Online learning analytics on social networking sites: how to tap the potential of data mining in research of educational technology

Qiang (Neo) Hao

Robert Maribe Branch

Questions to Answer by Text Mining in Education

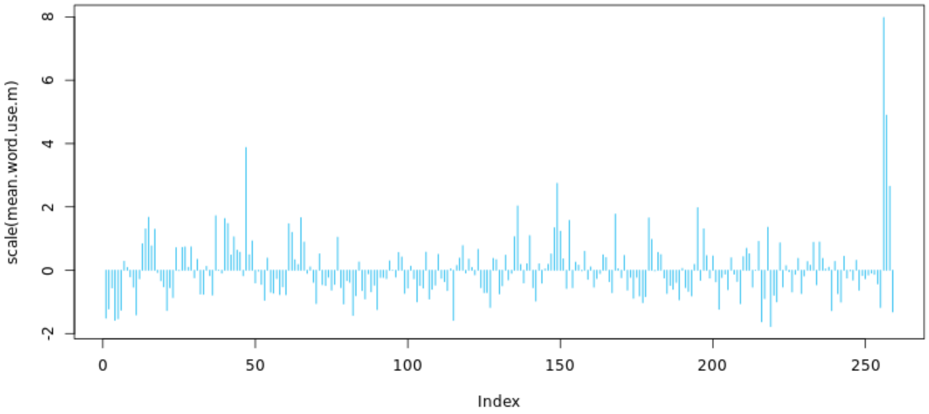
- What algorithm can score essays as teachers do?

Questions to Answer by Text Mining in Education

- What courses should we recommend students' based on their course reviews and engagement levels of their enrolled courses?

Questions to Answer by Text Mining in Education

- Does the treatment improve students' lexical variety in their writing?

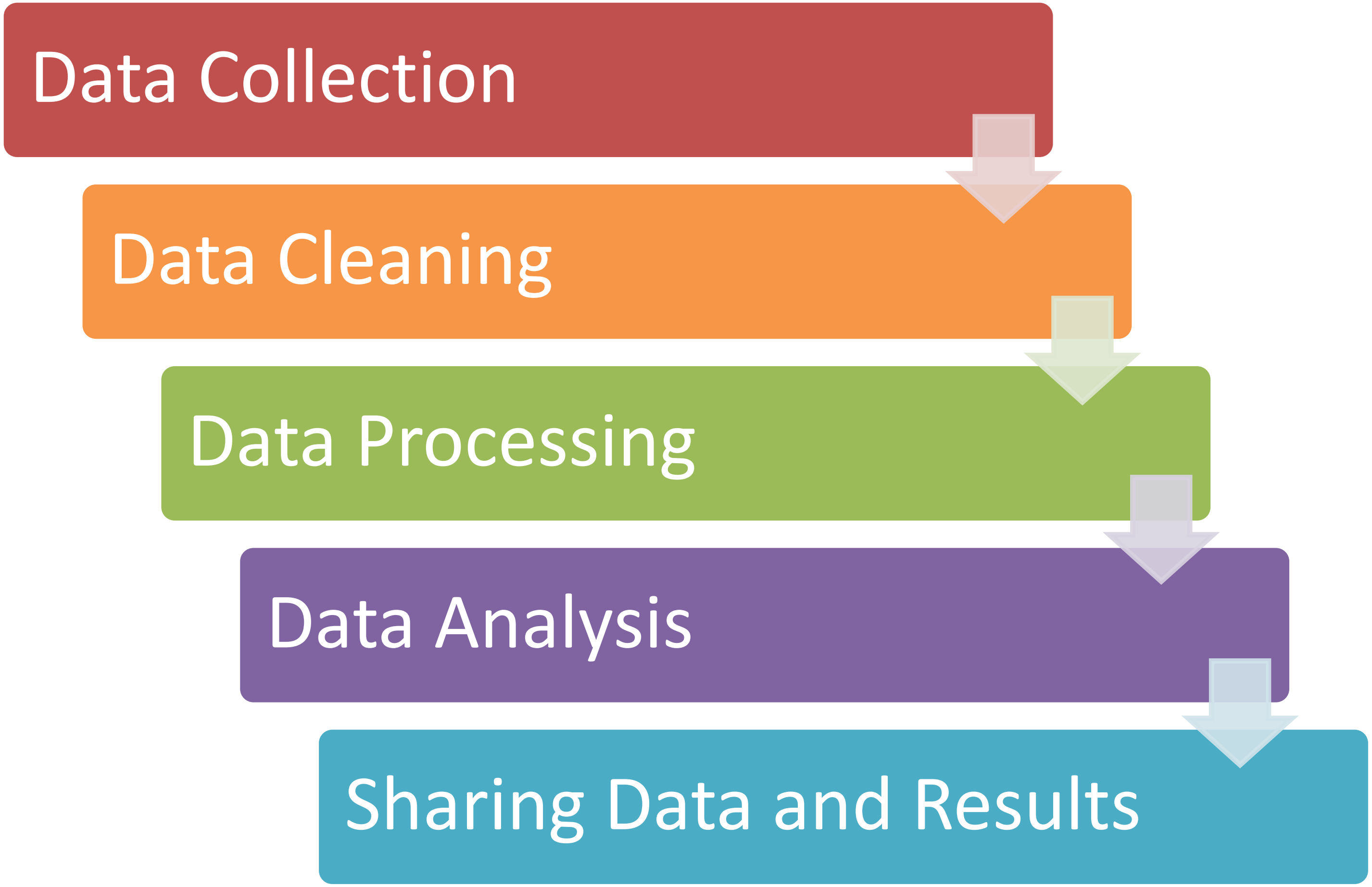


Questions to Answer by Text Mining in Education

- Are there different patterns in students' discussions; if so, are the patterns related to their academic performance?

stackoverflow

Research Pipeline



Data Collection

Data Cleaning

Data Processing

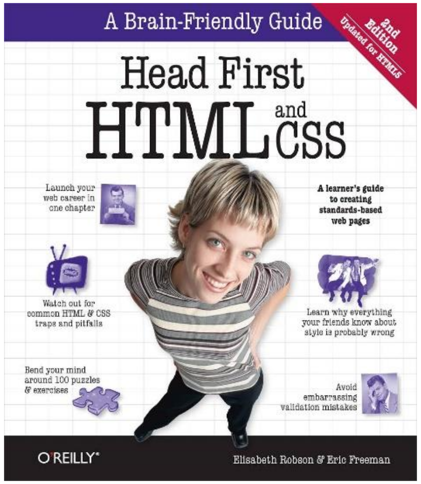
Data Analysis

Sharing Data and Results

Data Collection

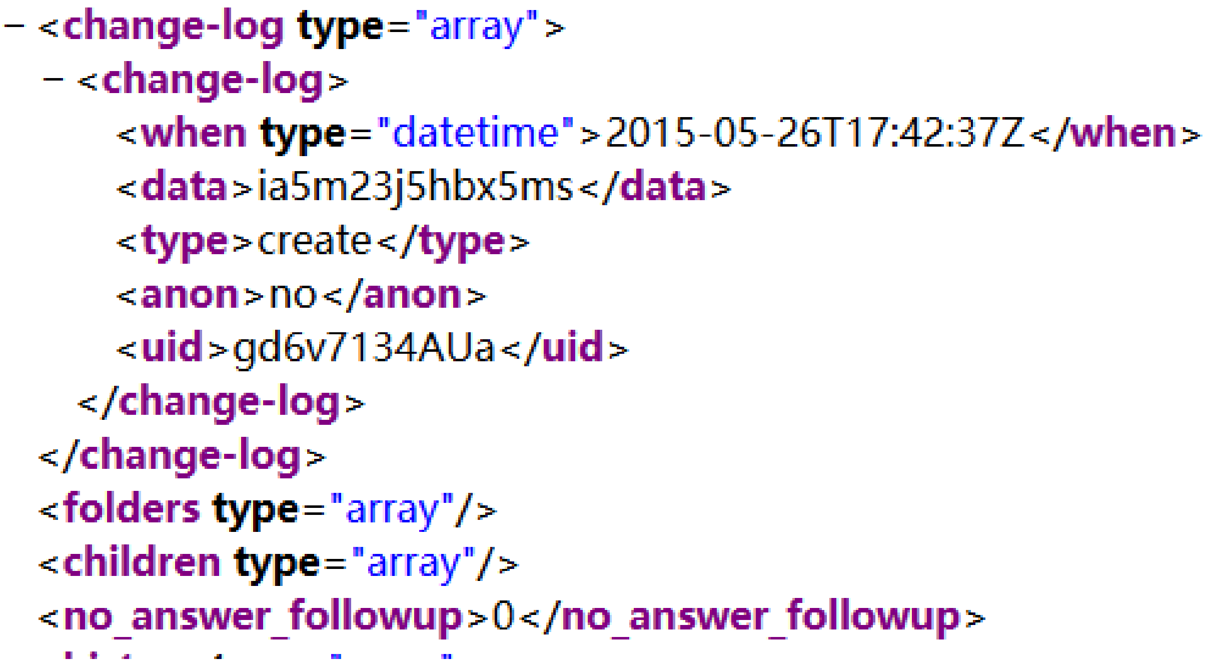
Scrapping data form static web pages:

1. **A good understanding of HTML & CSS**
2. **A good understanding of XML & JSON**


A Brain-Friendly Guide
2nd Edition
Updated for HTML5
Head First
HTML and CSS
Launch your web career in one chapter
A learner's guide to creating standards-based web pages
Watch out for common HTML & CSS traps and pitfalls
Learn why everything your friends know about style is probably wrong
Bend your mind around 100 puzzles & exercises
Avoid embarrassing validation mistakes
O'REILLY
Elisabeth Robson & Eric Freeman

Data Collection

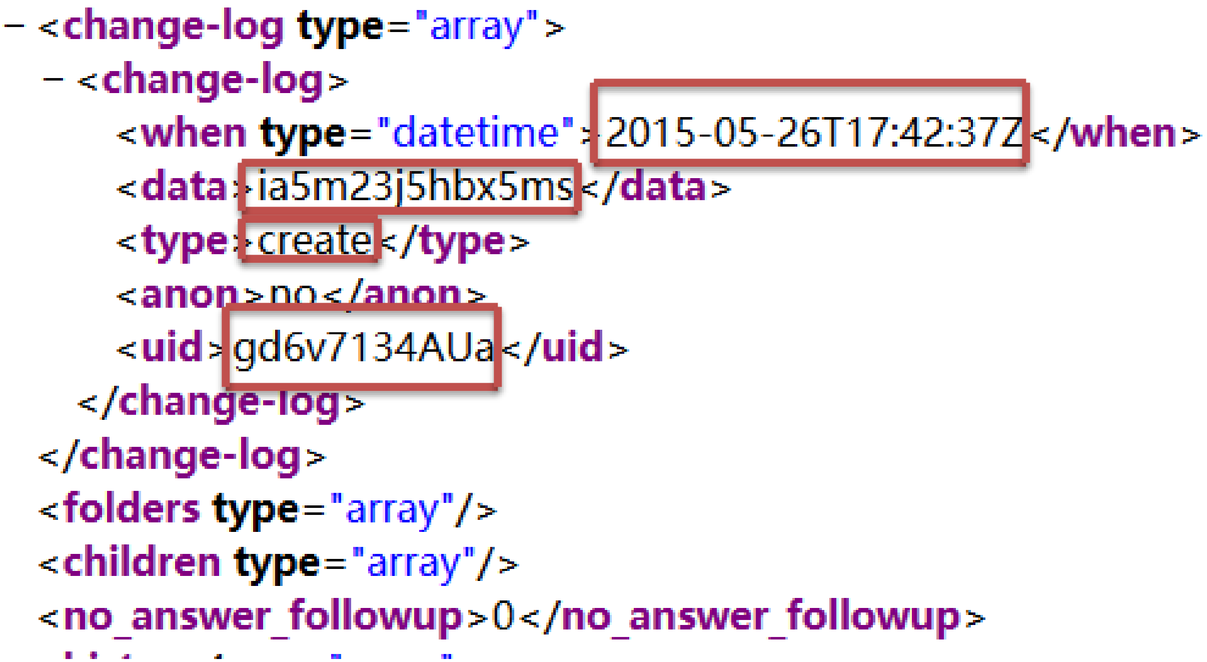
- **XML**



```
- <change-log type="array">
  - <change-log>
      <when type="datetime">2015-05-26T17:42:37Z</when>
      <data>ia5m23j5hbx5ms</data>
      <type>create</type>
      <anon>no</anon>
      <uid>gd6v7134AUa</uid>
    </change-log>
  </change-log>
  <folders type="array"/>
  <children type="array"/>
  <no_answer_followup>0</no_answer_followup>
```

Data Collection

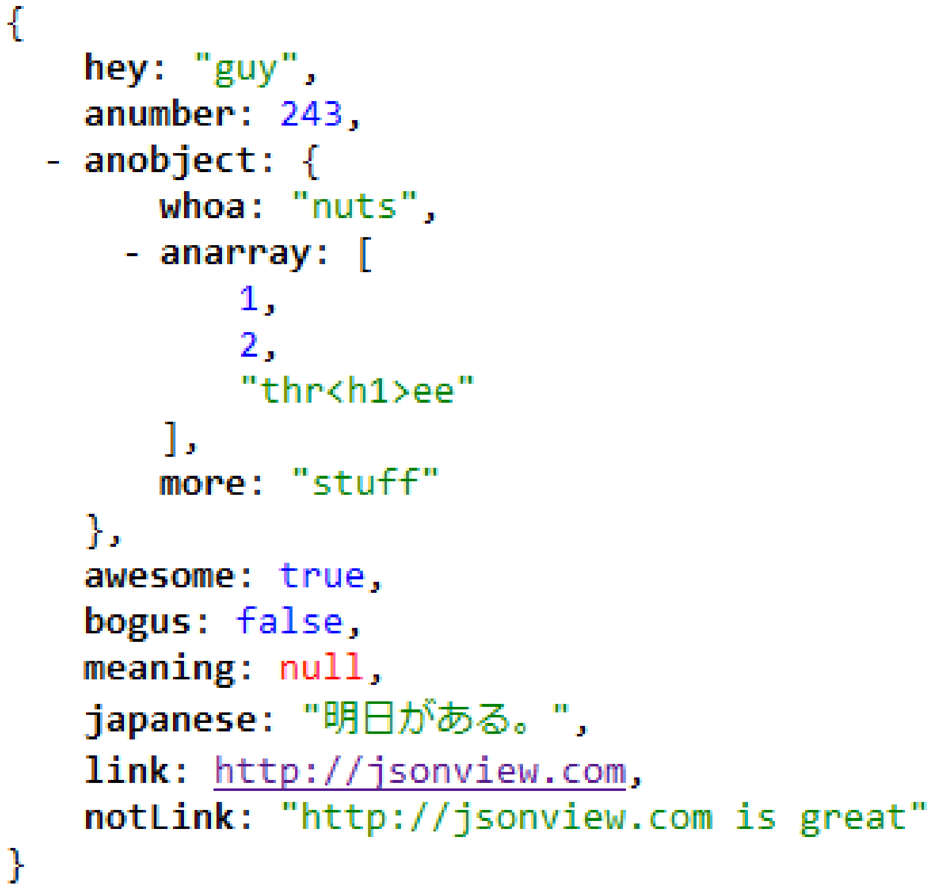
- **XML**



```
- <change-log type="array">
  - <change-log>
      <when type="datetime">2015-05-26T17:42:37Z</when>
      <data>ia5m23j5hbx5ms</data>
      <type>create</type>
      <anon>no</anon>
      <uid>gd6v7134AUa</uid>
    </change-log>
  </change-log>
  <folders type="array"/>
  <children type="array"/>
  <no_answer_followup>0</no_answer_followup>
```

Data Collection

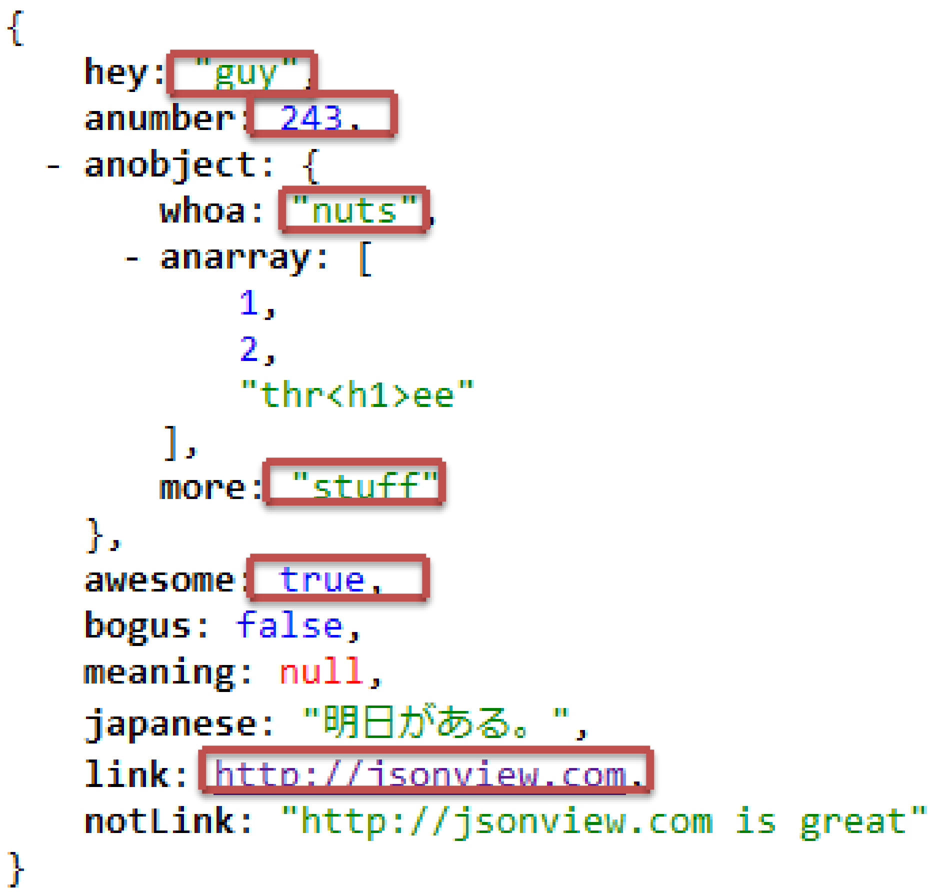
- **JSON**



```
{
    hey: "guy",
    anumber: 243,
  - anobject: {
        whoa: "nuts",
      - anarray: [
            1,
            2,
            "thr<h1>ee"
        ],
        more: "stuff"
    },
    awesome: true,
    bogus: false,
    meaning: null,
    japanese: "明日がある。",
    link: http://jsonview.com,
    notLink: "http://jsonview.com is great"
}
```

Data Collection

- **JSON**

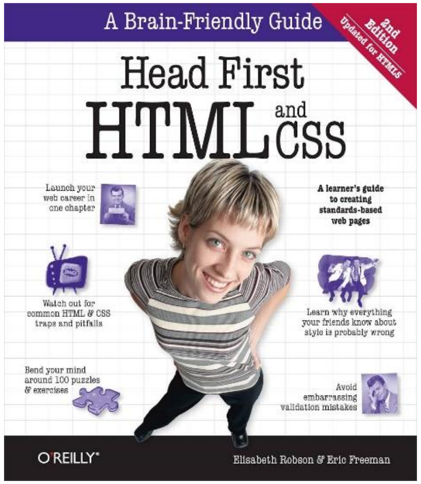


```
{
    hey: "guy",
    anumber: 243,
  - anobject: {
        whoa: "nuts",
      - anarray: [
            1,
            2,
            "thr<h1>ee"
        ],
        more: "stuff"
    },
    awesome: true,
    bogus: false,
    meaning: null,
    japanese: "明日がある。",
    link: http://jsonview.com,
    notLink: "http://jsonview.com is great"
}
```

Data Collection

Scrapping data form static web pages:

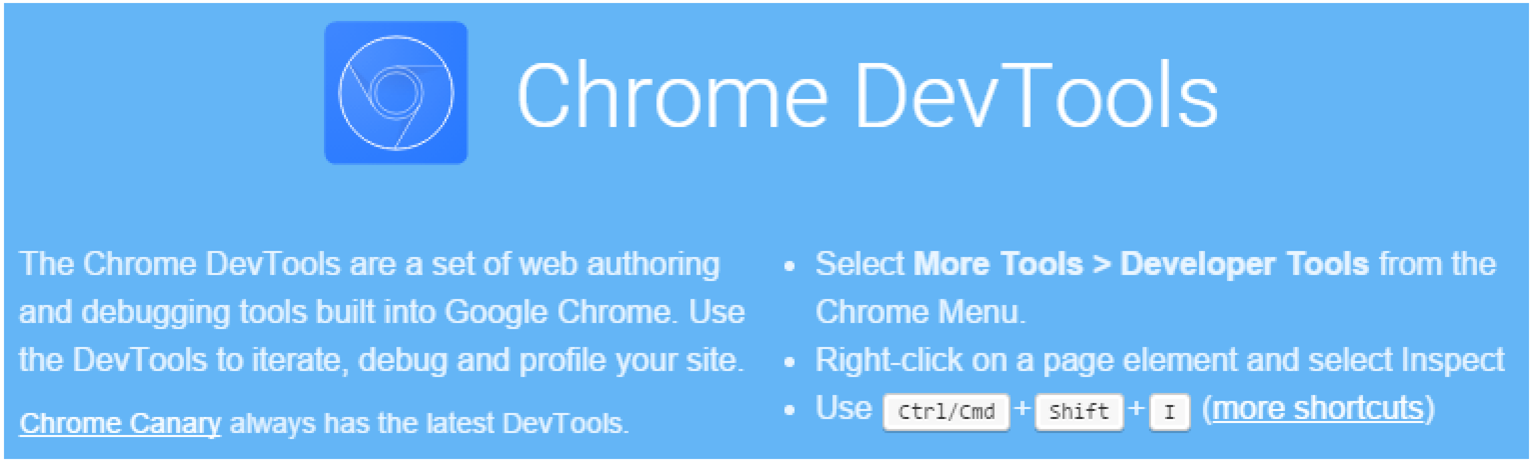
1. **A good understanding of HTML & CSS**
2. **A good understanding of XML & JSON**


A Brain-Friendly Guide
2nd Edition
Updated for HTML5
Head First
HTML and CSS
Launch your web career in one chapter
A learner's guide to creating standards-based web pages
Watch out for common HTML & CSS traps and pitfalls
Learn why everything your friends know about style is probably wrong
Bend your mind around 100 puzzles & exercises
Avoid embarrassing validation mistakes
O'REILLY
Elisabeth Robson & Eric Freeman

Data Collection

Scrapping data form static web pages:

1. **A good understanding of HTML & CSS**
2. **A good understanding of XML & JSON**
3. **Familiar with Development Tools of Browsers**



Chrome DevTools

The Chrome DevTools are a set of web authoring and debugging tools built into Google Chrome. Use the DevTools to iterate, debug and profile your site.

Chrome Canary always has the latest DevTools.

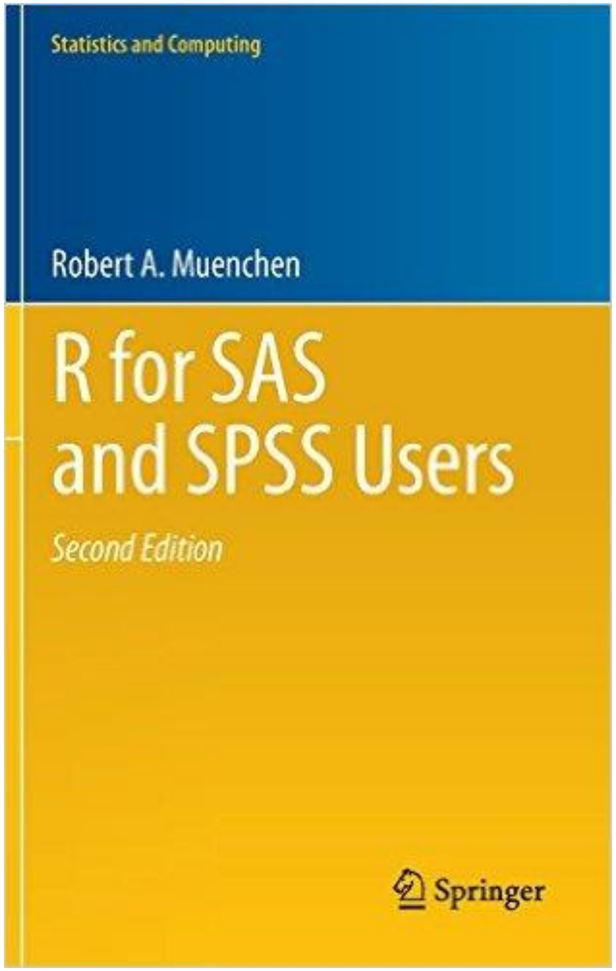
- Select **More Tools > Developer Tools** from the Chrome Menu.
- Right-click on a page element and select Inspect
- Use `Ctrl/Cmd` + `Shift` + `I` (more shortcuts)

Data Collection

Scrapping data form static web pages:

1. **A good understanding of HTML & CSS**
2. **A good understanding of XML & JSON**
3. **Familiar with Development Tools of Browsers**
4. **Familiar with R and package “XML”**

Data Collection

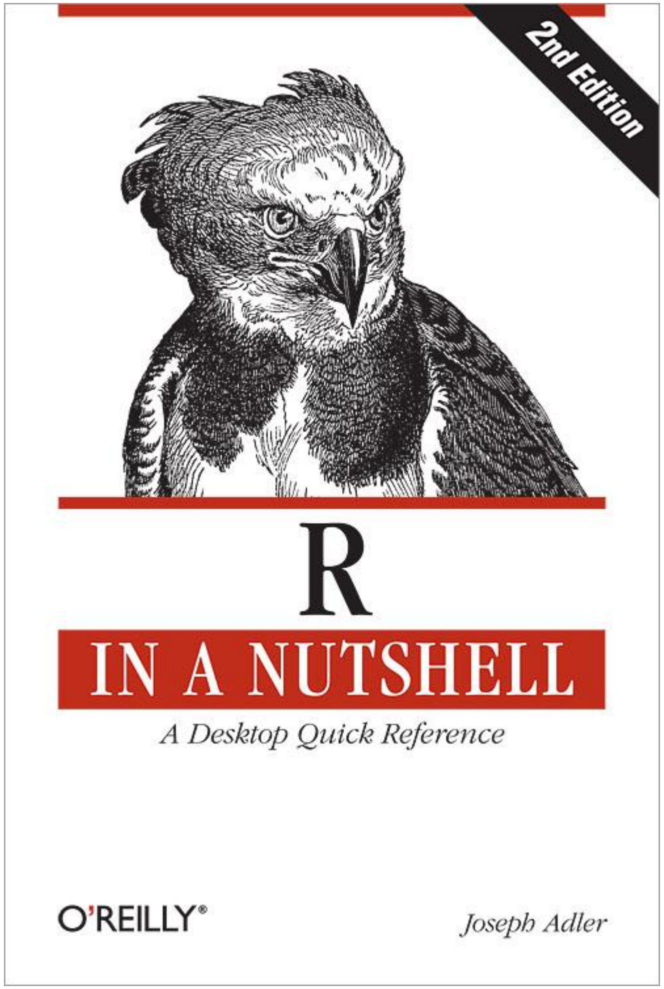


R for SAS and SPSS Users

Google *r xml package filetype:pdf*

Data Collection

R in a Nutshell



Data Collection

```
getwd()

setwd(“C:/Users/John/Analysis”)
setwd(“/home/Analysis”)
```

`setwd(“XXX/TwitterHashtagR/data”)`

Data Collection

```
install.packages(“XML”)
library(XML)
```

Data Collection

```
a <- 3
b <- c(1, 3, 7, 8)

c <- “Hello”
d <- c(“piggy1”, “piggy2”,
      “piggy3”)
```

Data Collection

Function

```
tweetCollectByUser <-
   function(username, numberOfTweets,
   nameOfFile) {
        ……
}

tweetCollectByUser( “aect”, 300,
“tweetsOfAect” )
```

Data Collection

Authentication

1. Register your own app
2. Keep your consumer keys and secrets
3. Go to *Data Collection/Authentication.R*
4. Replace consumer keys and secrets with yours
5. Run lines 1-42

Data Collection

Collect User Info

1. Go to *Data Collection/collectUsers.R*
2. Run lines 1-33
3. Practice: Find 5 twitter accounts that you would like to collect information about, and collect their basic information in a .csv file

Data Collection

Collect tweets of particular users

1. **Go to** ***Data Collection/getTweetsByUser.R***
2. **Run lines 1-24**

Data Collection

Collect tweets of particular users

1. **Go to *Data Collection/getTweetsByAllUser.R***
2. **Run lines 1-68**
3. **Practice: Get tweets from 2 different twitter accounts**

Data Collection

Collect tweets by Hashtag

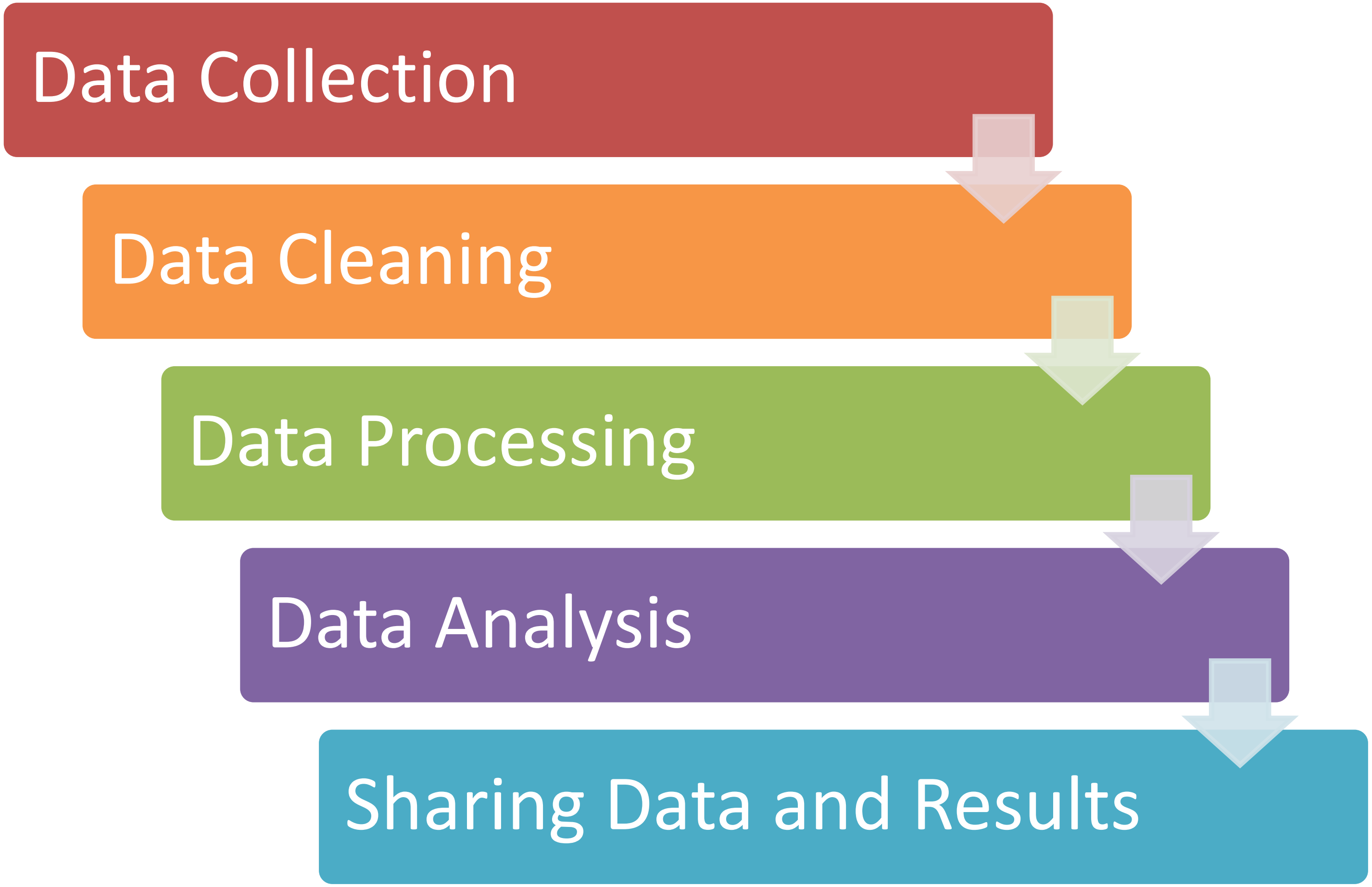
1. Go to *Data Collection/hashtagSearch.R*
2. Run lines 1-22
3. Practice: Get tweets with one hashtag you like

Data Collection

Collect tweets by Web Scrapping

1. Go to *Data Collection/parse_Tweets.R*
2. Run lines 1-34, 76-77
3. Practice: Do one web scrapping yourself
    1. *Search a hashtag using Twitter; keep scrolling down until you have all or enough number of tweets*
    2. *Download the HTML page*
    3. *……*

Research Pipeline



Data Collection

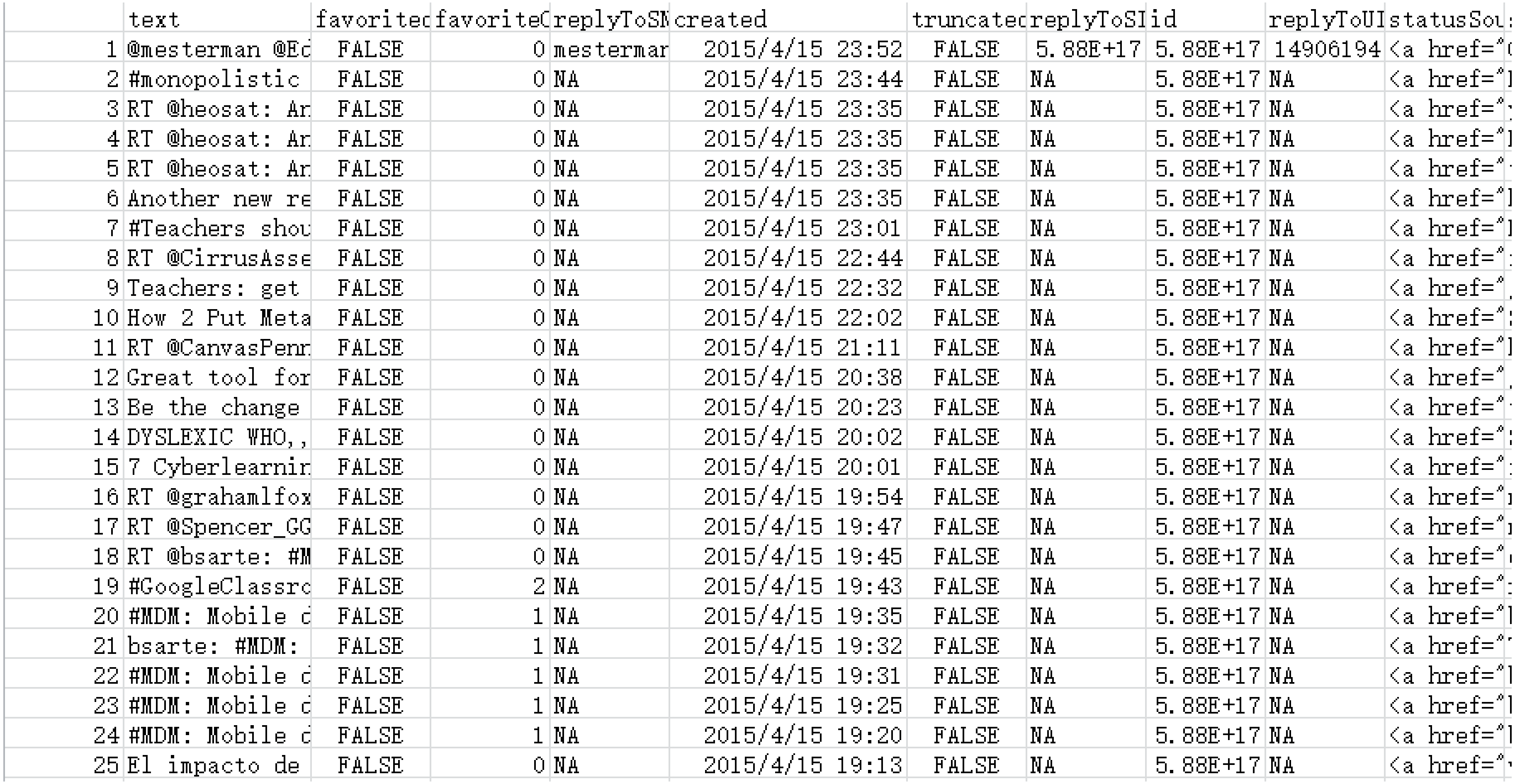
Data Cleaning

Data Processing

Data Analysis

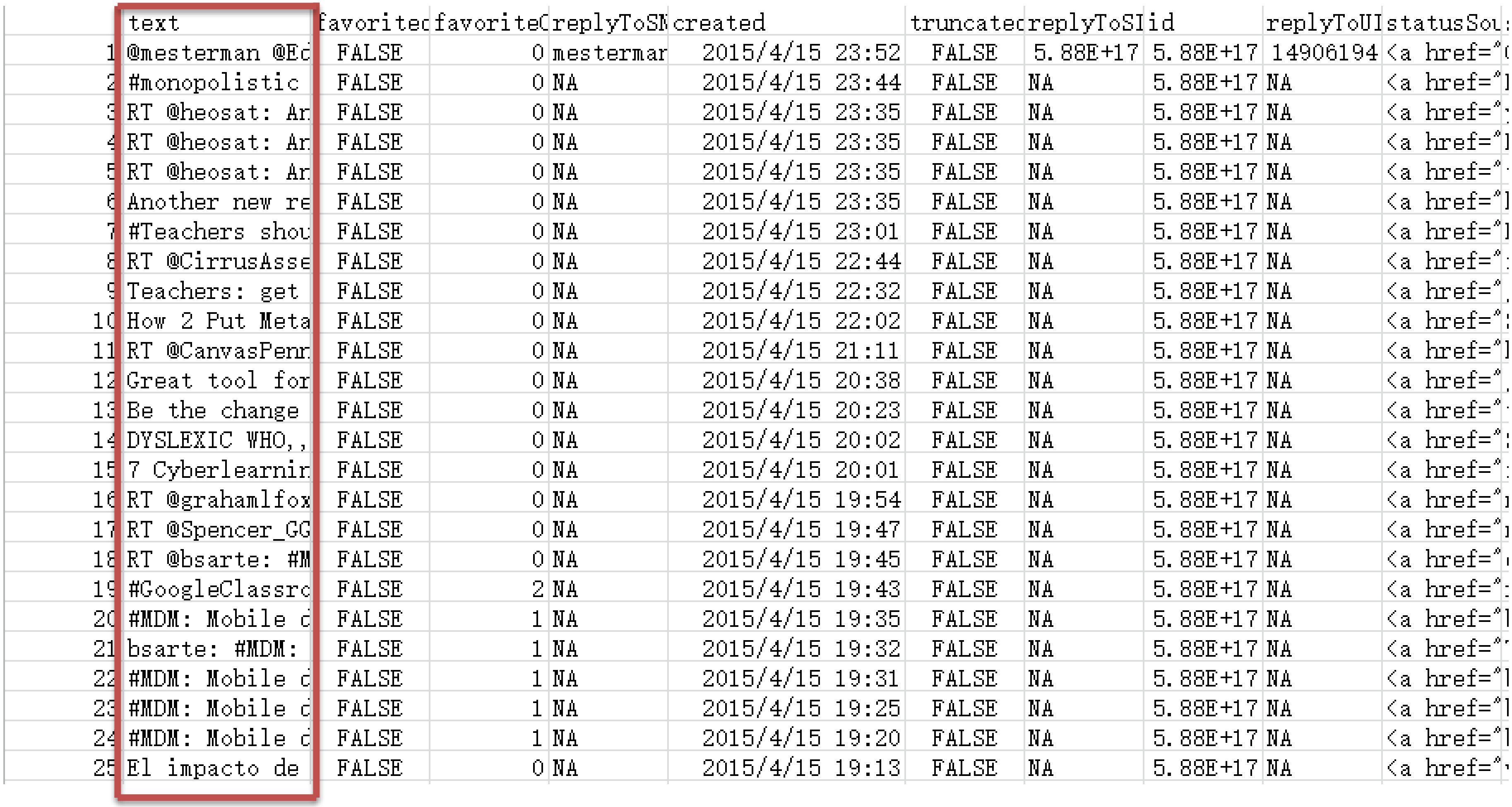
Sharing Data and Results

Data Cleaning



	text	favoritec	favoriteC	replyToSN	created	truncatec	replyToSI	id	replyToUI	statusSou
1	@mesterman @Ed	FALSE	0	mestermar	2015/4/15 23:52	FALSE	5.88E+17	5.88E+17	14906194	<a href="
2	#monopolistic	FALSE	0	NA	2015/4/15 23:44	FALSE	NA	5.88E+17	NA	<a href="
3	RT @heosat: Ar	FALSE	0	NA	2015/4/15 23:35	FALSE	NA	5.88E+17	NA	<a href="
4	RT @heosat: Ar	FALSE	0	NA	2015/4/15 23:35	FALSE	NA	5.88E+17	NA	<a href="
5	RT @heosat: Ar	FALSE	0	NA	2015/4/15 23:35	FALSE	NA	5.88E+17	NA	<a href="
6	Another new re	FALSE	0	NA	2015/4/15 23:35	FALSE	NA	5.88E+17	NA	<a href="
7	#Teachers shou	FALSE	0	NA	2015/4/15 23:01	FALSE	NA	5.88E+17	NA	<a href="
8	RT @CirrusAsse	FALSE	0	NA	2015/4/15 22:44	FALSE	NA	5.88E+17	NA	<a href="
9	Teachers: get	FALSE	0	NA	2015/4/15 22:32	FALSE	NA	5.88E+17	NA	<a href="
10	How 2 Put Meta	FALSE	0	NA	2015/4/15 22:02	FALSE	NA	5.88E+17	NA	<a href="
11	RT @CanvasPenn	FALSE	0	NA	2015/4/15 21:11	FALSE	NA	5.88E+17	NA	<a href="
12	Great tool for	FALSE	0	NA	2015/4/15 20:38	FALSE	NA	5.88E+17	NA	<a href="
13	Be the change	FALSE	0	NA	2015/4/15 20:23	FALSE	NA	5.88E+17	NA	<a href="
14	DYSLEXIC WHO,,	FALSE	0	NA	2015/4/15 20:02	FALSE	NA	5.88E+17	NA	<a href="
15	7 Cyberlearnin	FALSE	0	NA	2015/4/15 20:01	FALSE	NA	5.88E+17	NA	<a href="
16	RT @grahamlfox	FALSE	0	NA	2015/4/15 19:54	FALSE	NA	5.88E+17	NA	<a href="
17	RT @Spencer_GG	FALSE	0	NA	2015/4/15 19:47	FALSE	NA	5.88E+17	NA	<a href="
18	RT @bsarte: #M	FALSE	0	NA	2015/4/15 19:45	FALSE	NA	5.88E+17	NA	<a href="
19	#GoogleClassro	FALSE	2	NA	2015/4/15 19:43	FALSE	NA	5.88E+17	NA	<a href="
20	#MDM: Mobile d	FALSE	1	NA	2015/4/15 19:35	FALSE	NA	5.88E+17	NA	<a href="
21	bsarte: #MDM:	FALSE	1	NA	2015/4/15 19:32	FALSE	NA	5.88E+17	NA	<a href="
22	#MDM: Mobile d	FALSE	1	NA	2015/4/15 19:31	FALSE	NA	5.88E+17	NA	<a href="
23	#MDM: Mobile d	FALSE	1	NA	2015/4/15 19:25	FALSE	NA	5.88E+17	NA	<a href="
24	#MDM: Mobile d	FALSE	1	NA	2015/4/15 19:20	FALSE	NA	5.88E+17	NA	<a href="
25	El impacto de	FALSE	0	NA	2015/4/15 19:13	FALSE	NA	5.88E+17	NA	<a href="

Data Cleaning



	text	favoritec	favoriteC	replyToSN	created	truncatec	replyToSI	id	replyToUI	statusSou
1	@mesterman @Ed	FALSE	0	mestermar	2015/4/15 23:52	FALSE	5.88E+17	5.88E+17	14906194	<a href="
2	#monopolistic	FALSE	0	NA	2015/4/15 23:44	FALSE	NA	5.88E+17	NA	<a href="
3	RT @heosat: Ar	FALSE	0	NA	2015/4/15 23:35	FALSE	NA	5.88E+17	NA	<a href="
4	RT @heosat: Ar	FALSE	0	NA	2015/4/15 23:35	FALSE	NA	5.88E+17	NA	<a href="
5	RT @heosat: Ar	FALSE	0	NA	2015/4/15 23:35	FALSE	NA	5.88E+17	NA	<a href="
6	Another new re	FALSE	0	NA	2015/4/15 23:35	FALSE	NA	5.88E+17	NA	<a href="
7	#Teachers shou	FALSE	0	NA	2015/4/15 23:01	FALSE	NA	5.88E+17	NA	<a href="
8	RT @CirrusAsse	FALSE	0	NA	2015/4/15 22:44	FALSE	NA	5.88E+17	NA	<a href="
9	Teachers: get	FALSE	0	NA	2015/4/15 22:32	FALSE	NA	5.88E+17	NA	<a href="
10	How 2 Put Meta	FALSE	0	NA	2015/4/15 22:02	FALSE	NA	5.88E+17	NA	<a href="
11	RT @CanvasPenn	FALSE	0	NA	2015/4/15 21:11	FALSE	NA	5.88E+17	NA	<a href="
12	Great tool for	FALSE	0	NA	2015/4/15 20:38	FALSE	NA	5.88E+17	NA	<a href="
13	Be the change	FALSE	0	NA	2015/4/15 20:23	FALSE	NA	5.88E+17	NA	<a href="
14	DYSLEXIC WHO,,	FALSE	0	NA	2015/4/15 20:02	FALSE	NA	5.88E+17	NA	<a href="
15	7 Cyberlearnir	FALSE	0	NA	2015/4/15 20:01	FALSE	NA	5.88E+17	NA	<a href="
16	RT @grahamlfox	FALSE	0	NA	2015/4/15 19:54	FALSE	NA	5.88E+17	NA	<a href="
17	RT @Spencer_GG	FALSE	0	NA	2015/4/15 19:47	FALSE	NA	5.88E+17	NA	<a href="
18	RT @bsarte: #M	FALSE	0	NA	2015/4/15 19:45	FALSE	NA	5.88E+17	NA	<a href="
19	#GoogleClassrc	FALSE	2	NA	2015/4/15 19:43	FALSE	NA	5.88E+17	NA	<a href="
20	#MDM: Mobile d	FALSE	1	NA	2015/4/15 19:35	FALSE	NA	5.88E+17	NA	<a href="
21	bsarte: #MDM:	FALSE	1	NA	2015/4/15 19:32	FALSE	NA	5.88E+17	NA	<a href="
22	#MDM: Mobile d	FALSE	1	NA	2015/4/15 19:31	FALSE	NA	5.88E+17	NA	<a href="
23	#MDM: Mobile d	FALSE	1	NA	2015/4/15 19:25	FALSE	NA	5.88E+17	NA	<a href="
24	#MDM: Mobile d	FALSE	1	NA	2015/4/15 19:20	FALSE	NA	5.88E+17	NA	<a href="
25	El impacto de	FALSE	0	NA	2015/4/15 19:13	FALSE	NA	5.88E+17	NA	<a href="

Data Cleaning

Regular Expression

```
madam, baad, dad, gooffoog
```

Data Cleaning

Regular Expression

```
reg <- "([a-zA-Z0-9]+://)?([a-zA-Z0-
  9_]+:[a-zA-Z0-9_]+@)?([a-zA-Z0-9.-
  ]+\\.[A-Za-z]{2,4})(:[0-9]+)?(/.*)?«
```

Data Cleaning

Regular Expression

```
reg <- "([a-zA-Z0-
   9_]+:[a-zA-Z0-9_]+@)?([a-zA-Z0-9.-
   ]+\\.[A-Za-z]{2,4})(:[0-9]+)?(/.*)?«
```

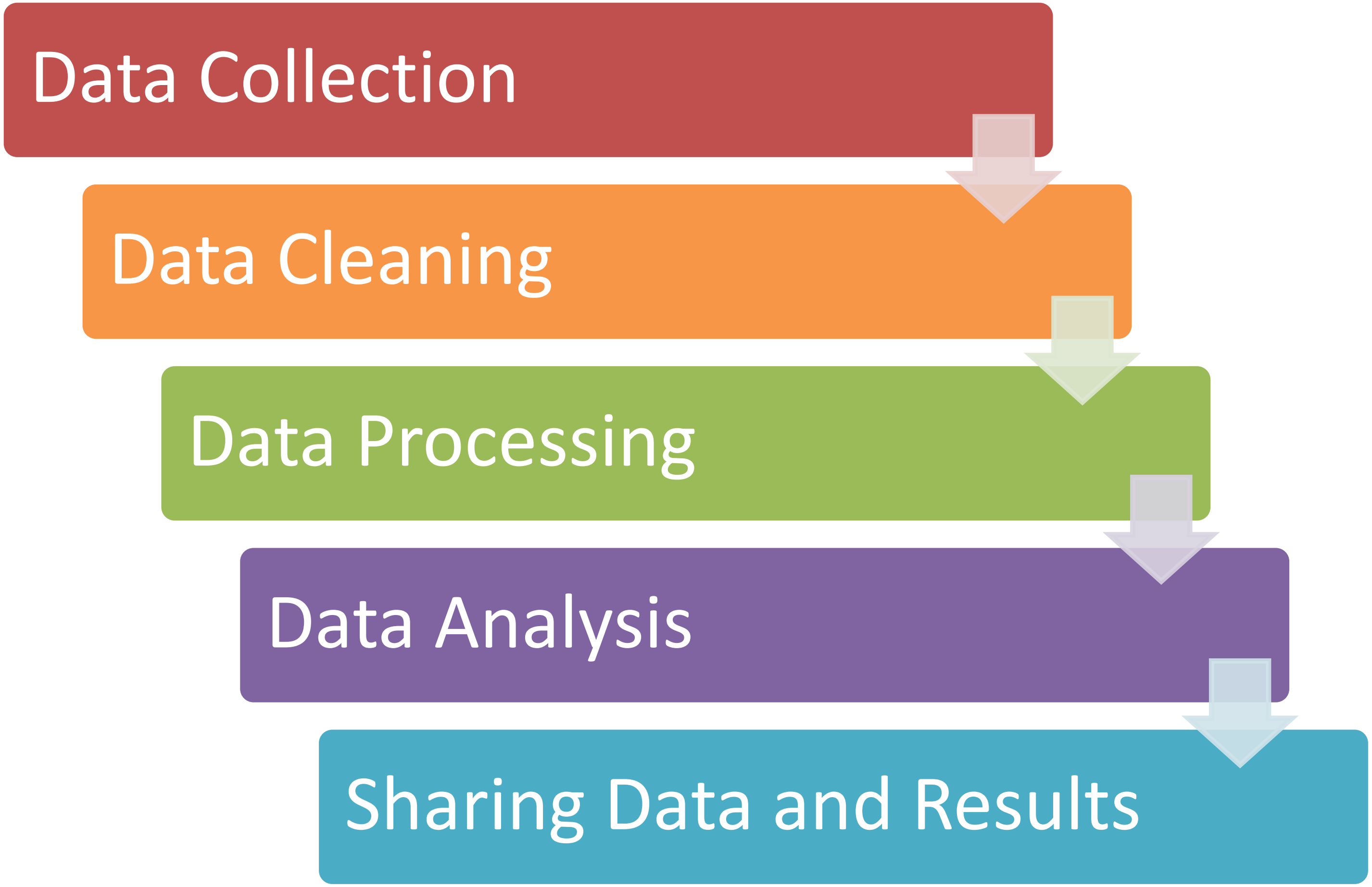
www.regular-expressions.info

Data Cleaning

Clean tweets

1. Go to *Data Cleaning/cleanData.R*
2. Run lines 1-57
3. Practice: Clean the diary data yourself

Research Pipeline



Data Collection

Data Cleaning

Data Processing

Data Analysis

Sharing Data and Results

Data Processing

Basic Procedures:

Data Processing

Basic Procedures:

1. **Remove punctuation**

Data Processing

Basic Procedures:

1. **Remove punctuation**
2. **Remove other non-characters**

Data Processing

Basic Procedures:

1. **Remove punctuation**
2. **Remove other non-characters**

!@#$%^&*()_+-~|\/<>

Data Processing

Basic Procedures:

1. **Remove punctuation**
2. **Remove other non-characters**
3. **Remove stop words**

Data Processing

Basic Procedures:

1. Remove punctuation
2. Remove other non-characters
3. Remove stop words

a, an, the, he, him, I, me, …

Data Processing

Basic Procedures:

1. Remove punctuation
2. Remove other non-characters
3. Remove stop words
4. Lowercases

Data Processing

Basic Procedures:

1. **Remove punctuation**
2. **Remove other non-characters**
3. **Remove stop words**
4. **Lowercases**
5. **Stem**

Data Processing

Basic Procedures:

1. **Remove punctuation**
2. **Remove other non-characters**
3. **Remove stop words**
4. **Lowercases**
5. **Stem**

do

does

did

Data Processing

Basic Procedures:

1. Remove punctuation
2. Remove other non-characters
3. Remove stop words
4. Lowercases
5. Stem

go

goes

went

Data Processing

Basic Procedures:

1. **Remove punctuation**
2. **Remove other non-characters**
3. **Remove stop words**
4. **Lowercases**
5. **Stem**

lie
lay
laid

Data Processing

Basic Procedures:

1. **Remove punctuation**
2. **Remove other non-characters**
3. **Remove stop words**
4. **Lowercases**
5. **Stem**

try

tries

tried

Data Processing

Assumption:

1. **Bag of words**

Data Processing

Assumption:

1. **Bag of words**

A dog bites a man.
A man bites a dog.

Data Processing

Assumption:

1. **Bag of words**

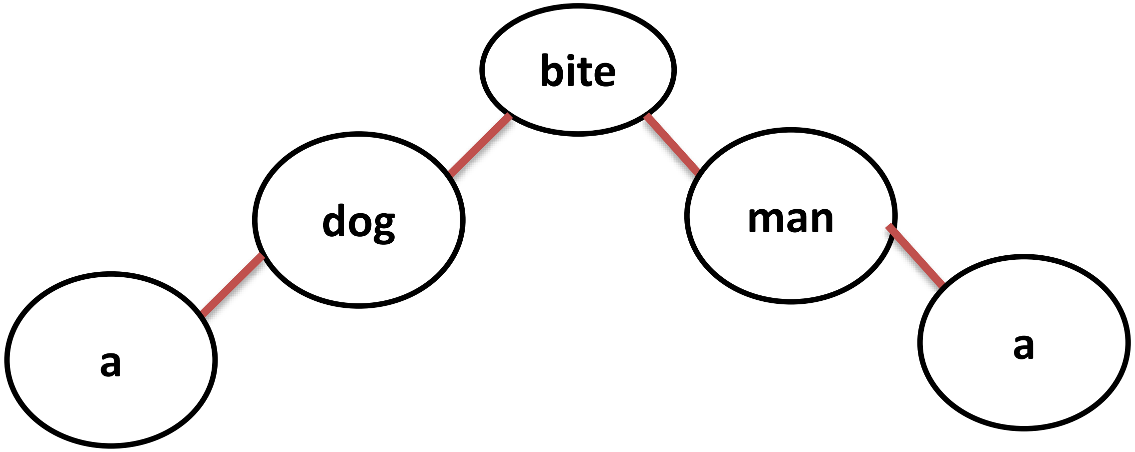
“a”, “man”, “dog”, “bites”

Data Processing

Assumption:

1. **Bag of words**

“a”, “man”, “dog”, “bites”



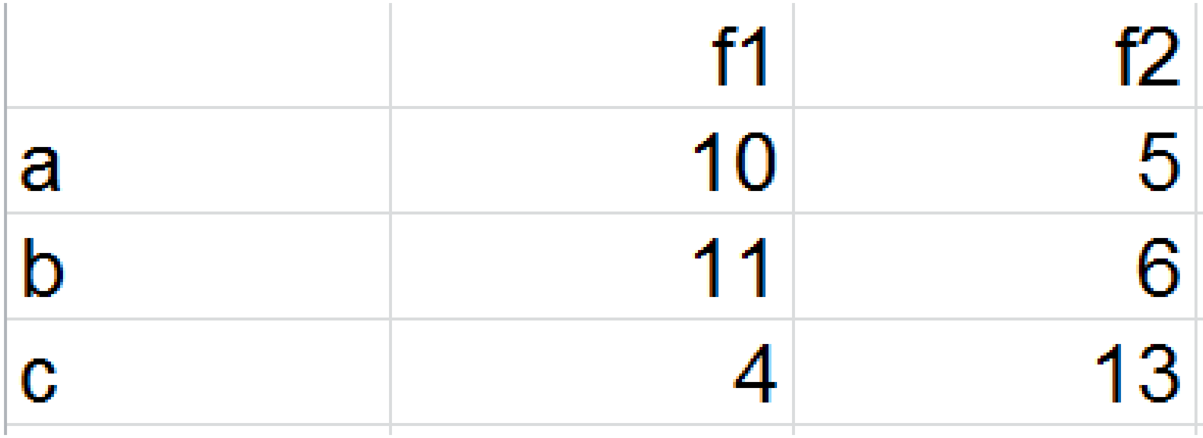
Data Processing

Assumption:

1. **Bag of words**
2. **Words as features**

Data Processing

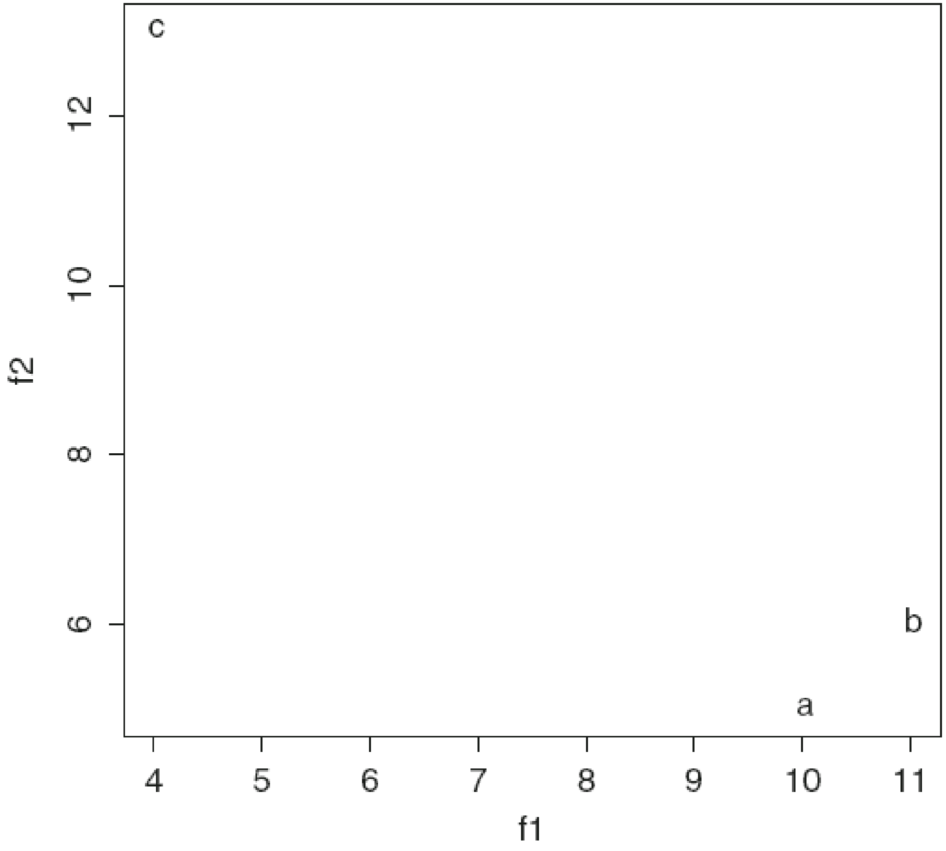
Explanation:



	f1	f2
a	10	5
b	11	6
c	4	13

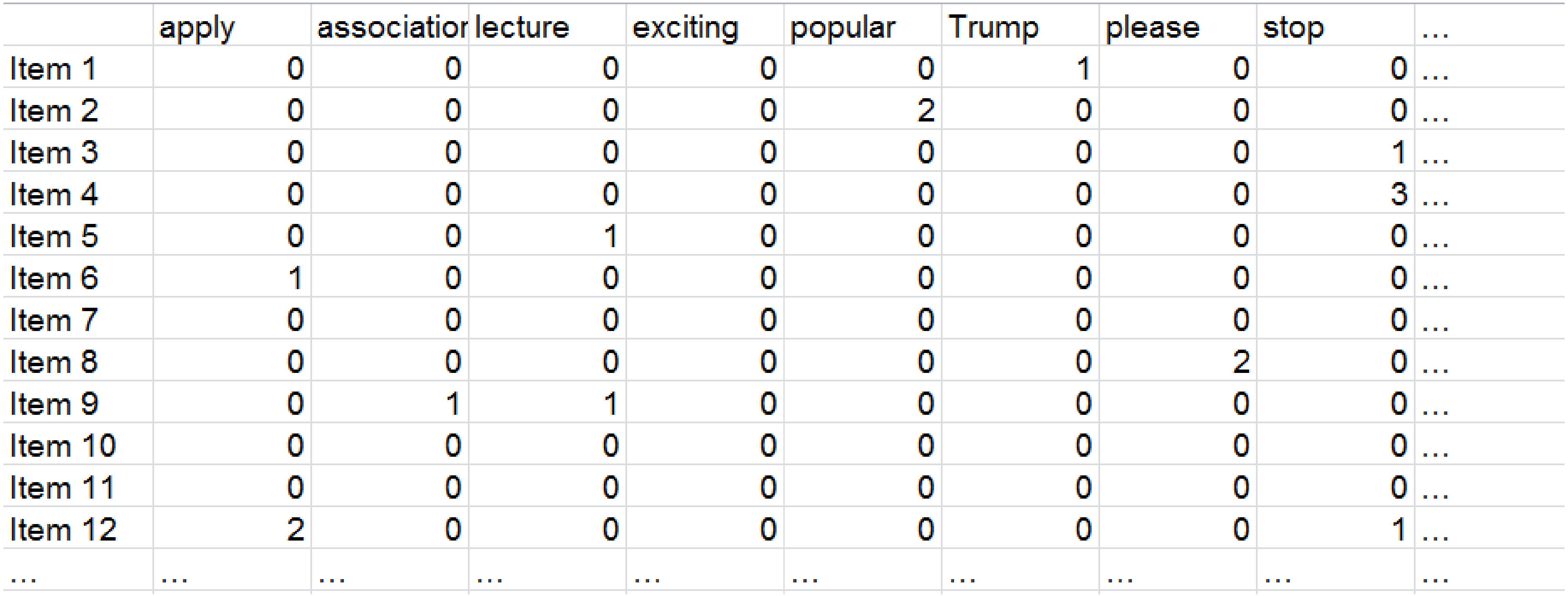
Data Processing

Explanation:



Data Processing

Goal:



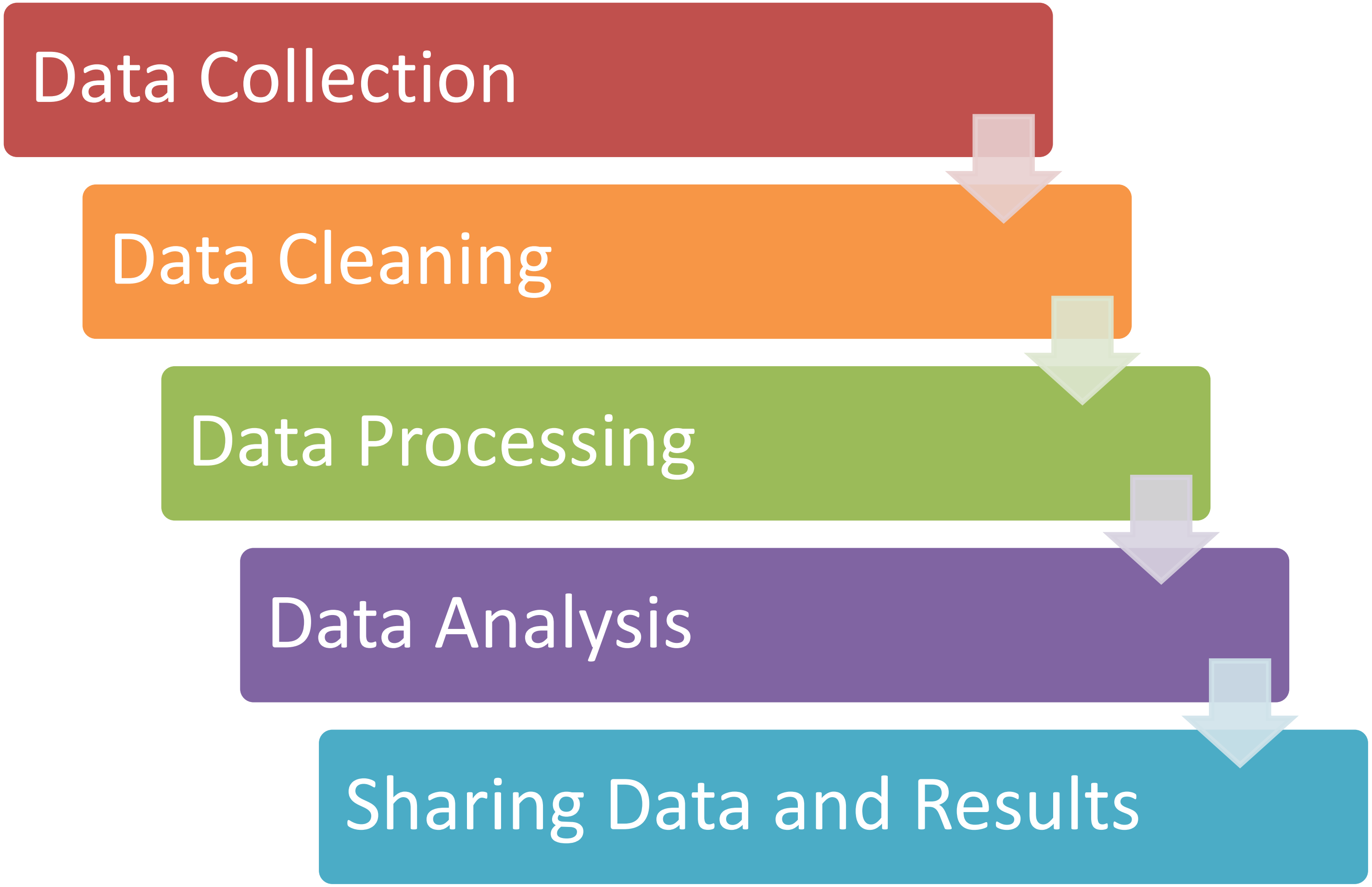
	apply	associatior	lecture	exciting	popular	Trump	please	stop	…
Item 1	0	0	0	0	0	1	0	0	…
Item 2	0	0	0	0	2	0	0	0	…
Item 3	0	0	0	0	0	0	0	1	…
Item 4	0	0	0	0	0	0	0	3	…
Item 5	0	0	1	0	0	0	0	0	…
Item 6	1	0	0	0	0	0	0	0	…
Item 7	0	0	0	0	0	0	0	0	…
Item 8	0	0	0	0	0	0	2	0	…
Item 9	0	1	1	0	0	0	0	0	…
Item 10	0	0	0	0	0	0	0	0	…
Item 11	0	0	0	0	0	0	0	0	…
Item 12	2	0	0	0	0	0	0	1	…
…	…	…	…	…	…	…	…	…	…

Data Processing

Data Processing

1. Go to *Data Processing/preProcess.R*
2. Run lines 1-45
3. Practice: Process the cleaned diary data yourself.

Research Pipeline



Data Collection

Data Cleaning

Data Processing

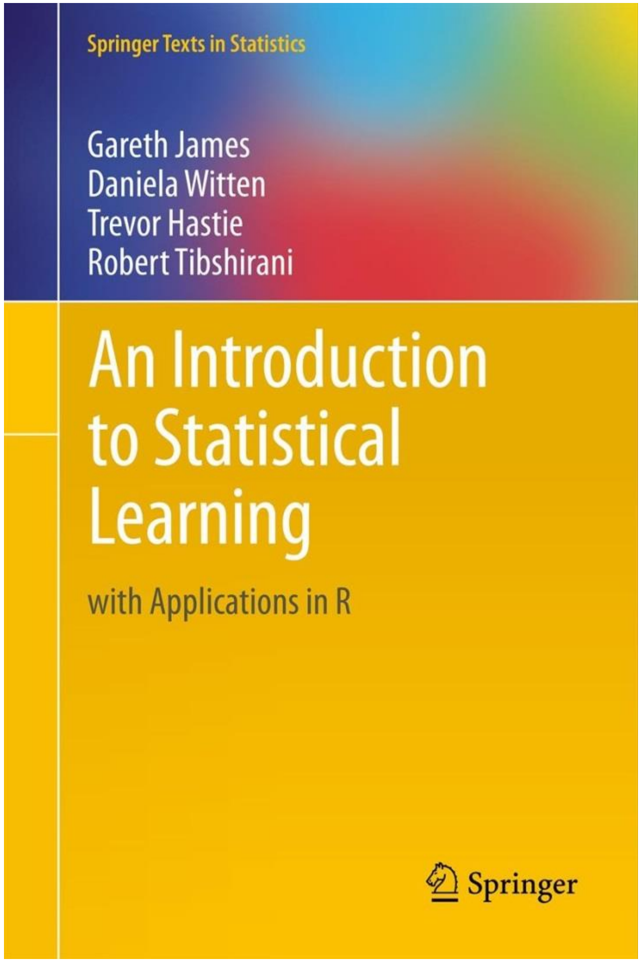
Data Analysis

Sharing Data and Results

Data Analysis

- Unsupervised Learning
    - Clustering Analysis
    - Sentimental Analysis
    - Latent Semantic Analysis
- Supervised Learning
    - Support Vector Machine
    - Random Forests
    - ……

Data Analysis



Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

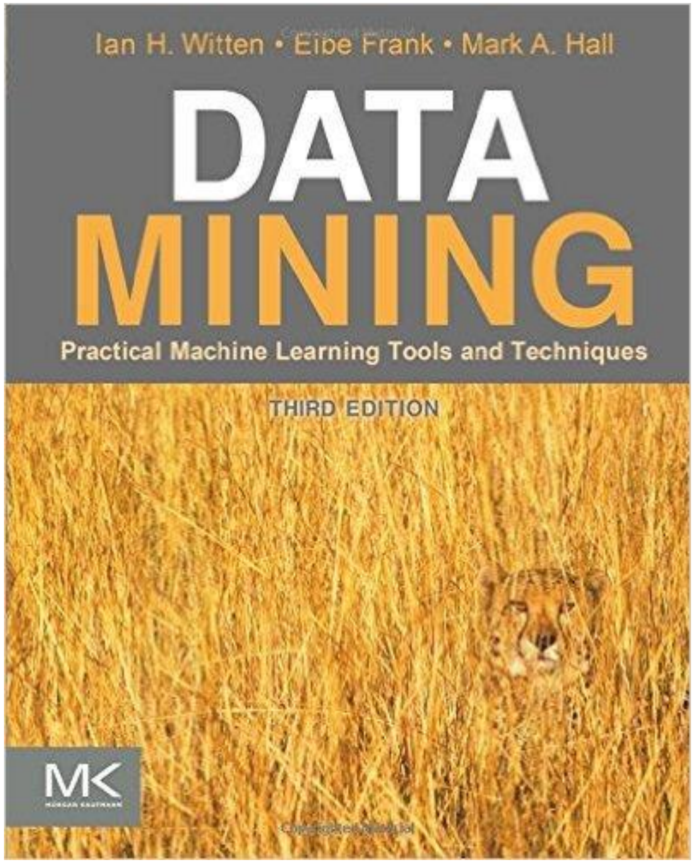
An Introduction to Statistical Learning

with Applications in R

Springer

An Introduction to Statistical Learning with Application in R

Data Analysis


Ian H. Witten • Eibe Frank • Mark A. Hall
DATA
MINING
Practical Machine Learning Tools and Techniques
THIRD EDITION
MK

Data Mining: Practical Machine Learning Tools and Techniques

Data Analysis

- Lexical Variety
- Sentimental Analysis
- Clustering Analysis

Data Analysis

Lexical Variety

Data Analysis

Lexical Variety

1. Vocabulary Richness = Number of unique words / Total number of words

2. Mean Word Frequency = Sum of unique Word Frequency / Total number of unique words

Data Analysis

Lexical Variety of students' diaries

1. **Find data at *Data/diary.csv***
2. ***Clean the data***
3. ***Process the data***
4. **Go to *Data Analysis/lexicalVar.R***

Data Analysis

Sentimental Analysis

Data Analysis

Lexical Variety of students' diaries

1. **Find data at *Data/diary.csv***
2. ***Clean the data***
3. ***Process the data***
4. **Go to *Data Analysis/sentiment.R***

Data Analysis

Clustering Analysis

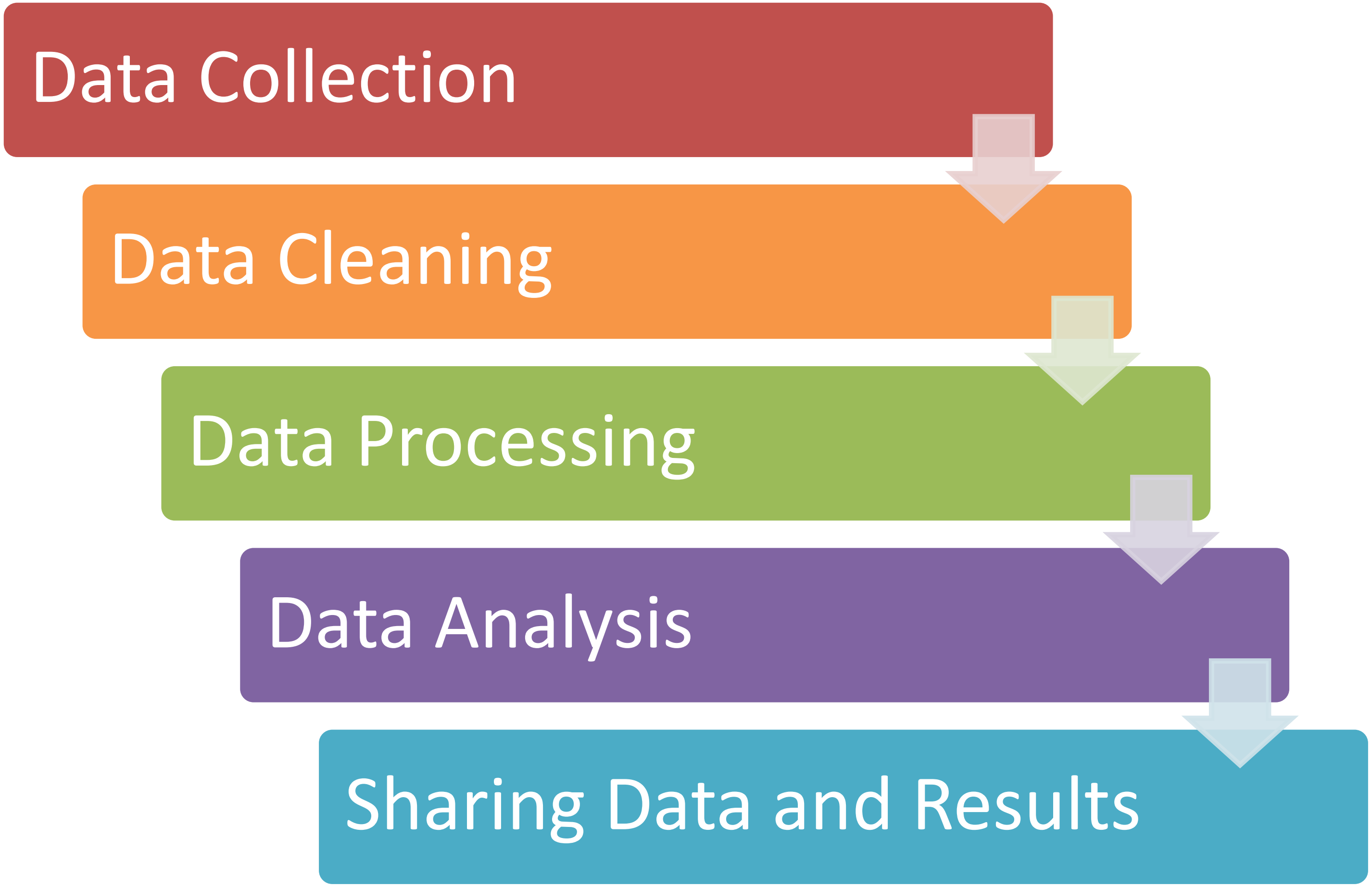
Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive science*, *21*(1), 1-29.

Data Analysis

Clustering Analysis

1. Go to *Data Analysis/hclusterofwords.R*

Research Pipeline



Data Collection

Data Cleaning

Data Processing

Data Analysis

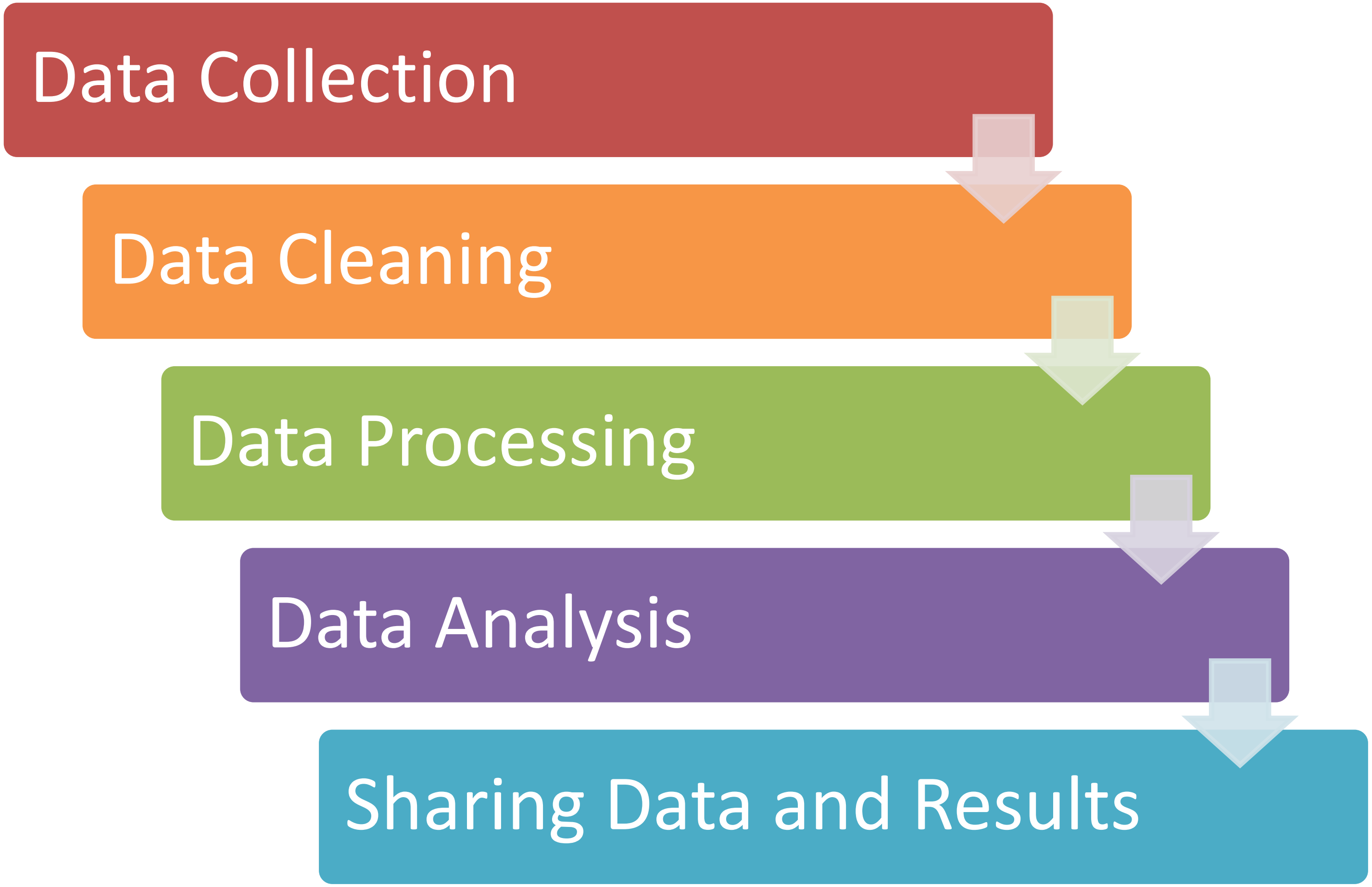
Sharing Data and Results

Sharing Data and Results

- Git + GitHub
  - Git: https://git-scm.com/downloads
  - GitHub https://github.com/Neo-Hao

- KnitR + Rpubs
  - Example: rpubs.com/neohao/online-help-seeking

Research Pipeline



Data Collection

Data Cleaning

Data Processing

Data Analysis

Sharing Data and Results

Thanks!

neohao@uga.edu