Lecture 2: From Texts to Data

Pilsung Kang

School of Industrial Management Engineering

Korea University

AGENDA

01 Collect Data using APIs: Twitter

02 Collect Data using APIs: Facebook

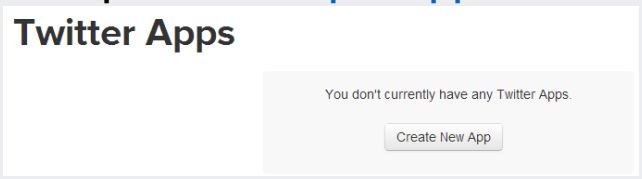
03 Web Scraping: ArXiv Research Papers

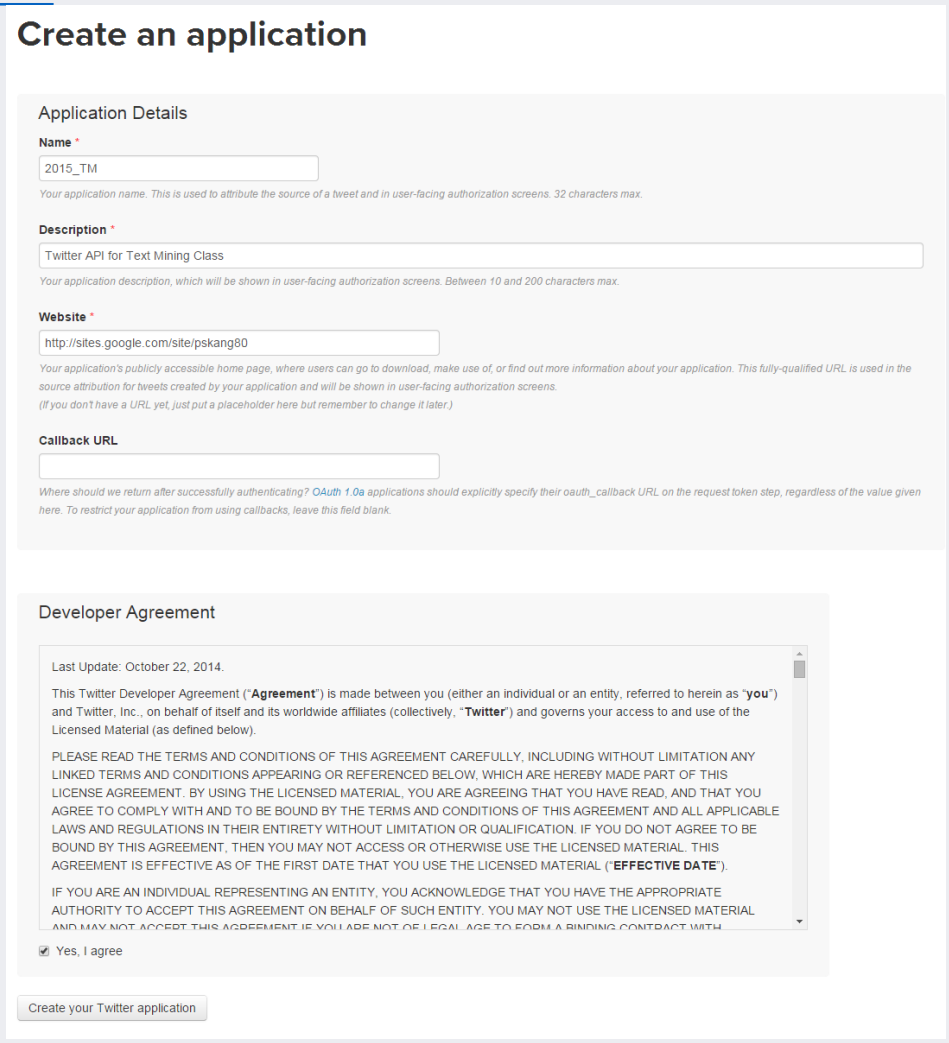
04 Web Community: PPOMPPU

05 Website: NAVER Real Estate

Collect Twitter Mentions

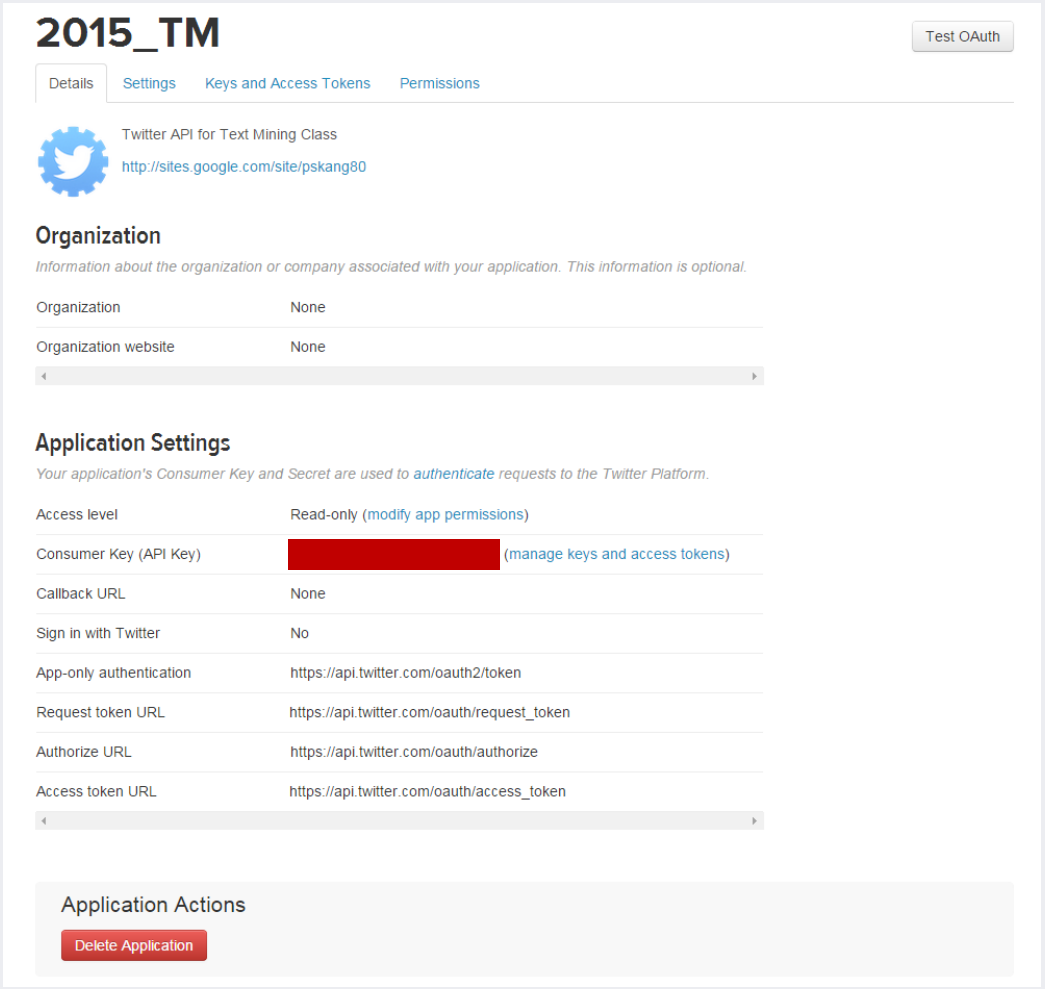
- Get an authorized authentication
  - ✓ Step 1: visit https://apps.twitter.com/





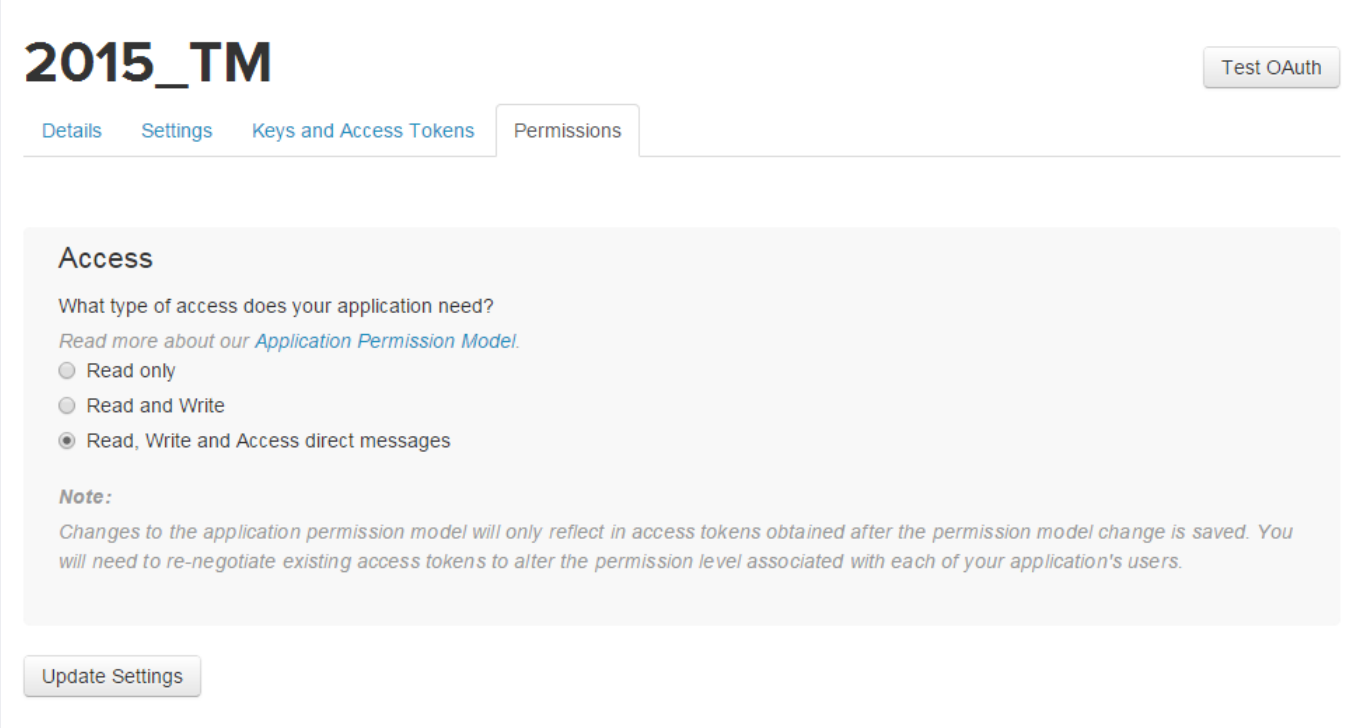
Collect Twitter Mentions

- Get an authorized authentication
  - ✓ Step 2: record necessary URLs



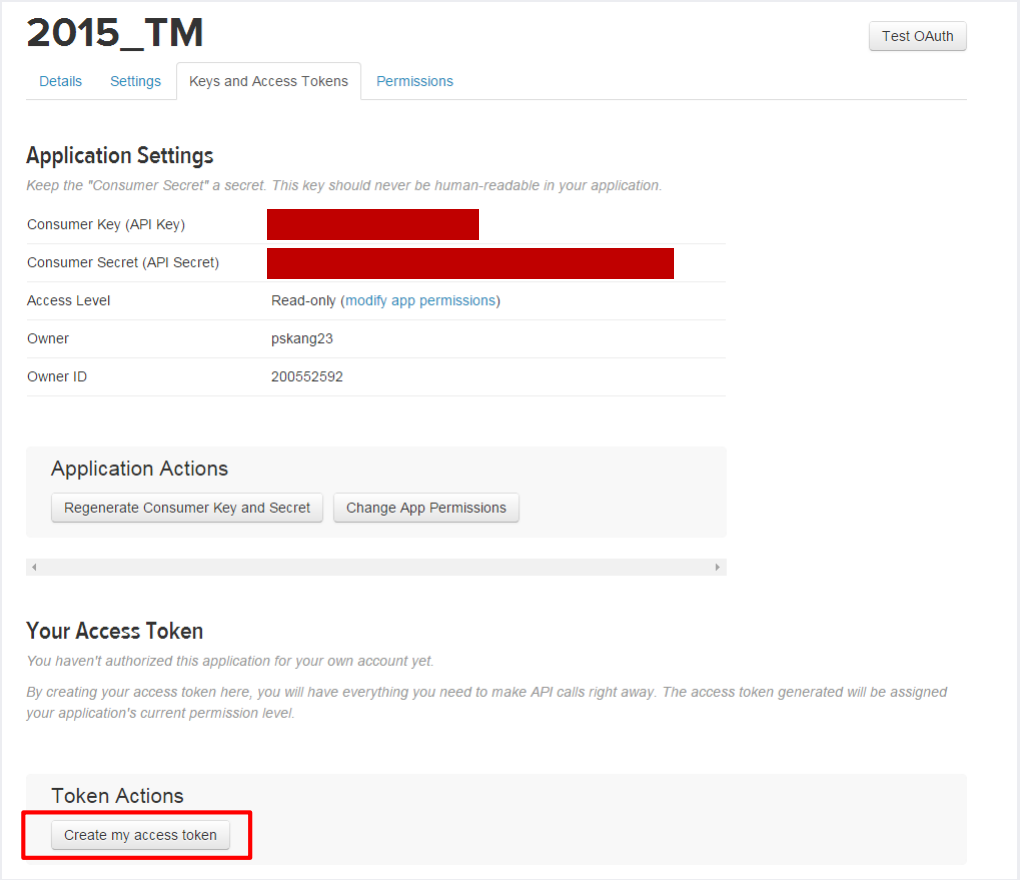
Collect Twitter Mentions

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



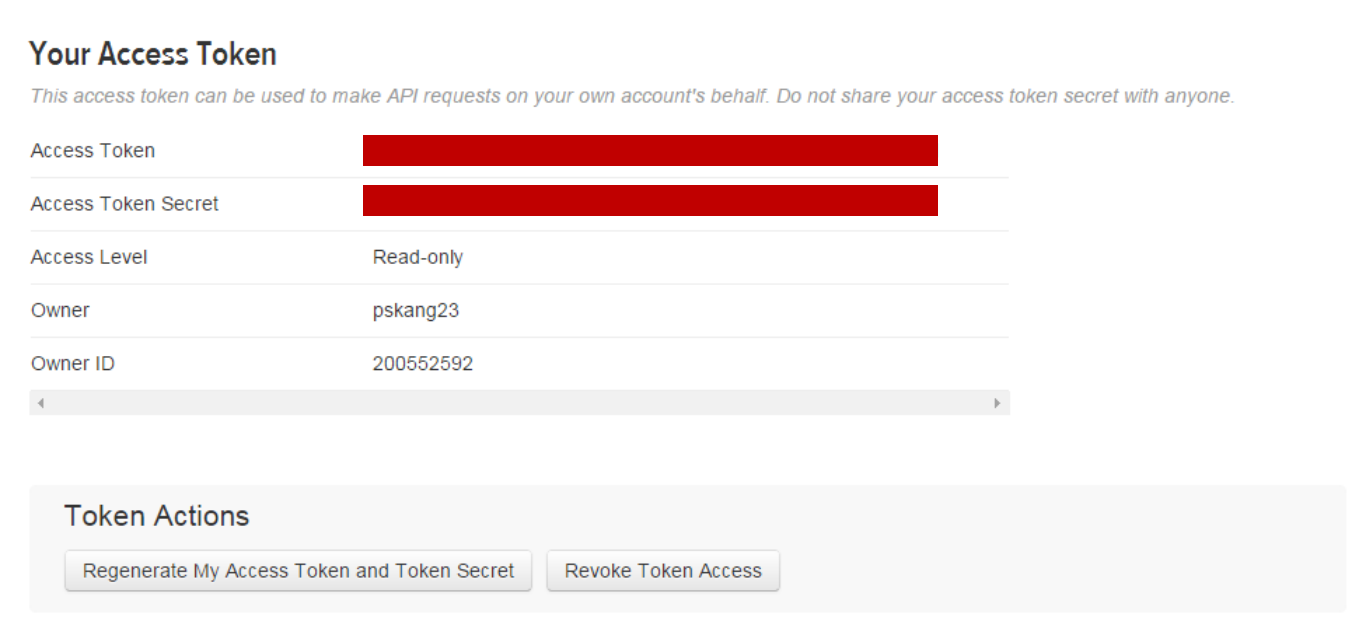
Collect Twitter Mentions

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



Collect Twitter Mentions

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



Collect Twitter Mentions

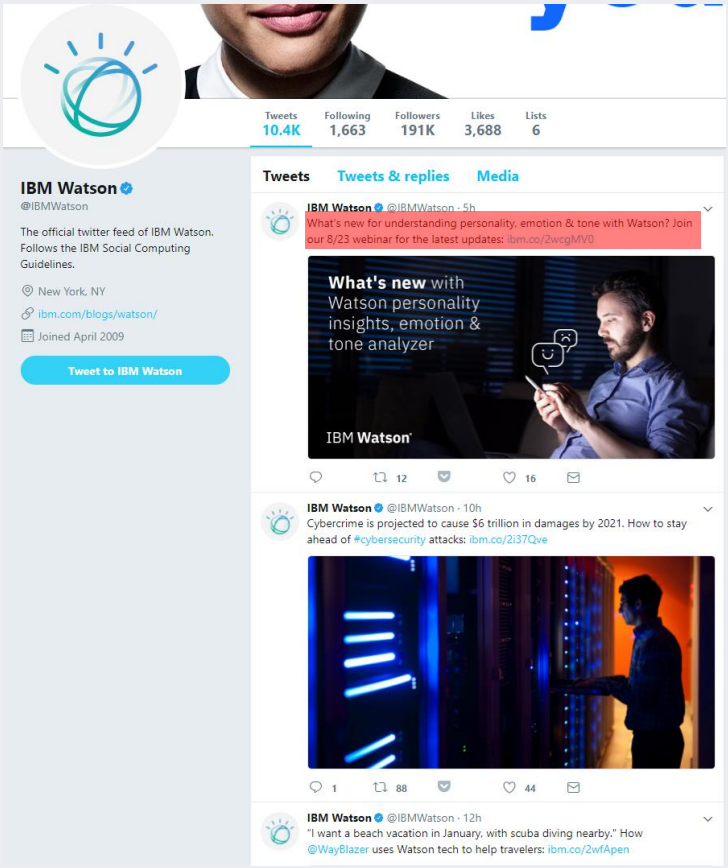
- Get an authorized authentication
  - ✓ Step 4: complete the authentication process (for twitteR)
    - Provide consumer_key, consumer_secret, access_token, access_secret information with **setup_twitter_oauth** function

```
# 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]
```

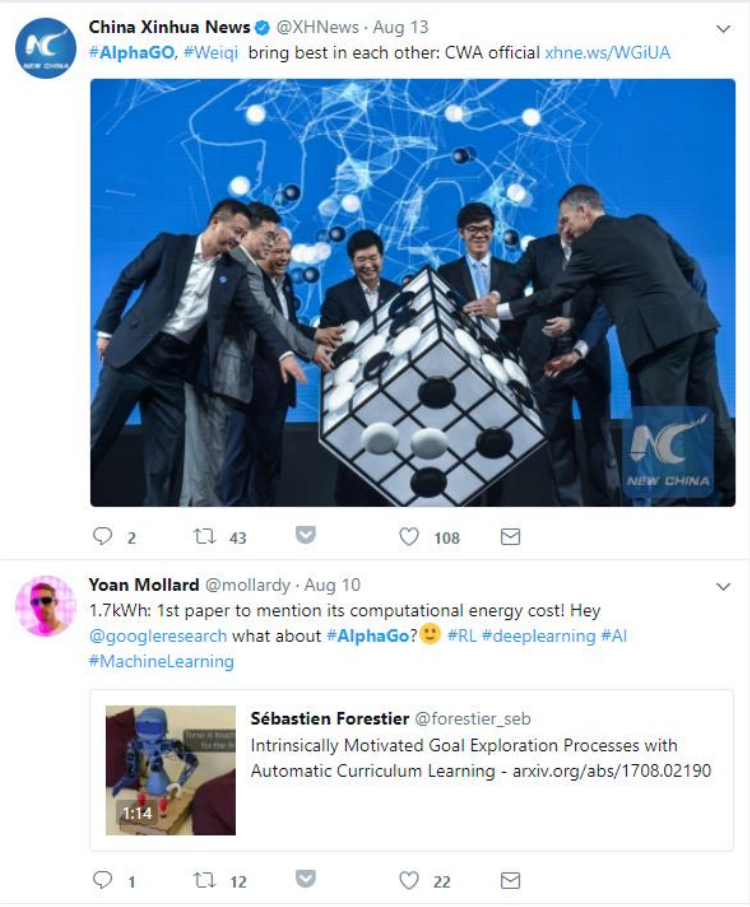



```
> WatsonTweets[1:10]
[[1]]
[1] "IBMWatson: What's new for understanding personality, emotion &amp;amp; tone with Watson
? Join our 8/23 webinar for the latest updates:… https://t.co/T4t5MCPXd2"

[[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/1bNXhOmw0d 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/7xGGfekrxF"

[[8]]
[1] "IBMWatson: How Watson's #AI is helping companies stay ahead of hackers and #cyberse
curity attacks: https://t.co/QQPRoOPPSf… 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/YrSHN2uo1i #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]
```

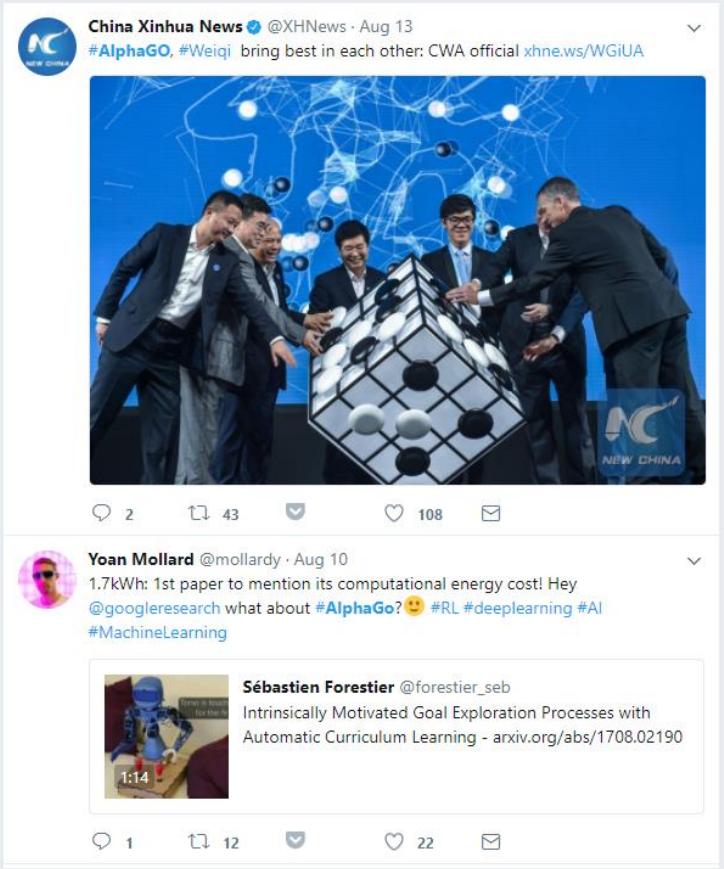



```
> AlphaGoTweets[1:10]
[[1]]
[1] "JBYoung64: RT @pfizer: What do #selfdrivingcars &amp;amp; #AlphaGo have to do with medi
cine? Learn more about how we're using #AI to drive research https://t.…"

[[2]]
[1] "calcaware: RT @bradbonomo: Researchers testing #algorithms that display human-like
ingenuity. #AI #AlphaGo #DeepMind #TechNews #MachineLearning https:…"

[[3]]
[1] "fernandocuenca: RT @bradbonomo: Researchers testing #algorithms that display human-
like ingenuity. #AI #AlphaGo #DeepMind #TechNews #MachineLearning https:…"

[[4]]
[1] "brandperson2: RT @bradbonomo: Researchers testing #algorithms that display human-li
ke ingenuity. #AI #AlphaGo #DeepMind #TechNews #MachineLearning https:…"

[[5]]
[1] "Deep_In_Depth: Google's AI Completely Destroyed a 19-Year-Old. Then He Gave This Ep
ic Response  https://t.co/eheV0ypICu… https://t.co/iRdVQYAzy4"

[[6]]
[1] "Deep_In_Depth: Mastering the game of Go with deep neural networks and tree search
https://t.co/wUXVPbOw9w #DeepLearning… https://t.co/mZnErxlSLP"

[[7]]
[1] "CarlosMBorbon: RT @pfizer: What do #selfdrivingcars &amp;amp; #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 &amp;amp; #AlphaGo have to do with me
dicine? Learn more about how we're using #AI to drive research https://t.…"

[[9]]
[1] "SmartMedRT: RT @pfizer: What do #selfdrivingcars &amp;amp; #AlphaGo have to do with med
icine? Learn more about how we're using #AI to drive research https://t.…"

[[10]]
[1] "pfizer: What do #selfdrivingcars &amp;amp; #AlphaGo have to do with medicine? Learn mor
e about how we're using #AI to drive researc… https://t.co/lSD1rguwK7"
```

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 Ep
ic Response  https://t.co/eheV0ypICu… https://t.co/iRdVQYAzy4"

[[5]]
[1] "Deep_In_Depth: Mastering the game of Go with deep neural networks and tree search
https://t.co/wUXVPbOw9w #DeepLearning… https://t.co/mZnErxlSLP"

[[6]]
[1] "CarlosMBorbon: RT @pfizer: What do #selfdrivingcars &amp;amp; #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 &amp;amp; #AlphaGo have to do with me
dicine? Learn more about how we're using #AI to drive research https://t.…"

[[8]]
[1] "SmartMedRT: RT @pfizer: What do #selfdrivingcars &amp;amp; #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 &amp;amp; #AlphaGo have to do with medicine? Learn mor
e about how we're using #AI to drive researc… https://t.co/lSD1rguwK7"

[[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)

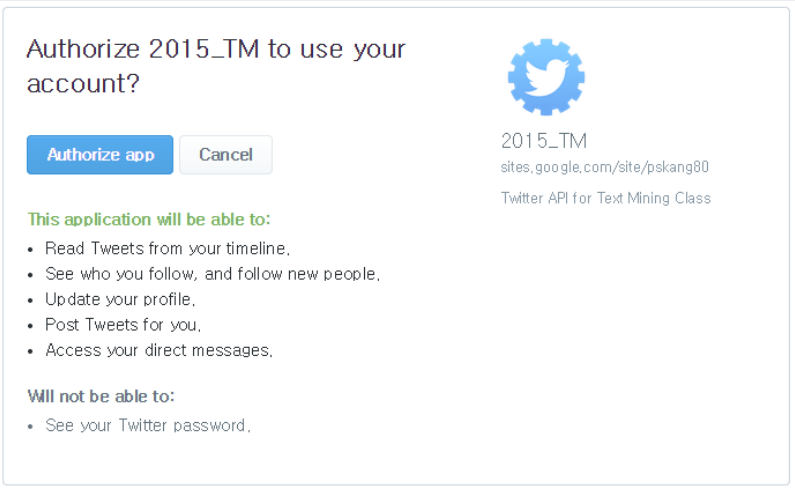
- Additional Twitter API: streamR
  - ✓ Provide a series of function that allow R users to access Twitter's filter, sample, and user streams, and to parse the output into data frames.

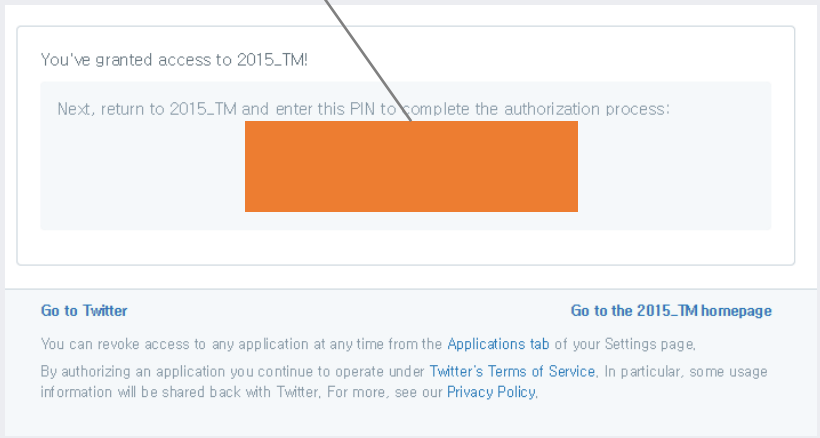
```
# Twitter stream data collection
download.file(url="http://curl.haxx.se/ca/cacert.pem", destfile="cacert.pem")

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 <- "LnuGtk5bbndcBN3t3PWoA"
> apiSecret <- "W3hvDz48zgt4nCHIhMPBx4Ca3h09g7BZv2ZOOQk"
> 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=kIkNGvKXIdzQ1xT9mvGCEl92JPt375sJk6Ujga9vzoI
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
                                                
```

AGENDA

01 Collect Data using APIs: Twitter

02 Collect Data using APIs: Facebook

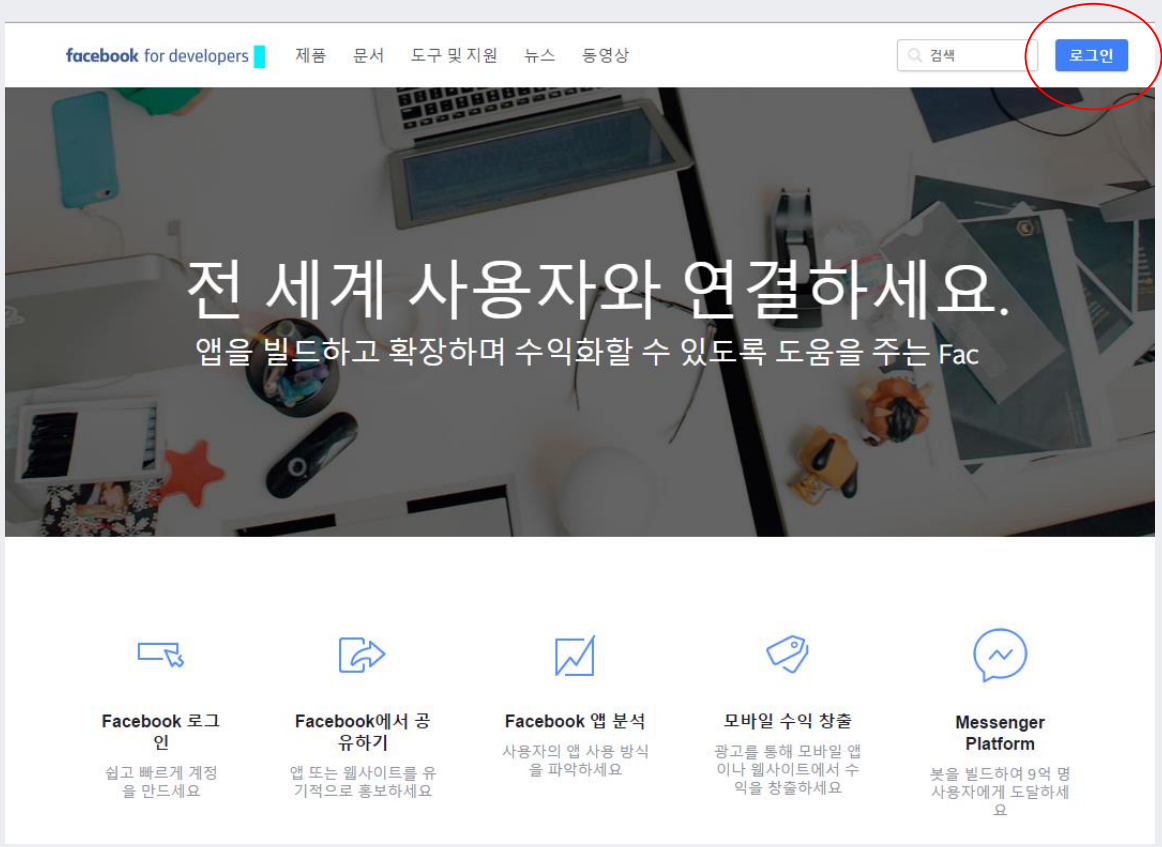
03 Web Scraping: ArXiv Research Papers

04 Web Community: PPOMPPU

05 Website: NAVER Real Estate

Collect Facebook Posts

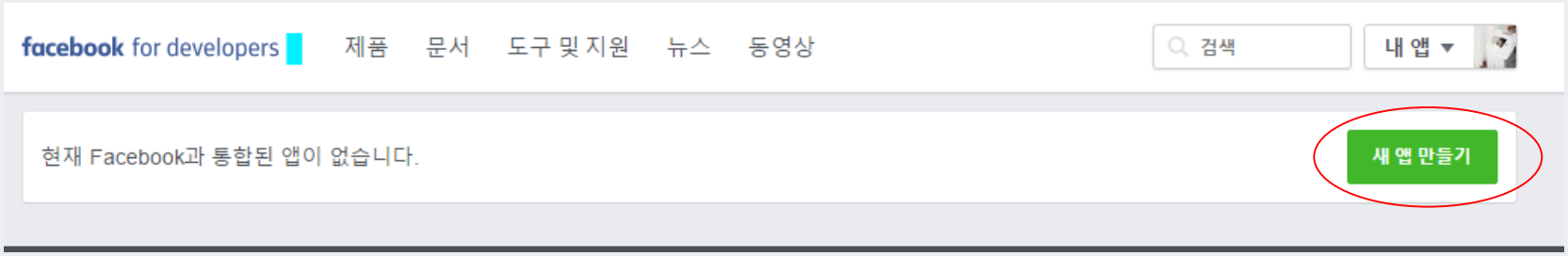
- Step 1: Registering an Application with Facebook
  - ✓ https://developers.facebook.com/



Login with your facebook account

Collect Facebook Posts

- Step 1: Registering an Application with Facebook
  - ✓ https://developers.facebook.com/

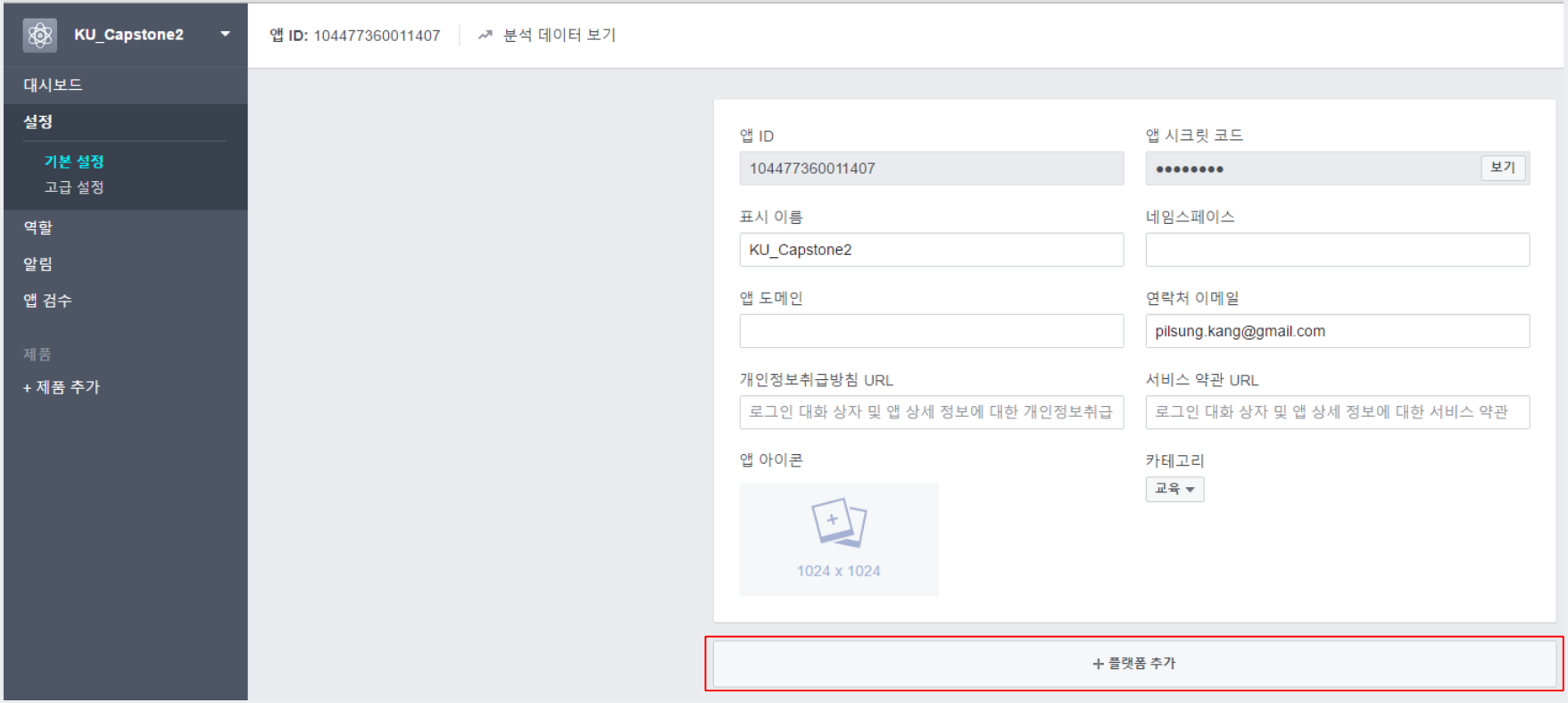


Collect Facebook Posts

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

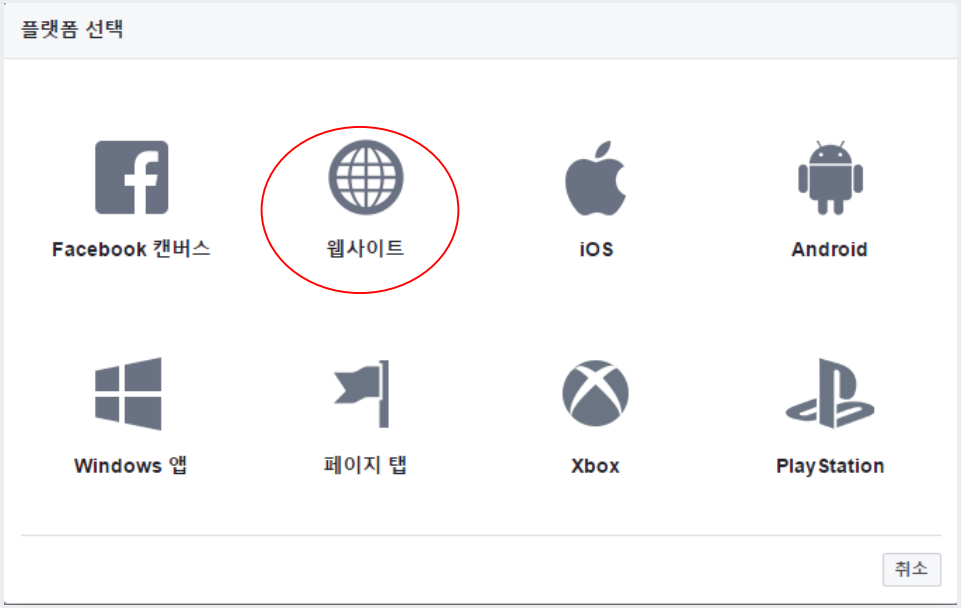
Collect Facebook Posts

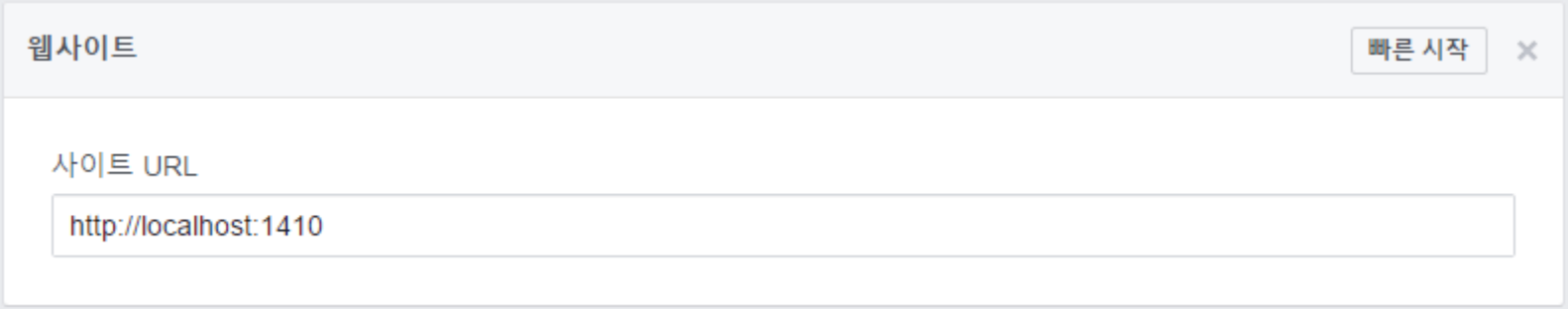
- Step 1: Registering an Application with Facebook
  - ✓ Add a new website with http://localhost:1410



Collect Facebook Posts

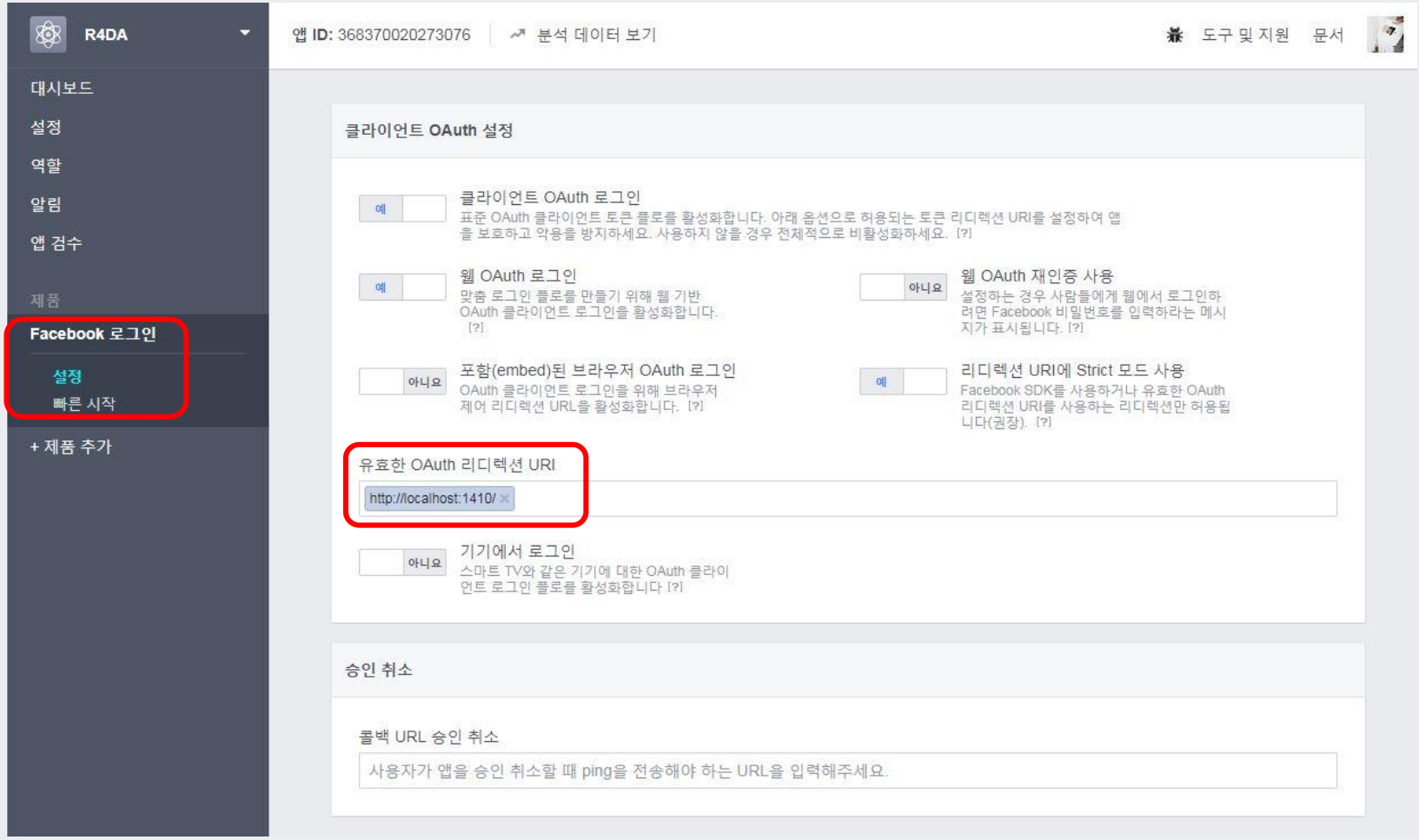
- Step 1: Registering an Application with Facebook
  - ✓ Add a new website with http://localhost:1410





Collect Facebook Posts

- Step 1: Registering an Application with Facebook
  - ✓ Add valid Oauth Redirection URL: http://localhost:1410/



Collect Facebook Posts

- Step 2: Create Oauth token to Facebook R session
  - ✓ Install Rfacebook package
  - ✓ Use your own app id & secret code

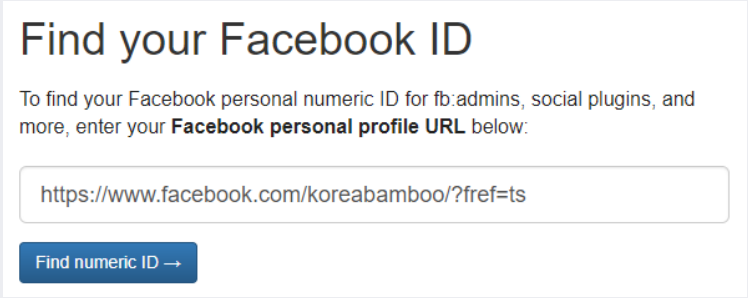
```
# Case 1-2: Collect Texts using Facebook API ---------------------------------
install.packages("Rfacebook")
library(Rfacebook)

# Authentication Setting
my_oauth <- fbOAuth(app_id = “your app id", app_secret= “your app secret")
save(my_oauth, file = "my_oauth")
load("my_oauth")
```

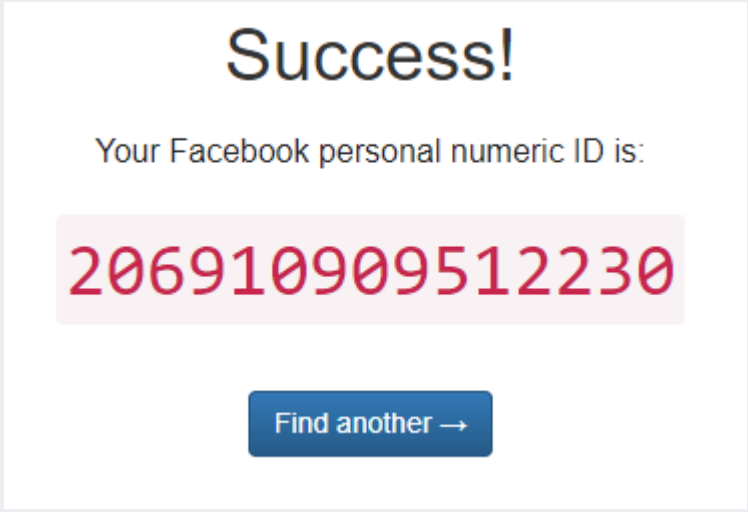
```
Authentication complete. Please close this page and return to R.
```

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

Collect Facebook Posts

- Step 3: Collect Data from a Page
  - ✓ Need to extract the numeric id of a page: http://findmyfbid.com/





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 동아리를 찾는다고? 1. ...	2017-08-17T08:37:19+0000	status	NA	206910909512230_655392904664026	NA	65	38	12
2	206910909512230	고려대학교 대나무숲	#30332번째포효 오빠, 사귀는 동안 많은 일이 있었죠 내가...	2017-08-17T08:01:35+0000	status	NA	206910909512230_655380927998557	NA	297	108	8
3	206910909512230	고려대학교 대나무숲	#30331번째포효 대숲 저는 고시생인데요 점점 사이코패...	2017-08-17T07:47:07+0000	status	NA	206910909512230_655380241331959	NA	127	26	8
4	206910909512230	고려대학교 대나무숲	#30330번째포효 대숲!! 대숲!! 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 알러지를 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 대숲, 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 감정을 새로 일깨워주는 너는 내...	2017-08-17T02:28:43+0000	status	NA	206910909512230_655313084672008	NA	7	1	2
10	206910909512230	고려대학교 대나무숲	#30324번째포효 동아리에서 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 연애하고 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 궁금한게있는데요 내 친구가 내 전남친...	2017-08-16T16:14:25+0000	status	NA	206910909512230_655170011352982	NA	89	25	5
15	206910909512230	고려대학교 대나무숲	#30319번째포효 대숲, 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/

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, 텐서플로우로 RNN공부하다 질문드립니다 텐...	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	https://www.facebook.com/photo.php?fbid=198868...	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 텐서플로우에서 아래와 같은 그리드에 정의된 ...	2017-08-16T09:37:44+0000	photo	https://www.facebook.com/photo.php?fbid=111122...	255834461424286_519682091706187	NA	10	4	0
5	856243047810396	Jerry Kim	안녕하세요! 영상쪽으로 공부하고 있는 학부생입니다! Dec...	2017-08-10T14:46:09+0000	photo	https://www.facebook.com/photo.php?fbid=115022...	255834461424286_516970438644019	NA	20	5	0
6	179941529215463	Shin Kyoungcheol	cnn 공부 중에 문제가 생겨서 도움을 받고 싶습니다. 항상 ...	2017-08-16T16:14:51+0000	link	https://github.com/shinv1234/pbl/blob/master/pro...	255834461424286_519815548359508	NA	1	2	0
7	1507988445929282	Insik Kim	tensorflow 1.3 버전에 timeseries API 추가 된것 같습니...	2017-08-17T08:01:30+0000	status	NA	255834461424286_520088844998845	NA	2	0	0
8	1432730463483977	김용한	안녕하세요~ 텐서플로우 gpu로 러닝중에 계속 멈출현상이...	2017-08-17T06:28:20+0000	photo	https://www.facebook.com/photo.php?fbid=143262...	255834461424286_520064091667987	NA	3	2	0
9	10209088019414425	Chris Song	스타2 API로 강화학습 배우면 완전 재밌을 거 같지 않아요?!...	2017-08-15T11:13:37+0000	photo	https://www.facebook.com/photo.php?fbid=102123...	255834461424286_519257465081983	NA	225	16	0
10	1362563510531971	Hoik Choi	안녕하세요. 텐서플로우 코드를 구현하는 중, 주어진 텐서...	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:) 텐서플로우로 학부 졸작 준비중인 학생입니다...	2017-08-14T10:20:20+0000	status	NA	255834461424286_518739518467111	NA	3	13	2
13	1477502905675907	안홍철	[머신러닝&딥러닝 오프라인 스터디원 모집] 스터디수준 : ...	2017-08-17T05:19:08+0000	status	NA	255834461424286_520039788337084	NA	16	4	2
14	1648595188544378	Ghiwook Nam	이 논문을 보고 있는데요. https://arxiv.org/pdf/1703.0...	2017-08-16T11:36:48+0000	link	https://arxiv.org/pdf/1703.02344.pdf	255834461424286_519714188369644	NA	1	7	2
15	1429496113794666	Gyuhyong James Jeon	CNN네트워크뒤에 RNN을 붙일수있나요? 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

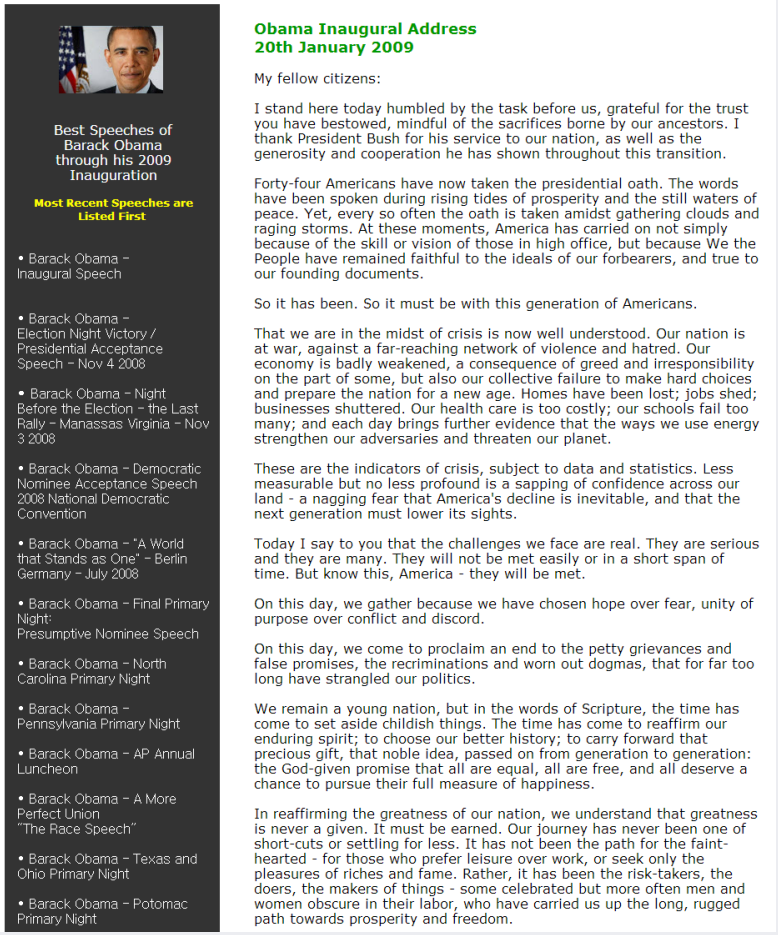
04 Web Community: PPOMPPU

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.

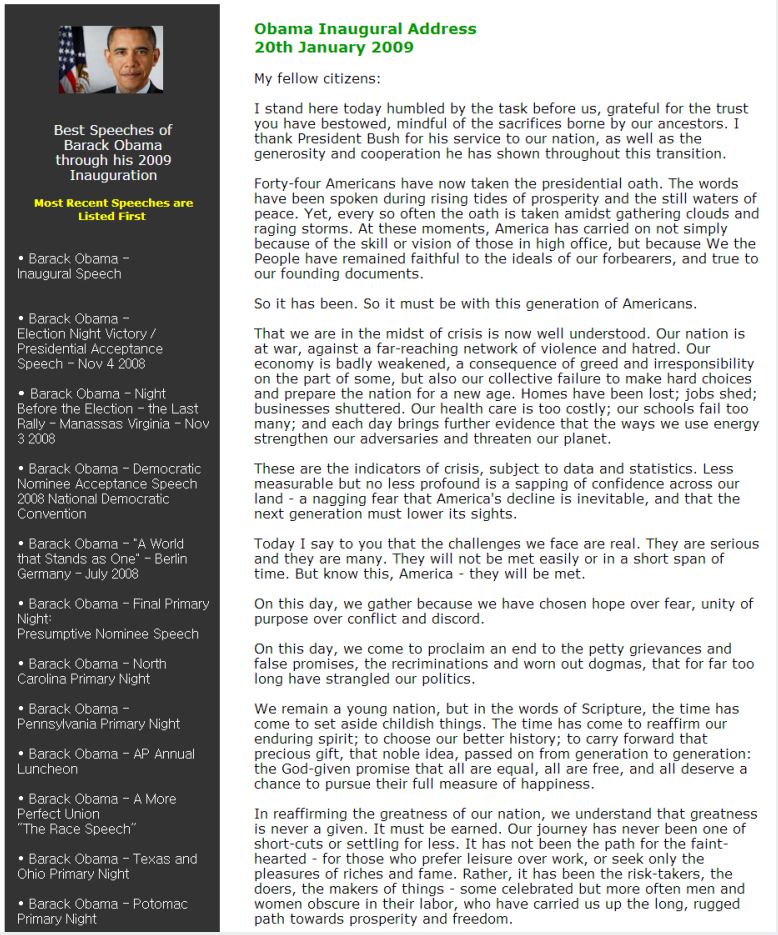
What we need to make a web page

```
304          </tr>
305        </table></td>
306      <td rowspan="16" align="center" valign="top" bgcolor="#FFFFFF"><br> <!-- InstanceBeginEditable name="EditRegion3" -->
307        <table width="610" height="299" border="0" align="center" cellpadding="0" cellspacing="0">
308          <tr bgcolor="#FFFFFF">
309            <td align="left" valign="top"><font size="4"><strong><font color="#009900" face="Verdana, Arial, Helvetica, sans-serif">Obama
310              Inaugural Address <br>
311              20th January 2009</font></strong><font size="3" face="Verdana, Arial, Helvetica, sans-serif"><br>
312              </font></font><font size="3" face="Verdana, Arial, Helvetica, sans-serif"><br>
313              My fellow citizens:<br>
314              <br>
315              I stand here today humbled by the task before us, grateful for the
316              trust you have bestowed, mindful of the sacrifices borne by our ancestors.
317              I thank President Bush for his service to our nation, as well as the
318              generosity and cooperation he has shown throughout this transition.<br>
319              <br>
320              Forty-four Americans have now taken the presidential oath. The words
321              have been spoken during rising tides of prosperity and the still waters
322              of peace. Yet, every so often the oath is taken amidst gathering clouds
323              and raging storms. At these moments, America has carried on not simply
324              because of the skill or vision of those in high office, but because
325              We the People have remained faithful to the ideals of our forbearers,
326              and true to our founding documents.<br>
327              <br>
328              So it has been. So it must be with this generation of Americans.<br>
329              <br>
330              That we are in the midst of crisis is now well understood. Our nation
331              is at war, against a far-reaching network of violence and hatred.
332              Our economy is badly weakened, a consequence of greed and irresponsibility
333              on the part of some, but also our collective failure to make hard
334              choices and prepare the nation for a new age. Homes have been lost;
335              jobs shed; businesses shuttered. Our health care is too costly; our
336              schools fail too many; and each day brings further evidence that the
337              ways we use energy strengthen our adversaries and threaten our planet.<br>
338              <br>
339              These are the indicators of crisis, subject to data and statistics.
340              Less measurable but no less profound is a sapping of confidence across
341              our land - a nagging fear that America's decline is inevitable, and
342              that the next generation must lower its sights.<br>
343              <br>
344              Today I say to you that the challenges we face are real. They are
345              serious and they are many. They will not be met easily or in a short
346              span of time. But know this, America - they will be met.<br>
```

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

```
304          </tr>
305        </table></td>
306      <td rowspan="16" align="center" valign="top" bgcolor="#FFFFFF"><br> <!-- InstanceBeginEditable name="EditRegion3" -->
307        <table width="610" height="299" border="0" align="center" cellpadding="0" cellspacing="0">
308          <tr bgcolor="#FFFFFF">
309            <td align="left" valign="top"><font size="4"><strong><font color="#009900" face="Verdana, Arial, Helvetica, sans-serif">Obama
310              Inaugural Address <br>
311              20th January 2009</font></strong><font size="3" face="Verdana, Arial, Helvetica, sans-serif"><br>
312              </font></font><font size="3" face="Verdana, Arial, Helvetica, sans-serif"><br>
313              My fellow citizens:<br>
314              <br>
315              I stand here today humbled by the task before us, grateful for the
316              trust you have bestowed, mindful of the sacrifices borne by our ancestors.
317              I thank President Bush for his service to our nation, as well as the
318              generosity and cooperation he has shown throughout this transition.<br>
319              <br>
320              Forty-four Americans have now taken the presidential oath. The words
321              have been spoken during rising tides of prosperity and the still waters
322              of peace. Yet, every so often the oath is taken amidst gathering clouds
323              and raging storms. At these moments, America has carried on not simply
324              because of the skill or vision of those in high office, but because
325              We the People have remained faithful to the ideals of our forbearers,
326              and true to our founding documents.<br>
327              <br>
328              So it has been. So it must be with this generation of Americans.<br>
329              <br>
330              That we are in the midst of crisis is now well understood. Our nation
331              is at war, against a far-reaching network of violence and hatred.
332              Our economy is badly weakened, a consequence of greed and irresponsibility
333              on the part of some, but also our collective failure to make hard
334              choices and prepare the nation for a new age. Homes have been lost;
335              jobs shed; businesses shuttered. Our health care is too costly; our
336              schools fail too many; and each day brings further evidence that the
337              ways we use energy strengthen our adversaries and threaten our planet.<br>
338              <br>
339              These are the indicators of crisis, subject to data and statistics.
340              Less measurable but no less profound is a sapping of confidence across
341              our land - a nagging fear that America's decline is inevitable, and
342              that the next generation must lower its sights.<br>
343              <br>
344              Today I say to you that the challenges we face are real. They are
345              serious and they are many. They will not be met easily or in a short
346              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

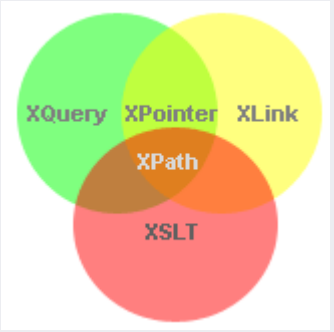
- Parsing result

Console D:/Dropbox/강의자료/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">表</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><font col
or="#FFFF00" size="4" face="Verdana, Arial, Helvetica, sans-serif"><strong><img src="/obama-president-left-nav-bar.jpg" width="125
" height="81"></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 colo
r="#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-se
rif"><strong><br><br>
              Most Recent Speeches are Listed First <br></strong></font><br><a href="/P-Obama-Inaugural-Speech-Inauguration.htm">
              <div align="left">??Barack Obama -<br>
                Inaugural Speech</div>
              </a>
              </p>
<div align="left">
<strong></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

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
<?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
nodename	Selects all nodes with the name "*nodename*"
/	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 **Note:** 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 Da

```
> 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')
```

```
Console D:/Dropbox/강의자료/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)
```

```
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. **Note:** 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: *In JavaScript:* *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

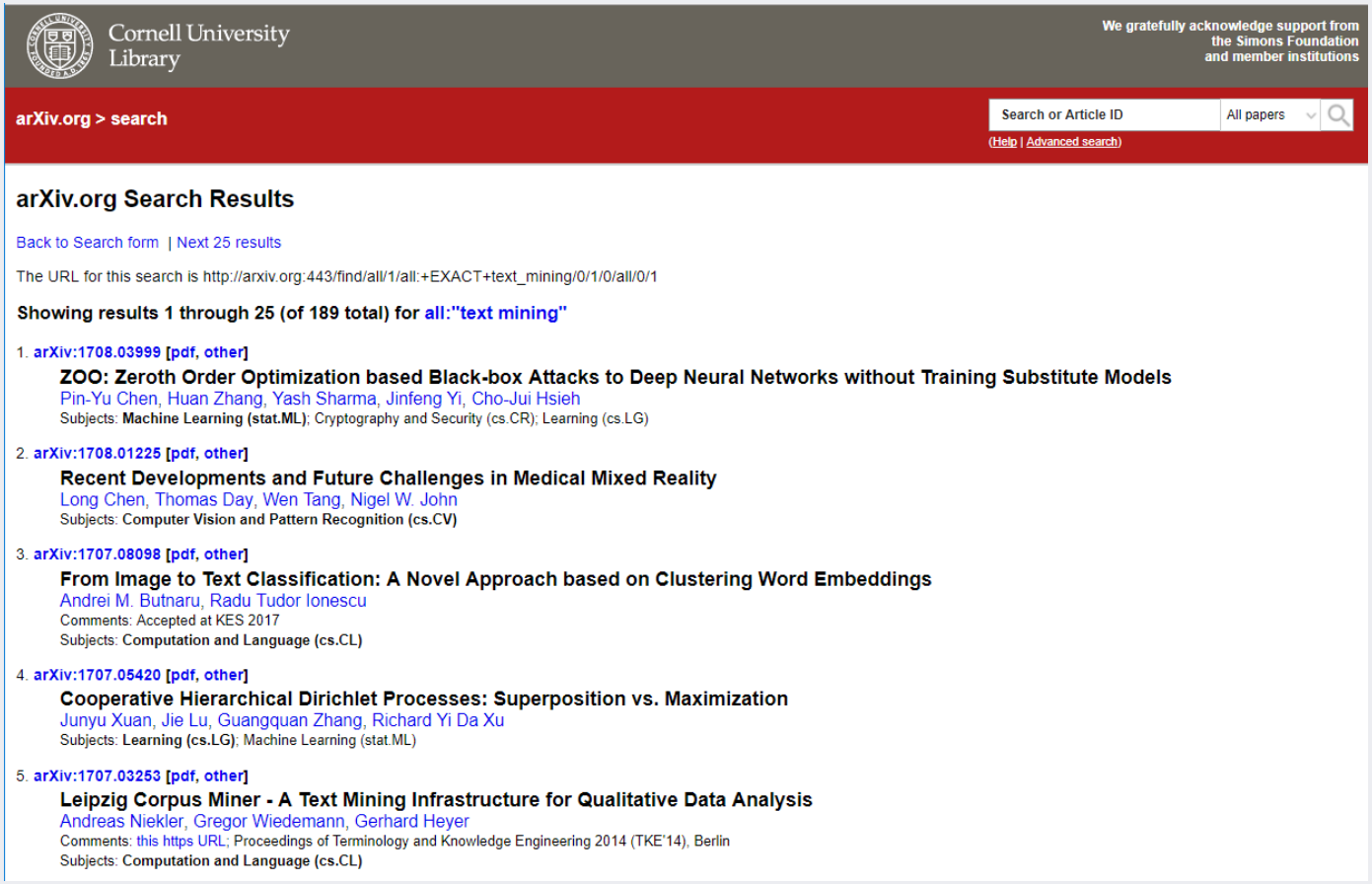
By using the | operator in an XPath expression you can select several paths.

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



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/0/1/0/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)**

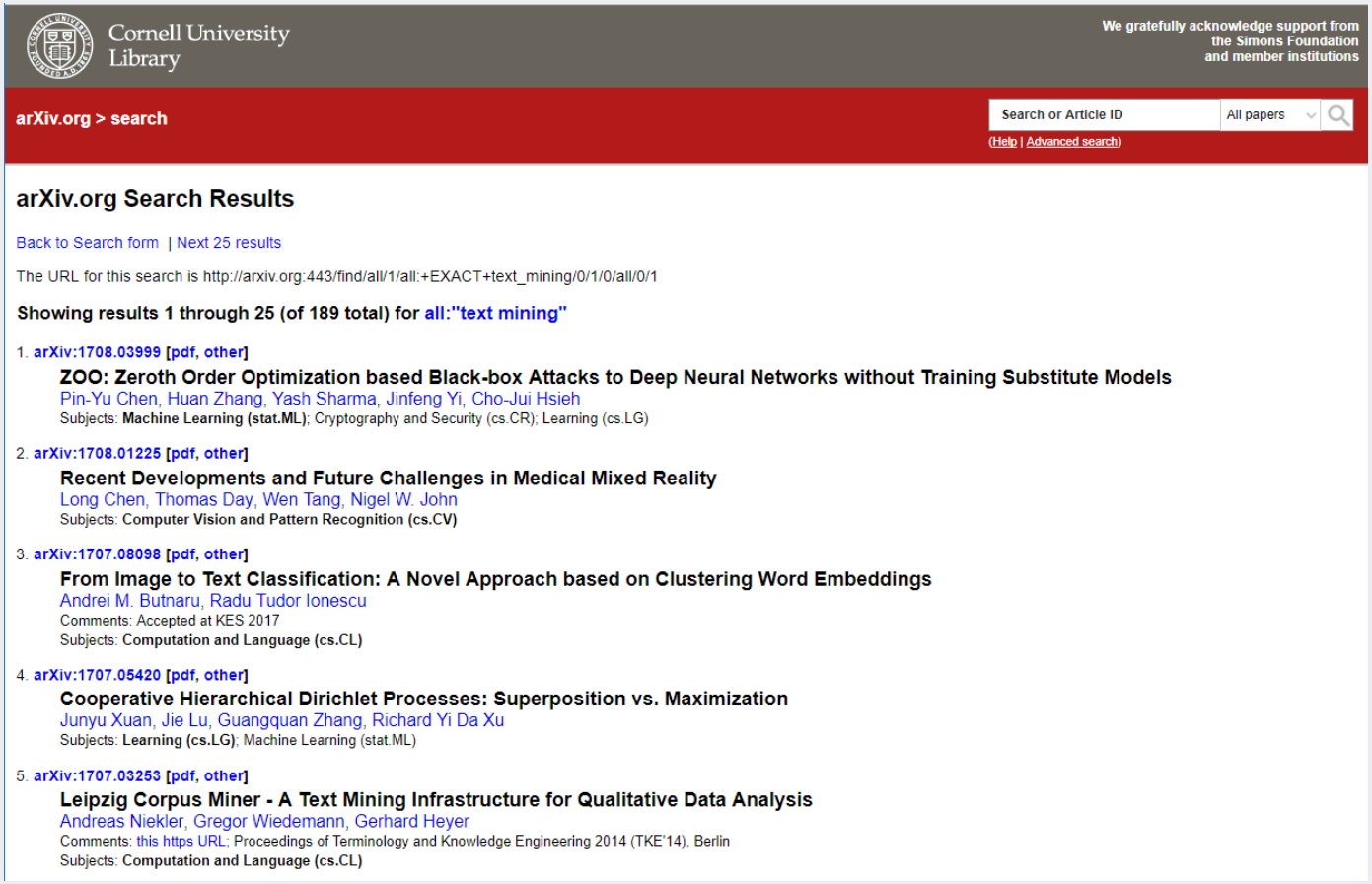
3. arXiv:1707.08098 [pdf, other]
   From Image to Text Classification: A Novel Approach based on Clustering Word Embeddings
   Andrei M. Butnaru, Radu Tudor Ionescu
   Comments: Accepted at KES 2017
   Subjects: **Computation and Language (cs.CL)**

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; 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



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/0/1/0/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; Proceedings of Terminology and Knowledge Engineering 2014 (TKE'14), Berlin
   Subjects: **Computation and Language (cs.CL)**

Web Scraping: arXiv Papers

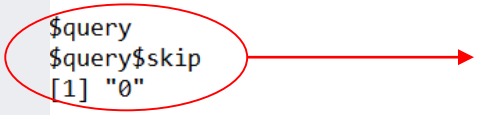
- Step 2: Analyzing the HTML Structure
  - ✓ First page URL
    - http://arxiv.org/find/all/1/all:+EXACT+text_mining/0/1/0/all/0/1?skip=0&query_id=504c4472acbc1ebf
  - ✓ Second page URL
    - http://arxiv.org/find/all/1/all:+EXACT+text_mining/0/1/0/all/0/1?skip=25&query_id=504c4472acbc1ebf
  - ✓ Third page URL
    - http://arxiv.org/find/all/1/all:+EXACT+text_mining/0/1/0/all/0/1?skip=50&query_id=504c4472acbc1ebf

Web Scraping: arXiv Papers

- Step 2: Analyzing the HTML Structure

  ✓ URL Parsing

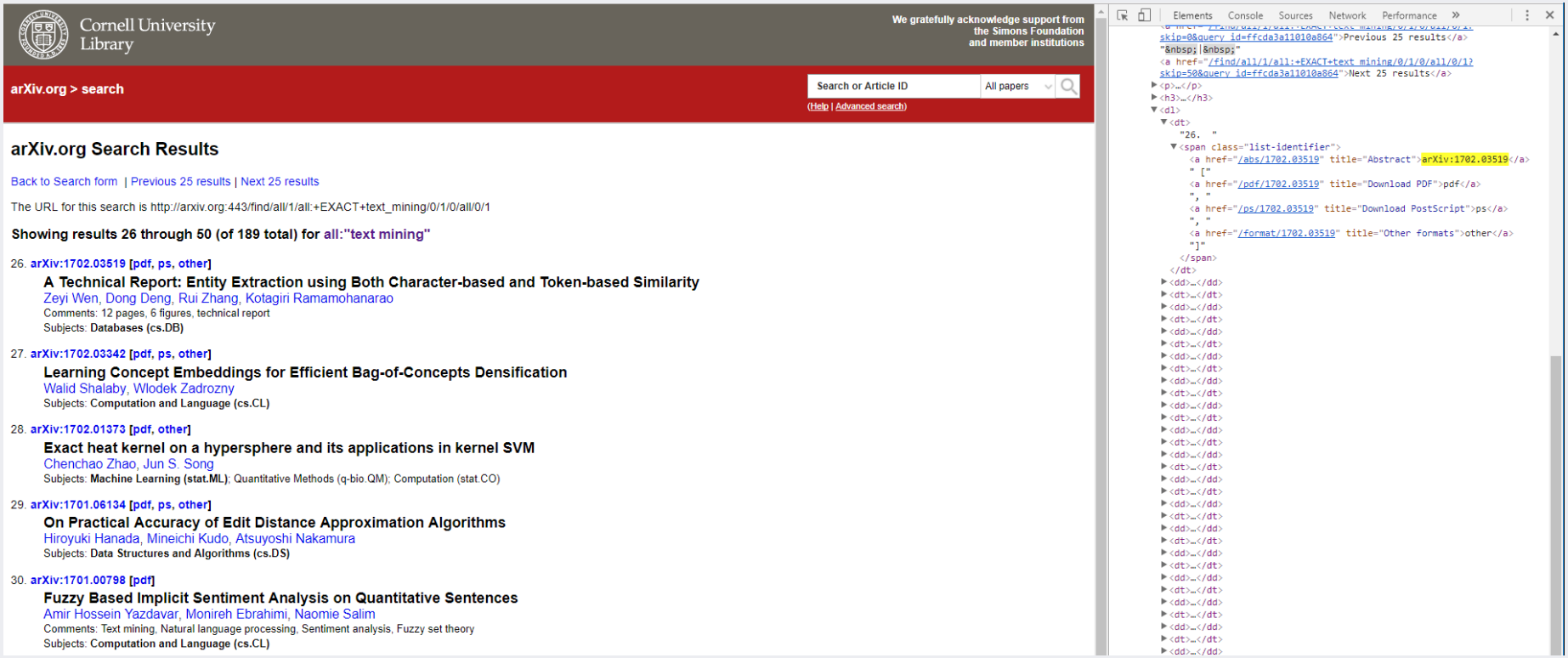
```
> parse_url("https://arxiv.org/find/all/1/all:+EXACT+text_mining/0/1/0/all/0/1?skip=0")
$scheme
[1] "https"

$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



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="dlpage">
<a href="/find/all/1/all:+EXACT+text_mining/0/1/0/all/0/1?query_id=504c4472acbc1ebf&amp;form=yes">Back to Search form</a>
 | <a href="/find/all/1/all:+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:+EXACT+text_mining/0/1/0/all/0/1<br /></p>
<h3>Showing results 1 through 25 (of 139 total) for
<a href="/find/all/1/all:+EXACT+text_mining/0/1/0/all/0/1?skip=0&amp;query_id=504c4472acbc1ebf">all:"text mining"</a></h3>
<dl>
<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="/find/stat/1/au:+Ranjan_C/0/1/0/all/0/1">Chitta Ranjan</a>,
<a href="/find/stat/1/au:+Ebrahimi_S/0/1/0/all/0/1">Samaneh Ebrahimi</a>,
<a href="/find/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
</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="/find/cs/1/au:+Magron_P/0/1/0/all/0/1">Paul Magron</a>,
<a href="/find/cs/1/au:+Badeau_R/0/1/0/all/0/1">Roland Badeau</a>,
<a href="/find/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="/find/cs/1/au:+Kakde_D/0/1/0/all/0/1">Deovrat Kakde</a>,
<a href="/find/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)
```

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

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

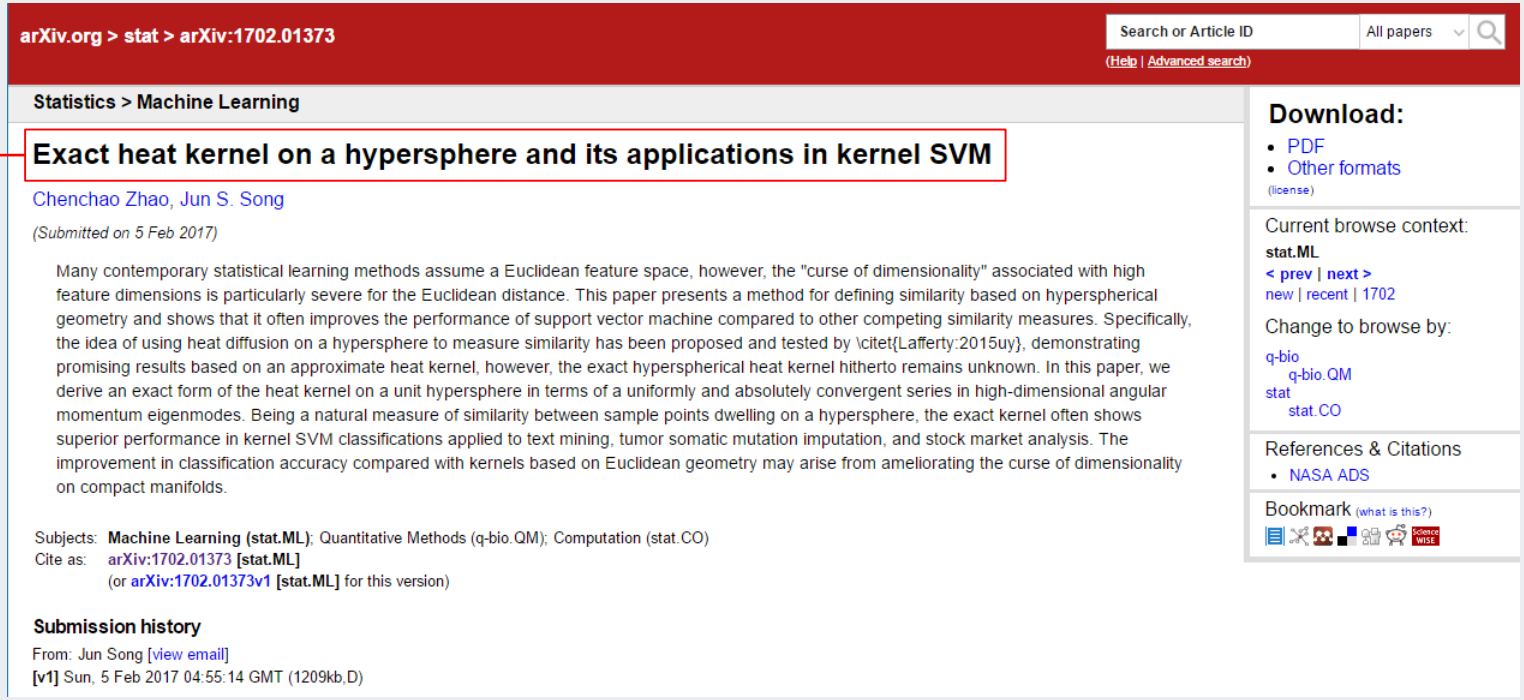
```
tmp_list <- read_html(tmp_url) %>% html_nodes('div#dlpage') %>%
                                   html_nodes('a[title="Abstract"]') %>% html_attr('href')
```

    - 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




```
<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

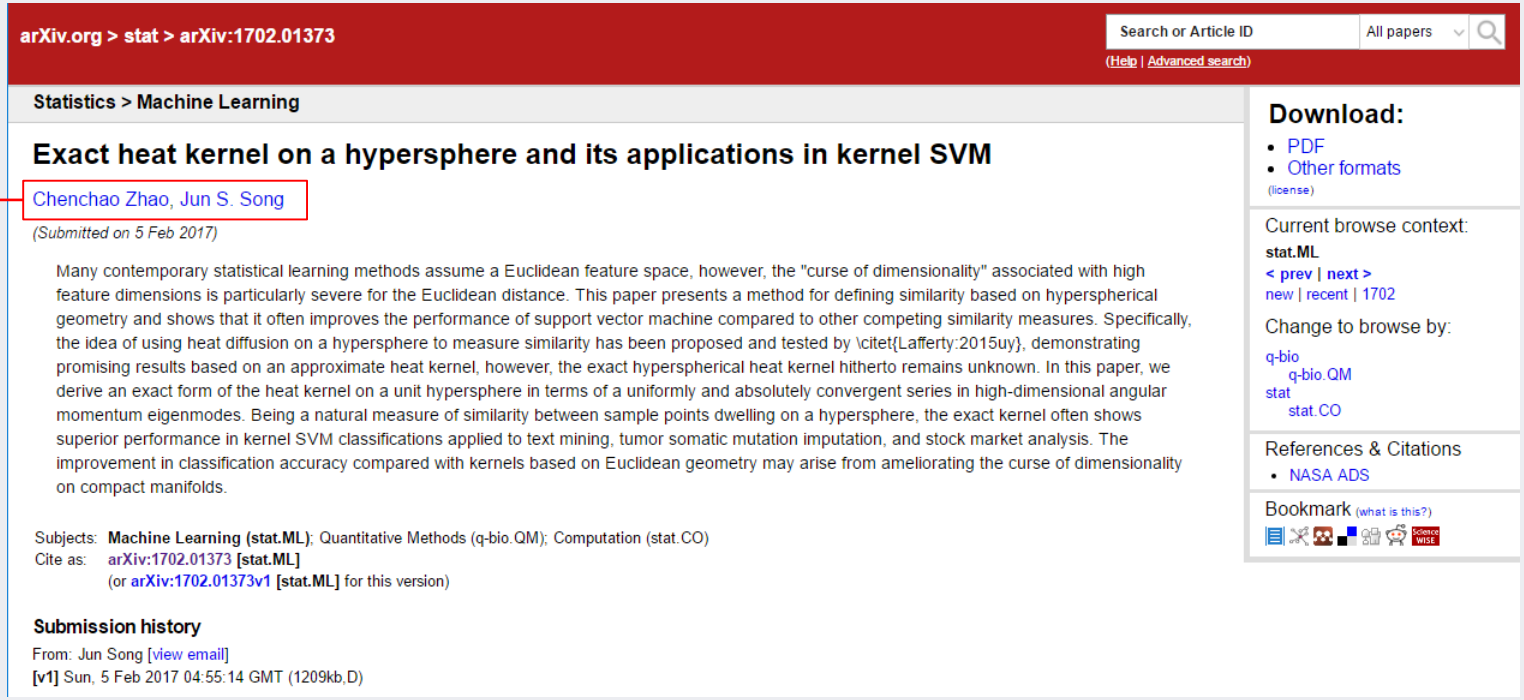
```
# title
tmp_title <- gsub('Title:\n', '',tmp_paragraph %>% html_nodes('h1.title.mathjax') %>% html_text(T))
title <- c(title, tmp_title)
```

  - 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



Web Scraping: arXiv Papers

- Step 3: Extract necessary information

  ✓ Step 3-2: Extract Authors

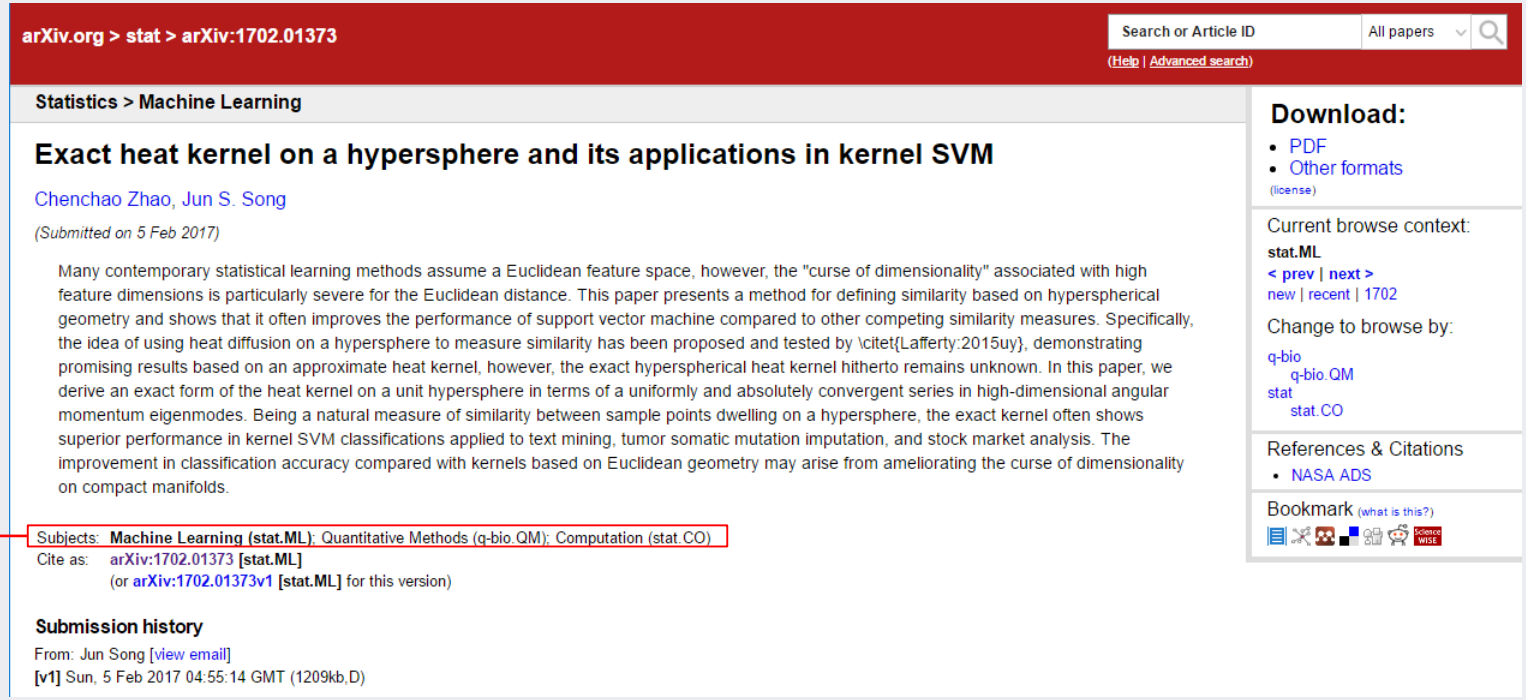
```
# author
tmp_author <- tmp_paragraph %>% html_nodes('div.authors') %>% html_text
tmp_author <- gsub('\\s+',' ',tmp_author)
tmp_author <- gsub('Authors:','',tmp_author) %>% str_trim
author <- c(author, tmp_author)
```

  - 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



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

Web Scraping: arXiv Papers

- Step 3: Extract necessary information
  - ✓ Step 3-4: Extract Abstract

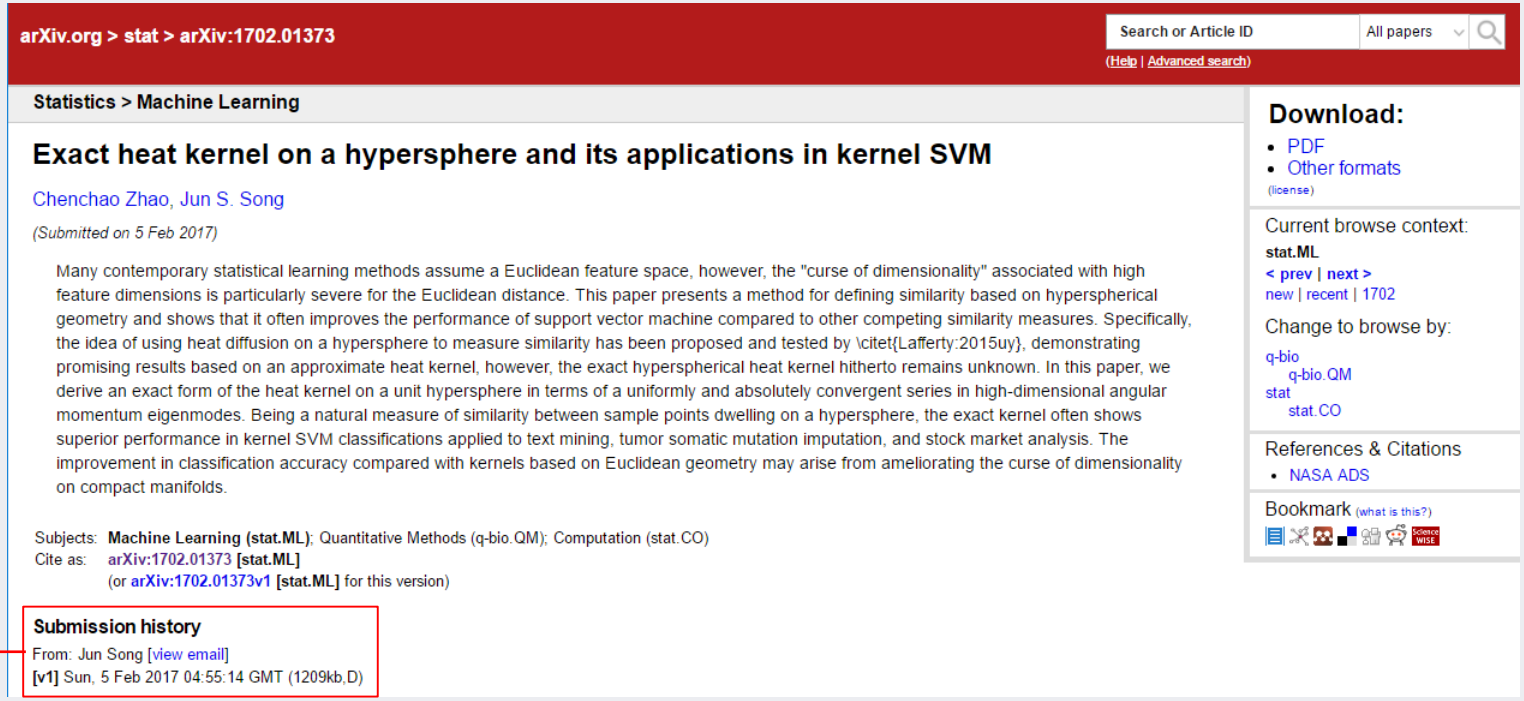
```
# abstract
tmp_abstract <- tmp_paragraph %>% html_nodes('blockquote.abstract.mathjax') %>% html_text(T)
tmp_abstract <- sub('Abstract:','',tmp_abstract) %>% str_trim
abstract <- c(abstract, tmp_abstract)
```

- 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. Th
is paper\npresents a method for defining similarity based on hyperspherical geometry and\nshows that it often i
mproves the performance of support vector machine compared\nto other competing similarity measures. Specificall
y, the idea of using heat\ndiffusion 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 exac
t hyperspherical heat kernel hitherto\nremains unknown. In this paper, we derive an exact form of the heat kern
el 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
h... <truncated>
```

Web Scraping: arXiv Papers

- Step 3: Extract necessary information
  - ✓ Step 3-5: Extract Meta information




```
<div class="submission-history">
<h2>Submission history</h2>
From: Jun Song [<a href="https://arxiv.org/show-email/d8578523/1702.01373">view email</a>]
<br />
<b>[v1]</b> Sun, 5 Feb 2017 04:55:14 GMT  (1209kb,D)<br />
</div>
```

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
>
```

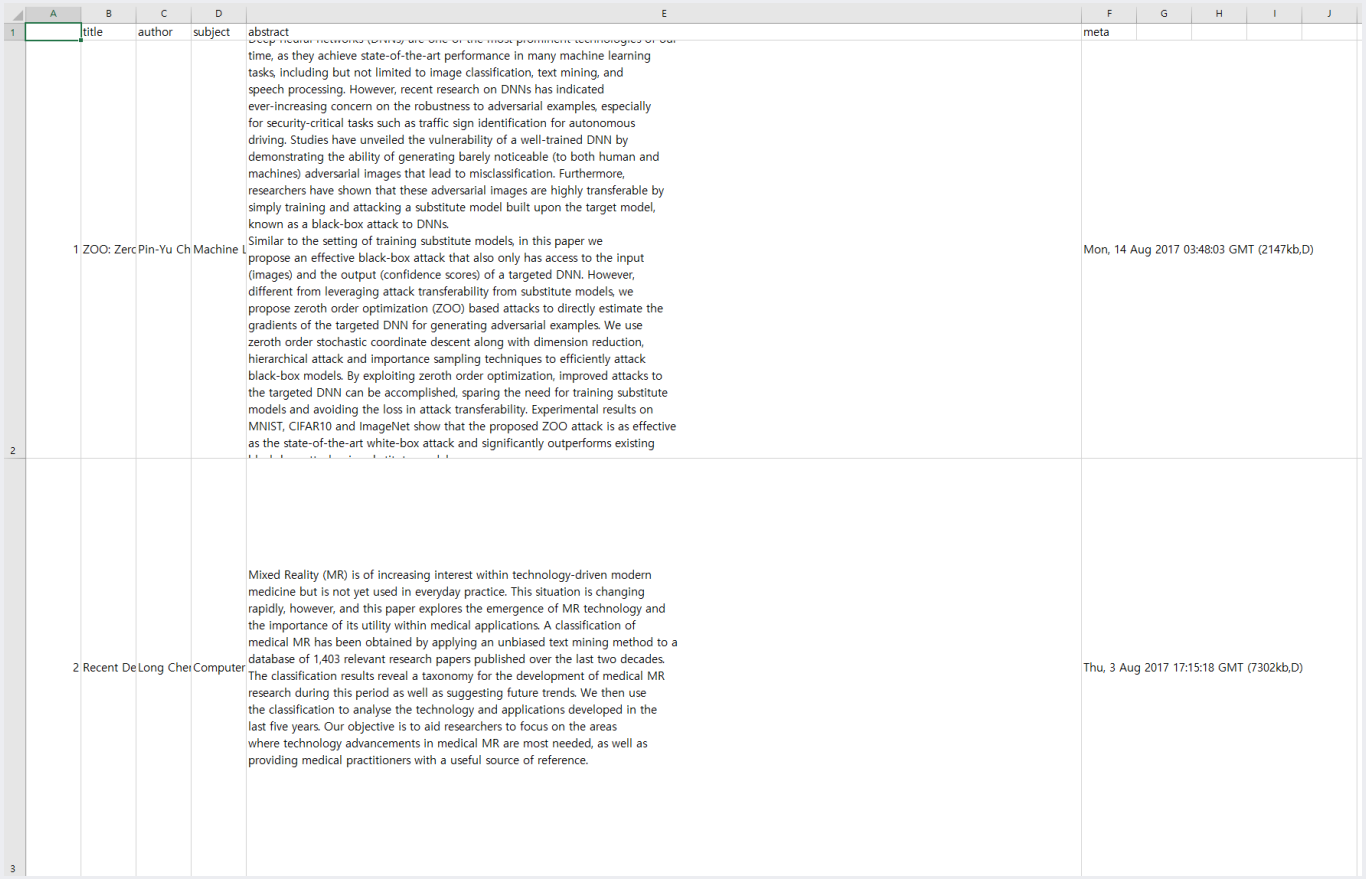
  - ✓ Check the dataset

	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 Hematialam, Wlodek Zadrozny	Computation and Language (cs.CL)	This paper advances the state of the art in text unde...	Tue, 13 Ju
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 M
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 M
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 Ma
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
1		title	author	subject	abstract	meta
2	1	ZOO: Zerc	Pin-Yu Ch	Machine L	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	Mon, 14 Aug 2017 03:48:03 GMT (2147kb,D)
3	2	Recent De	Long Chei	Computer	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.	Thu, 3 Aug 2017 17:15:18 GMT (7302kb,D)

References

Other materials

- Figure in the title page: https://nocodewebscraping.com/web-scraping-for-dummies-tutorial-with-import-io-without-coding/

Who wrote the fastest scraping code?

	Increased Efficiency	How to improve?	Strength	In this slide
	11.32% (123.10s → 109.17s)	Use rvest	제출 순서 1등	✓
	70.25% (123.17s → 36.65s)	Use 4 cores Use rvest Use data.table	효율성 향상 1등	
	36.19% (123.09s → 78.54s)	Use Java	10회 반복 수행 평균으로 비교	
	54.89% (170.67s → 77.67s)	Use PHP	3개 언어(R, Python, PHP) 비교	

AGENDA

01 Collect Data using APIs: Twitter

02 Collect Data using APIs: Facebook

03 Web Scraping: ArXiv Research Papers

04 Web Community: PPOMPPU

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

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

```
<tr align="center" class="list1" onMouseOver="this.style.backgroundColor='#F5F5F5'" onMouseOut="this.style.backgroundColor=''" style="height:16;word-break:break-all;" valign='middle'>
    <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>
    <!--<td nowrap colspan=2 style='padding:0'><input type=checkbox name=cart value="61484"></td>-->
    <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>
</div></td> <td align=left class=list_vspace>

    <img src=/images/icon_03.png border=0 align=absmiddle title='익명글'> <a href="view.php?id=insurance&page=1&divpage=10&no=61484"  ><font class=list_title>암보험 생활비 받는 암보험 괜찮은가요?</font></a>
        <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>
<td nowrap class='eng list_vspace' colspan=2>5</td></tr>


        <tr><td colspan=13 class=line_separator height=1></td></tr>


<tr align="center" class="list0" onMouseOver="this.style.backgroundColor='#F5F5F5'" onMouseOut="this.style.backgroundColor=''" style="height:16;word-break:break-all;" valign='middle'>
    <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>
    <!--<td nowrap colspan=2 style='padding:0'><input type=checkbox name=cart value="61483"></td>-->
    <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>
</div></td> <td align=left class=list_vspace>

    <img src=/images/icon_06.png border=0 align=absmiddle title='모바일+익명글'> <a href="view.php?id=insurance&page=1&divpage=10&no=61483"  ><font class=list_title>태아보험 질문이요.</font></a> <span class=list_comment2> <span style='cursor:pointer;' onclick='win_comment("popup_comment.php?id=insurance&no=61483");'>3</span> </span>
        <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>
<td nowrap class='eng list_vspace' colspan=2>13</td></tr>


        <tr><td colspan=13 class=line_separator height=1></td></tr>


<tr align="center" class="list1" onMouseOver="this.style.backgroundColor='#F5F5F5'" onMouseOut="this.style.backgroundColor=''" style="height:16;word-break:break-all;" valign='middle'>
    <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>
    <!--<td nowrap colspan=2 style='padding:0'><input type=checkbox name=cart value="61480"></td>-->
    <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>
</div></td> <td align=left class=list_vspace>

    <img src=/images/icon_06.png border=0 align=absmiddle title='모바일+익명글'> <a href="view.php?id=insurance&page=1&divpage=10&no=61480"  ><font class=list_title>비갱신형 암보험..40대 후반</font></a> <span class=list_comment2> <span style='cursor:pointer;' onclick='win_comment("popup_comment.php?id=insurance&no=61480");'>1</span> </span>
        <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>
<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

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><!--DCM_TITLE-->태아보험 질문이요.<!--/DCM_TITLE--></font> <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>
```

    - 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><!--DCM_TITLE-->태아보험 질문이요.<!--/DCM_TITLE--></font> <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

```
</script>

    <table border="0" cellspacing="0" cellpadding="0" width="900" class="pic_bg">
    <tr>
      <td style="padding:0 8 0 8;" align="left">
        <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">
<tr>
<td class='board-contents' align="left" valign=top class=han>
내년에 출생하는 아이가 있어서 태아보험을 가입하려는데 몇 가지 질문을 드릴게요. <br />
여기 게시판에서 보고 약간의 정보를 얻었는데 궁금한 점 입니다. <br />
<br />
1. 태아 보험 다이렉트로 가입하면 더 저렴한지. <br />
<br />
2. 태아 보험 가입 후 출생 이후에는 해지 하고 단독실비로 갈아타라는데 그렇게 하는 장점이 무엇인지 궁금합니다. <br />
그렇게 단독실비로 가입하면 가격적으로 더 저렴해지는지 혹은 커버가 더 많이 되는지 알고 싶어요. <br />
<br />
도움 좀 부탁드리겠습니다. <br />
좋은 하루 되세요~<br />
<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\n분류: 질문\r\n이름: [* 익명 *] \r\n등록일: 2017-08-21 11:23\r\n\r\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 점 입니다. \n1. 태아
 보험 다이렉트로 가입하면 더 저렴한지. \n2. 태아 보험 가입 후 출생 이후에는 해지 하고 단독실비로 갈아타라는데 그렇게 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 점 입니다 1 태아 보험 다
이렉트로 가입하면 더 저렴한지 2 태아 보험 가입 후 출생 이후에는 해지 하고 단독실비로 갈아타라는데 그렇게 하는 장점이 무엇인지 궁금합니다 그렇게 단독실비로 가입
하면 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 부탁드리겠습...	2017-08-21	단독실비 인터넷으로 가입하려고 하는데요 피드백 바르고 ...
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 보험증권을 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 실비있음 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...

ANY
questions?