



# Lecture 4: Data Collection from the Web

Pilsung Kang

School of Industrial Management Engineering  
Korea University

# AGENDA

01 **Collect Data using APIs: Twitter**

---

02 **Collect Data using APIs: Facebook**

---

03 **Web Scraping: ArXiv Research Papers**

---

04 **Web Community: PPOMPPU**

---

05 **Website: NAVER Real Estate**

---

# Collect Twitter Mentions

- Get an authorized authentication

✓ Step 1: visit <https://apps.twitter.com/>

## Twitter Apps

You don't currently have any Twitter Apps.

[Create New App](#)

## Create an application

### Application Details

#### Name \*

2015\_TM

Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.

#### Description \*

Twitter API for Text Mining Class

Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

#### Website \*

<http://sites.google.com/site/pskang80>

Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens.

(If you don't have a URL yet, just put a placeholder here but remember to change it later.)

#### Callback URL

Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth\_callback URL on the request token step, regardless of the value given here. To restrict your application from using callbacks, leave this field blank.

### Developer Agreement

Last Update: October 22, 2014.

This Twitter Developer Agreement ("Agreement") is made between you (either an individual or an entity, referred to herein as "you") and Twitter, Inc., on behalf of itself and its worldwide affiliates (collectively, "Twitter") and governs your access to and use of the Licensed Material (as defined below).

PLEASE READ THE TERMS AND CONDITIONS OF THIS AGREEMENT CAREFULLY, INCLUDING WITHOUT LIMITATION ANY LINKED TERMS AND CONDITIONS APPEARING OR REFERENCED BELOW, WHICH ARE HEREBY MADE PART OF THIS LICENSE AGREEMENT. BY USING THE LICENSED MATERIAL, YOU ARE AGREEING THAT YOU HAVE READ, AND THAT YOU AGREE TO COMPLY WITH AND TO BE BOUND BY THE TERMS AND CONDITIONS OF THIS AGREEMENT AND ALL APPLICABLE LAWS AND REGULATIONS IN THEIR ENTIRETY WITHOUT LIMITATION OR QUALIFICATION. IF YOU DO NOT AGREE TO BE BOUND BY THIS AGREEMENT, THEN YOU MAY NOT ACCESS OR OTHERWISE USE THE LICENSED MATERIAL. THIS AGREEMENT IS EFFECTIVE AS OF THE FIRST DATE THAT YOU USE THE LICENSED MATERIAL ("EFFECTIVE DATE").

IF YOU ARE AN INDIVIDUAL REPRESENTING AN ENTITY, YOU ACKNOWLEDGE THAT YOU HAVE THE APPROPRIATE AUTHORITY TO ACCEPT THIS AGREEMENT ON BEHALF OF SUCH ENTITY. YOU MAY NOT USE THE LICENSED MATERIAL AND MAY NOT ACCEPT THIS AGREEMENT IF YOU ARE NOT OF LEGAL AGE TO FORM A BINDING CONTRACT WITH

Yes, I agree

[Create your Twitter application](#)

# Collect Twitter Mentions

- Get an authorized authentication

- ✓ Step 2: record necessary URLs

**2015\_TM**

Details    Settings    Keys and Access Tokens    Permissions    Test OAuth

 Twitter API for Text Mining Class  
<http://sites.google.com/site/pskang80>

**Organization**  
*Information about the organization or company associated with your application. This information is optional.*

Organization	None
Organization website	<a href="#">None</a>

**Application Settings**  
*Your application's Consumer Key and Secret are used to authenticate requests to the Twitter Platform.*

Access level	Read-only ( <a href="#">modify app permissions</a> )
Consumer Key (API Key)	[REDACTED] ( <a href="#">manage keys and access tokens</a> )
Callback URL	None
Sign in with Twitter	No
App-only authentication	<a href="https://api.twitter.com/oauth2/token">https://api.twitter.com/oauth2/token</a>
Request token URL	<a href="https://api.twitter.com/oauth/request_token">https://api.twitter.com/oauth/request_token</a>
Authorize URL	<a href="https://api.twitter.com/oauth/authorize">https://api.twitter.com/oauth/authorize</a>
Access token URL	<a href="https://api.twitter.com/oauth/access_token">https://api.twitter.com/oauth/access_token</a>

**Application Actions**

[Delete Application](#)

# Collect Twitter Mentions

- Get an authorized authentication
  - ✓ Step 2: in the permissions tab

The screenshot shows the Twitter API v2 Application Settings page for an application named "2015\_TM". The "Permissions" tab is selected. The "Access" section asks for permission type, with "Read, Write and Access direct messages" selected. A note explains that changes will reflect in access tokens after saving. An "Update Settings" button is at the bottom.

2015\_TM

Details    Settings    Keys and Access Tokens    Permissions

Test OAuth

**Access**

What type of access does your application need?

*Read more about our [Application Permission Model](#).*

Read only

Read and Write

Read, Write and Access direct messages

**Note:**

*Changes to the application permission model will only reflect in access tokens obtained after the permission model change is saved. You will need to re-negotiate existing access tokens to alter the permission level associated with each of your application's users.*

Update Settings

# Collect Twitter Mentions

- Get an authorized authentication

- ✓ Step 3: get the access token

The screenshot shows the 'Application Settings' page for an application named '2015\_TM'. The page has tabs for 'Details', 'Settings', 'Keys and Access Tokens' (which is selected), and 'Permissions'. The 'Keys and Access Tokens' section displays the following information:

Setting	Value
Consumer Key (API Key)	[REDACTED]
Consumer Secret (API Secret)	[REDACTED]
Access Level	Read-only (modify app permissions)
Owner	pskang23
Owner ID	200552592

Below this, the 'Application Actions' section contains buttons for 'Regenerate Consumer Key and Secret' and 'Change App Permissions'. A navigation bar at the bottom shows the number '4'.

The 'Your Access Token' section notes that no token has been created yet. It explains that creating a token allows immediate API calls. A 'Token Actions' section features a button labeled 'Create my access token', which is highlighted with a red box.

# Collect Twitter Mentions

- Get an authorized authentication
  - ✓ Step 3: get the access token

**Your Access Token**

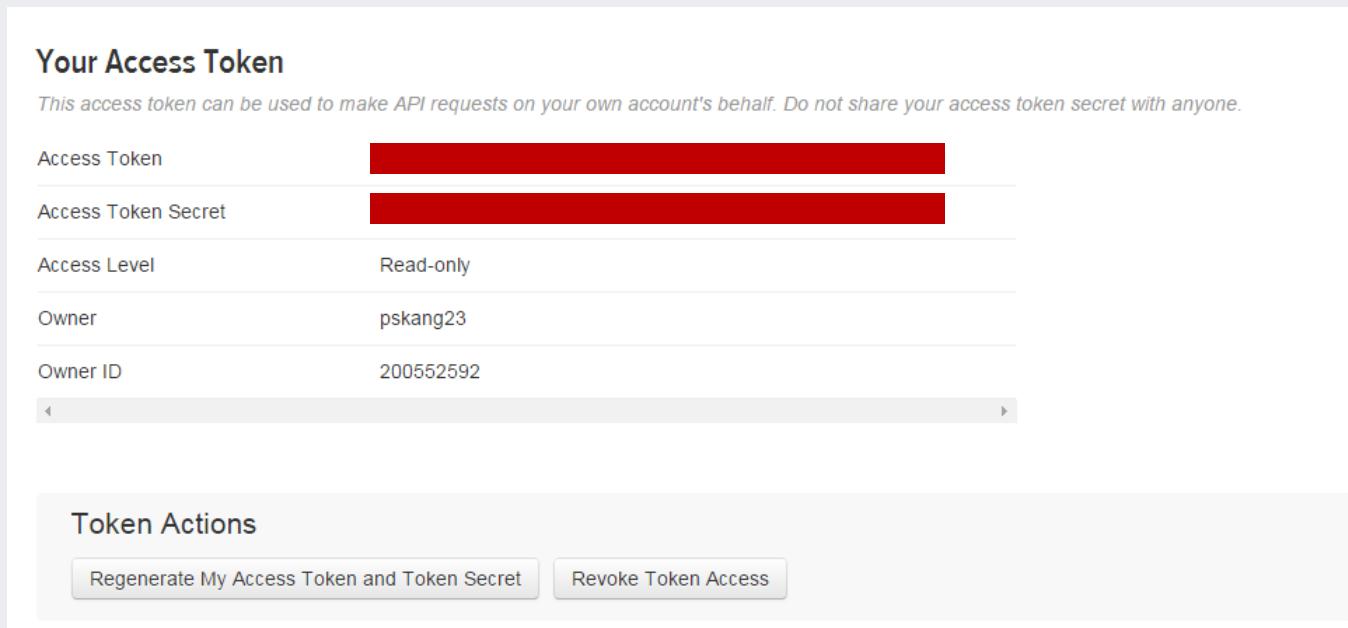
*This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.*

Access Token	[REDACTED]
Access Token Secret	[REDACTED]
Access Level	Read-only
Owner	pskang23
Owner ID	200552592

◀ ▶

**Token Actions**

[Regenerate My Access Token and Token Secret](#) [Revoke Token Access](#)



# Collect Twitter Mentions

- Get an authorized authentication
  - ✓ Step 4: complete the authentication process (for twitteR)
    - Provide consumer\_key, consumer\_secret, access\_token, access\_secret information with **setup\_twitter\_oauth** function

```
# Case 1-1: Collect Texts using Twitter API -----
install.packages("twitteR", "ROAuth", "RCurl", "streamR")
install.packages("rjson", "base64enc", "httr")

library(twitteR)
library(ROAuth)
library(RCurl)
library(streamR)
library(rjson)
library(base64enc)
library(httr)

consumer_key= "Your consumer_key"
consumer_secret= "Your consumer_secret"
access_token = "Your access_token"
access_secret = "Your access_secret"
setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)
```

# Collect Twitter Mentions

- Example 1: Get 100 recent mentions from @IBMWatson

```
# Retrieve the first 100 tweets (or all tweets if fewer than 100) # from the user timeline of  
@IBMWatson WatsonTweets <- userTimeline("IBMWatson", n=100)  
WatsonTweets[1:10]
```

The screenshot shows the official Twitter profile for IBM Watson (@IBMWatson). The profile picture is a stylized blue lightbulb icon. The bio reads: "The official twitter feed of IBM Watson. Follows the IBM Social Computing Guidelines." The header stats show 10.4K tweets, 1,663 following, 191K followers, 3,688 likes, and 6 lists. Below the header, there are three tweets:

- Tweet 1:** "What's new for understanding personality, emotion & tone with Watson? Join our 8/23 webinar for the latest updates: [ibm.co/2wcgMV0](https://ibm.co/2wcgMV0)" (with a photo of a man with a beard looking at a laptop).
- Tweet 2:** "Cybercrime is projected to cause \$6 trillion in damages by 2021. How to stay ahead of #cybersecurity attacks: [ibm.co/2l37Qve](https://ibm.co/2l37Qve)" (with a photo of a person in a server room).
- Tweet 3:** "'I want a beach vacation in January, with scuba diving nearby.' How @WayBlazer uses Watson tech to help travelers: [ibm.co/2wf4pen](https://ibm.co/2wf4pen)" (with a photo of a person in a server room).

```
> WatsonTweets[1:10]  
[[1]]  
[1] "IBMWatson: What's new for understanding personality, emotion & tone with Watson  
? Join our 8/23 webinar for the latest updates:... https://t.co/T4t5MCPxd2"  
  
[[2]]  
[1] "IBMWatson: Cybercrime is projected to cause $6 trillion in damages by 2021. How to  
stay ahead of #cybersecurity attacks:... https://t.co/HTar9F1UMY"  
  
[[3]]  
[1] "IBMWatson: "I want a beach vacation in January, with scuba diving nearby.\\" How @Wa  
yBlazer uses Watson tech to help travelers:... https://t.co/M5ew1HkHZE"  
  
[[4]]  
[1] "IBMWatson: 10 reasons why #AI-powered, automated customer service is the future: ht  
tps://t.co/wHTGStrchz https://t.co/OwkaQ4NWZq"  
  
[[5]]  
[1] "IBMWatson: A recent survey found developers are leading the charge in #AI. Check ou  
t more stats and findings about AI:... https://t.co/cWo86Gr3wv"  
  
[[6]]  
[1] "IBMWatson: 77% of companies using #cognitive tech and #AI use them to innovate prod  
ucts and services. https://t.co/lbnXh0mw0d https://t.co/ahVp19aqjx"  
  
[[7]]  
[1] "IBMWatson: Twitter quick hit: The top 25 people to follow in #AI. Get the full list  
here: https://t.co/UbQ2WYgg28 https://t.co/7xGGfekxF"  
  
[[8]]  
[1] "IBMWatson: How Watson's #AI is helping companies stay ahead of hackers and #cyberse  
curity attacks: https://t.co/QQPnRo0PPSf... https://t.co/Vkxd3z8sZ8"  
  
[[9]]  
[1] "IBMWatson: Video: How Watson Conversation can be quickly integrated into chat platf  
orms such as Facebook Messenger and @Slack https://t.co/JALybaRZYn"  
  
[[10]]  
[1] "IBMWatson: How Watson is poised to make hospital life a lot easier for patients and  
staff alike: https://t.co/YrSHN2uol1 #healthcare"
```

# Collect Twitter Mentions

- Example 2: Get 100 recent mentions with the hashtag #AlphaGo

```
# search research for the hashtag #AlphaGo
```

```
AlphaGoTweets <- searchTwitter("#AlphaGo", n=100)  
AlphaGoTweets[1:10]
```



China Xinhua News @XHNews · Aug 13  
#AlphaGO, #Weiqi bring best in each other: CWA official [xhne.ws/WGiUA](http://xhne.ws/WGiUA)

2 43 108



Yoan Mollard @mollardy · Aug 10  
1.7kWh: 1st paper to mention its computational energy cost! Hey @googleresearch what about #AlphaGo? 😊 #RL #deeplearning #AI #MachineLearning

1:14

1 12 22

```
> AlphaGoTweets[1:10]  
[[1]]  
[1] "JBYoung64: RT @pfizer: What do #selfdrivingcars & #AlphaGo have to do with medicine? Learn more about how we're using #AI to drive research https://t...."  
  
[[2]]  
[1] "calcaware: RT @bradbonomo: Researchers testing #algorithms that display human-like ingenuity. #AI #AlphaGo #DeepMind #TechNews #MachineLearning https://t...."  
  
[[3]]  
[1] "fernandocuenca: RT @bradbonomo: Researchers testing #algorithms that display human-like ingenuity. #AI #AlphaGo #DeepMind #TechNews #MachineLearning https://t...."  
  
[[4]]  
[1] "brandperson2: RT @bradbonomo: Researchers testing #algorithms that display human-like ingenuity. #AI #AlphaGo #DeepMind #TechNews #MachineLearning https://t...."  
  
[[5]]  
[1] "Deep_In_Depth: Google's AI Completely Destroyed a 19-Year-Old. Then He Gave This Epic Response https://t.co/eheV0ypICu... https://t.co/iRdVQYAz4y"  
  
[[6]]  
[1] "Deep_In_Depth: Mastering the game of Go with deep neural networks and tree search https://t.co/wUXVPb0w9w #DeepLearning... https://t.co/mZnErxtlSP"  
  
[[7]]  
[1] "CarlosMBorbon: RT @pfizer: What do #selfdrivingcars & #AlphaGo have to do with medicine? Learn more about how we're using #AI to drive research https://t...."  
  
[[8]]  
[1] "clairebotai: RT @pfizer: What do #selfdrivingcars & #AlphaGo have to do with medicine? Learn more about how we're using #AI to drive research https://t...."  
  
[[9]]  
[1] "SmartMedRT: RT @pfizer: What do #selfdrivingcars & #AlphaGo have to do with medicine? Learn more about how we're using #AI to drive research https://t...."  
  
[[10]]  
[1] "pfizer: What do #selfdrivingcars & #AlphaGo have to do with medicine? Learn more about how we're using #AI to drive research... https://t.co/lSDlrguwK7"
```

# Collect Twitter Mentions

- Example 3: Get all mentions with the hashtag #AlphaGo for a certain period

```
# search research for the hashtag #AlphaGo with time constraints
```

```
AlphaGoTweets2 <- searchTwitter("#AlphaGo", n=1000, since = '2017-08-01', until = '2017-08-17')  
AlphaGoTweets2[1:10]
```

```
> AlphaGoTweets2[1:10]  
[[1]]  
[1] "calcaware: RT @bradbonomo: Researchers testing #algorithms that display human-like  
ingenuity. #AI #AlphaGo #DeepMind #TechNews #MachineLearning https:..."  
  
[[2]]  
[1] "fernandocuenca: RT @bradbonomo: Researchers testing #algorithms that display human-  
like ingenuity. #AI #AlphaGo #DeepMind #TechNews #MachineLearning https:..."  
  
[[3]]  
[1] "brandperson2: RT @bradbonomo: Researchers testing #algorithms that display human-li-  
ke ingenuity. #AI #AlphaGo #DeepMind #TechNews #MachineLearning https:..."  
  
[[4]]  
[1] "Deep_In_Depth: Google's AI Completely Destroyed a 19-Year-Old. Then He Gave This Epic Response https://t.co/eheV0ypICu... https://t.co/iRdvQYAz4"  
  
[[5]]  
[1] "Deep_In_Depth: Mastering the game of Go with deep neural networks and tree search  
https://t.co/wUXVPb0w9w #DeepLearning... https://t.co/mZnErx1sLP"  
  
[[6]]  
[1] "CarlosMBorbon: RT @pfizer: What do #selfdrivingcars & #AlphaGo have to do with  
medicine? Learn more about how we're using #AI to drive research https://t...."  
  
[[7]]  
[1] "clairebotai: RT @pfizer: What do #selfdrivingcars & #AlphaGo have to do with me  
medicine? Learn more about how we're using #AI to drive research https://t...."  
  
[[8]]  
[1] "SmartMedRT: RT @pfizer: What do #selfdrivingcars & #AlphaGo have to do with med  
icine? Learn more about how we're using #AI to drive research https://t...."  
  
[[9]]  
[1] "pfizer: What do #selfdrivingcars & #AlphaGo have to do with medicine? Learn mor  
e about how we're using #AI to drive researc... https://t.co/lSDlrguwK7"  
  
[[10]]  
[1] "clairebotai: RT @MaheshwarLigade: #alphago #dota2 #AI will eSport be the next big b  
uzz? @elonmusk"
```

# Collect Twitter Mentions

- Twitter API limits

Rate Limits: Chart

Title	Resource family	Requests / 15-min window (user auth)	Requests / 15-min window (app auth)
GET application/rate_limit_status	application	180	180
GET favorites/list	favorites	15	15
GET followers/ids	followers	15	15
GET followers/list	followers	15	30
GET friends/ids	friends	15	15
GET friends/list	friends	15	30
GET friendships/show	friendships	180	15
GET help/configuration	help	15	15
GET help/languages	help	15	15
GET help/privacy	help	15	15
GET help/tos	help	15	15
GET lists/list	lists	15	15
GET lists/members	lists	180	15
GET lists/members/show	lists	15	15
GET lists/memberships	lists	15	15
GET lists/ownerships	lists	15	15
GET lists/show	lists	15	15
GET lists/statuses	lists	180	180

Rate Limits: Chart

Title	Resource family	Requests / 15-min window (user auth)	Requests / 15-min window (app auth)
GET lists/subscribers	lists	180	15
GET lists/subscribers/show	lists	15	15
GET lists/subscriptions	lists	15	15
GET search/tweets	search	180	450
GET statuses/lookup	statuses	180	60
GET statuses/retweeters/ids	statuses	15	60
GET statuses/retweets/:id	statuses	15	60
GET statuses/show/:id	statuses	180	180
GET statuses/user_timeline	statuses	180	300
GET trends/available	trends	15	15
GET trends/closest	trends	15	15
GET trends/place	trends	15	15
GET users/lookup	users	180	60
GET users/show	users	180	180
GET users/suggestions	users	15	15
GET users/suggestions/:slug	users	15	15
GET users/suggestions/:slug/members	users	15	15

<https://dev.twitter.com/rest/public/rate-limits>

# Collect Twitter Mentions (streamR)

- Additional Twitter API: streamR

- ✓ Provide a series of function that allow R users to access Twitter's filter, sample, and user streams, and to parse the output into data frames.

```
# Twitter stream data collection
download.file(url="http://curl.haxx.se/ca/cacert.pem", destfile="cacert.pem")

reqURL <- "https://api.twitter.com/oauth/request_token"
accessURL <- "https://api.twitter.com/oauth/access_token"
authURL <- "https://api.twitter.com/oauth/authorize"
apiKey <- "Your apiKey"
apiSecret <- "Your apiKecret"

twitCred <- OAuthFactory$new(consumerKey=apiKey, consumerSecret=apiSecret, requestURL=reqURL,
accessURL=accessURL, authURL=authURL)

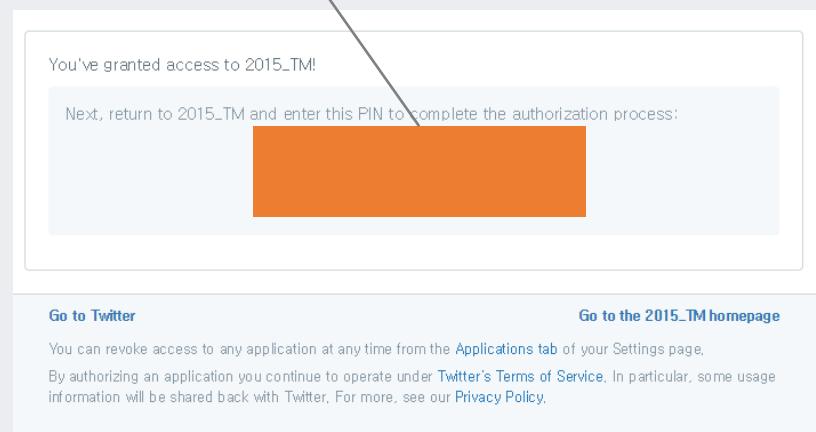
twitCred$handshake(cainfo ="cacert.pem")
```

# Collect Twitter Mentions (streamR)

- Get an authorized authentication

- ✓ Step 4: complete the authentication process

```
> options(RCurlOptions = list(cainfo = system.file("curlSSL", "cacert.pem", package = "RCurl")))
> reqURL <- "https://api.twitter.com/oauth/request_token"
> accessURL <- "https://api.twitter.com/oauth/access_token"
> authURL <- "https://api.twitter.com/oauth/authorize"
> apiKey <- "LnuGt5bbndCBN3t3PWoA"
> apiSecret <- "W3hvDz48zgt4nCHiHMPBx4Ca3h09g7Bzv2Z0oQk"
> twitCred <- OAuthFactory$new(consumerKey=apiKey, consumerSecret=apiSecret,
+ requestURL=reqURL, accessURL=accessURL, authURL=authURL)
> twitCred$handshake(cainfo = system.file("curlSSL", "cacert.pem", package = "RCurl"))
To enable the connection, please direct your web browser to:
https://api.twitter.com/oauth/authorize?oauth_token=kIKNGVKXIdzQ1xT9mvGCE192JPt375sJk6Ujga9vzoI
when complete, record the PIN given to you and provide it here: 5660683
```



# Collect Twitter Mentions (streamR)

- Additional Twitter API: streamR

- ✓ Provide a series of function that allow R users to access Twitter's filter, sample, and user streams, and to parse the output into data frames.

```
filterStream(file="AI.json", track="Artificial Intelligence", language = "en", timeout=30,
oauth=twitCred)

readFile <- file("AI.json", "r")
streamTweets <- readLines(readFile, -1L)

dfMentions <- data.frame()
for (i in 1:length(streamTweets)){
  dfMentions <- rbind(dfMentions, as.data.frame(fromJSON(streamTweets[i])$text))
}

> filterStream(file="AI.json", track="Artificial Intelligence", language = "en", timeout=30, oauth=twitCred)
Capturing tweets...
Connection to Twitter stream was closed after 30 seconds with up to 7 tweets downloaded.
```

# Collect Twitter Mentions (streamR)

- Additional Twitter API: streamR
  - ✓ Provide a series of function that allow R users to access Twitter's filter, sample, and user streams, and to parse the output into data frames.

```
> dfMentions  
fr  
omJSON(streamTweets[i])$text  
1      RT @wirelineio: U.S. #Blockchain Company @BitFuryGroup in Tie-Up on Medical Artificial Intelligence https://t.co/cFbajpkXrK  
2      RT @ForbesTech: Gartner Hype Cycle for emerging tech, 2017 including 5G, artificial general intelligence, deep learning:...  
3 The Most Important Question Underlying #ArtificialIntelligence Research <U+2013> Is #Math Real? via @Future  
ism\nhttps://t.co/h4TLn2dSSa
```

# AGENDA

- 01 Collect Data using APIs: Twitter
- 02 Collect Data using APIs: Facebook
- 03 Web Scraping: ArXiv Research Papers
- 04 Web Community: PPOMPPU
- 05 Website: NAVER Real Estate

# Collect Facebook Posts

- Step 1: Registering an Application with Facebook

✓ <https://developers.facebook.com/>



Login with your facebook account

facebook for developers 제품 문서 도구 및 지원 뉴스 동영상

로그인

전 세계 사용자와 연결하세요.

앱을 빌드하고 확장하며 수익화할 수 있도록 도움을 주는 Fac

Facebook 로그인

Facebook에서 공유하기

Facebook 앱 분석

모바일 수익 창출

Messenger Platform

# Collect Facebook Posts

- Step 1: Registering an Application with Facebook

✓ <https://developers.facebook.com/>

The screenshot shows the Facebook Developers website interface. At the top, there is a navigation bar with links for 'facebook for developers', '제품', '문서', '도구 및 지원', '뉴스', and '동영상'. To the right of the navigation bar are a search bar, a '내 앱' dropdown menu, and a user profile icon. Below the navigation bar, a message in Korean says '현재 Facebook과 통합된 앱이 없습니다.' (There are no apps integrated with Facebook currently). On the right side of the page, a green button labeled '새 앱 만들기' (Create New App) is circled in red. The main content area is titled '새 앱 ID 만들기' (Create New App ID) and contains instructions: 'Facebook을 앱이나 웹사이트로 통합합니다' (Integrate Facebook into your app or website). It has fields for '표시 이름' (Display Name) containing 'KU\_Capstone2', '연락처 이메일' (Contact Email) containing 'pilsung.kang@gmail.com' (which is highlighted in yellow), and '카테고리' (Category) with a dropdown menu showing '교육' (Education). At the bottom, there is a checkbox for accepting the 'Facebook 플랫폼 정책' (Facebook Platform Policy) and a blue '앱 ID 만들기' (Create App ID) button.

# Collect Facebook Posts

- Step 1: Registering an Application with Facebook
    - ✓ Check the AppID & Secret code

KU\_Capstone2

이 앱은 개발 모드 상태이므로 앱 관리자, 개발자, 테스터만 사용할 수 있습니다 [?]

API 버전 [?] 앱 ID

v2.7 104477360011407

앱 시크릿 코드

보기

# Collect Facebook Posts

- Step 1: Registering an Application with Facebook

✓ Add a new website with <http://localhost:1410>

The screenshot shows the Facebook App Dashboard for the application 'KU\_Capstone2'. The left sidebar has a dark theme with sections for 대시보드, 설정 (selected), 기본 설정 (highlighted in blue), 고급 설정, 역할, 알림, 앱 검수, 제품, and + 제품 추가. The main content area displays basic app settings:

앱 ID	104477360011407	앱 시크릿 코드	***** <a href="#">보기</a>
표시 이름	KU_Capstone2	네임스페이스	
앱 도메인		연락처 이메일	pilsung.kang@gmail.com
개인정보취급방침 URL	로그인 대화 상자 및 앱 상세 정보에 대한 개인정보취급	서비스 약관 URL	로그인 대화 상자 및 앱 상세 정보에 대한 서비스 약관
앱 아이콘		카테고리	교육 <a href="#">▼</a>

A red box highlights the '+ 플랫폼 추가' button at the bottom right.

# Collect Facebook Posts

- Step 1: Registering an Application with Facebook

✓ Add a new website with <http://localhost:1410>

The screenshot shows the 'Platfrom 선택' (Platform Selection) section of the Facebook App Review Platform. It displays various platform icons: Facebook Canvas (Facebook logo), Website (globe icon, circled in red), iOS (apple logo), Android (Android logo), Windows App (Windows logo), Page Tab (flag icon), Xbox (Xbox logo), and PlayStation (PlayStation logo). Below these icons is a '취소' (Cancel) button. The 'Website' icon is highlighted with a red circle.

웹사이트

사이트 URL

<http://localhost:1410>

빠른 시작

# Collect Facebook Posts

- Step 2: Create Oauth token to Facebook R session

- ✓ Install Rfacebook package
- ✓ Use your own app id & secret code

```
# Case 1-2: Collect Texts using Facebook API -----
install.packages("Rfacebook")
library(Rfacebook)

# Authentication Setting
my_oauth <- fbOAuth(app_id = "104477360011407", app_secret= "aeab0b819ddcbf4bca2decab0ba22878")
save(my_oauth, file = "my_oauth")
load("my_oauth")
```

The screenshot shows two windows. On the left is a Facebook consent screen for the 'KU\_Capstone2' app, asking for permission to access basic information like profile and friend lists. It includes a logo, a 'Continue' button, and a note about not sharing posting permissions. On the right is a terminal window with blue and red text providing instructions for authentication.

```
Authentication complete. Please close this page and return to R.  
When done, press any key to continue...  
Waiting for authentication in browser...  
Press Esc/Ctrl + C to abort  
Authentication complete.  
Authentication successful.
```

When done, press any key to continue...  
Waiting for authentication in browser...  
Press Esc/Ctrl + C to abort  
Authentication complete.  
Authentication successful.

# Collect Facebook Posts

- Step 3: Collect Data from a Page

- ✓ Need to extract the numeric id of a page: <http://findmyfbid.com/>

고려대학교 대나무숲 (@koreabamboo)

게시물

고려대학교 대나무숲님이 새로운 사진 9장을 추가했습니다.

7월 19일 오후 4:10

\*\*\*이 게시물은 고려대학교 대나무숲이 직접 작성하였습니다\*\*\*

"고대숲에 제보 어떻게 보내요?"  
"진짜 아무도 모르는 익명이에요?"  
"제 댓글이 삭제됐어요!"... 더 보기

망껏 소리칠 수 있는 공간  
고려대학교 대나무숲

제보방법  
페이지 우측의 <문의하기> 버튼 또는 정보란의 링크를 이용해주세요.  
사진 제보는 문의 페이버의 메시지로 보내주세요.  
\* 사진 제보의 링크를 원하시는 경우 제보자에게 사진 및 내용 전달하시면 됩니다.

문의하기

정보

사진

동영상

게시물

커뮤니티

페이지 만들기

## Find your Facebook ID

To find your Facebook personal numeric ID for fb:admins, social plugins, and more, enter your **Facebook personal profile URL** below:

`https://www.facebook.com/koreabamboo/?ref=ts`

Find numeric ID →

# Success!

Your Facebook personal numeric ID is:

**206910909512230**

Find another →

# Collect Facebook Posts

- Step 3: Collect Data from a Page

```
# Get data from a page (Bamboo Forest for KU Students) #
https://www.facebook.com/koreabamboo/?fref=ts
```

```
PageData <- getPage(206910909512230, token = my_oauth, n = 100)
write.csv(PageData, file = "Bamboo_KU.csv")
```

	from_id	from_name	message	created_time	type	link	id	story	likes_count	comments_count	shares_count
1	206910909512230	고려대학교 대나무숲	#30333번째포효 필참 안해도 되는 둘아리를 찾는다고? 1. ...	2017-08-17T08:37:19+0000	status	NA	206910909512230_655392904664026	NA	65	38	12
2	206910909512230	고려대학교 대나무숲	#30332번째포효 오빠, 시귀는 동안 많은 일이 있었죠 내가...	2017-08-17T08:01:35+0000	status	NA	206910909512230_655380927998557	NA	297	108	8
3	206910909512230	고려대학교 대나무숲	#30331번째포효 대술 저는 고시생인데요 점점 사이코파...	2017-08-17T07:47:07+0000	status	NA	206910909512230_655380241331959	NA	127	26	8
4	206910909512230	고려대학교 대나무숲	#30330번째포효 대술!! 대술!! 혹시 우리 가족만 택시운전...	2017-08-17T04:37:02+0000	status	NA	206910909512230_655343574668959	NA	62	48	0
5	206910909512230	고려대학교 대나무숲	#30329번째포효 저는 온갖 알려지를 다 갖고 태어났어요...	2017-08-17T04:21:02+0000	status	NA	206910909512230_655339751336008	NA	57	34	5
6	206910909512230	고려대학교 대나무숲	#30328번째포효 이 학교 들어올 때 내 미래 생각, 내가 진...	2017-08-17T03:21:49+0000	status	NA	206910909512230_655326311337352	NA	30	0	6
7	206910909512230	고려대학교 대나무숲	#30327번째포효 *돈으로 행복을 살 수는 없지만, 돈으로 ...	2017-08-17T03:00:41+0000	status	NA	206910909512230_655320354671281	NA	149	22	17
8	206910909512230	고려대학교 대나무숲	#30326번째포효 대술, 저 고민있어요. 사실 얼마 전 제 죽...	2017-08-17T02:43:37+0000	status	NA	206910909512230_655313514671965	NA	28	7	2
9	206910909512230	고려대학교 대나무숲	#30325번째포효 모든 내 갈정을 새로 일깨워주는 너는 내...	2017-08-17T02:28:43+0000	status	NA	206910909512230_655313084672008	NA	7	1	2
10	206910909512230	고려대학교 대나무숲	#30324번째포효 둘아리에서 내 외모 평가를 하는 남자 선...	2017-08-17T02:08:27+0000	status	NA	206910909512230_655305551339428	NA	342	33	16
11	206910909512230	고려대학교 대나무숲	#30323번째포효 엄마가 편찮으세요. 수술은 성공적으로 ...	2017-08-17T01:52:43+0000	status	NA	206910909512230_655304178006232	NA	13	1	1
12	206910909512230	고려대학교 대나무숲	#30322번째포효 대학오면 살빠지고 연예하고 그럴 수 있...	2017-08-17T00:46:28+0000	status	NA	206910909512230_655285378008112	NA	128	27	1
13	206910909512230	고려대학교 대나무숲	#30321번째포효 뭔가 불안정해요 특별히 눈에 보이는 문...	2017-08-17T00:32:10+0000	status	NA	206910909512230_655285331341450	NA	206	22	15
14	206910909512230	고려대학교 대나무숲	#30320번째포효 궁금한게 있는데요 내 친구가 내 전남친...	2017-08-16T16:14:25+0000	status	NA	206910909512230_655170011352982	NA	89	25	5
15	206910909512230	고려대학교 대나무숲	#30319번째포효 대술, 남자가 피어싱한게 그렇게 보기 안 ...	2017-08-16T16:00:20+0000	status	NA	206910909512230_655168964686420	NA	46	41	0

# Collect Facebook Posts

- Step 3: Collect Data from a Group

- ✓ Need to extract the numeric id of a page: <http://findmyfbid.com/>

The screenshot shows a Facebook group page for 'TensorFlow KR'. The page header includes the group name, a search bar, and navigation links for '강필성', '홈', '그룹', '메시지', '1', and 'FAQ'. On the left, there's a sidebar with links for '토론', '멤버', '이벤트', '동영상', '사진', and '파일', along with a search bar for the group. The main content area displays a post by '이지민님이 Sung Kim님, 권순선님과 함께 있습니다.' (posted on July 29, 2017) with a message about the ML Camp Jeju 2017. Below the post is a link to 'TensorFlowKR/MLJejuCamp' on GitHub. The right side of the screen shows a member list with 21,735 members, a '추천 멤버' section, and various group settings and statistics.

# Collect Facebook Posts

- Step 3: Collect Data from a Group

```
# Get data from a group (Deep learning group)
# https://www.facebook.com/groups/TensorFlowKR/
```

```
GroupData <- getGroup(255834461424286, token = my_oauth, n = 100)
write.csv(GroupData, file = "Tensorflow_KR.csv")
```

	from_id	from_name	message	created_time	type	link	id	story	likes_count	comments_count	shares_count
1	466676193690189	Kim Myungsi	안녕하세요, 엔서플로우로 RNN공부하다 질문드립니다 헨...	2017-08-17T09:11:02+0000	status	NA	255834461424286_520114564996273	NA	0	1	0
2	1988761951344754	박상훈	아주 간단한 질문이지만 ... 여쭈어봅니다 어떤 모델을 학습...	2017-08-17T07:11:26+0000	photo	<a href="https://www.facebook.com/photo.php?fbid=198868...">https://www.facebook.com/photo.php?fbid=198868...</a>	255834461424286_520073828333680	NA	0	2	0
3	1548471688544899	Ja-Keoung Koo	[증강현실 관련 PhD, 소프트웨어 개발자 모집] 안녕하세요, ...	2017-08-17T09:35:17+0000	status	NA	255834461424286_520119941662402	NA	4	0	1
4	898703886930372	박성진	안녕하세요 엔서플로우에서 아래와 같은 그림이 정의된 ...	2017-08-16T09:37:44+0000	photo	<a href="https://www.facebook.com/photo.php?fbid=111122...">https://www.facebook.com/photo.php?fbid=111122...</a>	255834461424286_519682091706187	NA	10	4	0
5	856243047810396	Jerry Kim	안녕하세요! 영상쪽으로 공부하고 있는 학부생입니다! Dec...	2017-08-10T14:46:09+0000	photo	<a href="https://www.facebook.com/photo.php?fbid=115022...">https://www.facebook.com/photo.php?fbid=115022...</a>	255834461424286_516970438644019	NA	20	5	0
6	179941529215463	Shin Kyounghcheol	cnn 풍부 충에 문제가 생겨서 도움을 받고 싶습니다. 항상 ...	2017-08-16T16:14:51+0000	link	<a href="https://github.com/shinv1234/pbl/blob/master/pro...">https://github.com/shinv1234/pbl/blob/master/pro...</a>	255834461424286_519815548359508	NA	1	2	0
7	1507988445929282	Insik Kim	tensorflow 1.3 버전에 timeseries API 추가 된 것 같습니...	2017-08-17T08:01:30+0000	status	NA	255834461424286_52008844098845	NA	2	0	0
8	1432730463483977	김용현	안녕하세요~ 엔서플로우 gpu로 러닝중에 계속 멈출현상이...	2017-08-17T06:28:20+0000	photo	<a href="https://www.facebook.com/photo.php?fbid=143262...">https://www.facebook.com/photo.php?fbid=143262...</a>	255834461424286_520064091667987	NA	3	2	0
9	10209088019414425	Chris Song	스티2 API로 강화학습 배우면 원전 재밌을 거 같지 않아요?...	2017-08-15T11:13:37+0000	photo	<a href="https://www.facebook.com/photo.php?fbid=102123...">https://www.facebook.com/photo.php?fbid=102123...</a>	255834461424286_519257465081983	NA	225	16	0
10	1362563510531971	Hoik Choi	안녕하세요. 엔서플로우 코드를 구현하는 중, 주어진 엔서...	2017-08-17T06:25:38+0000	status	NA	255834461424286_520063578334705	NA	2	0	0
11	470037260034348	Sanggun Kim	Object detection을 하고 싶은데 그래픽 카드가 GTX106...	2017-08-17T00:15:49+0000	status	NA	255834461424286_519956111678785	NA	0	2	0
12	1429884180465552	Young Sung Kim	안녕하세요:) 엔서플로우로 학부 졸작 준비중인 학생입니다...	2017-08-14T10:20:20+0000	status	NA	255834461424286_518739518467111	NA	3	13	2
13	1477502905675907	안홍길	[머신러닝&딥러닝 오프라인 스터디모임] 스터디수준 : ...	2017-08-17T05:19:08+0000	status	NA	255834461424286_520039788337084	NA	16	4	2
14	1648595188544378	Ghiwook Nam	이 논문을 보고 있는데요. <a href="https://arxiv.org/pdf/1703.02344.pdf">https://arxiv.org/pdf/1703.02344.pdf</a>	2017-08-16T11:36:48+0000	link	<a href="https://arxiv.org/pdf/1703.02344.pdf">https://arxiv.org/pdf/1703.02344.pdf</a>	255834461424286_519714188369644	NA	1	7	2
15	1429496113794666	Gyuhyong James Jeon	CNN네트워크위에 RNN을 볼일수있나요? 마지막 pooling...	2017-08-16T23:28:21+0000	status	NA	255834461424286_519944638346599	NA	1	1	2

# AGENDA

- 01 Collect Data using APIs: Twitter
- 02 Collect Data using APIs: Facebook
- 03 Web Scraping: ArXiv Research Papers
- 04 Web Community: PPOMPPU
- 05 Website: NAVER Real Estate

# Web Scraping

- Need to understand HTML/XML structures

## What we see with a browser



Best Speeches of Barack Obama through his 2009 Inauguration

Most Recent Speeches are Listed First

- Barack Obama - Inaugural Speech
- Barack Obama - Election Night Victory / Presidential Acceptance Speech - Nov 4 2008
- Barack Obama - Night Before the Election - The Last Rally - Manassas Virginia - Nov 3 2008
- Barack Obama - Democratic Nominee Acceptance Speech 2008 National Democratic Convention
- Barack Obama - "A World that Stands As One" - Berlin Germany - July 2008
- Barack Obama - Final Primary Night: Presumptive Nominee Speech
- Barack Obama - North Carolina Primary Night
- Barack Obama - Pennsylvania Primary Night
- Barack Obama - AP Annual Luncheon
- Barack Obama - A More Perfect Union "The Race Speech"
- Barack Obama - Texas and Ohio Primary Night
- Barack Obama - Potomac Primary Night

### Obama Inaugural Address 20th January 2009

My fellow citizens:

I stand here today humbled by the task before us, grateful for the trust you have bestowed, mindful of the sacrifices borne by our ancestors. I thank President Bush for his service to our nation, as well as the generosity and cooperation he has shown throughout this transition.

Forty-four Americans have now taken the presidential oath. The words have been spoken during rising tides of prosperity and the still waters of peace. Yet, every so often the oath is taken amidst gathering clouds and raging storms. At these moments, America has carried on not simply because of the skill or vision of those in high office, but because We the People have remained faithful to the ideals of our forbearers, and true to our founding documents.

So it has been. So it must be with this generation of Americans.

That we are in the midst of crisis is now well understood. Our nation is at war, against a far-reaching network of violence and hatred. Our economy is badly weakened, a consequence of greed and irresponsibility on the part of some, but also our collective failure to make hard choices and prepare the nation for a new age. Homes have been lost; jobs shed; businesses shuttered. Our health care is too costly; our schools fail too many; and each day brings further evidence that the ways we use energy strengthen our adversaries and threaten our planet.

These are the indicators of crisis, subject to data and statistics. Less measurable but no less profound is a sapping of confidence across our land - a nagging fear that America's decline is inevitable, and that the next generation must lower its sights.

Today I say to you that the challenges we face are real. They are serious and they are many. They will not be met easily or in a short span of time. But know this, America - they will be met.

On this day, we gather because we have chosen hope over fear, unity of purpose over conflict and discord.

On this day, we come to proclaim an end to the petty grievances and false promises, the recriminations and worn out dogmas, that for far too long have strangled our politics.

We remain a young nation, but in the words of Scripture, the time has come to set aside childish things. The time has come to reaffirm our enduring spirit; to choose our better history; to carry forward that precious gift, that noble idea, passed on from generation to generation: the God-given promise that all are equal, all are free, and all deserve a chance to pursue their full measure of happiness.

In reaffirming the greatness of our nation, we understand that greatness is never a given. It must be earned. Our journey has never been one of short-cuts or settling for less. It has not been the path for the faint-hearted - for those who prefer leisure over work, or seek only the pleasures of riches and fame. Rather, it has been the risk-takers, the doers, the makers of things - some celebrated but more often men and women obscure in their labor, who have carried us up the long, rugged path towards prosperity and freedom.

## What we need to make a web page

```
</tr>
</table></td>
<td rowspan="16" align="center" valign="top" bgcolor="#FFFFFF"><br> <!-- InstanceBeginEditable name="EditRegion3" -->
<table width="610" height="299" border="0" align="center" cellpadding="0" cellspacing="0">
<tr bgcolor="#FFFFFF">
<td align="left" valign="top"><font size="4"><strong><font color="#009900" face="Verdana, Arial, Helvetica, sans-serif">Obama
Inaugural Address <br>
20th January 2009</font></strong><font size="3" face="Verdana, Arial, Helvetica, sans-serif"><br>
</font></font><font size="3" face="Verdana, Arial, Helvetica, sans-serif"><br>
My fellow citizens:<br>
<br>
I stand here today humbled by the task before us, grateful for the
trust you have bestowed, mindful of the sacrifices borne by our ancestors.
I thank President Bush for his service to our nation, as well as the
generosity and cooperation he has shown throughout this transition.<br>
<br>
Forty-four Americans have now taken the presidential oath. The words
have been spoken during rising tides of prosperity and the still waters
of peace. Yet, every so often the oath is taken amidst gathering clouds
and raging storms. At these moments, America has carried on not simply
because of the skill or vision of those in high office, but because
We the People have remained faithful to the ideals of our forbearers,
and true to our founding documents.<br>
<br>
So it has been. So it must be with this generation of Americans.<br>
<br>
That we are in the midst of crisis is now well understood. Our nation
is at war, against a far-reaching network of violence and hatred.
Our economy is badly weakened, a consequence of greed and irresponsibility
on the part of some, but also our collective failure to make hard
choices and prepare the nation for a new age. Homes have been lost;
jobs shed; businesses shuttered. Our health care is too costly; our
schools fail too many; and each day brings further evidence that the
ways we use energy strengthen our adversaries and threaten our planet.<br>
<br>
These are the indicators of crisis, subject to data and statistics.
Less measurable but no less profound is a sapping of confidence across
our land - a nagging fear that America's decline is inevitable, and
that the next generation must lower its sights.<br>
<br>
Today I say to you that the challenges we face are real. They are
serious and they are many. They will not be met easily or in a short
span of time. But know this, America - they will be met.<br>
```

# Web Scraping

- Parsing

- ✓ The process of analyzing a string of symbols, either in natural language or in **computer languages (HTML/XML)**, conforming to the rules of a formal grammar

```
# Case 3: XPath with XML -----
install.packages("XML")
library("XML")

# XML/HTML parsing
obamaurl <- "http://www.obamaspeeches.com/"
obamaroot <- htmlParse(obamaurl) obamaroot
```

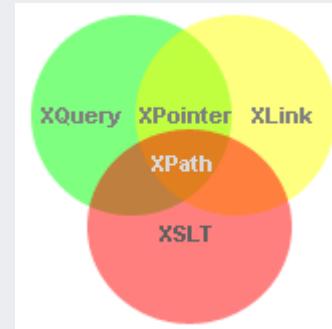
# Web Scraping

- Parsing result

```
Console D:/Dropbox/김의자료/고려대학교/학부 - 데이터 분석을 위한 프로그래밍 언어/04 Data Collection from the Web/ 
> obamaroot
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<html>
<!-- InstanceBegin template="Templates/ObamaSpeechesTemplate.dwt" codeOutsideHTMLIsLocked="false" --><head>
<meta name="description" content="Over 100 speeches by Barack Obama. Constantly updated. Complete and full text of each speech.">
<meta name="keywords" content="barack obama, speeches, barak, oboma">
<!-- InstanceBeginEditable name="doctitle" --><title>The Complete Text Transcripts of Over 100 Barack Obama Speeches</title>
<!-- InstanceEndEditable --><meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<!-- InstanceBeginEditable name="head" --><!-- InstanceEndEditable --><script language="JavaScript" type="text/JavaScript">
</script><script type="text/javascript" src="http://a.remarketstats.com/px/?c=1f5a08ecb0b8bde"></script>
</head>
<style type="text/css">
A:h1 { font-style: none; }
A:link {text-decoration: none;color:white}
A:visited {text-decoration: none; color:white}
A:active {text-decoration: none; background:#333333; color:white}
A:hover {background:yellow; color:blue}
#close {
border: thick dashed #cc0000;
padding: 15px;
margin: 15px;
}
</style>
<body>
<table width="950" border="0" align="center" cellpadding="0" cellspacing="0">
<tr bgcolor="#000000">
<td width="1" bgcolor="#333333">袁</td>
<td width="253" rowspan="16" align="left" valign="top" bgcolor="#333333">
<table width="250" border="0" align="left" cellpadding="10" cellspacing="0" bordercolor="#FFFF00"><tr>
<td height="22" align="left" valign="top">
<div align="center">
<p><font color="#FFFF00" size="2" face="Verdana, Arial, Helvetica, sans-serif"><strong><br></strong><font color="#FFFF00" size="4" face="Verdana, Arial, Helvetica, sans-serif"><strong></strong></font><font color="#FFFF00" size="2" face="Verdana, Arial, Helvetica, sans-serif"><strong>
<br><br><br></strong></font><font color="#FFFF00" size="4" face="Verdana, Arial, Helvetica, sans-serif"><font color="#FFFFFF" size="3">Best
Speeches of<br>
Barack Obama<br>
through his 2009 Inauguration</font></font><font color="#FFFF00" size="2" face="Verdana, Arial, Helvetica, sans-serif"><strong><br><br>
Most Recent Speeches are Listed First <br><strong></strong><br><a href="/P-Obama-Inaugural-Speech-Inauguration.htm">
<div align="left">??Barack Obama -<br>
Inaugural Speech</div>
</a>
</p>
<div align="left">
<strong><br><br><a href="/E11-Barack-Obama-Election-Night-Victory-Speech-Grant-Park-Illinois-November-4-2008.htm">??
```

# Web Scraping

- To extract information that we need from HTML/XML documents, we should also understand **Xpath** expressions
  - ✓ A syntax for defining parts of an XML document
  - ✓ Uses path expressions to navigate in XML documents
    - To select nodes or node-sets in an XML document
    - Path expressions look very much like the expressions you see when you work with a traditional computer file system
  - ✓ Contains a library of standard functions
    - Include over 100 built-in functions (string values, numeric values, date and time comparison, etc.)
  - ✓ For more information, visit [https://www.w3schools.com/xml/xpath\\_intro.asp](https://www.w3schools.com/xml/xpath_intro.asp)



# Web Scraping

- Xpath terminology

- ✓ Nodes: element, attribute, text, namespace, processing-instruction, comment, document
  - XML documents are treated as trees of nodes
  - Root node: the topmost element of the tree

- ✓ Atomic values: nodes with no children or parent

- ✓ Items: atomic values or nodes

Look at the following XML document:

```
<?xml version="1.0" encoding="UTF-8"?>  
  
<bookstore>  
  <book>  
    <title lang="en">Harry Potter</title>  
    <author>J K. Rowling</author>  
    <year>2005</year>  
    <price>29.99</price>  
  </book>  
</bookstore>
```

Example of nodes in the XML document above:

```
<bookstore> (root element node)  
  
<author>J K. Rowling</author> (element node)  
  
lang="en" (attribute node)
```

Example of atomic values:

```
J K. Rowling  
"en"
```

# Web Scraping

- Xpath terminology

- ✓ Relationship of Nodes: Parent, children, siblings, ancestors, descendants

## Parent

Each element and attribute has one parent.

In the following example; the book element is the parent of the title, author, year, and price:

```
<book>
  <title>Harry Potter</title>
  <author>J K. Rowling</author>
  <year>2005</year>
  <price>29.99</price>
</book>
```

## Children

Element nodes may have zero, one or more children.

In the following example; the title, author, year, and price elements are all children of the book element:

```
<book>
  <title>Harry Potter</title>
  <author>J K. Rowling</author>
  <year>2005</year>
  <price>29.99</price>
</book>
```

## Siblings

Nodes that have the same parent.

In the following example; the title, author, year, and price elements are all siblings:

```
<book>
  <title>Harry Potter</title>
  <author>J K. Rowling</author>
  <year>2005</year>
  <price>29.99</price>
</book>
```

## Ancestors

A node's parent, parent's parent, etc.

In the following example; the ancestors of the title element are the book element and the bookstore element:

```
<bookstore>
  <book>
    <title>Harry Potter</title>
    <author>J K. Rowling</author>
    <year>2005</year>
    <price>29.99</price>
  </book>
</bookstore>
```

## Descendants

A node's children, children's children, etc.

In the following example; descendants of the bookstore element are the book, title, author, year, and price elements:

```
<bookstore>
  <book>
    <title>Harry Potter</title>
    <author>J K. Rowling</author>
    <year>2005</year>
    <price>29.99</price>
  </book>
</bookstore>
```

# Web Scraping

- Xpath Syntax

- ✓ Example document:

```
<?xml version="1.0" encoding="UTF-8"?>

<bookstore>

    <book category="COOKING">
        <title lang="en">Everyday Italian</title>
        <author>Giada De Laurentiis</author>
        <year>2005</year>
        <price>30.00</price>
    </book>

    <book category="CHILDREN">
        <title lang="en">Harry Potter</title>
        <author>J K. Rowling</author>
        <year>2005</year>
        <price>29.99</price>
    </book>

    <book category="WEB">
        <title lang="en">XQuery Kick Start</title>
        <author>James McGovern</author>
        <author>Per Bothner</author>
        <author>Kurt Cagle</author>
        <author>James Linn</author>
        <author>Vaidyanathan Nagarajan</author>
        <year>2003</year>
        <price>49.99</price>
    </book>

    <book category="WEB">
        <title lang="en">Learning XML</title>
        <author>Erik T. Ray</author>
        <year>2003</year>
        <price>39.95</price>
    </book>

</bookstore>
```

# Web Scraping

- Xpath Syntax

- ✓ Example document:

```
# Xpath example
xmlfile <- "xml_example.xml"
tmpxml <- xmlParse(xmlfile)
root <- xmlRoot(tmpxml)
root
```

```
<?xml version="1.0" encoding="UTF-8"?>
<bookstore>
  <book category="COOKING">
    <title lang="en">Everyday Italian</title>
    <author>Giada De Laurentiis</author>
    <year>2005</year>
    <price>30.00</price>
  </book>

  <book category="CHILDREN">
    <title lang="en">Harry Potter</title>
    <author>J K. Rowling</author>
    <year>2005</year>
    <price>29.99</price>
  </book>

  <book category="WEB">
    <title lang="en">XQuery Kick Start</title>
    <author>James McGovern</author>
    <author>Per Bothner</author>
    <author>Kurt Cagle</author>
    <author>James Linn</author>
    <author>Vaidyanathan Nagarajan</author>
    <year>2003</year>
    <price>49.99</price>
  </book>

  <book category="WEB">
    <title lang="en">Learning XML</title>
    <author>Erik T. Ray</author>
    <year>2003</year>
    <price>39.95</price>
  </book>
</bookstore>
```

```
Console D:/Dropbox/강의자료/고려대학교/학부 - 데이터 분석을 위한 프로그래밍 언어/04 Data Collection from the Web/
> root
<bookstore>
  <book category="cooking">
    <title lang="en">Everyday Italian</title>
    <author>Giada De Laurentiis</author>
    <year>2005</year>
    <price>30.00</price>
  </book>
  <book category="children">
    <title lang="en">Harry Potter</title>
    <author>J K. Rowling</author>
    <year>2005</year>
    <price>29.99</price>
  </book>
  <book category="web">
    <title lang="en">XQuery Kick Start</title>
    <author>James McGovern</author>
    <author>Per Bothner</author>
    <author>Kurt Cagle</author>
    <author>James Linn</author>
    <author>Vaidyanathan Nagarajan</author>
    <year>2003</year>
    <price>49.99</price>
  </book>
  <book category="web">
    <title lang="en">Learning XML</title>
    <author>Erik T. Ray</author>
    <year>2003</year>
    <price>39.95</price>
  </book>
</bookstore>
```

# Web Scraping

- Xpath Syntax

- ✓ Selecting nodes with node index

```
# Select children node
xmlChildren(root)[[1]]

xmlChildren(xmlChildren(root)[[1]])[[1]]
xmlChildren(xmlChildren(root)[[1]])[[2]]
xmlChildren(xmlChildren(root)[[1]])[[3]]
xmlChildren(xmlChildren(root)[[1]])[[4]]
```

```
Console D:/Dropbox/강의자료/고려대학교/학부 - 데이터 분석을 위한 프로그래밍 언어/04 Data Collection from the Web/
> xmlChildren(root)[[1]]
<book category="cooking">
  <title lang="en">Everyday Italian</title>
  <author>Giada De Laurentiis</author>
  <year>2005</year>
  <price>30.00</price>
</book>
> xmlChildren(xmlChildren(root)[[1]])[[1]]
<title lang="en">Everyday Italian</title>
> xmlChildren(xmlChildren(root)[[1]])[[2]]
<author>Giada De Laurentiis</author>
> xmlChildren(xmlChildren(root)[[1]])[[3]]
<year>2005</year>
> xmlChildren(xmlChildren(root)[[1]])[[4]]
<price>30.00</price>
```

# Web Scraping

- Xpath Syntax

- ✓ Selecting nodes: some useful path expressions

Expression	Description
<i>nodename</i>	Selects all nodes with the name " <i>nodename</i> "
/	Selects from the root node
//	Selects nodes in the document from the current node that match the selection no matter where they are
.	Selects the current node
..	Selects the parent of the current node
@	Selects attributes

In the table below we have listed some path expressions and the result of the expressions:

Path Expression	Result
bookstore	Selects all nodes with the name "bookstore"
/bookstore	Selects the root element bookstore <b>Note:</b> If the path starts with a slash ( / ) it always represents an absolute path to an element!
bookstore/book	Selects all book elements that are children of bookstore
//book	Selects all book elements no matter where they are in the document
bookstore//book	Selects all book elements that are descendant of the bookstore element, no matter where they are under the bookstore element
//@lang	Selects all attributes that are named lang

# Web Scraping

- Xpath Syntax

- ✓ Selecting nodes: some useful path expressions

## # Selecting nodes

```
xpathSApply(root, "/bookstore/book[1]")
xpathSApply(root, "/bookstore/book[last()]")
xpathSApply(root, "/bookstore/book[last()-1]")
xpathSApply(root, "/bookstore/book[position()<3]")
```

```
Console D:/Dropbox/강의자료/고려대학교/학부 - 데이터 분석을 위한 프로그래밍 언어/04 Data Mining with R
> xpathSApply(root, "/bookstore/book[1]")
[[1]]
<book category="cooking">
  <title lang="en">Everyday Italian</title>
  <author>Giada De Laurentiis</author>
  <year>2005</year>
  <price>30.00</price>
</book>

> xpathSApply(root, "/bookstore/book[last()]")
[[1]]
<book category="web">
  <title lang="en">Learning XML</title>
  <author>Erik T. Ray</author>
  <year>2003</year>
  <price>39.95</price>
</book>

> xpathSApply(root, "/bookstore/book[last()-1]")
[[1]]
<book category="web">
  <title lang="en">XQuery Kick Start</title>
  <author>James McGovern</author>
  <author>Per Bothner</author>
  <author>Kurt Cagle</author>
  <author>James Linn</author>
  <author>Vaidyanathan Nagarajan</author>
  <year>2003</year>
  <price>49.99</price>
</book>
```

```
> xpathSApply(root, "/bookstore/book[position()<3]")
[[1]]
<book category="cooking">
  <title lang="en">Everyday Italian</title>
  <author>Giada De Laurentiis</author>
  <year>2005</year>
  <price>30.00</price>
</book>

[[2]]
<book category="children">
  <title lang="en">Harry Potter</title>
  <author>J. K. Rowling</author>
  <year>2005</year>
  <price>29.99</price>
</book>
```

# Web Scraping

- Xpath Syntax

- ✓ Selecting attributes: some useful path expressions

```
# Selecting attributes
xpathSApply(root, "//@category")
xpathSApply(root, "//@lang")
xpathSApply(root, "//book/title", xmlGetAttr, 'lang')
```



The screenshot shows the RStudio Console window. The title bar indicates the current directory is D:/Dropbox/감의자료/고려대학교/학부 - 데이터 분석을 위한 프로그래밍 언어/04 Data Collection from the Web/. The console output displays three R commands using the xpathSApply function to extract attributes from an XML document:

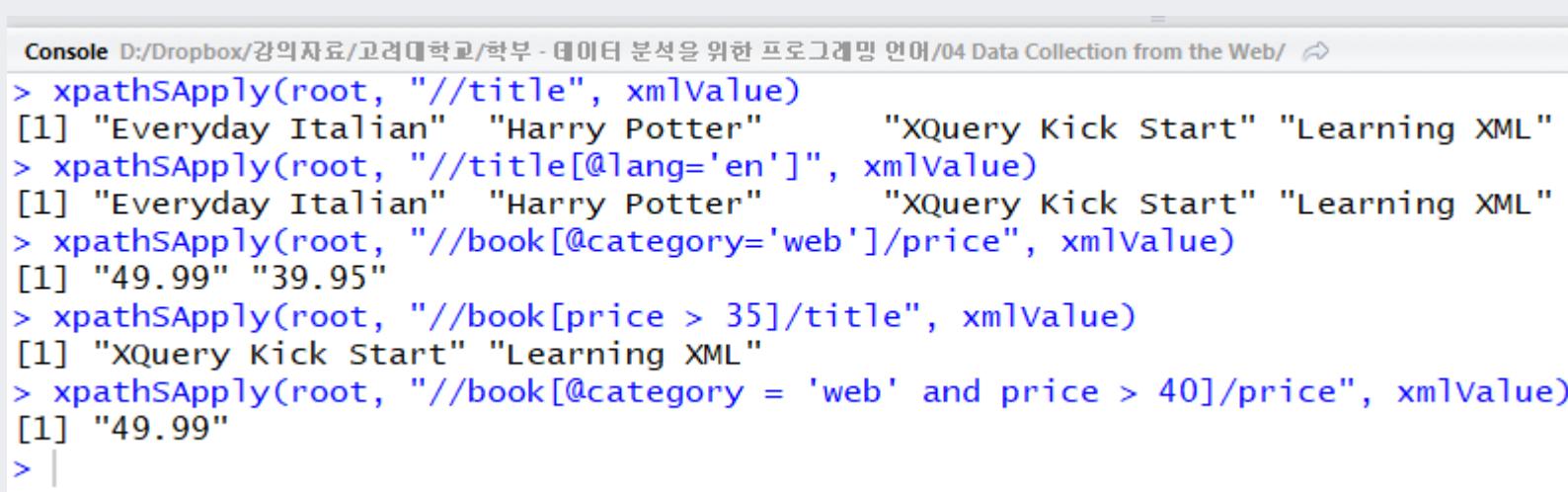
```
Console D:/Dropbox/감의자료/고려대학교/학부 - 데이터 분석을 위한 프로그래밍 언어/04 Data Collection from the Web/
> xpathSApply(root, "//@category")
  category    category    category    category
  "cooking"   "children"   "web"       "web"
> xpathSApply(root, "//@lang")
  lang    lang    lang    lang
  "en"   "en"   "en"   "en"
> xpathSApply(root, "//book/title", xmlGetAttr, 'lang')
[1] "en" "en" "en" "en"
>
```

# Web Scraping

- Xpath Syntax

- ✓ Selecting atomic values: some useful path expressions

```
# Selecting atomic values
xpathSApply(root, "//title", xmlValue)
xpathSApply(root, "//title[@lang='en']", xmlValue)
xpathSApply(root, "//book[@category='web']/price", xmlValue)
xpathSApply(root, "//book[price > 35]/title", xmlValue)
xpathSApply(root, "//book[@category = 'web' and price > 40]/price", xmlValue)
```



The screenshot shows the RStudio console window. The title bar indicates the current directory is D:/Dropbox/강의자료/고려대학교/학부 - 데이터 분석을 위한 프로그래밍 언어/04 Data Collection from the Web/. The console itself displays several examples of the xpathSApply function being used on an XML document root. The examples demonstrate how to select titles, titles in English, books in the 'web' category, titles of books priced above \$35, and books in the 'web' category priced above \$40.

```
Console D:/Dropbox/강의자료/고려대학교/학부 - 데이터 분석을 위한 프로그래밍 언어/04 Data Collection from the Web/
> xpathSApply(root, "//title", xmlValue)
[1] "Everyday Italian" "Harry Potter"      "XQuery Kick Start" "Learning XML"
> xpathSApply(root, "//title[@lang='en']", xmlValue)
[1] "Everyday Italian" "Harry Potter"      "XQuery Kick Start" "Learning XML"
> xpathSApply(root, "//book[@category='web']/price", xmlValue)
[1] "49.99" "39.95"
> xpathSApply(root, "//book[price > 35]/title", xmlValue)
[1] "XQuery Kick Start" "Learning XML"
> xpathSApply(root, "//book[@category = 'web' and price > 40]/price", xmlValue)
[1] "49.99"
> |
```

# Web Scraping

- Xpath Syntax

## ✓ Predicates, unknown nodes, and several paths

### Predicates

Predicates are used to find a specific node or a node that contains a specific value.

Predicates are always embedded in square brackets.

In the table below we have listed some path expressions with predicates and the result of the expressions:

Path Expression	Result
/bookstore/book[1]	Selects the first book element that is the child of the bookstore element. <b>Note:</b> In IE 5,6,7,8,9 first node is[0], but according to W3C, it is [1]. To solve this problem in IE, set the SelectionLanguage to XPath:  <i>In JavaScript:</i> xml.setProperty("SelectionLanguage","XPath");
/bookstore/book[last()]	Selects the last book element that is the child of the bookstore element
/bookstore/book[last()-1]	Selects the last but one book element that is the child of the bookstore element
/bookstore/book[position()<3]	Selects the first two book elements that are children of the bookstore element
//title[@lang]	Selects all the title elements that have an attribute named lang
//title[@lang='en']	Selects all the title elements that have an attribute named lang with a value of 'en'
/bookstore/book[price>35.00]	Selects all the book elements of the bookstore element that have a price element with a value greater than 35.00
/bookstore/book[price>35.00]/title	Selects all the title elements of the book elements of the bookstore element that have a price element with a value greater than 35.00

### Selecting Unknown Nodes

XPath wildcards can be used to select unknown XML elements.

Wildcard	Description
*	Matches any element node
@*	Matches any attribute node
node()	Matches any node of any kind

In the table below we have listed some path expressions and the result of the expressions:

Path Expression	Result
/bookstore/*	Selects all the child element nodes of the bookstore element
/*	Selects all elements in the document
//title[@*]	Selects all title elements which have at least one attribute of any kind

### Selecting Several Paths

By using the | operator in an XPath expression you can select several paths.

In the table below we have listed some path expressions and the result of the expressions:

Path Expression	Result
//book/title   //book/price	Selects all the title AND price elements of all book elements
//title   //price	Selects all the title AND price elements in the document
/bookstore/book/title   //price	Selects all the title elements of the book element of the bookstore element AND all the price elements in the document

# Web Scraping: arXiv Papers

- Web scraping example: arXiv papers about “Text Mining”
  - ✓ arXiv website: <http://arxiv.org/>
  - ✓ Collect Title, Authors, Subjects, Abstracts, and Meta Information

The screenshot shows the arXiv.org search results page for the query "text mining". The page has a dark header with the Cornell University Library logo and a note of gratitude to the Simons Foundation. The main content area is red and displays the search results. It includes a back link, a "Next 25 results" link, and a URL for the search. The results list 189 total items, showing five entries with titles, authors, subjects, and links to PDFs.

Cornell University  
Library

We gratefully acknowledge support from  
the Simons Foundation  
and member institutions

arXiv.org > search

Search or Article ID All papers 🔎

(Help | Advanced search)

### arXiv.org Search Results

Back to Search form | Next 25 results

The URL for this search is [http://arxiv.org:443/find/all/1/all:+EXACT+text\\_mining/0/1/all/0/1](http://arxiv.org:443/find/all/1/all:+EXACT+text_mining/0/1/all/0/1)

Showing results 1 through 25 (of 189 total) for all:"text mining"

- [arXiv:1708.03999 \[pdf, other\]](#)  
**ZOO: Zeroth Order Optimization based Black-box Attacks to Deep Neural Networks without Training Substitute Models**  
Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, Cho-Jui Hsieh  
Subjects: Machine Learning (stat.ML); Cryptography and Security (cs.CR); Learning (cs.LG)
- [arXiv:1708.01225 \[pdf, other\]](#)  
**Recent Developments and Future Challenges in Medical Mixed Reality**  
Long Chen, Thomas Day, Wen Tang, Nigel W. John  
Subjects: Computer Vision and Pattern Recognition (cs.CV)
- [arXiv:1707.08098 \[pdf, other\]](#)  
**From Image to Text Classification: A Novel Approach based on Clustering Word Embeddings**  
Andrei M. Butnaru, Radu Tudor Ionescu  
Comments: Accepted at KES 2017  
Subjects: Computation and Language (cs.CL)
- [arXiv:1707.05420 \[pdf, other\]](#)  
**Cooperative Hierarchical Dirichlet Processes: Superposition vs. Maximization**  
Junyu Xuan, Jie Lu, Guangquan Zhang, Richard Yi Da Xu  
Subjects: Learning (cs.LG); Machine Learning (stat.ML)
- [arXiv:1707.03253 \[pdf, other\]](#)  
**Leipzig Corpus Miner - A Text Mining Infrastructure for Qualitative Data Analysis**  
Andreas Niekler, Gregor Wiedemann, Gerhard Heyer  
Comments: this <https://https://URL/>; Proceedings of Terminology and Knowledge Engineering 2014 (TKE'14), Berlin  
Subjects: Computation and Language (cs.CL)

# Web Scraping: arXiv Papers

- Step 1: Understand the basic structure

- ✓ A total of 189 papers are returned (2017-08-18), each page contains 25 papers
- ✓ Each paper has a unique ID

The screenshot shows the arXiv.org search results page for the query "text mining". The page header includes the Cornell University Library logo and a note of thanks to the Simons Foundation and member institutions. The search bar allows for searching by article ID or title, with dropdown options for "All papers" and a search icon. Below the header, the search results are displayed under the heading "arXiv.org Search Results". The results are paginated, showing items 1 through 25 of 189 total. Each result entry includes a link to the PDF, the title of the paper, the authors, and the subject categories. The subjects are listed in parentheses, such as "(stat.ML)" for Machine Learning.

Cornell University  
Library

We gratefully acknowledge support from  
the Simons Foundation  
and member institutions

arXiv.org > search

Search or Article ID

All papers

(Help | Advanced search)

arXiv.org Search Results

Back to Search form | Next 25 results

The URL for this search is [http://arxiv.org:443/find/all/1/all:+EXACT+text\\_mining/0/1/all/0/1](http://arxiv.org:443/find/all/1/all:+EXACT+text_mining/0/1/all/0/1)

Showing results 1 through 25 (of 189 total) for all:"text mining"

1. [arXiv:1708.03999 \[pdf, other\]](#)  
**ZOO: Zeroth Order Optimization based Black-box Attacks to Deep Neural Networks without Training Substitute Models**  
Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, Cho-Jui Hsieh  
Subjects: Machine Learning (stat.ML); Cryptography and Security (cs.CR); Learning (cs.LG)
2. [arXiv:1708.01225 \[pdf, other\]](#)  
**Recent Developments and Future Challenges in Medical Mixed Reality**  
Long Chen, Thomas Day, Wen Tang, Nigel W. John  
Subjects: Computer Vision and Pattern Recognition (cs.CV)
3. [arXiv:1707.08098 \[pdf, other\]](#)  
**From Image to Text Classification: A Novel Approach based on Clustering Word Embeddings**  
Andrei M. Butnaru, Radu Tudor Ionescu  
Comments: Accepted at KES 2017  
Subjects: Computation and Language (cs.CL)
4. [arXiv:1707.05420 \[pdf, other\]](#)  
**Cooperative Hierarchical Dirichlet Processes: Superposition vs. Maximization**  
Junyu Xuan, Jie Lu, Guangquan Zhang, Richard Yi Da Xu  
Subjects: Learning (cs.LG); Machine Learning (stat.ML)
5. [arXiv:1707.03253 \[pdf, other\]](#)  
**Leipzig Corpus Miner - A Text Mining Infrastructure for Qualitative Data Analysis**  
Andreas Niekler, Gregor Wiedemann, Gerhard Heyer  
Comments: [this https URL](https://this https URL/); Proceedings of Terminology and Knowledge Engineering 2014 (TKE'14), Berlin  
Subjects: Computation and Language (cs.CL)

# Web Scraping: arXiv Papers

- Step 2: Analyzing the HTML Structure

- ✓ First page URL

- [http://arxiv.org/find/all/1/all:+EXACT+text\\_mining/0/1/0/all/0/1?skip=0&query\\_id=504c4472acbc1ebf](http://arxiv.org/find/all/1/all:+EXACT+text_mining/0/1/0/all/0/1?skip=0&query_id=504c4472acbc1ebf)

- ✓ Second page URL

- [http://arxiv.org/find/all/1/all:+EXACT+text\\_mining/0/1/0/all/0/1?skip=25&query\\_id=504c4472acbc1ebf](http://arxiv.org/find/all/1/all:+EXACT+text_mining/0/1/0/all/0/1?skip=25&query_id=504c4472acbc1ebf)

- ✓ Third page URL

- [http://arxiv.org/find/all/1/all:+EXACT+text\\_mining/0/1/0/all/0/1?skip=50&query\\_id=504c4472acbc1ebf](http://arxiv.org/find/all/1/all:+EXACT+text_mining/0/1/0/all/0/1?skip=50&query_id=504c4472acbc1ebf)

# Web Scraping: arXiv Papers

- Step 2: Analyzing the HTML Structure

- ✓ URL Parsing

```
> parse_url("https://arxiv.org/find/all/1/all:+EXACT+text_mining/0/1/0/all/0/1?skip=0")
$scheme
[1] "https"

$hostname
[1] "arxiv.org"

$port
NULL

$path
[1] "find/all/1/all:+EXACT+text_mining/0/1/0/all/0/1"

$query
$query$skip
[1] "0"

$params
NULL

$fragment
NULL

$username
NULL

$password
NULL

attr("class")
[1] "url"
```

The only part that actually changes

# Web Scraping: arXiv Papers

- Step 2: Analyzing the HTML Structure (Press F12 in Chrome browser)
  - ✓ Find the node that contains the necessary links

Cornell University Library

We gratefully acknowledge support from the Simons Foundation and member institutions

arXiv.org > search

Search or Article ID All papers

(Help | Advanced search)

### arXiv.org Search Results

Back to Search form | Previous 25 results | Next 25 results

The URL for this search is [http://arxiv.org:443/find/all/1/all:+EXACT+text\\_mining/0/1/0/all/0/1](http://arxiv.org:443/find/all/1/all:+EXACT+text_mining/0/1/0/all/0/1)

Showing results 26 through 50 (of 189 total) for all:"text mining"

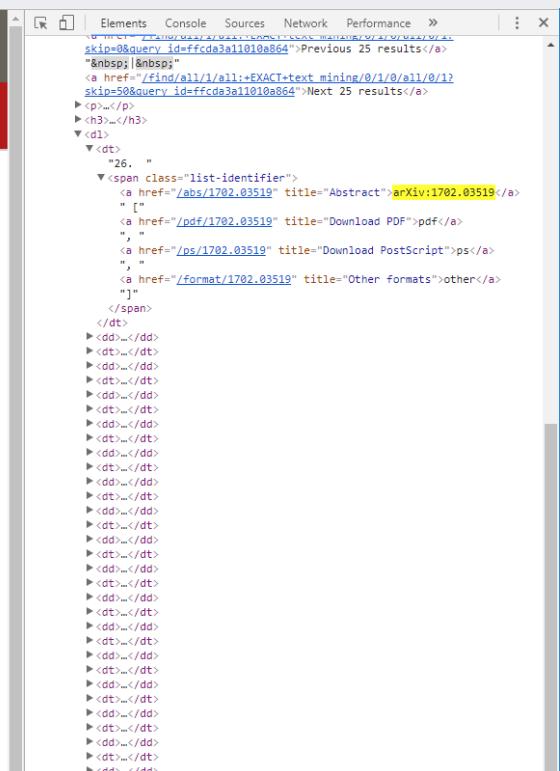
26. [arXiv:1702.03519 \[pdf, ps, other\]](#)  
**A Technical Report: Entity Extraction using Both Character-based and Token-based Similarity**  
Zeyi Wen, Dong Deng, Rui Zhang, Kotagiri Ramamohanarao  
Comments: 12 pages, 6 figures, technical report  
Subjects: Databases (cs.DB)

27. [arXiv:1702.03342 \[pdf, ps, other\]](#)  
**Learning Concept Embeddings for Efficient Bag-of-Concepts Densification**  
Walid Shalaby, Wlodek Zadrozny  
Subjects: Computation and Language (cs.CL)

28. [arXiv:1702.01373 \[pdf, other\]](#)  
**Exact heat kernel on a hypersphere and its applications in kernel SVM**  
Chenchao Zhao, Jun S. Song  
Subjects: Machine Learning (stat.ML); Quantitative Methods (q-bio.QM); Computation (stat.CO)

29. [arXiv:1701.06134 \[pdf, ps, other\]](#)  
**On Practical Accuracy of Edit Distance Approximation Algorithms**  
Hiroyuki Hanada, Mineichi Kudo, Atsuyoshi Nakamura  
Subjects: Data Structures and Algorithms (cs.DS)

30. [arXiv:1701.00798 \[pdf\]](#)  
**Fuzzy Based Implicit Sentiment Analysis on Quantitative Sentences**  
Amir Hossein Yazdavar, Monireh Ebrahimi, Naomie Salim  
Comments: Text mining, Natural language processing, Sentiment analysis, Fuzzy set theory  
Subjects: Computation and Language (cs.CL)



The screenshot shows the browser's developer tools (Elements tab) with the HTML code for the arXiv search results page. The code is heavily nested with `<dd>` and `<dt>` elements, typical of a structured list or table format. The URL in the address bar is [http://arxiv.org:443/find/all/1/all:+EXACT+text\\_mining/0/1/0/all/0/1](http://arxiv.org:443/find/all/1/all:+EXACT+text_mining/0/1/0/all/0/1).

```
<ul style="list-style-type: none; padding-left: 0; margin: 0; border-top: 1px solid black; border-bottom: 1px solid black; background-color: #f9f9f9; font-size: 0.9em; font-weight: bold; color: #333; position: relative; z-index: 1; border-radius: 5px; padding: 5px; margin-bottom: 10px; border-left: 1px solid black; border-right: 1px solid black; ">
    <li><a href="#">26. " </a><br/>
        <span class="list-identifier"><a href="/abs/1702.03519" title="Abstract">arXiv:1702.03519</a>
        " [<span>[</span>]<br/>
            <a href="/pdf/1702.03519" title="Download PDF">pdf</a>
            " [<span>[</span>]<br/>
            <a href="/ps/1702.03519" title="Download PostScript">ps</a>
            " [<span>[</span>]<br/>
            <a href="/format/1702.03519" title="Other formats">other</a>
            " [<span>[</span>]<br/>
        </span>
    </li>
    <li><a href="#">27. " </a><br/>
        <span class="list-identifier"><a href="/abs/1702.03342" title="Abstract">arXiv:1702.03342</a>
        " [<span>[</span>]<br/>
            <a href="/pdf/1702.03342" title="Download PDF">pdf</a>
            " [<span>[</span>]<br/>
            <a href="/ps/1702.03342" title="Download PostScript">ps</a>
            " [<span>[</span>]<br/>
            <a href="/format/1702.03342" title="Other formats">other</a>
            " [<span>[</span>]<br/>
        </span>
    </li>
    <li><a href="#">28. " </a><br/>
        <span class="list-identifier"><a href="/abs/1702.01373" title="Abstract">arXiv:1702.01373</a>
        " [<span>[</span>]<br/>
            <a href="/pdf/1702.01373" title="Download PDF">pdf</a>
            " [<span>[</span>]<br/>
            <a href="/ps/1702.01373" title="Download PostScript">ps</a>
            " [<span>[</span>]<br/>
            <a href="/format/1702.01373" title="Other formats">other</a>
            " [<span>[</span>]<br/>
        </span>
    </li>
    <li><a href="#">29. " </a><br/>
        <span class="list-identifier"><a href="/abs/1701.06134" title="Abstract">arXiv:1701.06134</a>
        " [<span>[</span>]<br/>
            <a href="/pdf/1701.06134" title="Download PDF">pdf</a>
            " [<span>[</span>]<br/>
            <a href="/ps/1701.06134" title="Download PostScript">ps</a>
            " [<span>[</span>]<br/>
            <a href="/format/1701.06134" title="Other formats">other</a>
            " [<span>[</span>]<br/>
        </span>
    </li>
    <li><a href="#">30. " </a><br/>
        <span class="list-identifier"><a href="/abs/1701.00798" title="Abstract">arXiv:1701.00798</a>
        " [<span>[</span>]<br/>
            <a href="/pdf/1701.00798" title="Download PDF">pdf</a>
            " [<span>[</span>]<br/>
            <a href="/ps/1701.00798" title="Download PostScript">ps</a>
            " [<span>[</span>]<br/>
            <a href="/format/1701.00798" title="Other formats">other</a>
            " [<span>[</span>]<br/>
        </span>
    </li>

```

# Web Scraping: arXiv Papers

- Step 2: Analyzing the HTML Structure (Press F12 in Chrome browser)

- ✓ Find the node that contains the necessary links

```
<div id="content">
<h2>arXiv.org Search Results</h2>
<div id="dpage">
<a href="/find/all/1/all/1+EXACT+text_mining/0/1/0/all/0/1?query_id=504c4472acbc1ebf&form=yes">Back to Search form</a>
&nbsp;&nbsp;&nbsp;<a href="/find/all/1/all/1+EXACT+text_mining/0/1/0/all/0/1?skip=25&amp;query_id=504c4472acbc1ebf">Next 25 results</a><p>The URL for this search is http://arxiv.org/find/all/1/all/1+EXACT+text_mining/0/1/0/all/0/1<br /></p>
<dt>Showing results 1 through 25 (of 139 total) for
<a href="/find/all/1/all/1+EXACT+text_mining/0/1/0/all/0/1?skip=0&amp;query_id=504c4472acbc1ebf">all:+text mining</a></h3>
<dd>
<dt>1. <span class="list-identifier"><a href="/abs/1608.03533" title="Abstract">arXiv:1608.03533</a> [<a href="/pdf/1608.03533" title="Download PDF">pdf</a>, <a href="/format/1608.03533" title="Other formats">other</a>]</span></dt>
<dd>
<div class="meta">
<div class="list-title mathjax">
<span class="descriptor">Title:</span> Sequence Graph Transform (SGT): A Feature Extraction Function for Sequence Data Mining
</div>
<div class="list-authors">
<span class="descriptor">Authors:</span>
<a href="/ind/stat/1/au:+Ranjan_C/0/1/0/all/0/1">Chitta Ranjan</a>,
<a href="/ind/stat/1/au:+Ebrahimi_S/0/1/0/all/0/1">Samaneh Ebrahimi</a>,
<a href="/ind/stat/1/au:+Paynabar_K/0/1/0/all/0/1">Kamran Paynabar</a>
</div>
<div class="list-subjects">
<span class="descriptor">Subjects:</span> <span class="primary-subject">Machine Learning (stat.ML)</span>; Learning (cs.LG)
</div>
</div>
</dd>
<dt>2. <span class="list-identifier"><a href="/abs/1608.01844" title="Abstract">arXiv:1608.01844</a> [<a href="/pdf/1608.01844" title="Download PDF">pdf</a>, <a href="/ps/1608.01844" title="Download PostScript">ps</a>, <a href="/format/1608.01844" title="Other formats">other</a>]</span></dt>
<dd>
<div class="meta">
<div class="list-title mathjax">
<span class="descriptor">Title:</span> Separation of nonnegative alpha-stable sources
</div>
<div class="list-authors">
<span class="descriptor">Authors:</span>
<a href="/ind/cs/1/au:+Magron_P/0/1/0/all/0/1">Paul Magron</a>,
<a href="/ind/cs/1/au:+Badeau_R/0/1/0/all/0/1">Roland Badeau</a>,
<a href="/ind/cs/1/au:+Liutkus_A/0/1/0/all/0/1">Antoine Liutkus</a>
</div>
<div class="list-subjects">
<span class="descriptor">Subjects:</span> <span class="primary-subject">Sound (cs.SD)</span>
</div>
</div>
</dd>
<dt>3. <span class="list-identifier"><a href="/abs/1607.07745" title="Abstract">arXiv:1607.07745</a> [<a href="/pdf/1607.07745" title="Download PDF">pdf</a>]</span></dt>
<dd>
<div class="meta">
<div class="list-title mathjax">
<span class="descriptor">Title:</span> Leveraging Unstructured Data to Detect Emerging Reliability Issues
</div>
<div class="list-authors">
<span class="descriptor">Authors:</span>
<a href="/ind/cs/1/au:+Kakde_D/0/1/0/all/0/1">Deovrat Kakde</a>,
<a href="/ind/cs/1/au:+Chaudhuri_A/0/1/0/all/0/1">Arin Chaudhuri</a>
</div>
<div class="list-subjects">
<span class="descriptor">Subjects:</span> <span class="primary-subject">Artificial Intelligence (cs.AI)</span>; Applications (stat.AP); Methodology (stat.ME); Machine Learning (stat.ML)
</div>
</div>
```

# Web Scraping: arXiv Papers

- Step 2: Analyzing the HTML Structure

- ✓ Extract the link information
- ✓ Should be familiar to the usage of CSS Selector

- [http://www.w3schools.com/cssref/css\\_selectors.asp](http://www.w3schools.com/cssref/css_selectors.asp)

## CSS Selectors

In CSS, selectors are patterns used to select the element(s) you want to style.

Use our [CSS Selector Tester](#) to demonstrate the different selectors.

The "CSS" column indicates in which CSS version the property is defined (CSS1, CSS2, or CSS3).

Selector	Example	Example description	CSS
<code>.class</code>	<code>.intro</code>	Selects all elements with class="intro"	1
<code>#id</code>	<code>#firstname</code>	Selects the element with id="firstname"	1
<code>*</code>	<code>*</code>	Selects all elements	2
<code>element</code>	<code>p</code>	Selects all <p> elements	1
<code>element,element</code>	<code>div, p</code>	Selects all <div> elements and all <p> elements	1
<code>element element</code>	<code>div p</code>	Selects all <p> elements inside <div> elements	1
<code>element&gt;element</code>	<code>div &gt; p</code>	Selects all <p> elements where the parent is a <div> element	2
<code>element+element</code>	<code>div + p</code>	Selects all <p> elements that are placed immediately after <div> elements	2
<code>element1~element2</code>	<code>p ~ ul</code>	Selects every <ul> element that are preceded by a <p> element	3
<code>[attribute]</code>	<code>[target]</code>	Selects all elements with a target attribute	2
<code>[attribute=value]</code>	<code>[target=_blank]</code>	Selects all elements with target="_blank"	2
<code>[attribute~=value]</code>	<code>[title~=flower]</code>	Selects all elements with a title attribute containing the word "flower"	2
<code>[attribute =value]</code>	<code>[lang =en]</code>	Selects all elements with a lang attribute value starting with "en"	2
<code>[attribute^=value]</code>	<code>a[href^="https"]</code>	Selects every <a> element whose href attribute value begins with "https"	3
<code>[attribute\$=value]</code>	<code>a[href\$=".pdf"]</code>	Selects every <a> element whose href attribute value ends with ".pdf"	3
<code>[attribute*=value]</code>	<code>a[href*="w3schools"]</code>	Selects every <a> element whose href attribute value contains the substring "w3schools"	3

# Web Scraping: arXiv Papers

- Step 2: Analyzing the HTML Structure

- ✓ Extract the link information

```
tmp_list <- read_html(tmp_url) %>% html_nodes('div#dlpage') %>%  
  html_nodes('a[title="Abstract"]') %>% html_attr('href')
```

- find the node (div id = “dlpage”) → find the node title attribute is Abstract → Store the attribute value of ‘href’ to the tmp\_list

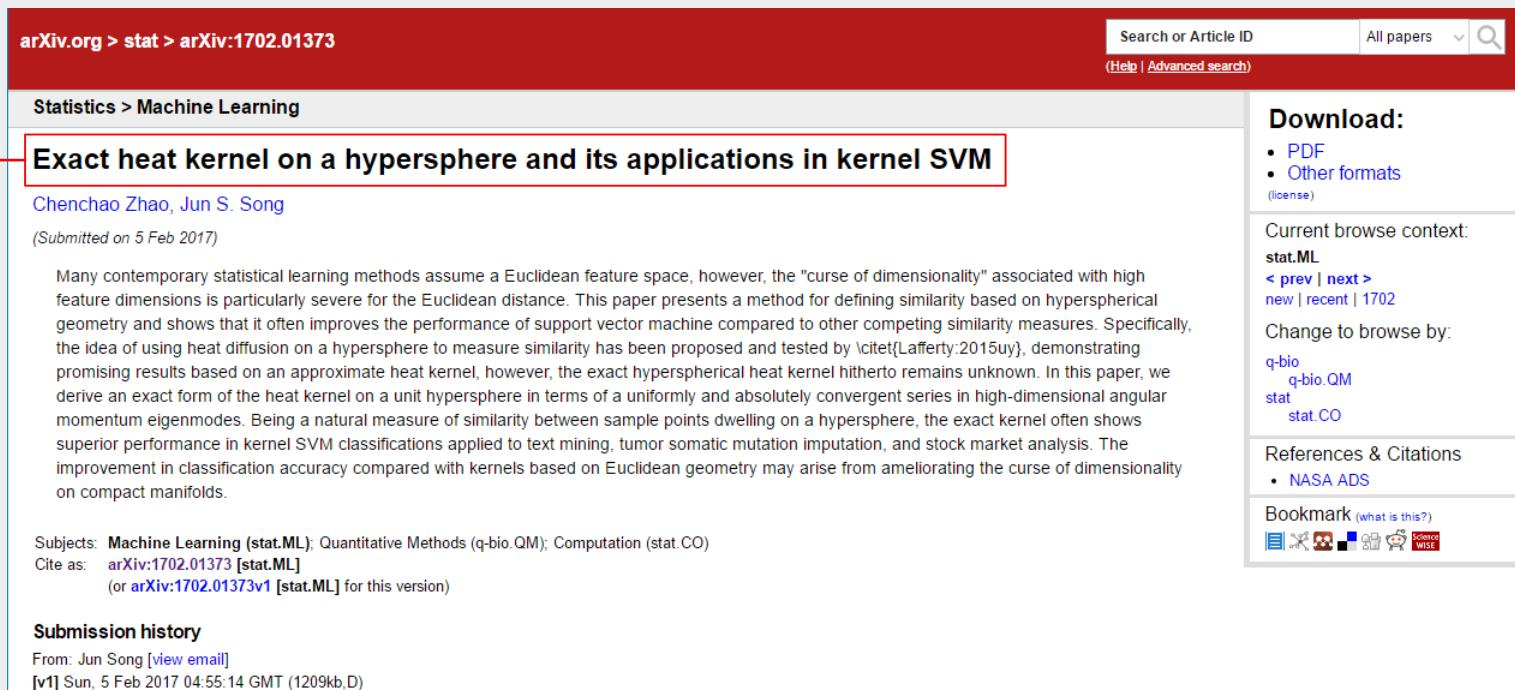
- ✓ Values that are stored in the “tmp\_list”

```
> tmp_list  
[1] "/abs/1701.06134" "/abs/1701.00798" "/abs/1701.00487" "/abs/1612.09535"  
[5] "/abs/1612.08913" "/abs/1612.07630" "/abs/1612.07215" "/abs/1612.04112"  
[9] "/abs/1612.03409" "/abs/1612.01556" "/abs/1611.05204" "/abs/1611.04822"  
[13] "/abs/1611.03660" "/abs/1611.02101" "/abs/1611.00315" "/abs/1610.06370"  
[17] "/abs/1610.01891" "/abs/1609.09154" "/abs/1609.09019" "/abs/1609.07585"  
[21] "/abs/1609.07302" "/abs/1608.03533" "/abs/1608.01844" "/abs/1607.07745"  
[25] "/abs/1606.09636"
```

# Web Scraping: arXiv Papers

- Step 3: Extract necessary information

- ✓ Step 3-1: Extract Title



The screenshot shows a detailed view of an arXiv paper page. At the top, the URL is arXiv.org > stat > arXiv:1702.01373. The search bar includes fields for 'Search or Article ID' and 'All papers' with a dropdown menu, and a magnifying glass icon. Below the header, the paper's category is listed as 'Statistics > Machine Learning'. The title 'Exact heat kernel on a hypersphere and its applications in kernel SVM' is prominently displayed in a large, bold, black font, enclosed in a red rectangular box. Below the title, the authors are Chenchao Zhao, Jun S. Song, and the submission date is (Submitted on 5 Feb 2017). The abstract discusses the curse of dimensionality and the performance of support vector machines. The subjects listed are Machine Learning (stat.ML), Quantitative Methods (q-bio.QM), and Computation (stat.CO). The citation identifier is arXiv:1702.01373 [stat.ML]. The submission history shows it was submitted by Jun Song on Sun, 5 Feb 2017 at 04:55:14 GMT. On the right side of the page, there is a sidebar titled 'Download:' which includes links for PDF and other formats, along with a license link. It also shows the current browse context as stat.ML and provides links for previous and next documents, as well as recent submissions. There are also links to change the browse category to q-bio, q-bio.QM, stat, or stat.CO. A 'References & Citations' section lists NASA ADS, and a 'Bookmark' section includes links for Mendeley, ResearchGate, and ScienceWISE.

```
<div class="leftcolumn">
<div class="subheader">
<h1>Statistics > Machine Learning</h1>
</div>
<h1 class="title mathjax"><span class="descriptor">Title:</span>
Exact heat kernel on a hypersphere and its applications in kernel SVM</h1>
```

# Web Scraping: arXiv Papers

- Step 3: Extract necessary information

- ✓ Step 3-1: Extract Title

```
# title  
tmp_title <- gsub('Title:\n', '', tmp_paragraph %>% html_nodes('h1.title.mathjax') %>% html_text(T))  
title <- c(title, tmp_title)
```

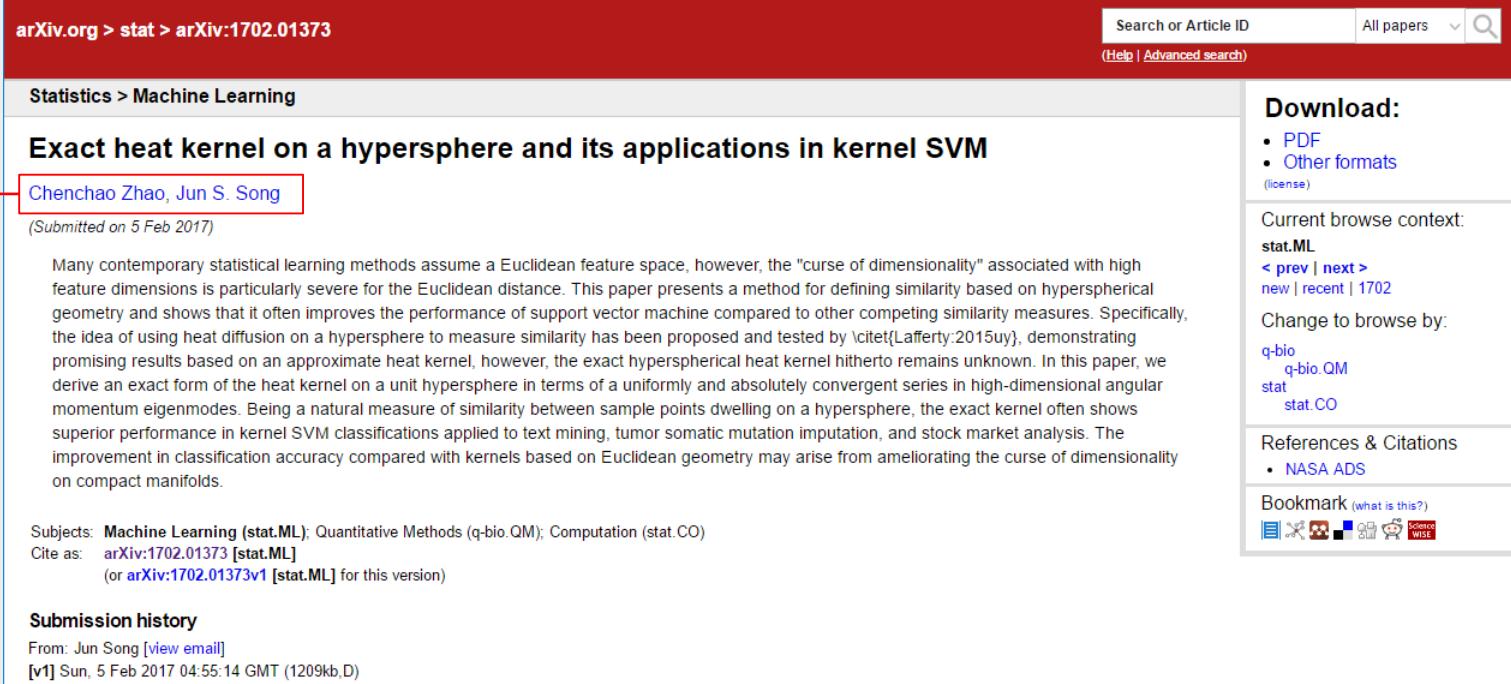
- From tmp\_paragraph → find the node whose h1 class name is “title mathjax” → extract the html text and store in to tmp\_title

```
> tmp_title  
[1] "Exact heat kernel on a hypersphere and its applications in kernel SVM"
```

# Web Scraping: arXiv Papers

- Step 3: Extract necessary information

## ✓ Step 3-2: Extract Authors



The screenshot shows a detailed view of an arXiv paper page. At the top, the URL is arXiv.org > stat > arXiv:1702.01373. The search bar contains "Search or Article ID" and "All papers". Below the header, the paper title is "Exact heat kernel on a hypersphere and its applications in kernel SVM". The authors' names, "Chencho Zhao, Jun S. Song", are highlighted with a red box. The text "(Submitted on 5 Feb 2017)" is below the title. The abstract discusses the curse of dimensionality and the performance of support vector machines. The subjects listed are Machine Learning (stat.ML), Quantitative Methods (q-bio.QM), and Computation (stat.CO). The submission history shows it was submitted on Sun, 5 Feb 2017 at 04:55:14 GMT. The right sidebar includes sections for Download (PDF, Other formats, license), Current browse context (stat.ML), Change to browse by (q-bio, q-bio.QM, stat, stat.CO), References & Citations (NASA ADS), and Bookmark (with icons for various services). A large red arrow points from the bottom left towards the highlighted authors' names.

```
<div class="authors"><span class="descriptor">Authors:</span>
<a href="/find/stat/1/au:+Zhao_C/0/1/0/all/0/1">Chencho Zhao</a>,
<a href="/find/stat/1/au:+Song_J/0/1/0/all/0/1">Jun S. Song</a></div>
```

# Web Scraping: arXiv Papers

- Step 3: Extract necessary information

- ✓ Step 3-2: Extract Authors

```
# author
tmp_author <- tmp_paragraph %>% html_nodes('div.authors') %>% html_text
tmp_author <- gsub('\\s+', ' ', tmp_author)
tmp_author <- gsub('Authors:', '', tmp_author) %>% str_trim
author <- c(author, tmp_author)
```

- From tmp\_paragraph → Select node whose div class = “authors” → Store the html text
- Replace various spaces (space, tab, etc.) by a single space
- Remove ‘Authors:’ and trim the string

```
> tmp_author
[1] "Chenchao Zhao, Jun S. Song"
```

# Web Scraping: arXiv Papers

- Step 3: Extract necessary information

## ✓ Step 3-3: Extract Subjects

The screenshot shows a detailed view of an arXiv paper page. At the top, there's a red header bar with the arXiv.org logo, the category 'stat', and the ID 'arXiv:1702.01373'. To the right of the header are search fields ('Search or Article ID', dropdown 'All papers', and a magnifying glass icon), and links for 'Help | Advanced search'.

The main content area has a white background. It starts with a breadcrumb navigation 'arXiv.org > stat > arXiv:1702.01373' and a category 'Statistics > Machine Learning'. The title of the paper is 'Exact heat kernel on a hypersphere and its applications in kernel SVM' by 'Chencho Zhao, Jun S. Song' (submitted on 5 Feb 2017). The abstract discusses the curse of dimensionality and the performance of support vector machines compared to other similarity measures, mentioning the exact heat kernel on a hypersphere.

A red rectangular box highlights the 'Subjects' section, which lists 'Machine Learning (stat.ML); Quantitative Methods (q-bio.QM); Computation (stat.CO)'. Below this, it says 'Cite as: arXiv:1702.01373 [stat.ML] (or arXiv:1702.01373v1 [stat.ML] for this version)'.

The 'Submission history' section shows the paper was submitted by 'Jun Song' on 'Sun, 5 Feb 2017 04:55:14 GMT (1209kB,D)'.

To the right of the main content, there's a sidebar with a grey header 'Download:' followed by a list of download options: PDF, Other formats, and a link to the license. Below that is 'Current browse context: stat.ML' with links for 'prev | next', 'new | recent | 1702'. There's also a 'Change to browse by:' section with links for 'q-bio', 'q-bio.QM', 'stat', and 'stat.CO'. Further down are sections for 'References & Citations' (with a NASA ADS link) and 'Bookmark' (with links for Mendeley, ResearchGate, and ScienceWISE).

At the bottom of the page, there's a snippet of the HTML code for the 'Subjects' section, showing the structure of the table cell and the subject terms.

```
<td class="tablecell_label">Subjects:</td>
<td class="tablecell_subjects"><span class="primary-subject">Machine Learning (stat.ML)</span>; Quantitative Methods (q-bio.QM); Computation (stat.CO)</td>
```

# Web Scraping: arXiv Papers

- Step 3: Extract necessary information

- ✓ Step 3-3: Extract Subjects

```
# subject
tmp_subject <- tmp_paragraph %>% html_nodes('span.primary-subject') %>% html_text(T)
subject <- c(subject, tmp_subject)
```

- From tmp\_paragraph → find the node whose span class = “primary-subject” → store the html text to tmp\_subject

```
> tmp_subject
[1] "Machine Learning (stat.ML)"
```

# Web Scraping: arXiv Papers

- Step 3: Extract necessary information

- ✓ Step 3-4: Extract Abstract

The screenshot shows a web browser displaying an arXiv paper. The URL in the address bar is `arXiv.org > stat > arXiv:1702.01373`. The page title is "Exact heat kernel on a hypersphere and its applications in kernel SVM". The authors are Chenchao Zhao, Jun S. Song. The submission date is (Submitted on 5 Feb 2017). The abstract text is highlighted with a red box:

Many contemporary statistical learning methods assume a Euclidean feature space, however, the "curse of dimensionality" associated with high feature dimensions is particularly severe for the Euclidean distance. This paper presents a method for defining similarity based on hyperspherical geometry and shows that it often improves the performance of support vector machine compared to other competing similarity measures. Specifically, the idea of using heat diffusion on a hypersphere to measure similarity has been proposed and tested by \citet{Lafferty:2015uy}, demonstrating promising results based on an approximate heat kernel, however, the exact hyperspherical heat kernel hitherto remains unknown. In this paper, we derive an exact form of the heat kernel on a unit hypersphere in terms of a uniformly and absolutely convergent series in high-dimensional angular momentum eigenmodes. Being a natural measure of similarity between sample points dwelling on a hypersphere, the exact kernel often shows superior performance in kernel SVM classifications applied to text mining, tumor somatic mutation imputation, and stock market analysis. The improvement in classification accuracy compared with kernels based on Euclidean geometry may arise from ameliorating the curse of dimensionality on compact manifolds.

Subjects: Machine Learning (stat.ML); Quantitative Methods (q-bio.QM); Computation (stat.CO)  
Cite as: arXiv:1702.01373 [stat.ML]  
(or arXiv:1702.01373v1 [stat.ML] for this version)

Submission history  
From: Jun Song [view email]  
[v1] Sun, 5 Feb 2017 04:55:14 GMT (1209kb,D)

Download:  
• PDF  
• Other formats  
(license)

Current browse context:  
stat.ML  
< prev | next >  
new | recent | 1702

Change to browse by:  
q-bio  
q-bio.QM  
stat  
stat.CO

References & Citations  
• NASA ADS

Bookmark (what is this?)

57/89

# Web Scraping: arXiv Papers

- Step 3: Extract necessary information

- ✓ Step 3-4: Extract Abstract

```
# abstract
tmp_abstract <- tmp_paragraph %>% html_nodes('blockquote.abstract.mathjax') %>% html_text(T)
tmp_abstract <- sub('Abstract:', '', tmp_abstract) %>% str_trim
abstract <- c(abstract, tmp_abstract)
```

- From tmp\_paragraph → find the node whose blockquote class = “abstract mathjax” → Store the html text to tmp\_abstract
- Remove “Abstract:“ and trim the text

```
> tmp_abstract
[1] "Many contemporary statistical learning methods assume a Euclidean feature\nspace, however, the \"curse of dimensionality\" associated with high feature\ndimensions is particularly severe for the Euclidean distance. This paper\npresents a method for defining similarity based on hyperspherical geometry and\nshows that it often improves the performance of support vector machine compared\nto other competing similarity measures. Specifically, the idea of using heat\nand diffusion on a hypersphere to measure similarity has been proposed and tested\nby \\citet{Lafferty:2015uy}, demonstrating promising results based on an\napproximate heat kernel, however, the exact hyperspherical heat kernel hitherto\nremains unknown. In this paper, we derive an exact form of the heat kernel on a\nunit hypersphere in terms of a uniformly and absolutely convergent series in\nhigh-dimensional angular momentum eigenmodes. Being a natural measure of\nsimilarity between sample points dwelling on a hypersphere, t\n... <truncated>
```

# Web Scraping: arXiv Papers

- Step 3: Extract necessary information

- ✓ Step 3-5: Extract Meta information

The screenshot shows a web browser displaying an arXiv paper page. The URL in the address bar is `arXiv.org > stat > arXiv:1702.01373`. The page title is "Exact heat kernel on a hypersphere and its applications in kernel SVM" by Chenchao Zhao, Jun S. Song. The abstract discusses the curse of dimensionality and the performance of support vector machines. The "Submission history" section is highlighted with a red box and an arrow, showing the email from Jun Song and the submission date. The right sidebar provides download options (PDF, Other formats), browse context (stat.ML), change to browse by (q-bio, q-bio.QM, stat, stat.CO), references & citations (NASA ADS), and bookmarking options.

arXiv.org > stat > arXiv:1702.01373

Search or Article ID All papers

(Help | Advanced search)

Statistics > Machine Learning

## Exact heat kernel on a hypersphere and its applications in kernel SVM

Chenchao Zhao, Jun S. Song  
(Submitted on 5 Feb 2017)

Many contemporary statistical learning methods assume a Euclidean feature space, however, the "curse of dimensionality" associated with high feature dimensions is particularly severe for the Euclidean distance. This paper presents a method for defining similarity based on hyperspherical geometry and shows that it often improves the performance of support vector machine compared to other competing similarity measures. Specifically, the idea of using heat diffusion on a hypersphere to measure similarity has been proposed and tested by \citet{Lafferty:2015uy}, demonstrating promising results based on an approximate heat kernel, however, the exact hyperspherical heat kernel hitherto remains unknown. In this paper, we derive an exact form of the heat kernel on a unit hypersphere in terms of a uniformly and absolutely convergent series in high-dimensional angular momentum eigenmodes. Being a natural measure of similarity between sample points dwelling on a hypersphere, the exact kernel often shows superior performance in kernel SVM classifications applied to text mining, tumor somatic mutation imputation, and stock market analysis. The improvement in classification accuracy compared with kernels based on Euclidean geometry may arise from ameliorating the curse of dimensionality on compact manifolds.

Subjects: Machine Learning (stat.ML); Quantitative Methods (q-bio.QM); Computation (stat.CO)  
Cite as: arXiv:1702.01373 [stat.ML]  
(or arXiv:1702.01373v1 [stat.ML] for this version)

**Submission history**

From: Jun Song [view email]  
[v1] Sun, 5 Feb 2017 04:55:14 GMT (1209kb,D)

```
<div class="submission-history">
<h2>Submission history</h2>
From: Jun Song [<a href="https://arxiv.org/show-email/d8e78523/1702.01373">view email</a>]
<br />
<b>[v1]</b> Sun, 5 Feb 2017 04:55:14 GMT (1209kb,D)<br />
</div>
```

Download:

- PDF
- Other formats

(license)

Current browse context:  
stat.ML  
< prev | next >  
new | recent | 1702

Change to browse by:  
q-bio  
q-bio.QM  
stat  
stat.CO

References & Citations  
• NASA ADS

Bookmark (what is this?)

# Web Scraping: arXiv Papers

- Step 3: Extract necessary information

- ✓ Step 3-5: Extract Meta information

```
# meta
tmp_meta <- tmp_paragraph %>% html_nodes('div.submission-history') %>% html_text
tmp_meta <- lapply(strsplit(gsub('\\s+', ' ',tmp_meta), '[v1]', fixed = T), '[',2) %>% unlist %>% str_trim
meta <- c(meta, tmp_meta)
```

- From tmp\_paragraph → find the node whose div class name is “submission-history” →  
Store the html text to tmp\_meta

```
> tmp_meta
[1] "\nSubmission history\nFrom: Jun Song [view email]\n[v1] Sun, 5 Feb 2017 04:55:14 GMT (1209kb,D)"
```

- Replace all spaces by a single space → Split the text (split point = [v1]) → Take the second element → Unlist it → trim the text

```
> tmp_meta
[1] "Sun, 5 Feb 2017 04:55:14 GMT (1209kb,D)"
```

# Web Scraping: arXiv Papers

- Step 4: Repeat the process and export the data

- ✓ Elapsed time for data collection

```
> end - start # Total Elapsed Time  
사용자 시스템 elapsed  
8.97 0.32 287.00  
>
```

- ✓ Check the dataset

	title	author	subject	abstract	meta
1	ZOO: Zeroth Order Optimization based Black-box Att...	Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, C...	Machine Learning (stat.ML)	Deep neural networks (DNNs) are one of the most pro...	Mon, 14 Au...
2	Recent Developments and Future Challenges in Med...	Long Chen, Thomas Day, Wen Tang, Nigel W. John	Computer Vision and Pattern Recognition (cs.CV)	Mixed Reality (MR) is of increasing interest within tec...	Thu, 3 Aug...
3	From Image to Text Classification: A Novel Approach ...	Andrei M. Butnaru, Radu Tudor Ionescu	Computation and Language (cs.CL)	In this paper, we propose a novel approach for text c...	Tue, 25 Jul...
4	Cooperative Hierarchical Dirichlet Processes: Super...	Junyu Xuan, Jie Lu, Guangquan Zhang, Richard Yi Da Xu	Learning (cs.LG)	The cooperative hierarchical structure is a common ...	Tue, 18 Jul...
5	Leipzig Corpus Miner - A Text Mining Infrastructure f...	Andreas Niekler, Gregor Wiedemann, Gerhard Heyer	Computation and Language (cs.CL)	This paper presents the "Leipzig Corpus Miner", a te...	Tue, 11 Jul...
6	A Brief Survey of Text Mining: Classification, Clusteri...	Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, S...	Computation and Language (cs.CL)	The amount of text that is generated every day is in...	Mon, 10 Ju...
7	Identifying Condition-Action Statements in Medical G...	Hossein Hematalam, Wlodek Zadrozny	Computation and Language (cs.CL)	This paper advances the state of the art in text unde...	Tue, 13 Jun...
8	Joint Workshop on Bibliometric-enhanced Information...	Muthu Kumar Chandrasekaran, Kokil Jaidka, Philipp Mayr	Digital Libraries (cs.DL)	The large scale of scholarly publications poses a ch...	Thu, 8 Jun...
9	Max-Cosine Matching Based Neural Models for Reco...	Zhipeng Xie, Junfeng Hu	Computation and Language (cs.CL)	Recognizing textual entailment is a fundamental task...	Thu, 25 Ma...
10	Towards Interrogating Discriminative Machine Learni...	Wenbo Guo, Kaixuan Zhang, Lin Lin, Sui Huang, Xinyu...	Learning (cs.LG)	It is oftentimes impossible to understand how machi...	Tue, 23 Ma...
11	Social Media-based Substance Use Prediction	Tao Ding, Warren K. Bickel, Shimei Pan	Computation and Language (cs.CL)	In this paper, we demonstrate how the state-of-the-ar...	Tue, 16 Ma...
12	Testing Reading Tactics for Automated Reading Assi...	Zhe Yu, Tim Menzies	Software Engineering (cs.SE)	Given the growing number of new publications appe...	Mon, 15 Ma...
13	ResumeVis: A Visual Analytics System to Discover Se...	Chen Zhang, Hao Wang, Yingcai Wu	Human-Computer Interaction (cs.HC)	Massive public resume data emerging on the WWW in...	Mon, 15 Ma...
14	OncoScore: an R package to measure the oncogenic...	Daniele Ramazzotti, Luca De Sano, Roberta Spinelli, ...	Genomics (q-bio.GN)	Motivation: We here present OncoScore, an open-so...	Mon, 8 May...
15	ChestX-ray8: Hospital-scale Chest X-ray Database an...	Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Moha...	Computer Vision and Pattern Recognition (cs.CV)	The chest X-ray is one of the most commonly access...	Fri, 5 May...

# Web Scraping: arXiv Papers

- Step 4: Repeat the process and export the data
    - ✓ Elapsed time for data collection

```
# Export the result  
write.csv(final, file = "Text Mining arxiv papers.csv")
```

A	B	C	D	E	F	G	H	I	J
1	title	author	subject	abstract Deep neural networks (DNNs) are one of the most prominent technologies of our time, as they achieve state-of-the-art performance in many machine learning tasks, including but not limited to image classification, text mining, and speech processing. However, recent research on DNNs has indicated ever-increasing concern on the robustness to adversarial examples, especially for security-critical tasks such as traffic sign identification for autonomous driving. Studies have unveiled the vulnerability of a well-trained DNN by demonstrating the ability of generating barely noticeable (to both human and machines) adversarial images that lead to misclassification. Furthermore, researchers have shown that these adversarial images are highly transferable by simply training and attacking a substitute model built upon the target model, known as a black-box attack to DNNs.  Similar to the setting of training substitute models, in this paper we propose an effective black-box attack that also only has access to the input (images) and the output (confidence scores) of a targeted DNN. However, different from leveraging attack transferability from substitute models, we propose zeroth order optimization (ZOO) based attacks to directly estimate the gradients of the targeted DNN for generating adversarial examples. We use zeroth order stochastic coordinate descent along with dimension reduction, hierarchical attack and importance sampling techniques to efficiently attack black-box models. By exploiting zeroth order optimization, improved attacks to the targeted DNN can be accomplished, sparing the need for training substitute models and avoiding the loss in attack transferability. Experimental results on MNIST, CIFAR10 and ImageNet show that the proposed ZOO attack is as effective as the state-of-the-art white-box attack and significantly outperforms existing	meta				
2				Mixed Reality (MR) is of increasing interest within technology-driven modern medicine but is not yet used in everyday practice. This situation is changing rapidly, however, and this paper explores the emergence of MR technology and the importance of its utility within medical applications. A classification of medical MR has been obtained by applying an unbiased text mining method to a database of 1,403 relevant research papers published over the last two decades. The classification results reveal a taxonomy for the development of medical MR research during this period as well as suggesting future trends. We then use the classification to analyse the technology and applications developed in the last five years. Our objective is to aid researchers to focus on the areas where technology advancements in medical MR are most needed, as well as providing medical practitioners with a useful source of reference.	Mon, 14 Aug 2017 03:48:03 GMT (2147kb,D)				

# AGENDA

- 01 Collect Data using APIs: Twitter
- 02 Collect Data using APIs: Facebook
- 03 Web Scraping: ArXiv Research Papers
- 04 Web Community: PPOMPPU
- 05 Website: NAVER Real Estate

# Web Scraping: Open Forum

- Web scraping: Collect data from an open forum

✓ <http://www.ppomppu.co.kr/zboard/zboard.php?id=insurance>

✓ Date, Title, and Contents

The screenshot shows the homepage of PPOMPPU, a Korean online community platform. The header features the site's logo 'PPOMPPU' in large, colorful letters, followed by the tagline '사람이 좋아 함께하는 곳.. 뽐뿌!' (A place where people like to be together.. PPOMPPU!). A banner on the right side says '사기에는 비싸고.. 사용은 하고 싶고.. 고민될 때! 렌탈상담실!' (When it's expensive.. I want to use it.. When you're worried.. Rental consultation service!). Below the header is a navigation bar with categories: 뽐뿌 (Main), 이벤트, 정보, 커뮤니티, 갤러리, 장터, 포럼, 뉴스, and 상담실.

The main content area displays a grid of forum categories. Each category has a thumbnail image, a title, and a small icon indicating its status or type. The categories include:

카테고리	제목	상세설명
게시글검색	휴대폰/가전	스포츠/레저
모바일 실용주의	휴대폰포럼	골프포럼
8,69	구입개통수령	낚시포럼
이걸 대	휴대폰질문	제테크포럼
쇼핑몰특	마이폰포럼	소셜포럼
[다나와] 노	아이폰포럼	결혼포럼
[GSO아이수퍼]	아이패드포럼	고민포럼
[삼성와닷컴]	안드로이드	게임포럼
[11번가] 디	안드로이드탭	TV/드라마
[CJ몰] BLA	윈도우태블릿	DIV포럼
[롯데닷컴]	기타스마트폰	독서/e-book
[신세계몰] 텔	가전포럼	드론포럼
[마리오마을]	음향기기	40+ 포럼
[티몬] 씨게	컴퓨터포럼	과학포럼
	NAS포럼	문화포럼
	맥포럼	40+ 포럼
	자동차포럼	개발자포럼
	자전거포럼	문구포럼
	축구포럼	문서/서식
	캠핑포럼	뷰티/케어
	테니스포럼	군대포럼
	건강/헬스	문서포럼

Icons at the bottom right of the grid indicate: (B) 베타포럼, (R) 리뉴얼, and (Q) 포럼검색.

# Web Scraping: Open Forum

- Step 1: Check the structure of the URL

- ✓ Check the part that changes with regard to the pages

- <http://www.ppomppu.co.kr/zboard/zboard.php?id=insurance&page=1&divpage=10>
    - <http://www.ppomppu.co.kr/zboard/zboard.php?id=insurance&page=2&divpage=10>
    - <http://www.ppomppu.co.kr/zboard/zboard.php?id=insurance&page=3&divpage=10>
    - ...

# Web Scraping: Open Forum

- Step 1: Check the structure of the URL

## ✓ Check the URL to each page

```
015 <tr align="center" class="list1" onMouseOver="this.style.backgroundColor='#F5F5F5'" onMouseOut="this.style.backgroundColor=''" style="height:16;word-break:break-all;" valign="middle">
016   <td class="eng list_vspace" colspan=2>45876</td>   <td class="han4 list_vspace" nowrap colspan=2><nobr class='han4 list_vspace'>일반</nobr></td>
017   <!--<td nowrap colspan=2 style="padding:0"><input type=checkbox name=cart value='61484'></td-->
018   <td colspan=2 class="list_vspace" align="left"><div style="width:80px;overflow:hidden;text-overflow:ellipsis" class="list_name"><nobr class='list_vspace'>[* 익명 *]</nobr>
019   </div></td> <td align="left" class="list_vspace">
020     <img src=/images/icon_03.png border=0 align=absmiddle title="익명을 ">&ampnbsp<a href="view.php?id=insurance&page=1&divpage=10&no=61484" ><font class=list_title>암보험 생활비 받는 암
021     보험 괜찮은가요?</font></a>
022     <td nowrap class="eng list_vspace" colspan=2 title="17.08.21 11:38:13"><nobr class="eng list_vspace">11:38:13</nobr></td> <td nowrap class="eng list_vspace" colspan=2></td>
023   <td nowrap class="eng list_vspace" colspan=2>5</td></tr>
024
025
026
027 <tr align="center" class="list0" onMouseOver="this.style.backgroundColor='#F5F5F5'" onMouseOut="this.style.backgroundColor=''" style="height:16;word-break:break-all;" valign="middle">
028   <td class="eng list_vspace" colspan=2>45875</td>   <td class="han4 list_vspace" nowrap colspan=2><nobr class='han4 list_vspace'>질문</nobr></td>
029   <!--<td nowrap colspan=2 style="padding:0"><input type=checkbox name=cart value='61483'></td-->
030   <td colspan=2 class="list_vspace" align="left"><div style="width:80px;overflow:hidden;text-overflow:ellipsis" class="list_name"><nobr class='list_vspace'>[* 익명 *]</nobr>
031   </div></td> <td align="left" class="list_vspace">
032     <img src=/images/icon_06.png border=0 align=absmiddle title="모바일+익명을 ">&ampnbsp<a href="view.php?id=insurance&page=1&divpage=10&no=61483" ><font class=list_title>대아보험 질문이
033     요.</font></a>&ampnbsp<span class=list_comment></span> <span style="cursor:pointer;" onclick="win_comment('popup_comment.php?id=insurance&no=61483');">3</span> </span>
034     <td nowrap class="eng list_vspace" colspan=2 title="17.08.21 11:23:59"><nobr class="eng list_vspace">11:23:59</nobr></td> <td nowrap class="eng list_vspace" colspan=2></td>
035   <td nowrap class="eng list_vspace" colspan=2>13</td></tr>
036
037
038
039 <tr align="center" class="list1" onMouseOver="this.style.backgroundColor='#F5F5F5'" onMouseOut="this.style.backgroundColor=''" style="height:16;word-break:break-all;" valign="middle">
040   <td class="eng list_vspace" colspan=2>45874</td>   <td class="han4 list_vspace" nowrap colspan=2><nobr class='han4 list_vspace'>질문</nobr></td>
041   <!--<td nowrap colspan=2 style="padding:0"><input type=checkbox name=cart value='61480'></td-->
042   <td colspan=2 class="list_vspace" align="left"><div style="width:80px;overflow:hidden;text-overflow:ellipsis" class="list_name"><nobr class='list_vspace'>[* 익명 *]</nobr>
043   </div></td> <td align="left" class="list_vspace">
044     <img src=/images/icon_06.png border=0 align=absmiddle title="모바일+익명을 ">&ampnbsp<a href="view.php?id=insurance&page=1&divpage=10&no=61480" ><font class=list_title>비갱신형 암보
045     험..40대 후반</font></a>&ampnbsp<span class=list_comment></span> <span style="cursor:pointer;" onclick="win_comment('popup_comment.php?id=insurance&no=61480');">1</span> </span>
046     <td nowrap class="eng list_vspace" colspan=2 title="17.08.21 05:19:07"><nobr class="eng list_vspace">05:19:07</nobr></td> <td nowrap class="eng list_vspace" colspan=2></td>
047   <td nowrap class="eng list_vspace" colspan=2>43</td></tr>
```

<http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61484>

<http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61483>

<http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61480>

# Web Scraping: Open Forum

- Step 1: Check the structure of the URL

- ✓ Collect the URLs to the individual posts

```
# Extract the link of each post (for first 10 pages)
for( i in c(1:10)){
  tryCatch({ tmp_url <- paste(url, i, '&divpage=10', sep="")
    tmp_list <- read_html(tmp_url) %>% html_nodes('tr.list1') %>% html_nodes('a') %>
      html_attr('href')
  tmp_list <- paste0('http://www.ppomppu.co.kr/zboard/',tmp_list)
}
> tmp_list
[1] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61484"
[2] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61480"
[3] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61476"
[4] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61473"
[5] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61470"
[6] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61468"
[7] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61465"
[8] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61463"
[9] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61461"
[10] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61458"
[11] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61454"
[12] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61452"
[13] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61450"
[14] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61448"
[15] "http://www.ppomppu.co.kr/zboard/view.php?id=insurance&page=1&divpage=10&no=61445"
```

# Web Scraping: Open Forum

- Step 2: Collect the information
  - ✓ Title, date, and content from each post

The screenshot shows a Korean forum website with a navigation bar at the top. The main content area displays a post by a user named '태아보험 질문이요.' (Taemabohm Question). The post asks about joining Taemabohm insurance directly and whether it's cheaper than joining after birth. The user also wants to know if the single premium is higher than the monthly premium. The post has 13 likes and 0 replies. Below the post, there is a summary of the question and two numbered answers.

내년에 출생하는 아이가 있어서 태아보험을 가입하려는데 몇 가지 질문을 드릴게요.  
여기 게시판에서 보고 약간의 정보를 얻었는데 궁금한 점입니다.

1. 태아 보험 다이렉트로 가입하면 더 저렴한지.
2. 태아 보험 가입 후 출생 이후에는 해지하고 단독실비로 갈아타라는데 그렇게 하는 장점이 무엇인지 궁금합니다.  
그렇게 단독실비로 가입하면 가격적으로 더 저렴해지는지 혹은 커버가 더 많이 되는지 알고 싶어요.

도움 좀 부탁드리겠습니다.  
좋은 하루 되세요~

# Web Scraping: Open Forum

- Step 2: Collect the information
  - ✓ Title, date, and content from each post

- Title

```
653 <td valign=top nowrap style="padding-left:6px;line-height:140%;" class=han>
654 <font class=view_title2><!--DOM_TITLE-->태아보험 질문이요.<!--DOM_TITLE--></font>&nbsp;<sup><span class=list_comment>3</span></sup><br>
655 분류: <font class=view_cate>질문</font><br>
656 이름: <span title="">[* 익명 *]</b></span><br><img src=skin/DQ_Revolution_BBS/t.gif height=5 width=5 border=0><br>
657 <img src=skin/DQ_Revolution_BBS/t.gif height=5 width=5 border=0><br>
658 등록일: 2017-08-21 11:23<br>
```

- Date

```
649 <td valign=top class=han width=100 align=right><img src='/images/no_face.jpg' "></td>
650 <td width="6"></td>
651 <td class="separator2" width="3"></td>
652 <td width=3></td>
653 <td valign=top nowrap style="padding-left:6px;line-height:140%;" class=han>
654 <font class=view_title2><!--DOM_TITLE-->태아보험 질문이요.<!--DOM_TITLE--></font>&nbsp;<sup><span class=list_comment>3</span></sup><br>
655 분류: <font class=view_cate>질문</font><br>
656 이름: <span title="">[* 익명 *]</b></span><br><img src=skin/DQ_Revolution_BBS/t.gif height=5 width=5 border=0><br>
657 <img src=skin/DQ_Revolution_BBS/t.gif height=5 width=5 border=0><br>
658 등록일: 2017-08-21 11:23<br>
```

# Web Scraping: Open Forum

- Step 2: Collect the information
  - ✓ Title, date, and content from each post
    - Content

```
862 </script>
863
864 <table border="0" cellspacing="0" cellpadding="0" width="900" class="pic_bg">
865 <tr>
866   <td style="padding:0 8 0 8;" align="left">
867     <table width="100%" style="word-break:break-all;"><tbody><tr><td><!--DCM_BODY--><table border=0 cellspacing=0 cellpadding=0 width=100% style="table-layout:fixed;" align="left">
868 <tr>
869 <td class='board-contents' align="left" valign=top class=han>
870 내년에 출생하는 아이가 있어서 태아보험을 가입하려는데 몇 가지 질문을 드릴게요. <br />
871 여기 게시판에서 보고 약간의 정보를 얻었는데 궁금한 점입니다. <br />
872 <br />
873 1. 태아 보험 다이렉트로 가입하면 더 저렴한지. <br />
874 <br />
875 2. 태아 보험 가입 후 출생 이후에는 해지하고 단독실비로 갈마타라는데 그렇게 하는 장점이 무엇인지 궁금합니다. <br />
876 그렇게 단독실비로 가입하면 가격적으로 더 저렴해지는지 혹은 커버가 더 많이 되는지 알고 싶어요. <br />
877 <br />
878 도움 좀 부탁드리겠습니다. <br />
879 좋은 하루 되세요~<br />
880 <br />
```

# Web Scraping: Open Forum

- Step 2: Collect the information
  - ✓ Visit each page and collect the title, date, and content
  - ✓ To skip unexpected errors, use tryCatch function
  - ✓ Ex:
    - Do the instruction inside the tryCatch
    - If there is an error, store NULL to the title

```
# title
tryCatch({
  tmp_title <- repair_encoding(tmp_paragraph %>% html_nodes('font.view_title2') %>% html_text(T))
}, error = function(e){tmp_title <- NULL})
```

```
Best guess: UTF-8 (100% confident)
> tmp_title
[1] "태아보험 질문이요."
```

# Web Scraping: Open Forum

- Step 2: Collect the information
  - ✓ Visit each page and collect the title, date, and content

```
# date
tryCatch({
  tmp_date <- repair_encoding(tmp_paragraph %>% html_nodes('td.han') %>% html_text(T))[2]
  date_start_idx <- gregexpr(pattern = '등록일', tmp_date)[[1]][1] tmp_date <- substr(tmp_date,
  date_start_idx+5, date_start_idx+14)
}, error = function(e){tmp_date <- NULL})
```

- ✓ Before preprocessing

```
> tmp_date
[1] "태아보험 질문이요.?3\r\n\n분류: 질문\r\n\n이름: [* 익명 *] \r\n\n등록일: 2017-08-21 11:23\r\n\n\n조회수: 21 / 추천수: 0"
```

- ✓ After preprocessing

```
> tmp_date
[1] "2017-08-21"
```

# Web Scraping: Open Forum

- Step 2: Collect the information

- ✓ Visit each page and collect the title, date, and content

```
# contents
tryCatch({
  tmp_contents <- repair_encoding(tmp_paragraph %>% html_nodes('td.board-contents') %>%
  html_text(T))
  tmp_contents <- gsub("[[:punct:]]", " ", tmp_contents)
  tmp_contents <- gsub("[[:space:]]", " ", tmp_contents)
  tmp_contents <- gsub("\\s+", " ", tmp_contents)
  tmp_contents <- str_trim(tmp_contents, side = "both")
}, error = function(e){tmp_contents <- NULL})
```

- ✓ Before preprocessing

```
> tmp_contents
[1] "내년에 출생하는 아이가 있어서 태아보험을 가입하려는데 몇 가지 질문을 드릴게요. \n여기 게시판에서 보고 약간의 정보를 얻었는데 궁금한 점입니다. \n1. 태아 보험 다이렉트로 가입하면 더 저렴한지. \n2. 태아 보험 가입 후 출생 이후에는 해지하고 단독실비로 갈아타라는데 그럴게 하는 장점이 무엇인지 궁금합니다. \n그렇게 단독실비로 가입하면 가격적으로 더 저렴해지는지 혹은 커버가 더 많이 되는지 알고 싶어요. \n도움 좀 부탁드리겠습니다. \n좋은 하루 되세요~"
```

- ✓ After preprocessing

```
> tmp_contents
[1] "내년에 출생하는 아이가 있어서 태아보험을 가입하려는데 몇 가지 질문을 드릴게요 여기 게시판에서 보고 약간의 정보를 얻었는데 궁금한 점입니다 1 태아 보험 다이렉트로 가입하면 더 저렴한지 2 태아 보험 가입 후 출생 이후에는 해지하고 단독실비로 갈아타라는데 그럴게 하는 장점이 무엇인지 궁금합니다 그럴게 단독실비로 가입하면 가격적으로 더 저렴해지는지 혹은 커버가 더 많이 되는지 알고 싶어요 도움 좀 부탁드리겠습니다 좋은 하루 되세요"
```

# Web Scraping: Open Forum

- Step 2: Collect the information

- ✓ Store the information in the dataframe and export it to a CSV file

```
ppomppu_insurance[Npost,1] <- tmp_title  
ppomppu_insurance[Npost,2] <- tmp_date  
ppomppu_insurance[Npost,3] <- tmp_contents  
Npost <- Npost + 1
```

```
# Export the result  
write.csv(ppomppu_insurance, file = "ppomppu_insurance.csv")
```

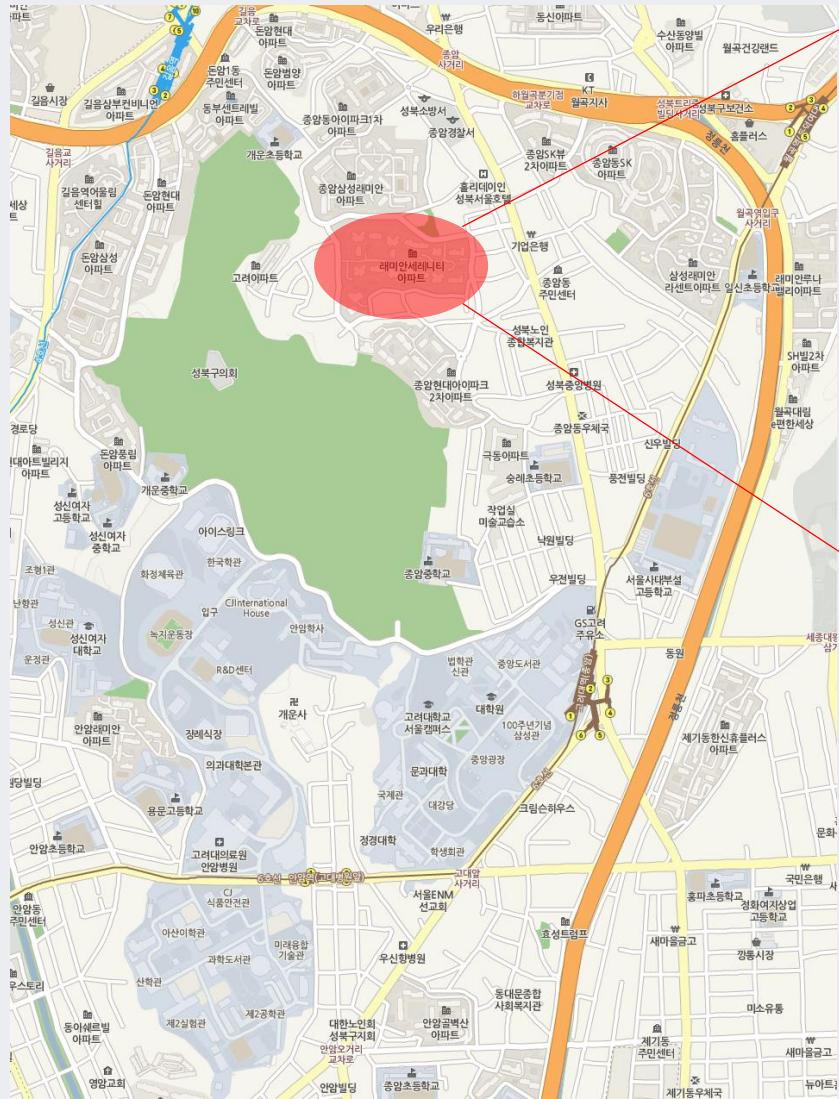
	V1	V2	V3
1	태아보험 질문이요.	2017-08-21	내년에 출생하는 아이가 있어서 태아보험을 가입하려는데 ...
2	단독실비 가입하려고 하는데 보험사 추천좀 부탁드리겠습니다...	2017-08-21	단독실비 인터넷으로 가입하려고 하는데요 피드백 빠르고 ...
3	일상생활배상책임 질문이요	2017-08-20	휴대폰액정을 깨트려서 제 실비에 있는 일상배상책임으로 ...
4	보험가입하려합니다.	2017-08-20	87년생 남자 직장인 실비보험 암보험 갱신형 가입금액 최...
5	30대중반 가장 생명보험 가입 문의	2017-08-20	안녕하세요 아침에 보험증권들 정리하다보니 제가 사망시 ...
6	주택화재보험 가입 시 보험사의 좋고 나쁨이 있나요?(가입 ...)	2017-08-19	안녕하세요 주택화재보험을 알아 보고 있습니다 이제 시작 ...
7	치아보험 문의 드립니다	2017-08-19	치아보험 보장 중에 충치치료도 받을 수 있는게 있나요 저...
8	보험 추천 부탁드립니다.	2017-08-18	만 30세 남 살비있음 만 31세 여 두명 암 심장 뇌 같은 보험 ...
9	여행자 보험 질문	2017-08-18	교환학생으로 인도네시아에 6개월 가량 체류할 예정입니다...
10	교보 생명 보험에 관해 몇 자 여쭙습니다.	2017-08-18	10년정도 납입한 교보생명 보험이 있습니다 지인분 추천으...

# AGENDA

- 01 Collect Data using APIs: Twitter
- 02 Collect Data using APIs: Facebook
- 03 Web Scraping: ArXiv Research Papers
- 04 Web Community: PPOMPPU
- 05 Website: NAVER Real Estate

# Web Scraping: Portal Site

- Step 0: Decide the target apartment



# Web Scraping: Portal Site

## • Step 0: Decide the target apartment

- ✓ [http://land.naver.com/article/articleList.nhn?rletTypeCd=A01&tradeTypeCd=A1&rletNo=25827&cortarNo=1129013500&hscpTypeCd=A01%3AA03%3AA04&mapX=&mapY=&mapLevel=&page=1&articlePage=&ptpNo=&rltrId=&mnex=&bildNo=&articleOrderCode=&cplId=&period=&prodTab=&atclNo=&atclRletTypeCd=&location=2160&bbs\\_tp\\_cd=&sort=&siteOrderCode=&schlCd=&tradYy=&exclsSpc=&splSpCnR=&cmplYn=#\\_content\\_list\\_target](http://land.naver.com/article/articleList.nhn?rletTypeCd=A01&tradeTypeCd=A1&rletNo=25827&cortarNo=1129013500&hscpTypeCd=A01%3AA03%3AA04&mapX=&mapY=&mapLevel=&page=1&articlePage=&ptpNo=&rltrId=&mnex=&bildNo=&articleOrderCode=&cplId=&period=&prodTab=&atclNo=&atclRletTypeCd=&location=2160&bbs_tp_cd=&sort=&siteOrderCode=&schlCd=&tradYy=&exclsSpc=&splSpCnR=&cmplYn=#_content_list_target)

[종암동] 래미안세레나티 관심단지등록   단지매물전체							
매물	시세	실거래가	단지정보	평면도	학군정보	관리비	커뮤니티
● 전체	● 매매	● 전세	● 월세	● 단기임대			
<b>확인매물</b> [?] 거래완료 매물 포함 <input checked="" type="checkbox"/>							
거래	확인일자	매물명	면적(m <sup>2</sup> )	동	층	매물가(만원)	연락처
매매	<span>확인마물 17.08.17.</span>	<span>현장 래미안세레나티</span>	84C/59	203동	2/21	<b>48,500</b>	에이스공인중개…
		확장형 상태아주깨끗 이사일협의				매경부동산	02-921-3100
매매	<span>확인마물 17.08.18.</span>	<span>점주인 래미안세레나티</span>	111B/84	214동	3/24	<b>56,000</b>	에이스공인중개…
		상태최상 시스템에어컨 인테리어된확장형				매경부동산	02-921-3100
매매	<span>확인마물 17.08.16.</span>	<span>점주인 래미안세레나티</span>	110A/84	205동	6/18	<b>59,500</b>	심순례공인중개사
		남향의 아담한 정원이 있는집, 선호하는…				매경부동산	02-953-6633
매매	<span>확인마물 17.08.14.</span>	<span>점주인 래미안세레나티</span>	143/114	210동	14/24	<b>70,500</b>	우리공인중개사
		풀확장 깊깨끗 임대승계 만기18년4월보…				매경부동산	02-953-1116
매매	<span>확인마물 17.08.14.</span>	<span>점주인 래미안세레나티</span>	143/114	206동	1/18	<b>64,000</b>	부동산114삼광…
		정남향 1층 확장 아이들키우기좋음 채광굿				매경부동산	02-921-7600
매매	<span>거리원료 17.08.17.</span>	<span>점주인 래미안세레나티</span>	109C/84	207동	24/24	<b>57,000</b>	심순례공인중개사
		탑층, 전망 좋고 환기 잘됨.				매경부동산	02-953-6633
매매	<span>거리원료 17.08.17.</span>	<span>점주인 래미안세레나티</span>	109C/84	207동	24/24	<b>↓ 57,000</b>	에이스공인중개…
		탁트인조망 시원한통풍 탑층으로조용 상…				매경부동산	02-921-3100

# Web Scraping: Portal Site

- Step 1: Check the structure of the URL

- ✓ Check the part that changes with regard to the pages

- [http://land.naver.com/article/articleList.nhn?rletTypeCd=A01&tradeTypeCd=A1&rletNo=25827&cortarNo=1129013500&hscpTypeCd=A01%3AA03%3AA04&mapX=&mapY=&mapLevel=&page=1&articlePage=&ptpNo=&rltrId=&mnex=&bildNo=&articleOrderCode=&cpld=&period=&prodTab=&atclNo=&atclRletTypeCd=&location=2160&bbs\\_tp\\_cd=&sort=&siteOrderCode=&schlCd=&tradYy=&exclsPpc=&splySpcR=&cmplyYn=#\\_content\\_list\\_target](http://land.naver.com/article/articleList.nhn?rletTypeCd=A01&tradeTypeCd=A1&rletNo=25827&cortarNo=1129013500&hscpTypeCd=A01%3AA03%3AA04&mapX=&mapY=&mapLevel=&page=1&articlePage=&ptpNo=&rltrId=&mnex=&bildNo=&articleOrderCode=&cpld=&period=&prodTab=&atclNo=&atclRletTypeCd=&location=2160&bbs_tp_cd=&sort=&siteOrderCode=&schlCd=&tradYy=&exclsPpc=&splySpcR=&cmplyYn=#_content_list_target)
    - [http://land.naver.com/article/articleList.nhn?rletTypeCd=A01&tradeTypeCd=A1&rletNo=25827&cortarNo=1129013500&hscpTypeCd=A01%3AA03%3AA04&mapX=&mapY=&mapLevel=&page=2&articlePage=&ptpNo=&rltrId=&mnex=&bildNo=&articleOrderCode=&cpld=&period=&prodTab=&atclNo=&atclRletTypeCd=&location=2160&bbs\\_tp\\_cd=&sort=&siteOrderCode=&schlCd=&tradYy=&exclsPpc=&splySpcR=&cmplyYn=#\\_content\\_list\\_target](http://land.naver.com/article/articleList.nhn?rletTypeCd=A01&tradeTypeCd=A1&rletNo=25827&cortarNo=1129013500&hscpTypeCd=A01%3AA03%3AA04&mapX=&mapY=&mapLevel=&page=2&articlePage=&ptpNo=&rltrId=&mnex=&bildNo=&articleOrderCode=&cpld=&period=&prodTab=&atclNo=&atclRletTypeCd=&location=2160&bbs_tp_cd=&sort=&siteOrderCode=&schlCd=&tradYy=&exclsPpc=&splySpcR=&cmplyYn=#_content_list_target)
    - [http://land.naver.com/article/articleList.nhn?rletTypeCd=A01&tradeTypeCd=A1&rletNo=25827&cortarNo=1129013500&hscpTypeCd=A01%3AA03%3AA04&mapX=&mapY=&mapLevel=&page=3&articlePage=&ptpNo=&rltrId=&mnex=&bildNo=&articleOrderCode=&cpld=&period=&prodTab=&atclNo=&atclRletTypeCd=&location=2160&bbs\\_tp\\_cd=&sort=&siteOrderCode=&schlCd=&tradYy=&exclsPpc=&splySpcR=&cmplyYn=#\\_content\\_list\\_target](http://land.naver.com/article/articleList.nhn?rletTypeCd=A01&tradeTypeCd=A1&rletNo=25827&cortarNo=1129013500&hscpTypeCd=A01%3AA03%3AA04&mapX=&mapY=&mapLevel=&page=3&articlePage=&ptpNo=&rltrId=&mnex=&bildNo=&articleOrderCode=&cpld=&period=&prodTab=&atclNo=&atclRletTypeCd=&location=2160&bbs_tp_cd=&sort=&siteOrderCode=&schlCd=&tradYy=&exclsPpc=&splySpcR=&cmplyYn=#_content_list_target)

# Web Scraping: Portal Site

## • Step 1: Check the structure of the URL

- ✓ Check the URL of each apartment in the page

```
tmp_url <- modify_url(url, query = list(page = i))
tmp_list <- read_html(tmp_url) %>% html_nodes('a[href^="/article"]') %>% html_attr('href')
tmp_list <- paste0('http://land.naver.com', tmp_list)
```

Screenshot of a real estate portal website showing a list of apartments for sale. The page includes filters for location, price, and type. A specific listing is highlighted with a red border.

거래	확인일자	매물명	면적(m <sup>2</sup> )	동	층	매물가(만원)	연락처
매매	17.08.17	현대 래미안세레나티	84C/59	203동	2/21	48,500	에이스공인중개…
		확장형 상태아주깨끗 미사일협의				매경부동산	02-921-3100
매매	17.08.18	집주인 래미안세레나티	111B/84	214동	3/24	56,000	에이스공인중개…
		상태최상 시스템에어컨 인테리어된확장형				매경부동산	02-921-3100
매매	17.08.16	집주인 래미안세레나티	110A/84	205동	6/18	59,500	심순례공인중개사…
		남향의 아담한 정원이 있는집, 선호하는 …				매경부동산	02-953-6633
매매	17.08.14	집주인 래미안세레나티	143/114	206동	1/18	64,000	부동산114삼광…
		정남향 1층 확장 아이들키우기좋음 채광굿				매경부동산	02-921-7600
매매	17.08.14	집주인 래미안세레나티	143/114	210동	14/24	70,500	우리공인중개사…
		출화장 집깨끗 입대승계 만기18년4월 보…				매경부동산	02-953-1116
매매	17.08.17	집주인 래미안세레나티	109C/84	207동	24/24	57,000	심순례공인중개사…
		탑층, 전망 좋고 환기 잘됨.				매경부동산	02-953-6633

HTML code for the highlighted listing (Row 1):

```
<a href="/article/articleDetailInfo.nhn?atclNo=1711225034&atclRletTypeCd=A01&rletTypeCd=A01&rateTypeCd=A1" class="NP1=a:t!listnew,r:1,i:1711225034 btn_newpage" target="_blank"></a>
```

```
</td>
<td class="num" onmouseover="$Element($.getSingle('div.inner',this)).css('zIndex', '10');$Element($.getSingle('div.rate_layer',this)).show();" onmouseout="$Element($.getSingle('div.inner',this)).css('zIndex', '0');$Element($.getSingle('div.rate_layer',this)).hide();">
<div class="inner" tabIndex="0">
  84C/59
  <div class="rate_layer" style="display:none;">
    <div class="layer">
      <p class="calc_area">
        공급면적 84.88m2
      <br>
      전용면적 59.92m2
      <br>
    </p>
  </div>
</div>
</td>
<td class="num2" title="203동"><div class="inner">203동</div></td>
<td class="num2"><div class="inner" tabIndex="0"><span>2/21</span></div></td>
<td class="num align_r">
<div class="inner">
  <strong title="48,500">48,500</strong>
</div>
</td>
```

# Web Scraping: Portal Site

## • Step 1: Check the structure of the URL

- ✓ Check the URL of each apartment in the page

```
tmp_url <- modify_url(url, query = list(page = i))
tmp_list <- read_html(tmp_url) %>% html_nodes('a[href^="/article"]') %>% html_attr('href')
tmp_list <- paste0('http://land.naver.com', tmp_list)
```

[종암동] 래미안세레나티 관심단지등록 | 단지매물전체

매물	시세	설거래가	단지정보	평면도	학군정보	관리비	커뮤니티	대출
<input type="radio"/> 전체	<input checked="" type="radio"/> 매매	<input type="radio"/> 전세	<input type="radio"/> 월세	<input type="radio"/> 단기임대				

확인매물 ? 거래완료 매물 포함

거래	확인일자	매물명	면적(m <sup>2</sup> )	동	층	매물가(만원)	연락처
매매	17.08.17	현장 래미안세레나티	84C/59	203동	2/21	48,500	에이스공인중개…
		확장형 상태아주깨끗 이사일협의				매경부동산	02-921-3100
매매	17.08.18	집주인 래미안세레나티	111B/84	214동	3/24	56,000	에이스공인중개…
		상태최상 시스템에어컨 인테리어된확장형				매경부동산	02-921-3100

매매 17.08.16 집주인 래미안세레나티 110A/84 205동 6/18 59,500 심순례공인중개사

남향의 아담한 정원이 있는집, 선호하는 …

매경부동산 02-953-6633

매매 17.08.14 집주인 래미안세레나티 143/114 206동 1/18 64,000 부동산114삼광…

정남향 1층 확장 아이들키우기좋음 채광굿

매경부동산 02-921-7600

매매 17.08.14 집주인 래미안세레나티 143/114 210동 14/24 70,500 우리공인중개사

출확장 짐깨끗 임대승계 만기18년4월 보…

매경부동산 02-953-1116

매매 17.08.17 집주인 래미안세레나티 109C/84 207동 24/24 57,000 심순례공인중개사

탑층, 전망 좋고 환기 잘됨.

매경부동산 02-953-6633

```
<a href="/article/articleDetailInfo.nhn?atcINo=1711298355&atcIRletTypeCd=A01&rletTypeCd=A01&rateTypeCd=A1" class="NPI=a:listnew_r2_i:1711298355 btn_newpage" target="_blank"></a>
</td>
<td class="num" onmouseover="$Element($$.getSingle('div.inner',this)).css('zIndex', '10');$Element($$.getSingle('div.rate_layer',this)).show();" onmouseout="$Element($$.getSingle('div.inner',this)).css('zIndex', '0');$Element($$.getSingle('div.rate_layer',this)).hide();">
<div class="inner" tabIndex="0">
  111B/84
  <div class="rate_layer" style="display:none;">
    <div class="layer">
      <p class="calc_area">
        공급면적 111.11m2
        <br>
        전용면적 84.99m2
        <br>
      </p>
    </div>
  </div>
</div>
<td class="num2" title="214동"><div class="inner">214동</div></td>
<td class="num2"><div class="inner" tabIndex="0"><span>3/24</span></div></td>
<td class="num align_r">
  <div class="inner">
    <strong title="56,000">56,000</strong>
  </div>
</td>
```

# Web Scraping: Portal Site

- Step 2: Collect the information

- ✓ Building No., Top-floor, Floor, Area I(provided), Area 2(Used), Price

Scraped data from a real estate portal site (www.kakao.com) showing a listing for a building.

**Building Details:**

관심마을등록	위치매출신고
[인쇄]	
[날짜]	
[현장확인]	
[부동산마켓검증센터]	
[2017년 08월 17일]	
[매물번호: 203동 단지별물건 전체]	
[현장확인]	
[부동산마켓검증센터에서 2017년 08월 17일 현장확인한 매물입니다.]	
[공급/전용면적 (전용률 70.8%) [단위: m <sup>2</sup> ]]	
[84.88 / 59.92 m <sup>2</sup> ]	
[매매가]	
[48,500 만원 (1,887만원/3.3평)]	

**Transaction Status:**

해당층/총층	2/21층
방수/욕실수	3/2개
용자금	- 만원
입주가능일	2개월 이내
기보증금/월세	-/- 만원
특징	확장형 상대아주깨끗 이사임업의
증개업소	에이스공인중개사사무소 02-921-3100   소재지 - 대표: 김복희 010-5230-2733

**Location Map:**

위치: 평택도 현장확인 단지사진 (1/21)

서울 성북구 종암동 133 도로명 네이버지도 지적면집도 일반 위성

지도내 주요 건물과 위치 표시. 예상 거리와 방향 표시.

**Building Information:**

층개보수(VAT별도)	최대 194 만원
상한율	0.4%
• 층개보수는 실제 금액과 다를 수 있습니다.	
[보기]	
[현장확인: 3 매물: 25]	
[위치보기]	

**Financial Summary:**

취득세	485 만원
지방교육세	49 만원
농어촌특별세	- 만원
등기신청 수수료 및 수입인지	165,000 원
국민주택채권	약 189,150 원
공과금 합계	약 5,694,150 원
• 공과금은 실제 금액과 다를 수 있습니다.	
[보기]	

**Additional Features:**

AD: 등기건적 무료로 조회하고 할인받자!

초등학교 학군정보 [보기]

# Web Scraping: Portal Site

- Step 2: Collect the information

✓ Building No., Top-floor, Floor, Area I (provided), Area 2(Used), Price

```
# Building
tryCatch({
  tmp_building <- repair_encoding(tmp_paragraph %>%
  html_nodes('div.info_area.info_area_v2.first') %>%
  html_nodes('div.inner') %>% html_text(T))
  tmp_building <- as.numeric(substr(tmp_building[2], 1, 3))
}, error = function(e){tmp_title <- NA})
```

```
<div class="p_area _js_image_layout image360_wrapper NE=a:map" style="height:508px;z-index:20"></div>
<div class="info_area info_area_v2 first">
  <div class="info_wrap">
    <div class="sale_detail">
      <h3 class="t_end t_sale"><span>매물세부정보</span></h3>
      <div class="tbl_end">
        <table cellspacing="0" border="1" summary="매물세부정보">
          <caption>매물세부정보</caption>
          <colgroup>
            <col width="111"><col width="">
          </colgroup>
          <tbody>
            <tr>
              <th scope="row"><div class="inner">해당동</div></th>
              <td><div class="inner">203동</div></td>
            </tr>
          </tbody>
        </table>
      </div>
    </div>
  </div>
</div>
```

# Web Scraping: Portal Site

- Step 2: Collect the information

- ✓ Building No., Top-floor, Floor, Area 1 (provided), Area 2(Used), Price

```
# Floor tryCatch({  
  tmp_info <- repair_encoding(tmp_paragraph %>% html_nodes('div.detail_view.detail_view_v2') %>%  
    html_nodes('div.inner') %>% html_text(T))  
  tmp_floor_info <- strsplit(tmp_info[2], "/")  
  tmp_floor <- tmp_floor_info[[1]][1]  
}, error = function(e){tmp_floor <- NA})  
  
# Top_Floor  
tryCatch({  
  tmp_info <- repair_encoding(tmp_paragraph %>% html_nodes('div.detail_view.detail_view_v2') %>%  
    html_nodes('div.inner') %>% html_text(T))  
  tmp_floor_info <- strsplit(tmp_info[2], "/")  
  tmp_top_floor <- as.numeric(substr(tmp_floor_info[[1]][2], 1, nchar(tmp_floor_info[[1]][2])-1))  
}, error = function(e){tmp_top_floor <- NA})
```



```
<!-- [D] 시세 있을 경우 detail_view_v2 추가 -->  
<div class="detail_view detail_view_v2">  
  <div class="view_info">  
    <table cellspacing="0" border="1" summary="매물정보">  
      <caption>매물정보</caption>  
      <colgroup>  
        <col width="90"><col><col width="90"><col><col width="90"><col>  
      </colgroup>  
      <tbody>  
        <tr>  
          <th scope="row"><div class="inner">해당층/총층</div></th>  
          <td>  
            <div class="inner">2/21층</div>
```

# Web Scraping: Portal Site

- Step 2: Collect the information

✓ Building No., Top-floor, Floor, Area 1(provided), Area 2(Used), Price

```
<script language="JavaScript" type="text/javascript">
var jsonPageData = {
  isIpad : "false",
  imgUriService : "http://static.land.naver.net/static/service/20170810",
  tradeTypeCd : "A1",
  rletTypeCd : "A01",
/* location : "", */
  atclNo : "1711225034",
  hscp_yn : "Y",
  cityNo : "1100000000",
  dvsnNo : "1129000000",
  secNo : "1129013500",
  cityNm : "서울시",
  dvsnNm : "성북구",
  secNm : "종암동",
  pricePeriod : 1,
  priceTradTpCd : "A1",
  atclPrice : "48500",
  cortarNo : "1129013500",
  rletNo : "25827",
  defaultRltrMbrIds : "",
  totalCount : "",
  curCortarNo : "1129013500",
  cortarMapX : "127.0383",
  cortarMapY : "37.5966",
  query : "",
  Hcortar_no : "",
  Hhscp_nm : "",
  Hhscp_no : "",
  Harticle_cnt : "",
  Hcomplex_cnt : "", */
  atclNm : "래미안세레나티 203동",
  roadAddr : "종암로23길 35",
  rltrNm : "에이스공인중개사사무소",
  cpid : "bizmk",
  complexDongHo : '',
  complexDongList : '',
  complexDongPtpList : ''
};
```

# Web Scraping: Portal Site

- Step 2: Collect the information

- ✓ Building No., Top-floor, Floor, Area I (provided), Area 2(Used), Price

```
# Price
tryCatch({
  tmp_price <- tmp_paragraph %>% html_nodes(xpath = "//script[@language='JavaScript']") %>%
  html_text(T)
  price_start_idx <- gregexpr(pattern = 'atclPrice', tmp_price)[[1]][1]
  tmp_price <- substr(tmp_price, price_start_idx+13, price_start_idx+17)
}, error = function(e){tmp_price <- NA})
```

```
> tmp_price
[1] "var jsonPageData = {\n\t\tisIpad : \"false\", \n\t\timgUrlService : \"http://static.land.naver.net/static/service/20170810\", \n\t\ttradeTypeCd : \"A1\", \n\t\trelTypeCd : \"A01\", \n\t/* \tlocation : \"\", */\n\t\tatclNo : \"1711225034\", \n\t\thcp_yn : \"Y\", \n\t\tcityNo : \"1100000000\", \n\t\tdvsnNo : \"1129000000\", \n\t\tsecNo : \"1129013500\", \n\t\tcityNm : \"서울시\", \n\t\tdvsnNm : \"성북구\", \n\t\tsecNm : \"종암동\", \n\t\tpricePeriod : 1, \n\t\tpriceTradTpCd : \"A1\", \n\t\tatclPrice : \"48500\", \n\t\tcortarNo : \"1129013500\", \n\t\trelNo : \"25827\", \n\t\tdefaultRltrMbrIds : \"\", \n\t\ttotalCount : \"\", \n\t\tcurCortarNo : \"1129013500\", \n\t\tcortarMapX : \"127.0333\", \n\t\tcortarMapY : \"37.5966\", \n\t\tquery : \"\", \n\t/* \tHcortar_no : \"\", \n\t\tHscp_nm : \"\", \n\t\tHhcp_no : \"\", \n\t\tHarticle_cnt : \"\", \n\t\tHcomplex_cnt : \"\", \n\t\tatclNm : \"래미안세레니티 203동\", \n\t\troadAddr : \"종암로23길 35\", \n\t\trelNm : \"에이스공인증개사사무소\", \n\t\tcpid : \"bizmk\", \n\t\tcomplexDongHo : ''... <truncated
```

# Web Scraping: Portal Site

- Step 2: Collect the information

- ✓ Building No., Top-floor, Floor, Area 1 (provided), Area 2(Used), Price

```
<div class="ly_tb">
  <table cellspacing="0" border="1" summary="평형정보">
    <caption>평형정보</caption>
    <colgroup>
      <col width="84"><col width="58"><col width="58">
    </colgroup>
    <thead>
      <tr>
        <th scope="col" class="frst"><strong>면적단위</strong></th>

        <th scope="col">공급면적</th>

        <th scope="col">전용면적</th>
      </tr>
    </thead>
    <tbody>
      <tr >
        <td class="frst "><strong>평</strong></td>

        <td><strong>25.68</strong></td>

        <td><strong>18.13</strong></td>
    
```

# Web Scraping: Portal Site

- Step 2: Collect the information

- ✓ Building No., Top-floor, Floor, Area 1 (provided), Area 2(Used), Price

```
# Areas
tryCatch({
  tmp_area <- tmp_paragraph %>% html_nodes('div.ar_area') %>% html_nodes('div.ly_tbl') %>%
  html_text(T)
  tmp_area <- gsub("\t", "", tmp_area)
  tmp_area <- gsub("\n", " ", tmp_area)
  tmp_area <- strsplit(tmp_area, " ")[[1]]
  del_idx <- which(nchar(tmp_area) == 0)
  tmp_area <- tmp_area[-del_idx]
  tmp_Area1 <- as.numeric(tmp_area[6])
}, error = function(e){tmp_Area1 <- NA})

tryCatch({
  tmp_area <- tmp_paragraph %>% html_nodes('div.ar_area') %>% html_nodes('div.ly_tbl') %>%
  html_text(T)
  tmp_area <- gsub("\t", "", tmp_area)
  tmp_area <- gsub("\n", " ", tmp_area)
  tmp_area <- strsplit(tmp_area, " ")[[1]]
  del_idx <- which(nchar(tmp_area) == 0)
  tmp_area <- tmp_area[-del_idx]
  tmp_Area2 <- as.numeric(tmp_area[7])
}, error = function(e){tmp_Area2 <- NA})
```

# Web Scraping: Portal Site

- Step 2: Collect the information

- ✓ Store the information in the dataframe and export it to a CSV file

```
APT <- data.frame(Building, Top_Floor, Floor, Area1, Area2, Price)
```

```
# Export the result
```

```
write.csv(APT, file = "APT_Price.csv")
```

	Building	Top_Floor	Floor	Area1	Area2	Price
1	203	21	2	25.68	18.13	48500
2	214	24	3	33.61	25.71	56000
3	205	18	6	33.50	25.69	59500
4	210	24	14	43.50	34.78	70500
5	206	18	1	43.50	34.78	64000
6	207	24	24	33.26	25.71	57000
7	207	24	24	33.26	25.71	57000
8	205	18	12	43.50	34.78	68000
9	210	24	24	43.50	34.78	69000
10	214	24	5	33.26	25.71	57000
11	207	24	24	33.26	25.71	57000
12	207	24	14	33.61	25.71	58500
13	207	24	14	33.61	25.71	58500
14	214	24	5	33.26	25.71	57000
15	207	24	14	33.61	25.71	58500



ANY  
questions?