAE of funding it.

## Russian president s invitation

* On Wednesday, the Emir Tamim bin Hamad Al Thani met the President of the Russian Republic of Ingushetia Yunus bek Yevkurov in Doha.

* During the meeting, they reviewed the relations between Qatar and Ingushetia and discussed aspects of cooperation in economic, trade and investment, as

well as means of developing them.

* President Yunus bek Yevkurov handed a written message from Russian President Vladimir Putin, including an invitation to visit Russia.

## Calls to end the blockade

* On Tuesday, local media reported that Dr Ali bin Smaikh al Marri, Chairman of the National Human Rights Committee (NHRC), called on the European Union countries and some other nations to join international human rights organisations demanding an immediate end to the blockade imposed on Qatar.

## Louvre Map

* On Monday, Qatar s National Human Rights Committee reported that the Louvre museum apologised and opened an official inquiry into the incident where Abu Dhabi s Louvre Museum map omitted the Qatari Peninsula.

## Bilateral agreements

* Anti terrorism  On Monday, both countries held a session of official talks and signed a memorandum of understanding (MoU) on the "fight against terrorism" and organised crime."

* Sierra Leone  On Monday, both countries signed an agreement encouraging and protecting mutual investments and an agreement on air services between the two governments.

* Sierra Leone was not among the eight countries that downgraded ties with Qatar at the start of the crisis.

June 5, 2017   February 02, 2018 For all previous developments click here Al Jazeera and news agencies

In [9]:

```
#Remove punctuation - replace punctuation marks with " "
article.corpus <- tm_map(article.corpus, removePunctuation)

#Good practice to check after each step.
writeLines(as.character(article.corpus[[30]]))
```

# Qatar Gulf crisis  All the latest updates

630 words
17 February 2018
Al Jazeera English
AJAZEN
English

Here are all the latest updates of the Qatar Gulf crisis now in its eighth month

Seven months ago an air sea and land blockade was imposed on Qatar by neighbouring countries Here are the latest developments as of Saturday February 17

Munich Security Conference

It has been a futile crisis manufactured by our neighbours Qatari Emir Sheikh Tamim bin Hamad Al Thani told a major security conference held in Germany on Friday By defusing the impact of the illegal and aggressive measures imposed on our people Qatar has preserved its sovereignty

Those aggressive actors wish to use smaller states as pawns within their power games and sectarian conflicts It is vital to the interests of the people of the Middle East to guarantee the sovereignty of states like Qatar

Tillerson in Kuwait

On Tuesday the US Secretary of State Rex Tillerson said that the restoration of Arab Gulf unity was in the best interest of all parties in the region

Tillerson made the assertion at a press conference held in Kuwait where he is attending a high level meeting between members of a US led coalition agains t the Islamic State of Iraq and the Levant ISIL also known as ISIS

## Asian Championships League

On Monday Al Gharafa of Qatar opened its Asian Championships League campaign in Abu Dhabi against Al Jazira of United Arab Emirates

UAE requested to play those games in a third country but the idea was rejected by the Asian Football Confederation which organises the tournament featuring 32 teams split into eight groups of four

Clubs from Qatar Saudi Arabia and the United Arab Emirates should be played on a home and away basis in 2018 as per the AFC regulations the AFC said in a recent statement

The soccer federations of the UAE and Saudi Arabia accepted the decision though expressed reservations about how it had been made

## Anti Qatar conference

On Sunday a report on Buzzfeed revealed that Daniel Kawczynski a British parliamentarian was paid 15000 British pounds 20700 to help organise an anti Qatar conference in London

At the time analysts described the conference as an attempt to gather support for a coup in Qatar and accused Saudi Arabia and the UAE of funding it

## Russian president s invitation

On Wednesday the Emir Tamim bin Hamad Al Thani met the President of the Russian Republic of Ingushetia Yunus bek Yevkurov in Doha

During the meeting they reviewed the relations between Qatar and Ingushetia and discussed aspects of cooperation in economic trade and investment as well as means of developing them

President Yunus bek Yevkurov handed a written message from Russian President Vladimir Putin including an invitation to visit Russia

Calls to end the blockade

On Tuesday local media reported that Dr Ali bin Smaikh al Marri Chairman of the National Human Rights Committee NHRC called on the European Union countries and some other nations to join international human rights organisations demanding an immediate end to the blockade imposed on Qatar

Louvre Map

On Monday Qatar s National Human Rights Committee reported that the Louvre museum apologised and opened an official inquiry into the incident where Abu Dhabi s Louvre Museum map omitted the Qatari Peninsula

Bilateral agreements

Anti terrorism  On Monday both countries held a session of official talks and signed a memorandum of understanding MoU on the fight against terrorism and organised crime

Sierra Leone  On Monday both countries signed an agreement encouraging and protecting mutual investments and an agreement on air services between the two governments

Sierra Leone was not among the eight countries that downgraded ties with Qatar at the start of the crisis

June 5 2017   February 02 2018 For all previous developments click here Al Jazeera and news agencies

In [10]:

```
#Transform to lower case
article.corpus <- tm_map(article.corpus,content_transformer(tolower))

#Strip digits
article.corpus <- tm_map(article.corpus, removeNumbers)

#Remove stopwords from standard stopword list
article.corpus <- tm_map(article.corpus, removeWords, stopwords("english"))

#inspect output
writeLines(as.character(article.corpus[[30]]))
```

qatar gulf crisis latest updates

words<br>
february<br>
al jazeera english<br>
ajazen<br>
english<br>

latest updates qatar gulf crisis now eighth month

seven months ago air sea land blockade imposed qatar neighbouring countries latest developments saturday february

munich security conference

futile crisis manufactured neighbours qatari emir sheikh tamim bin hamad al thani told major security conference held germany friday defusing impact illegal aggressive measures imposed people qatar preserved sovereignty

aggressive actors wish use smaller states pawns within power games sectarian conflicts vital interests people middle east guarantee sovereignty states like qatar

tillerson kuwait

tuesday us secretary state rex tillerson said restoration arab gulf unity best interest parties region

tillerson made assertion press conference held kuwait attending high level meeting members us led coalition islamic state iraq levant isil also known isis

## asian championships league

monday al gharafa qatar opened asian championships league campaign abu dhabi al jazira united arab emirates

uae requested play games third country idea rejected asian football confederation organises tournament featuring teams split eight groups four

clubs qatar saudi arabia united arab emirates played home away basis per afc regulations afc said recent statement

soccer federations uae saudi arabia accepted decision though expressed reservations made

## anti qatar conference

sunday report buzzfeed revealed daniel kawczynski british parliamentarian paid british pounds help organise anti qatar conference london

time analysts described conference attempt gather support coup qatar accused saudi arabia uae funding

## russian president s invitation

wednesday emir tamim bin hamad al thani met president russian republic ingushetia yunus bek yevkurov doha

meeting reviewed relations qatar ingushetia discussed aspects cooperation economic trade investment well means developing

president yunus bek yevkurov handed written message russian president vladimir putin including invitation visit russia

## calls end blockade

```
  tuesday local media reported  dr ali bin smaikh a
l marri chairman   national human rights committee
nhrc called   european union countries    nations
join international human rights organisations deman
ding  immediate end   blockade imposed  qatar

louvre map

  monday qatar s national human rights committee re
ported   louvre museum apologised  opened  official
inquiry   incident  abu dhabi s louvre museum map o
mitted  qatari peninsula

bilateral agreements

 anti terrorism   monday  countries held  session
official talks  signed  memorandum  understanding m
ou   fight  terrorism  organised crime

 sierra leone   monday  countries signed  agreement
encouraging  protecting mutual investments   agreem
ent  air services   two governments

 sierra leone   among  eight countries  downgraded
ties  qatar   start   crisis

june     february     previous developments click
al jazeera  news agencies
```

Stopwords (https://github.com/arc12/Text-Mining-Weak-Signals/wiki/Standard-set-of-english-stopwords)

In [11]:

```
#define and eliminate all custom stopwords
myStopwords <- c("monday")
article.corpus <- tm_map(article.corpus, removeWords, myStopwords)

#Strip whitespace (cosmetic?)
article.corpus <- tm_map(article.corpus, stripWhitespace)

#inspect output
writeLines(as.character(article.corpus[[30]]))
```

qatar gulf crisis latest updates words february al jazeera english ajazen english copyright al jazeera english latest updates qatar gulf crisis now eighth month seven months ago air sea land blockade imposed qatar neighbouring countries latest developments saturday february munich security conference futile crisis manufactured neighbours qatari emir sheikh tamim bin hamad al thani told major security conference held germany friday defusing impact illegal aggressive measures imposed people qatar preserved sovereignty aggressive actors wish use smaller states pawns within power games sectarian conflicts vital interests people middle east guarantee sovereignty states like qatar tillerson kuwait tuesday us secretary state rex tillerson said restoration arab gulf unity best interest parties region tillerson made assertion press conference held kuwait attending high level meeting members us led coalition islamic state iraq levant isil also known isis asian championships league al gharafa qatar opened asian championships league campaign abu dhabi al jazira united arab emirates uae requested play games third country idea rejected asian football confederation organises tournament featuring teams split eight groups four clubs qatar saudi arabia united arab emirates played home away basis per afc regulations afc said recent statement soccer federations uae saudi arabia accepted decision though expressed reservations made anti qatar conference sunday report buzzfeed revealed daniel kawczynski british parliamentarian paid british pounds help organise anti qatar conference london time analysts described conference attempt gather support coup qatar accused saudi arabia uae funding russian president s invitation wednesday emir tamim bin hamad al thani met president russian republic ingushetia yunus bek yevkurov doha meeting reviewed relations qatar ingushetia discussed aspects cooperation economic trade investment well means developing president yunus bek yevkurov handed written message russian president vladimir putin including invitation visit russia calls end blockade tuesday local media reported dr ali bin smaikh al marri chairman national human rights committee nhrc calle

```
d european union countries nations join internation
al human rights organisations demanding immediate e
nd blockade imposed qatar louvre map qatar s nation
al human rights committee reported louvre museum ap
ologised opened official inquiry incident abu dhabi
s louvre museum map omitted qatari peninsula bilate
ral agreements anti terrorism countries held sessio
n official talks signed memorandum understanding mo
u fight terrorism organised crime sierra leone coun
tries signed agreement encouraging protecting mutua
l investments agreement air services two government
s sierra leone among eight countries downgraded tie
s qatar start crisis june february previous develop
ments click al jazeera news agencies
```

Word Stemming (http://www.omegahat.net/Rstem/stemming.pdf)

In [12]:

```
#Stem document
article.corpus <- tm_map(article.corpus,stemDocument)

#inspect output
writeLines(as.character(article.corpus[[30]]))
```

qatar gulf crisi latest updat word februari al jazeera english ajazen english copyright al jazeera english latest updat qatar gulf crisi now eighth month seven month ago air sea land blockad impos qatar neighbour countri latest develop saturday februari munich secur confer futil crisi manufactur neighbour qatari emir sheikh tamim bin hamad al thani told major secur confer held germani friday defus impact illeg aggress measur impos peopl qatar preserv sovereignti aggress actor wish use smaller state pawn within power game sectarian conflict vital interest peopl middl east guarante sovereignti state like qatar tillerson kuwait tuesday us secretari state rex tillerson said restor arab gulf uniti best interest parti region tillerson made assert press confer held kuwait attend high level meet member us led coalit islam state iraq levant isil also known isi asian championship leagu al gharafa qatar open asian championship leagu campaign abu dhabi al jazira unit arab emir uae request play game third countri idea reject asian footbal confeder organis tournament featur team split eight group four club qatar saudi arabia unit arab emir play home away basi per afc regul afc said recent statement soccer feder uae saudi arabia accept decis though express reserv made anti qatar confer sunday report buzzfe reveal daniel kawczynski british parliamentarian paid british pound help organis anti qatar confer london time analyst describ confer attempt gather support coup qatar accus saudi arabia uae fund russian presid s invit wednesday emir tamim bin hamad al thani met presid russian republ ingushetia yunus bek yevkurov doha meet review relat qatar ingushetia discuss aspect cooper econom trade invest well mean develop presid yunus bek yevkurov hand written messag russian presid vladimir putin includ invit visit russia call end blockad tuesday local media report dr ali bin smaikh al marri chairman nation human right committe nhrc call european union countri nation join intern human right organis demand immedi end blockad impos qatar louvr map qatar s nation human right committe report louvr museum apologis open offici inquiri incid abu dhabi s louvr museum map omit qatari peninsul

```
a bilater agreement anti terror countri held sessio
n offici talk sign memorandum understand mou fight
terror organis crime sierra leon countri sign agree
ment encourag protect mutual invest agreement air s
ervic two govern sierra leon among eight countri do
wngrad tie qatar start crisi june februari previous
develop click al jazeera news agenc
```

### Prepare for Analysis - create word counts

In [13]:

```
#Create document-term matrix
dtm <- DocumentTermMatrix(article.corpus)

dtm
```

```
<<DocumentTermMatrix (documents: 71, terms: 3593)>>
Non-/sparse entries: 15956/239147
Sparsity           : 94%
Maximal term length: 16
Weighting          : term frequency (tf)
```

In [14]:

```
#inspect segment of document term matrix
inspect(dtm[15:16,100:105])
```

```
<<DocumentTermMatrix (documents: 2, terms: 6)>>
Non-/sparse entries: 9/3
Sparsity           : 25%
Maximal term length: 7
Weighting          : term frequency (tf)
Sample             :
    Terms
Docs isil islam jazeera known last least
  15    2     1       4     1    2     0
  16    2     1       3     1    0     0
```

In [15]:

```
#collapse matrix by summing over columns - this gets total counts (over all docs) for each term
freq <- colSums(as.matrix(dtm))
#length should be total number of terms
length(freq)
```

```
3593
```

In [16]:

```
#create sort order (descending)
ord <- order(freq,decreasing=TRUE)
#inspect most frequently occurring terms
freq[head(ord)]
#inspect least frequently occurring terms
freq[tail(ord)]

#List all terms in decreasing order of freq and write to disk
write.csv(freq[ord],"word_freq.csv")
```

```
said
348
syria
347
syrian
299
jazeera
243
state
217
english
214


rampant
1
riyadh
1
secessionist
1
stabl
1
starvat
1
takeov
1
```

In [17]:

```
#alterantive: remove very frequent and very rare words
dtmr <-DocumentTermMatrix(article.corpus, control=list(wordLen
gths=c(4, 20),
                                  bounds = list(global = c(3,2
7))))

dtmr

freqr <- colSums(as.matrix(dtmr))
#length should be total number of terms
length(freqr)

#create sort order (desc)
ordr <- order(freqr,decreasing=TRUE)
#inspect most frequently occurring terms
freqr[head(ordr)]
#inspect least frequently occurring terms
freqr[tail(ordr)]
```

```
<<DocumentTermMatrix (documents: 71, terms: 1254)>>
Non-/sparse entries: 9263/79771
Sparsity           : 90%
Maximal term length: 15
Weighting          : term frequency (tf)
```

1254

**assad**

116

**turkish**

103

**israel**

103

**kurdish**

89

**russia**

78

**rebel**

74

**weaponri**

3

**tribe**

3

**revolt**

3

**technic**

3

**event**

3

**diplomaci**

3

In [18]:

```
#list most frequent terms. Lower bound specified as second argument
findFreqTerms(dtmr,lowfreq=60)
```

```
'eastern'  'right'  'turkish'  'kurdish'  'rebel'  'iran'
'qatar'  'assad'  'children'  'isra'  'israel'  'russia'
```

Now that we have the most frequently occurring terms in hand, we can check for correlations between some of these and other terms that occur in the corpus. In this context, correlation is a quantitative measure of the co-occurrence of words in multiple documents.

In [ ]:

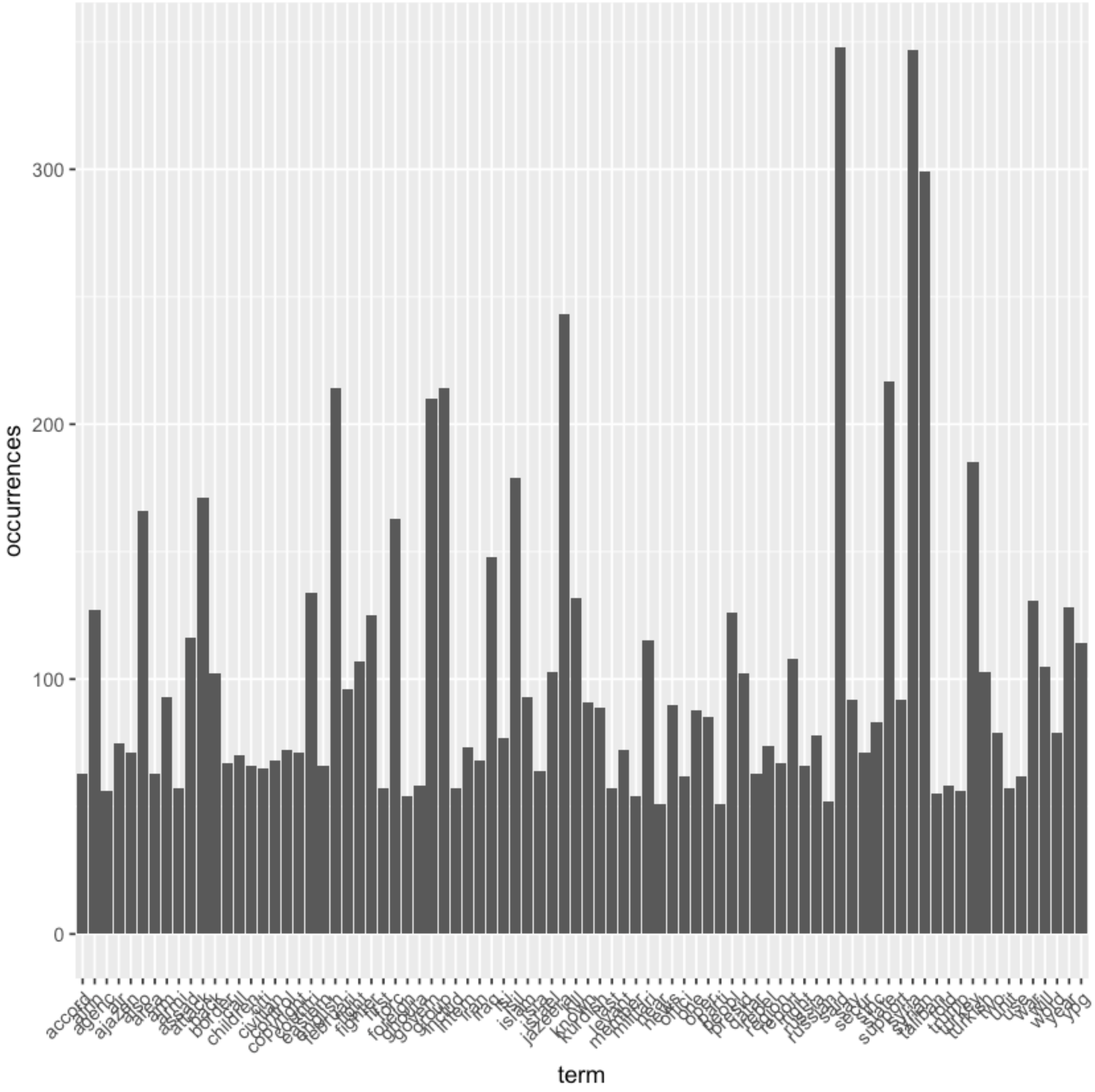
```
#correlations
findAssocs(dtm,"turkish",0.5)
#findAssocs(dtm,"children",0.5)
```

One needs to specify the DTM, the term of interest and the correlation limit. The latter is a number between 0 and 1 that serves as a lower bound for the strength of correlation between the search and result terms. For example, if the correlation limit is 1, findAssocs() will return only those words that always co-occur with the search term. A correlation limit of 0.5 will return terms that have a search term co-occurrence of at least 50% and so on.

### Visualizations

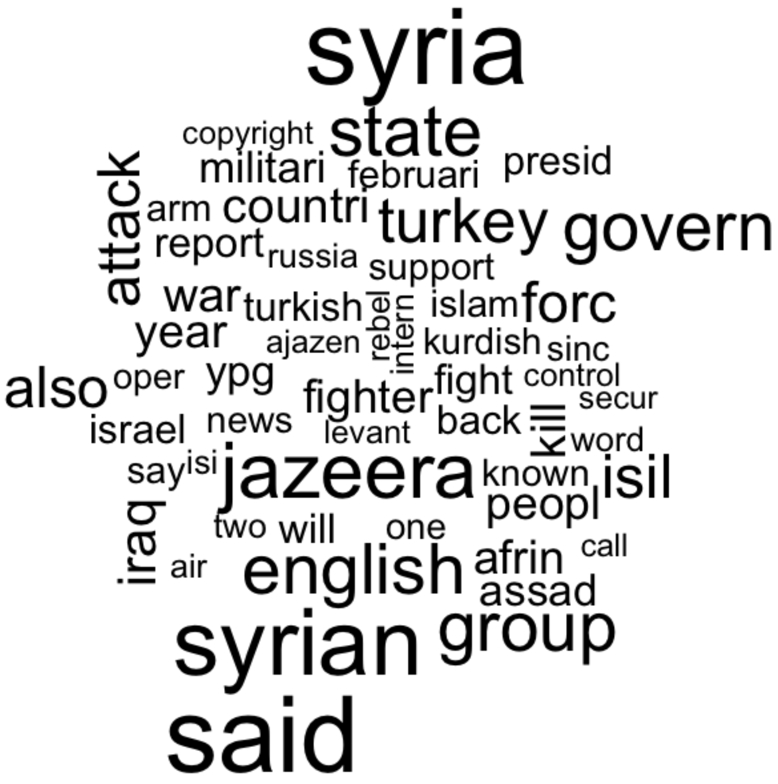
In [20]:

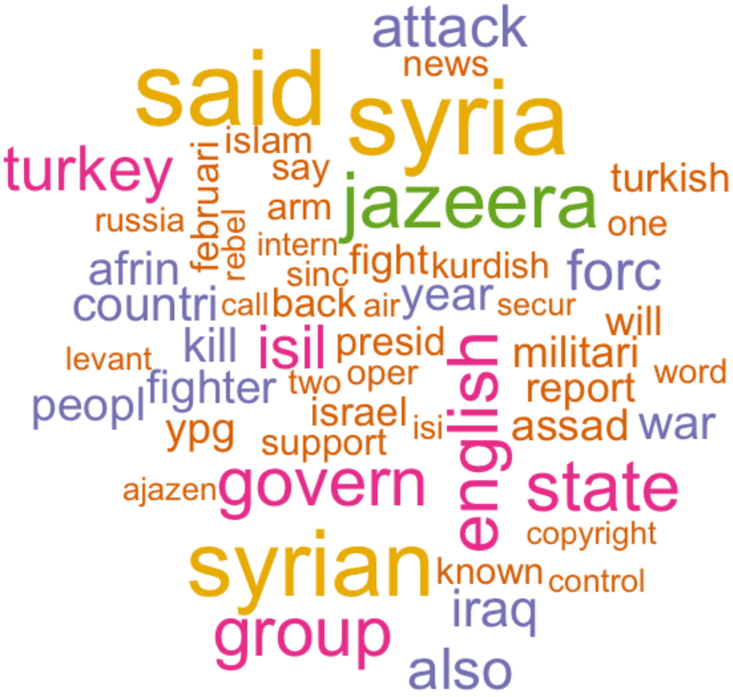
```
#Basic graphics
#histogram
wf=data.frame(term=names(freq),occurrences=freq)

p <- ggplot(subset(wf, freq>50), aes(term, occurrences))
p <- p + geom_bar(stat="identity")
p <- p + theme(axis.text.x=element_text(angle=45, hjust=1))
p
```



In [21]:

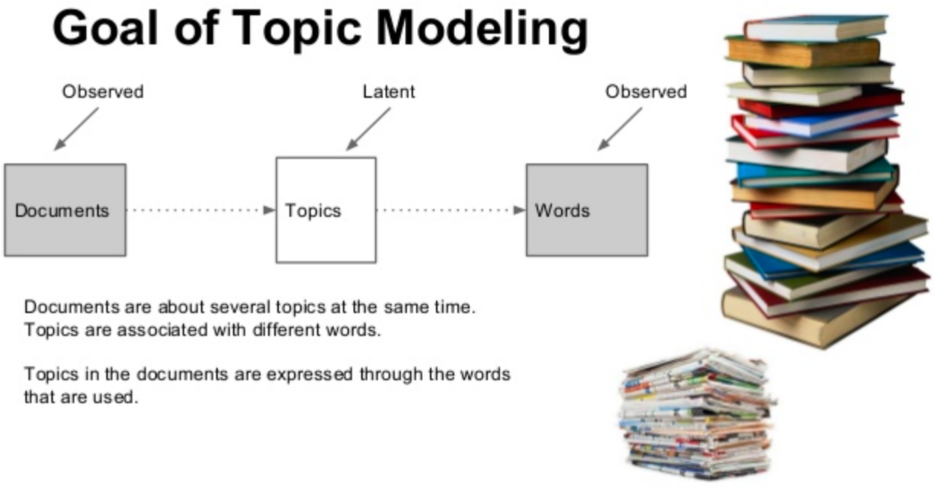
```
#wordcloud
#setting the same seed each time ensures consistent look acros
s clouds
set.seed(42)
#limit words by specifying min frequency
wordcloud(names(freq),freq, min.freq=70)
#...add color
wordcloud(names(freq),freq,min.freq=70,colors=brewer.pal(6,"Da
rk2"))
```


syria
copyright
state
militari
februari
presid
attack
arm
countri
turkey
govern
report
russia
support
war
turkish
rebel
intern
islam
forc
year
ajazen
kurdish
sinc
also
oper
ypg
fight
control
fighter
secur
israel
news
levant
back
kill
word
say
isi
jazeera
known
isil
iraq
peopl
two
will
one
call
air
english
afrin
assad
syrian
group
said


attack
news
said
syria
islam
turkey
februari
say
turkish
jazeera
arm
russia
rebel
one
intern
fight
kurdish
afrin
sinc
forc
countri
call
back
air
year
secur
will
kill
isil
presid
militari
levant
english
word
oper
two
fighter
report
peopl
israel
isi
assad
war
ypg
support
ajazen
govern
state
copyright
syrian
known
control
iraq
group
also

# What is Topic Modeling

- deals with the problem of automatically classifying sets of documents into themes



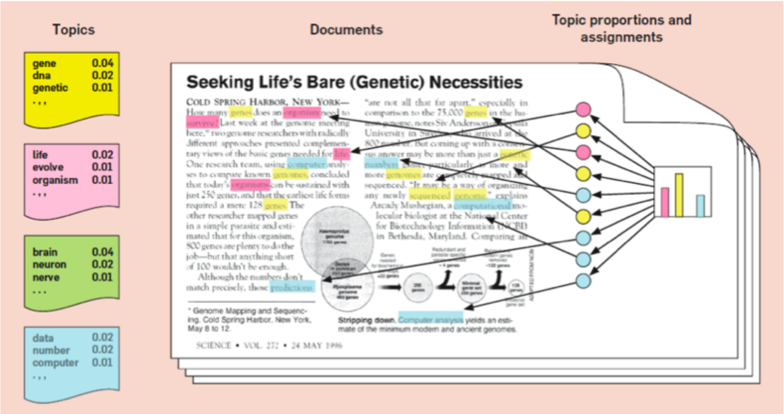
## Goal of Topic Modeling

Documents are about several topics at the same time.
Topics are associated with different words.

Topics in the documents are expressed through the words that are used.

# What is behind Topic Modeling?

- Latent Dirichlet Allocation (LDA) …
- … assumes that each of the documents in a collection consist of a mixture of collection-wide topics
- in reality we observe only documents and words, not topics – the latter are part of the hidden (or latent) structure of documents
- goal is to infer the latent topic structure given the words and document - LDA does this by recreating the documents in the corpus by adjusting the relative importance of topics in documents and words in topics iteratively



In [22]:

```
#Topic modeling
#Set parameters for Gibbs sampling
burnin <- 4000
iter <- 2000
thin <- 500
seed <-list(2003,5,63,100001,765)
nstart <- 5
best <- TRUE


#Number of topics
k <- 5

#Run LDA using Gibbs sampling
ldaOut <-LDA(dtm,k, method="Gibbs", control=list(nstart=nstart
, seed = seed, best=best, burnin = burnin, iter = iter, thin=t
hin))
```

In [23]:

```
#have a look at the model and some output
ldaOut
topics(ldaOut)
as.matrix(terms(ldaOut,6))

#write out results
#docs to topics
ldaOut.topics <- as.matrix(topics(ldaOut))
write.csv(ldaOut.topics,file=paste("LDAGibbs",k,"DocsToTopics.csv"))

#top 6 terms in each topic
ldaOut.terms <- as.matrix(terms(ldaOut,6))
write.csv(ldaOut.terms,file=paste("LDAGibbs",k,"TopicsToTerms.csv"))
```

```
A LDA_Gibbs topic model with 5 topics.
```

**1**

4

**2**

4

**3**

2

**4**

3

**5**

5

**6**

4

**7**

5

**8**

3

**9**

5

**10**

4

**11**

5

**12**

3

**13**

1

**14**

3

**15**

4

**16**

5

**17**

2

**18**

5

**19**

3

**20**

3

**21**

4

**22**

5

**23**

3

**24**

4

**25**

4

**26**

4

**27**

2

**28**

2

**29**

3

**30**

2

**31**

2

**32**

4

**33**

2

**34**

1

**35**

4

**36**

5

**37**

2

**38**

3

**39**

4

**40**

2

**41**

5

**42**

4

**43**

2

**44**

2

**45**

2

**46**

3

**47**

2

**48**

3

**49**

4

**50**

4

**51**

2

**52**

3

**53**

4

**54**

4

**55**

4

**56**

3

**57**

4

**58**

1

**59**

4

**60**

3

**61**

5

**62**

1

**63**

1

**64**

3

**65**

2

**66**

4

**67**

4

**68**

3

**69**

5

**70**

1

**71**

2

| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
|---|---|---|---|---|
| presid | iraq | said | turkey | syrian |
| state | english | attack | syria | syria |
| right | jazeera | kill | afrin | govern |
| trump | isil | jazeera | said | group |
| two | countri | year | ypg | war |
| israel | qatar | children | forc | assad |