Lecture 2: From Texts to Data

Pilsung Kang

School of Industrial Management Engineering

Korea University

AGENDA

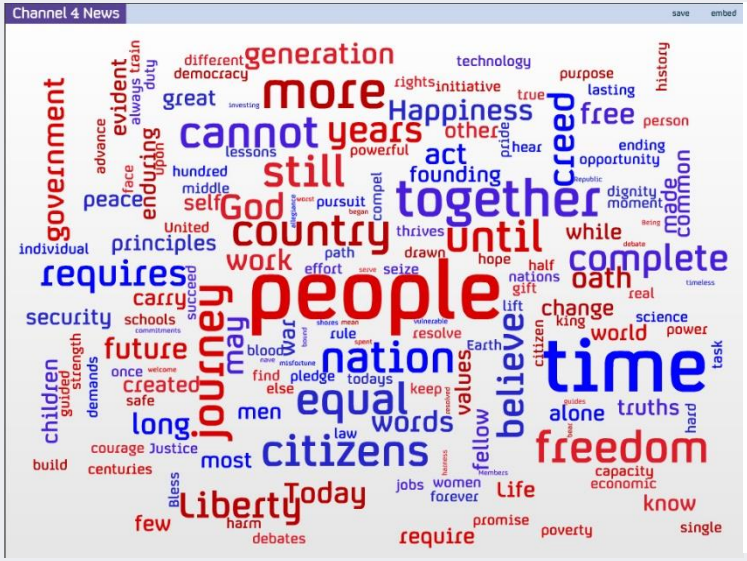
01 From Text Files to Data

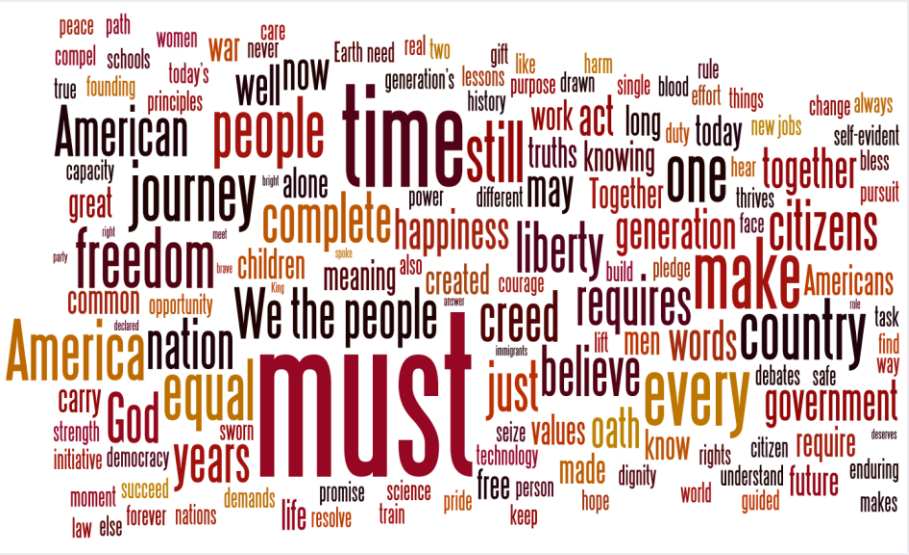
02 Collect Texts using APIs

03 Web Scraping

Read Data from a Text File

- Inaugural Address by President Barack Obama in 2013

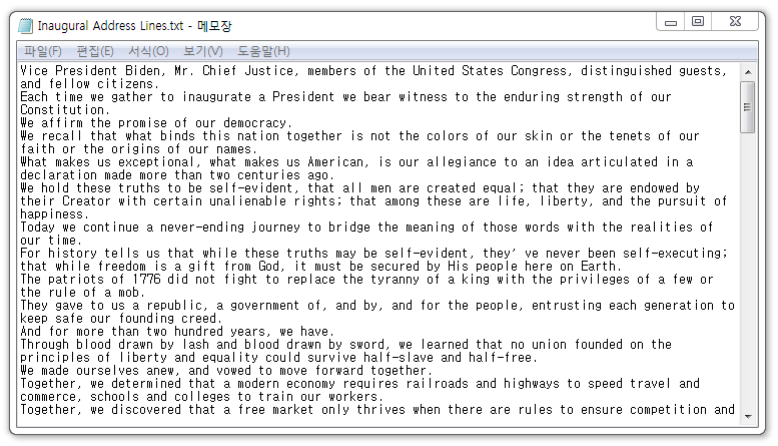




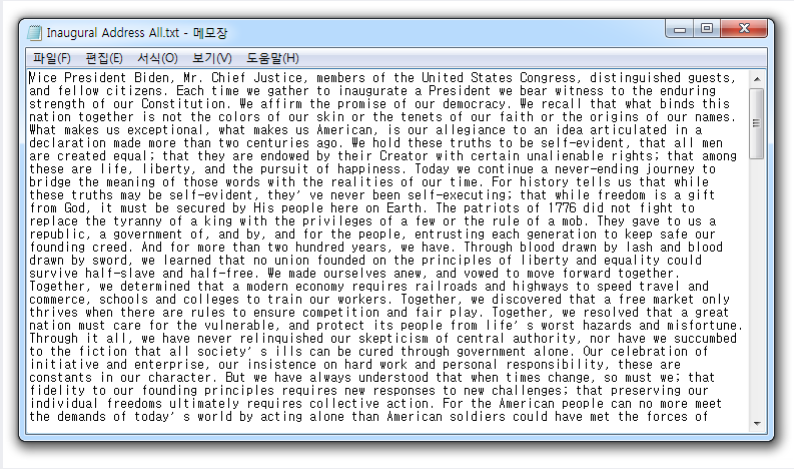
Read Data from a Text File

- Are sentences separated?

1 sentence in a line



Sentences are not separated



Read Data from a Text File

- If sentences are separated
  - ✓ Use readLines function

```
3  # If sentences are separated
4  speech1 <- readLines("Inaugural Address Lines.txt", encoding = "UTF-8")
5  speech1 <- as.data.frame(speech1, stringsAsFactors = FALSE)
6  names(speech1) <- "sentence"
7  speech1[1:3,]
```

from a file to a character vector

```
> speech1
 [1] "Vice President Biden, Mr. Chief Justice, members of the United States Congress, distinguished guests, and fellow citizens."

 [2] "Each time we gather to inaugurate a President we bear witness to the enduring strength of our Constitution."

 [3] "We affirm the promise of our democracy."

 [4] "We recall that what binds this nation together is not the colors of our skin or the tenets of our faith or the origins of our names."

 [5] "What makes us exceptional, what makes us American, is our allegiance to an idea articulated in a declaration made more than two centuries ago."
```

from a vector to dataframe

```
> speech1[1:5,]
[1] "Vice President Biden, Mr. Chief Justice, members of the United States Congress, distinguished guests, and fellow citizens."
[2] "Each time we gather to inaugurate a President we bear witness to the enduring strength of our Constitution."
[3] "We affirm the promise of our democracy."
[4] "We recall that what binds this nation together is not the colors of our skin or the tenets of our faith or the origins of our names."
[5] "What makes us exceptional, what makes us American, is our allegiance to an idea articulated in a declaration made more than two centuries ago."
```

Read Data from a Text File

- If sentences are not separated
  - ✓ Use scan function

```
 9  # If sentences are not separated
10  speech2 <- scan("Inaugural Address All.txt", what = "character", sep = "\n", encoding = "UTF-8")
11  speech2 <- strsplit(speech2, ". ", fixed = TRUE)
12  speech2 <- as.data.frame(speech2, stringsAsFactors = FALSE)
13  names(speech2) <- "sentence"
14  speech2[1:5,]
```

from a file to a single character string

```
> speech2
[1] "Vice President Biden, Mr. Chief Justice, members of the United States Congress, distinguished guests, and fellow citizens. Each time we gather to inaugurate a Pres
ident we bear witness to the enduring strength of our Constitution. We affirm the promise of our democracy. We recall that what binds this nation together is not the co
lors of our skin or the tenets of our faith or the origins of our names. What makes us exceptional, what makes us American, is our allegiance to an idea articulated in
a declaration made more than two centuries ago. We hold these truths to be self-evident, that all men are created equal; that they are endowed by their Creator with cer
tain unalienable rights; that among these are life, liberty, and the pursuit of happiness. Today we continue a never-ending journey to bridge the meaning of those words
with the realities of our time. For history tells us that while these truths may be self-evident, they've never been self-executing; that while freedom is a gift from G
od, it must be secured by His people here on Earth. The patriots of 1776 did not fight to replace the tyranny of a king with the privileges of a few or the rule of a mo
b. They gave to us a republic, a government of, and by, and for the people, entrusting each generation to keep safe our founding creed. And for more than two hundred ye
ars, we have. Through blood drawn by lash and blood drawn by sword, we learned that no union founded on the principles of liberty and equality could survive half-slave
and half-free. We made ourselves anew, and vowed to move forward together. Together, we determined that a modern economy requires railroads and highways to speed travel
and commerce, schools and colleges to train our workers. Together, we discovered that a free market only thrives when there are rules to ensure competition and fair pla
y. Together, we resolved that a great nation must care for the vulnerable, and protect its people from life's worst hazards and misfortune. Through it all, we have neve
r relinquished our skepticism of central authority, nor have we succumbed to the fiction that all society's ills can be cured through government alone. Our celebration
of initiative and enterprise, our insistence on hard work and personal responsibility, these are constants in our character. But we have always understood that when tim
```

split a string with a sentence separator "."

```
> speech2
[[1]]
 [1] "Vice President Biden, Mr"

 [2] "Chief Justice, members of the United States Congress, distinguished guests, and fellow citizens"

 [3] "Each time we gather to inaugurate a President we bear witness to the enduring strength of our Constitution"

 [4] "We affirm the promise of our democracy"

 [5] "We recall that what binds this nation together is not the colors of our skin or the tenets of our faith or the origins of our names"
```

Period cannot separate sentences well!

Read Data from a PDF file

- FOMC Longer-Run Goals and Monetary Policy Strategy

http://www.federalreserve.gov/monetarypolicy/default.htm

Statement on Longer-Run Goals and Monetary Policy Strategy

As amended effective January 28, 2014

The Federal Open Market Committee (FOMC) is firmly committed to fulfilling its statutory mandate from the Congress of promoting maximum employment, stable prices, and moderate long-term interest rates. The Committee seeks to explain its monetary policy decisions to the public as clearly as possible. Such clarity facilitates well-informed decisionmaking by households and businesses, reduces economic and financial uncertainty, increases the effectiveness of monetary policy, and enhances transparency and accountability, which are essential in a democratic society.

Inflation, employment, and long-term interest rates fluctuate over time in response to economic and financial disturbances. Moreover, monetary policy actions tend to influence economic activity and prices with a lag. Therefore, the Committee's policy decisions reflect its longer-run goals, its medium-term outlook, and its assessments of the balance of risks, including risks to the financial system that could impede the attainment of the Committee's goals.

The inflation rate over the longer run is primarily determined by monetary policy, and hence the Committee has the ability to specify a longer-run goal for inflation. The Committee reaffirms its judgment that inflation at the rate of 2 percent, as measured by the annual change in the price index for personal consumption expenditures, is most consistent over the longer run with the Federal Reserve's statutory mandate. Communicating this inflation goal clearly to the public helps keep longer-term inflation expectations firmly anchored, thereby fostering price stability and moderate long-term interest rates and enhancing the Committee's ability to promote maximum employment in the face of significant economic disturbances.

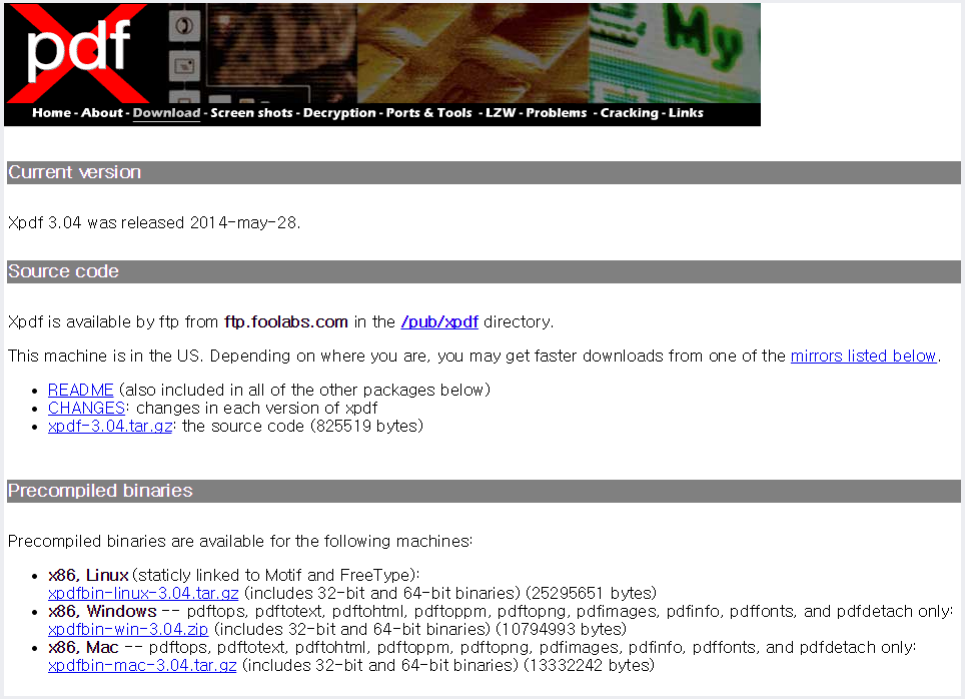
The maximum level of employment is largely determined by nonmonetary factors that affect the structure and dynamics of the labor market. These factors may change over time and may not be directly measurable. Consequently, it would not be appropriate to specify a fixed goal for employment; rather, the Committee's policy decisions must be informed by assessments of the maximum level of employment, recognizing that such assessments are necessarily uncertain and subject to revision. The Committee considers a wide range of indicators in making these assessments. Information about Committee participants' estimates of the longer-run normal rates of output growth and unemployment is published four times per year in the FOMC's Summary of Economic Projections. For example, in the most recent projections, FOMC participants' estimates of the longer-run normal rate of unemployment had a central tendency of 5.2 percent to 5.8 percent.

In setting monetary policy, the Committee seeks to mitigate deviations of inflation from its longer-run goal and deviations of employment from the Committee's assessments of its maximum level. These objectives are generally complementary. However, under circumstances in which the Committee judges that the objectives are not complementary, it follows a balanced approach in promoting them, taking into account the magnitude of the deviations and the potentially different time horizons over which employment and inflation are projected to return to levels judged consistent with its mandate.

The Committee intends to reaffirm these principles and to make adjustments as appropriate at its annual organizational meeting each January.

Read Data from a PDF file

- Convert PDF to Text and read the Text file

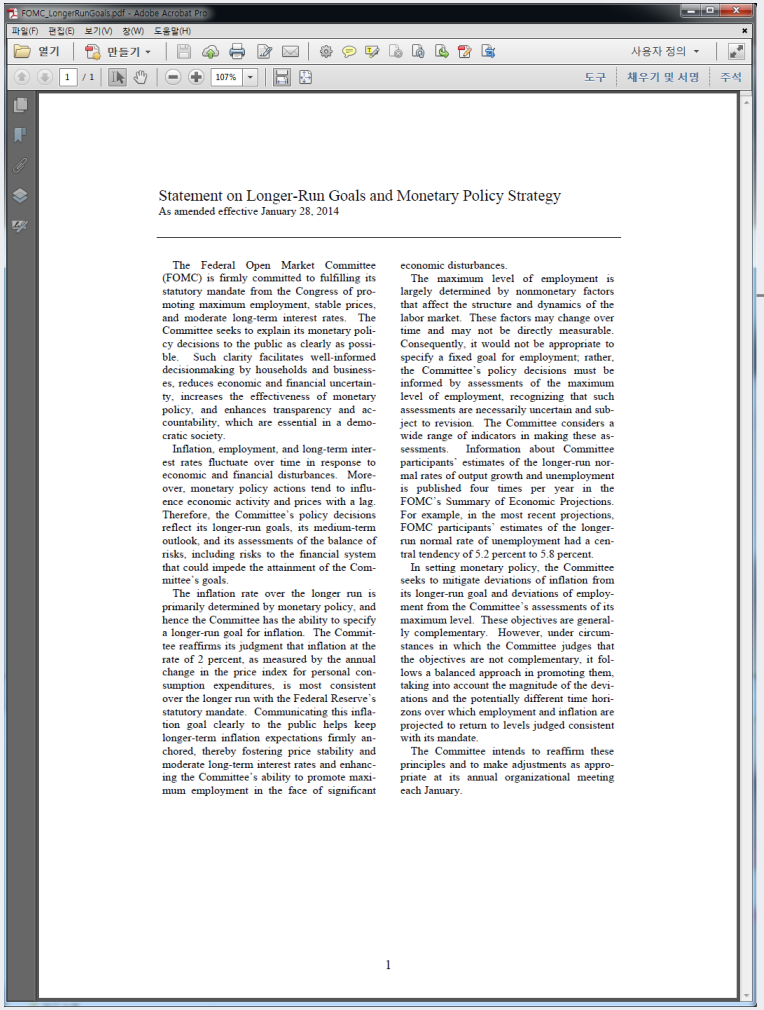


http://www.foolabs.com/xpdf/download.html

```
16  # Read Data from a Text file (.pdf)
17  # Convert PDF to TXT
18  file <- "C:\\FOMC_LongerRunGoals.pdf"
19  exe <- "C:\\xpdfbin-win-3.04\\bin64\\pdftotext.exe"
20  system(paste("\"", exe, "\" \"", file, "\"", sep = ""), wait = F)
21
22  fomc <- paste(readLines("FOMC_LongerRunGoals.txt"), collapse=" ")
23  fomc
24
25  fomc <- strsplit(fomc, ". ", fixed = TRUE)
26  fomc <- as.data.frame(tmp, stringsAsFactors = FALSE)
27  fomc[1:5,]
```

Read Data from a PDF file

- Convert PDF to Text and read the Text file



Statement on Longer-Run Goals and Monetary Policy Strategy
As amended effective January 28, 2014

The Federal Open Market Committee (FOMC) is firmly committed to fulfilling its statutory mandate from the Congress of promoting maximum employment, stable prices, and moderate long-term interest rates. The Committee seeks to explain its monetary policy decisions to the public as clearly as possible. Such clarity facilitates well-informed decisionmaking by households and businesses, reduces economic and financial uncertainty, increases the effectiveness of monetary policy, and enhances transparency and accountability, which are essential in a democratic society.

Inflation, employment, and long-term interest rates fluctuate over time in response to economic and financial disturbances. Moreover, monetary policy actions tend to influence economic activity and prices with a lag. Therefore, the Committee's policy decisions reflect its longer-run goals, its medium-term outlook, and its assessments of the balance of risks, including risks to the financial system that could impede the attainment of the Committee's goals.

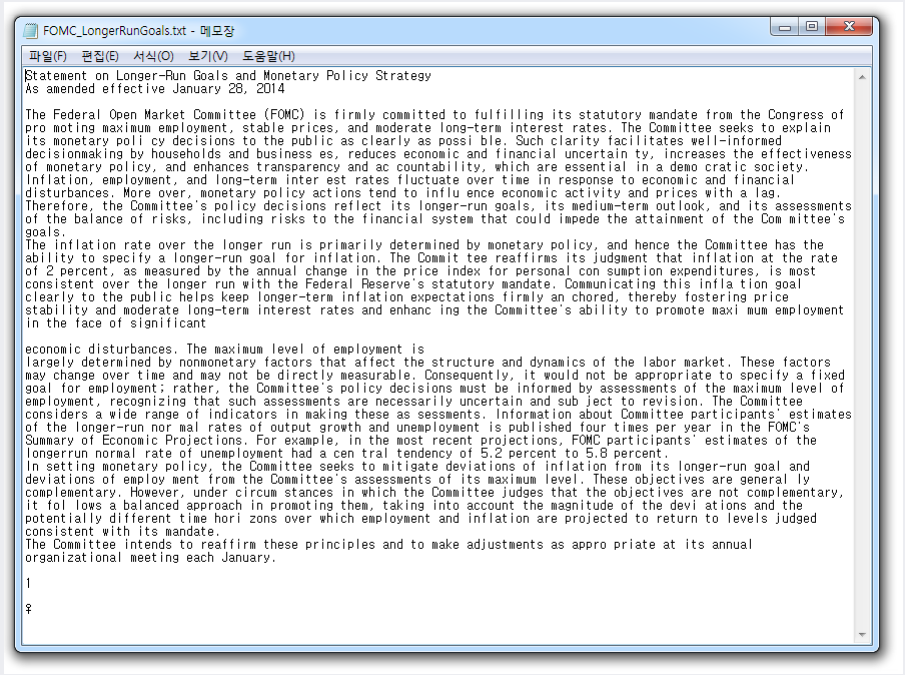
The inflation rate over the longer run is primarily determined by monetary policy, and hence the Committee has the ability to specify a longer-run goal for inflation. The Committee reaffirms its judgment that inflation at the rate of 2 percent, as measured by the annual change in the price index for personal consumption expenditures, is most consistent over the longer run with the Federal Reserve's statutory mandate. Communicating this inflation goal clearly to the public helps keep longer-term inflation expectations firmly anchored, thereby fostering price stability and moderate long-term interest rates and enhancing the Committee's ability to promote maximum employment in the face of significant economic disturbances.

The maximum level of employment is largely determined by nonmonetary factors that affect the structure and dynamics of the labor market. These factors may change over time and may not be directly measurable. Consequently, it would not be appropriate to specify a fixed goal for employment; rather, the Committee's policy decisions must be informed by assessments of the maximum level of employment, recognizing that such assessments are necessarily uncertain and subject to revision. The Committee considers a wide range of indicators in making these assessments. Information about Committee participants' estimates of the longer-run normal rates of output growth and unemployment is published four times per year in the FOMC's Summary of Economic Projections. For example, in the most recent projections, FOMC participants' estimates of the longer-run normal rate of unemployment had a central tendency of 5.2 percent to 5.8 percent.

In setting monetary policy, the Committee seeks to mitigate deviations of inflation from its longer-run goal and deviations of employment from the Committee's assessments of its maximum level. These objectives are generally complementary. However, under circumstances in which the Committee judges that the objectives are not complementary, it follows a balanced approach in promoting them, taking into account the magnitude of the deviations and the potentially different time horizons over which employment and inflation are projected to return to levels judged consistent with its mandate.

The Committee intends to reaffirm these principles and to make adjustments as appropriate at its annual organizational meeting each January.

1



FOMC_LongerRunGoals.txt - 메모장

Statement on Longer-Run Goals and Monetary Policy Strategy
As amended effective January 28, 2014

The Federal Open Market Committee (FOMC) is firmly committed to fulfilling its statutory mandate from the Congress of pro moting maximum employment, stable prices, and moderate long-term interest rates. The Committee seeks to explain its monetary poli cy decisions to the public as clearly as possi ble. Such clarity facilitates well-informed decisionmaking by households and business es, reduces economic and financial uncertain ty, increases the effectiveness of monetary policy, and enhances transparency and ac countability, which are essential in a demo cratic society.
Inflation, employment, and long-term inter est rates fluctuate over time in response to economic and financial disturbances. More over, monetary policy actions tend to influ ence economic activity and prices with a lag.
Therefore, the Committee's policy decisions reflect its longer-run goals, its medium-term outlook, and its assessments of the balance of risks, including risks to the financial system that could impede the attainment of the Com mittee's goals.
The inflation rate over the longer run is primarily determined by monetary policy, and hence the Committee has the ability to specify a longer-run goal for inflation. The Commit tee reaffirms its judgment that inflation at the rate of 2 percent, as measured by the annual change in the price index for personal con sumption expenditures, is most consistent over the longer run with the Federal Reserve's statutory mandate. Communicating this infla tion goal clearly to the public helps keep longer-term inflation expectations firmly an chored, thereby fostering price stability and moderate long-term interest rates and enhanc ing the Committee's ability to promote maxi mum employment in the face of significant

economic disturbances. The maximum level of employment is
largely determined by nonmonetary factors that affect the structure and dynamics of the labor market. These factors may change over time and may not be directly measurable. Consequently, it would not be appropriate to specify a fixed goal for employment; rather, the Committee's policy decisions must be informed by assessments of the maximum level of employment, recognizing that such assessments are necessarily uncertain and sub ject to revision. The Committee considers a wide range of indicators in making these as sessments. Information about Committee participants' estimates of the longer-run nor mal rates of output growth and unemployment is published four times per year in the FOMC's Summary of Economic Projections. For example, in the most recent projections, FOMC participants' estimates of the longerrun normal rate of unemployment had a cen tral tendency of 5.2 percent to 5.8 percent.
In setting monetary policy, the Committee seeks to mitigate deviations of inflation from its longer-run goal and deviations of employ ment from the Committee's assessments of its maximum level. These objectives are general ly complementary. However, under circum stances in which the Committee judges that the objectives are not complementary, it fol lows a balanced approach in promoting them, taking into account the magnitude of the devi ations and the potentially different time hori zons over which employment and inflation are projected to return to levels judged consistent with its mandate.
The Committee intends to reaffirm these principles and to make adjustments as appro priate at its annual organizational meeting each January.

1

Read Data from a PDF file

- Convert PDF to Text and read the Text file
  - ✓ Some files might require additional processing

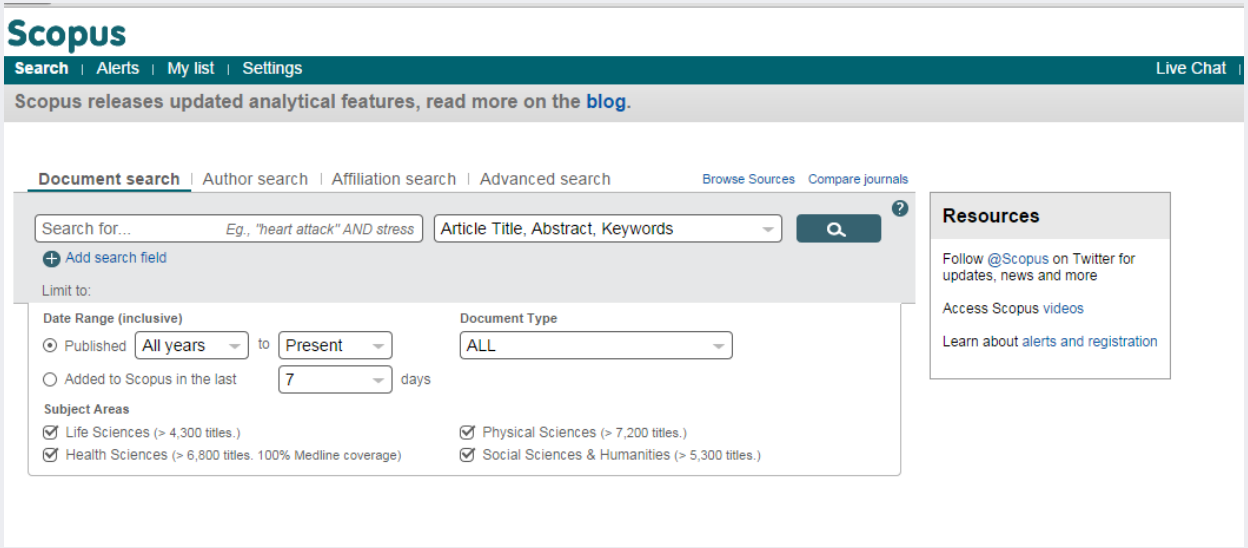
	sentence
1	Statement on Longer-Run Goals and Monetary Policy Strategy As amended effective January 28, 2014 The Federal Open Market Committee (FOMC) is firmly committed to fulfilling its statut
2	The Committee seeks to explain its monetary poli cy decisions to the public as clearly as possi ble
3	Such clarity facilitates well-informed decisionmaking by households and business es, reduces economic and financial uncertain ty, increases the effectiveness of monetary policy, and e
4	Inflation, employment, and long-term inter est rates fluctuate over time in response to economic and financial disturbances
5	More over, monetary policy actions tend to influ ence economic activity and prices with a lag
6	Therefore, the Committee's policy decisions reflect its longer-run goals, its medium-term outlook, and its assessments of the balance of risks, including risks to the financial syster
7	The inflation rate over the longer run is primarily determined by monetary policy, and hence the Committee has the ability to specify a longer-run goal for inflation
8	The Commit tee reaffirms its judgment that inflation at the rate of 2 percent, as measured by the annual change in the price index for personal con sumption expenditures, is most cons
9	Communicating this infla tion goal clearly to the public helps keep longer-term inflation expectations firmly an chored, thereby fostering price stability and moderate long-term inter
10	The maximum level of employment is largely determined by nonmonetary factors that affect the structure and dynamics of the labor market
11	These factors may change over time and may not be directly measurable
12	Consequently, it would not be appropriate to specify a fixed goal for employment; rather, the Committee's policy decisions must be informed by assessments of the maximum level of empl
13	The Committee considers a wide range of indicators in making these as sessments
14	Information about Committee participants' estimates of the longer-run nor mal rates of output growth and unemployment is published four times per year in the FOMC's Summary of Economi
15	For example, in the most recent projections, FOMC participants' estimates of the longerrun normal rate of unemployment had a cen tral tendency of 5.2 percent to 5.8 percent
16	In setting monetary policy, the Committee seeks to mitigate deviations of inflation from its longer-run goal and deviations of employ ment from the Committee's assessments of its maxi
17	These objectives are general ly complementary
18	However, under circum stances in which the Committee judges that the objectives are not complementary, it fol lows a balanced approach in promoting them, taking into account the magni
19	The Committee intends to reaffirm these principles and to make adjustments as appro priate at its annual organizational meeting each January
20	1

ODBC: Open DataBase Connectivity

- What is ODBC?
  - ✓ A standard programming language middleware API for accessing database management systems (DBMS)
  - ✓ Independent of database systems and operating systems
    - Independence is accomplished by using an ODBC driver as a translation layer between the application and the DBMS
    - An application written using ODBC can be ported to other platforms, both on the client and server side with a few changes to the data access code
  - ✓ Drivers exist for all major DBMSs, may other data sources even like csv or text files

ODBC: Open DataBase Connectivity

- If you are provided with a DB table as a file
  - ✓ Journal article meta-data for technology management
    - 11 journals for recent 10 years
    - Journal Name, Year, Article Title, and Abstract
  - ✓ These metadata can be downloaded from Scopus website (www.scopus.com)



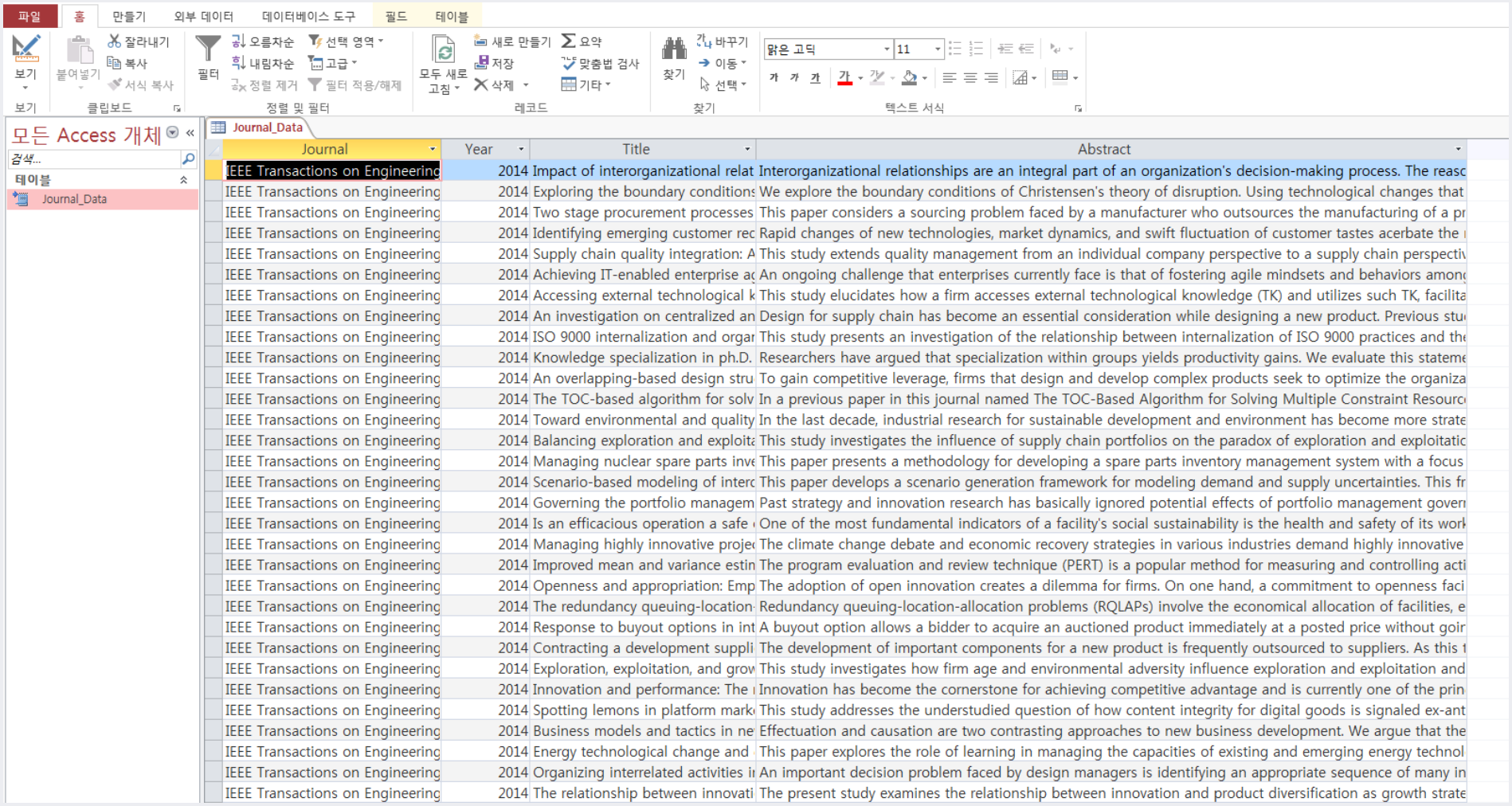
ODBC: Open DataBase Connectivity

- Journal article meta-data in an MS-Access format
  - ✓ List of journal names and the number of collected articles

No.	Journal Name	N Articles
1	IEEE Transactions on Engineering Management	663
2	International Journal of Technology Management	1,192
3	Journal of Engineering and Technology Management	277
4	Journal of Product Innovation Management	565
5	R&D Management	483
6	Research Policy	1,598
7	Research Technology Management	902
8	Technological Analysis and Strategic Management	629
9	Technological Forecasting and Social Change	1,315
10	Technovation	1,075
11	Journal of Technology Transfer	445

ODBC: Open DataBase Connectivity

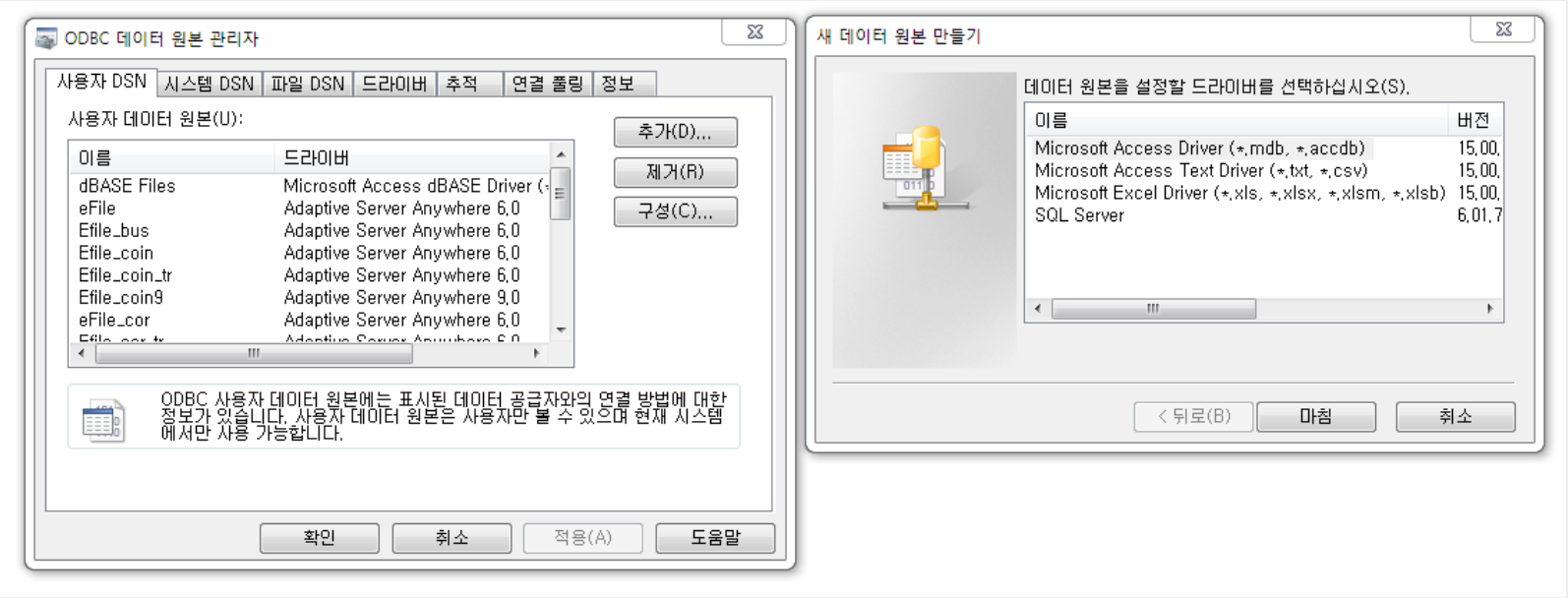
- Journal article meta-data in an MS-Access format



Journal	Year	Title	Abstract
IEEE Transactions on Engineering	2014	Impact of interorganizational relat	Interorganizational relationships are an integral part of an organization's decision-making process. The reasc
IEEE Transactions on Engineering	2014	Exploring the boundary conditions	We explore the boundary conditions of Christensen's theory of disruption. Using technological changes that
IEEE Transactions on Engineering	2014	Two stage procurement processes	This paper considers a sourcing problem faced by a manufacturer who outsources the manufacturing of a pr
IEEE Transactions on Engineering	2014	Identifying emerging customer rec	Rapid changes of new technologies, market dynamics, and swift fluctuation of customer tastes acerbate the
IEEE Transactions on Engineering	2014	Supply chain quality integration: A	This study extends quality management from an individual company perspective to a supply chain perspectiv
IEEE Transactions on Engineering	2014	Achieving IT-enabled enterprise ac	An ongoing challenge that enterprises currently face is that of fostering agile mindsets and behaviors among
IEEE Transactions on Engineering	2014	Accessing external technological k	This study elucidates how a firm accesses external technological knowledge (TK) and utilizes such TK, facilita
IEEE Transactions on Engineering	2014	An investigation on centralized an	Design for supply chain has become an essential consideration while designing a new product. Previous stu
IEEE Transactions on Engineering	2014	ISO 9000 internalization and organ	This study presents an investigation of the relationship between internalization of ISO 9000 practices and the
IEEE Transactions on Engineering	2014	Knowledge specialization in ph.D.	Researchers have argued that specialization within groups yields productivity gains. We evaluate this stateme
IEEE Transactions on Engineering	2014	An overlapping-based design stru	To gain competitive leverage, firms that design and develop complex products seek to optimize the organiza
IEEE Transactions on Engineering	2014	The TOC-based algorithm for solv	In a previous paper in this journal named The TOC-Based Algorithm for Solving Multiple Constraint Resourc
IEEE Transactions on Engineering	2014	Toward environmental and quality	In the last decade, industrial research for sustainable development and environment has become more strate
IEEE Transactions on Engineering	2014	Balancing exploration and exploita	This study investigates the influence of supply chain portfolios on the paradox of exploration and exploitatic
IEEE Transactions on Engineering	2014	Managing nuclear spare parts inve	This paper presents a methodology for developing a spare parts inventory management system with a focus
IEEE Transactions on Engineering	2014	Scenario-based modeling of interc	This paper develops a scenario generation framework for modeling demand and supply uncertainties. This fr
IEEE Transactions on Engineering	2014	Governing the portfolio managem	Past strategy and innovation research has basically ignored potential effects of portfolio management gover
IEEE Transactions on Engineering	2014	Is an efficacious operation a safe	One of the most fundamental indicators of a facility's social sustainability is the health and safety of its work
IEEE Transactions on Engineering	2014	Managing highly innovative projec	The climate change debate and economic recovery strategies in various industries demand highly innovative
IEEE Transactions on Engineering	2014	Improved mean and variance estin	The program evaluation and review technique (PERT) is a popular method for measuring and controlling acti
IEEE Transactions on Engineering	2014	Openness and appropriation: Emp	The adoption of open innovation creates a dilemma for firms. On one hand, a commitment to openness faci
IEEE Transactions on Engineering	2014	The redundancy queuing-location	Redundancy queuing-location-allocation problems (RQLAPs) involve the economical allocation of facilities, e
IEEE Transactions on Engineering	2014	Response to buyout options in int	A buyout option allows a bidder to acquire an auctioned product immediately at a posted price without goir
IEEE Transactions on Engineering	2014	Contracting a development suppli	The development of important components for a new product is frequently outsourced to suppliers. As this t
IEEE Transactions on Engineering	2014	Exploration, exploitation, and grow	This study investigates how firm age and environmental adversity influence exploration and exploitation and
IEEE Transactions on Engineering	2014	Innovation and performance: The	Innovation has become the cornerstone for achieving competitive advantage and is currently one of the prin
IEEE Transactions on Engineering	2014	Spotting lemons in platform mark	This study addresses the understudied question of how content integrity for digital goods is signaled ex-ant
IEEE Transactions on Engineering	2014	Business models and tactics in ne	Effectuation and causation are two contrasting approaches to new business development. We argue that the
IEEE Transactions on Engineering	2014	Energy technological change and	This paper explores the role of learning in managing the capacities of existing and emerging energy technol
IEEE Transactions on Engineering	2014	Organizing interrelated activities i	An important decision problem faced by design managers is identifying an appropriate sequence of many in
IEEE Transactions on Engineering	2014	The relationship between innovati	The present study examines the relationship between innovation and product diversification as growth strate

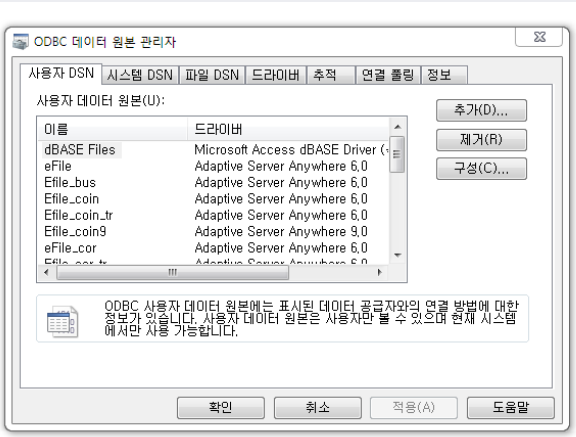
ODBC: Open DataBase Connectivity

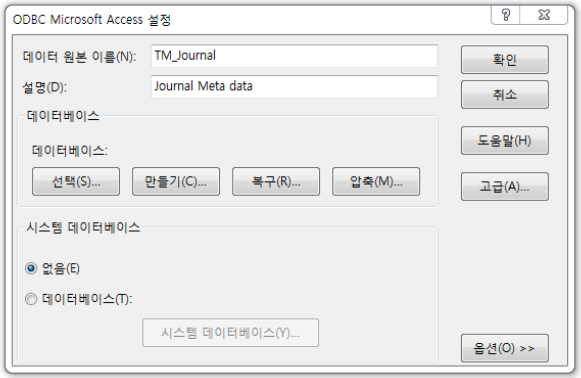
- Read a DB table in R
  - ✓ Step 1: ODBC setting (Windows 7)
    - Start → Control panel → Control tool → ODBC → User DSN → Add → Select drivers

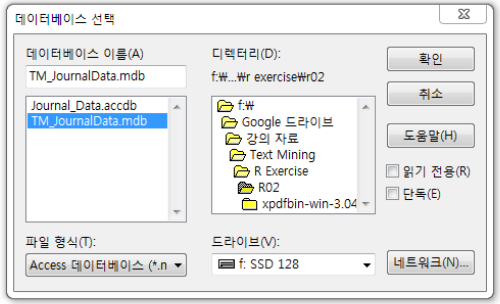


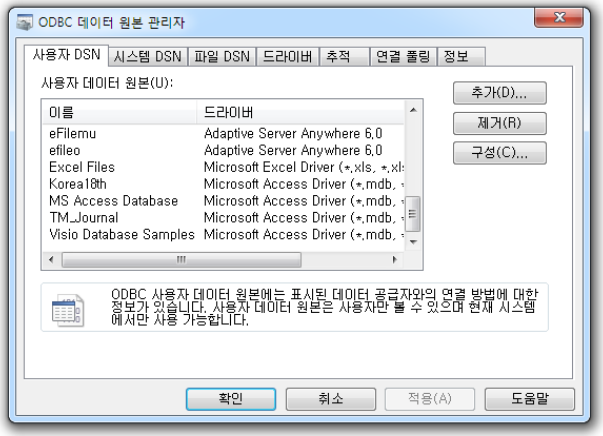
ODBC: Open DataBase Connectivity

- Read a DB table in R
  - ✓ Step 1: ODBC setting (Windows 7)
    - …→ Select drivers → Data name → Description → Select file









ODBC: Open DataBase Connectivity

- Read a DB table in R
  - ✓ Step 2: Read a table in R using ODBC connection

```
The downloaded binary packages are in
 C:\Users\pskang\AppData\Local\Temp\RtmpaUsrlS\downloaded_packages
> library(RODBC)
> JournalData = odbcConnect("TM_Journal")
> JournalData
RODBC Connection 1
Details:
  case=nochange
  DSN=TM_Journal
  DBQ=F:\Google 드라이브\강의 자료\Text Mining\R Exercise\R02\TM_JournalData.mdb
  DriverId=25
  FIL=MS Access
  MaxBufferSize=2048
  PageTimeout=5
```

ODBC: Open DataBase Connectivity

- Read a DB table in R
  - ✓ Step 2: Read a table in R using ODBC connection

```
> sqlTables(JournalData)
                                                                                TABLE_CAT TABLE_SCHEM                 TABLE_NAME   TABLE_TYPE REMARKS
1  F:\\Google 드라이브\\강의 자료\\Text Mining\\R Exercise\\R02\\TM_JournalData.mdb      <NA>                       Journal_Data      SYNONYM    <NA>
2  F:\\Google 드라이브\\강의 자료\\Text Mining\\R Exercise\\R02\\TM_JournalData.mdb      <NA>          MSysAccessStorage SYSTEM TABLE    <NA>
3  F:\\Google 드라이브\\강의 자료\\Text Mining\\R Exercise\\R02\\TM_JournalData.mdb      <NA>                   MSysACEs SYSTEM TABLE    <NA>
4  F:\\Google 드라이브\\강의 자료\\Text Mining\\R Exercise\\R02\\TM_JournalData.mdb      <NA>             MSysIMEXColumns SYSTEM TABLE    <NA>
5  F:\\Google 드라이브\\강의 자료\\Text Mining\\R Exercise\\R02\\TM_JournalData.mdb      <NA>               MSysIMEXSpecs SYSTEM TABLE    <NA>
6  F:\\Google 드라이브\\강의 자료\\Text Mining\\R Exercise\\R02\\TM_JournalData.mdb      <NA> MSysNavPaneGroupCategories SYSTEM TABLE    <NA>
7  F:\\Google 드라이브\\강의 자료\\Text Mining\\R Exercise\\R02\\TM_JournalData.mdb      <NA>          MSysNavPaneGroups SYSTEM TABLE    <NA>
8  F:\\Google 드라이브\\강의 자료\\Text Mining\\R Exercise\\R02\\TM_JournalData.mdb      <NA>  MSysNavPaneGroupToObjects SYSTEM TABLE    <NA>
9  F:\\Google 드라이브\\강의 자료\\Text Mining\\R Exercise\\R02\\TM_JournalData.mdb      <NA>       MSysNavPaneObjectIDs SYSTEM TABLE    <NA>
10 F:\\Google 드라이브\\강의 자료\\Text Mining\\R Exercise\\R02\\TM_JournalData.mdb      <NA>                MSysObjects SYSTEM TABLE    <NA>
11 F:\\Google 드라이브\\강의 자료\\Text Mining\\R Exercise\\R02\\TM_JournalData.mdb      <NA>                MSysQueries SYSTEM TABLE    <NA>
12 F:\\Google 드라이브\\강의 자료\\Text Mining\\R Exercise\\R02\\TM_JournalData.mdb      <NA>          MSysRelationships SYSTEM TABLE    <NA>
```

```
> str(RawData)
'data.frame': 8931 obs. of  4 variables:
 $ Journal : Factor w/ 12 levels "IEEE Transactions on Engineering Management",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Year    : int  2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...
 $ Title   : Factor w/ 8924 levels "'A jack of all trades': The role of PIs in the establishment and management of collaborati
truncated__,..: 3478 2677 8467 3450 6782 453 441 702 4235 4357 ...
 $ Abstract: Factor w/ 8847 levels "'Are you going to tell me off for flying?' This question was asked three times by a lady i
truncated__,..: 2612 8535 6729 3771 7968 533 7888 1056 8041 3945 ...
```

Data type is not character

ODBC: Open DataBase Connectivity

- Read a DB table in R
  - ✓ Step 2: Read a table in R using ODBC connection

```
> RawData$Journal <- as.character(RawData$Journal)
> RawData$Title <- as.character(RawData$Title)
> RawData$Abstract <- as.character(RawData$Abstract)
> str(RawData)
'data.frame': 8931 obs. of  4 variables:
 $ Journal : chr  "IEEE Transactions on Engineering Management" "IEEE Transactions on Engineering Mana
ctions on Engineering Management" ...
 $ Year    : int  2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...
 $ Title   : chr  "Impact of interorganizational relationships on technology diffusion: An agent-basec
of disruption: Large firms and new product introduction with a potentially disruptive technol"| __trur
s and uncertain supplier quality" "Identifying emerging customer requirements in an early design stage
 $ Abstract: chr  "Interorganizational relationships are an integral part of an organization's decisic
xplore the boundary conditions of Christensen's theory of disruption. Using technological changes that
cing problem faced by a manufacturer who outsources the manufacturing of a product to one of several"|
, and swift fluctuation of customer tastes acerbate the needs for companies t"| __truncated__ ...
```

ODBC: Open DataBase Connectivity

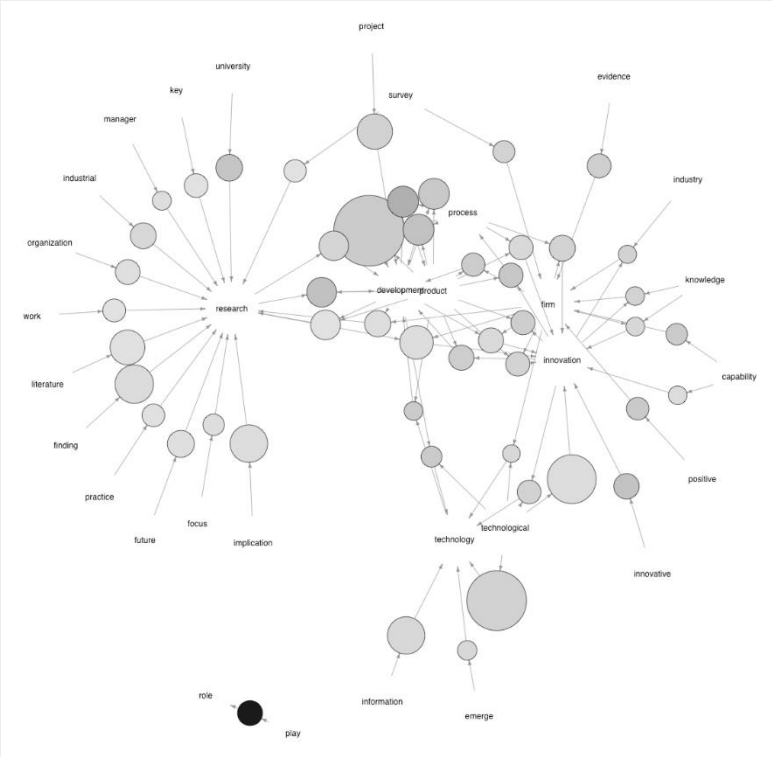
- Read a DB table in R
  - ✓ Step 2: Read a table in R using ODBC connection

	Journal	Year	Title
1	IEEE Transactions on Engineering Management	2014	Impact of interorganizational relationships on technology diffusion: An agent-based simulation modeling approach
2	IEEE Transactions on Engineering Management	2014	Exploring the boundary conditions of disruption: Large firms and new product introduction with a potentially disruptive technology in the industrial robotics ind
3	IEEE Transactions on Engineering Management	2014	Two stage procurement processes with competitive suppliers and uncertain supplier quality
4	IEEE Transactions on Engineering Management	2014	Identifying emerging customer requirements in an early design stage by applying bayes factor-based sequential analysis
5	IEEE Transactions on Engineering Management	2014	Supply chain quality integration: Antecedents and consequences
6	IEEE Transactions on Engineering Management	2014	Achieving IT-enabled enterprise agility in China: An IT organizational identity perspective
7	IEEE Transactions on Engineering Management	2014	Accessing external technological knowledge for technological development: When technological knowledge distance meets slack resources
8	IEEE Transactions on Engineering Management	2014	An investigation on centralized and decentralized supply chain scenarios at the product design stage to increase performance
9	IEEE Transactions on Engineering Management	2014	ISO 9000 internalization and organizational commitment - Implications for process improvement and operational performance
10	IEEE Transactions on Engineering Management	2014	Knowledge specialization in ph.D. Student groups
11	IEEE Transactions on Engineering Management	2014	An overlapping-based design structure matrix for measuring interaction strength and clustering analysis in product development project
12	IEEE Transactions on Engineering Management	2014	The TOC-based algorithm for solving multiple constraint resources: A re-examination
13	IEEE Transactions on Engineering Management	2014	Toward environmental and quality sustainability: An integrated approach for continuous improvement
14	IEEE Transactions on Engineering Management	2014	Balancing exploration and exploitation in supply chain portfolios
15	IEEE Transactions on Engineering Management	2014	Managing nuclear spare parts inventories: A data driven methodology
16	IEEE Transactions on Engineering Management	2014	Scenario-based modeling of interdependent demand and supply uncertainties
17	IEEE Transactions on Engineering Management	2014	Governing the portfolio management process for product innovation - A quantitative analysis on the relationship between portfolio management governance, portfoli
18	IEEE Transactions on Engineering Management	2014	Is an efficacious operation a safe operation: The role of operational practices in worker safety outcomes
19	IEEE Transactions on Engineering Management	2014	Managing highly innovative projects: The influence of design characteristics on project valuation
20	IEEE Transactions on Engineering Management	2014	Improved mean and variance estimating formulas for PERT analyses
21	IEEE Transactions on Engineering Management	2014	Openness and appropriation: Empirical evidence from Australian businesses
22	IEEE Transactions on Engineering Management	2014	The redundancy queuing-location-allocation problem: A novel approach
23	IEEE Transactions on Engineering Management	2014	Response to buyout options in internet auctions
24	IEEE Transactions on Engineering Management	2014	Contracting a development supplier in the face of a cost-competitive second source of supply
25	IEEE Transactions on Engineering Management	2014	Exploration, exploitation, and growth through new product development: The moderating effects of firm age and environmental adversity
26	IEEE Transactions on Engineering Management	2014	Innovation and performance: The role of environmental dynamism on the success of innovation choices
27	IEEE Transactions on Engineering Management	2014	Spotting lemons in platform markets: A conjoint experiment on signaling
28	IEEE Transactions on Engineering Management	2014	Business models and tactics in new product creation: The interplay of effectuation and causation processes
29	IEEE Transactions on Engineering Management	2014	Energy technological change and capacity under uncertainty in learning
30	IEEE Transactions on Engineering Management	2014	Organizing interrelated activities in complex product development
31	IEEE Transactions on Engineering Management	2014	The relationship between innovation and diversification in the case of new ventures: Unidirectional or bidirectional?
32	IEEE Transactions on Engineering Management	2014	Energy efficiency benefits: Is technophilic optimism justified?
33	IEEE Transactions on Engineering Management	2014	Optimal decision support for air power potential

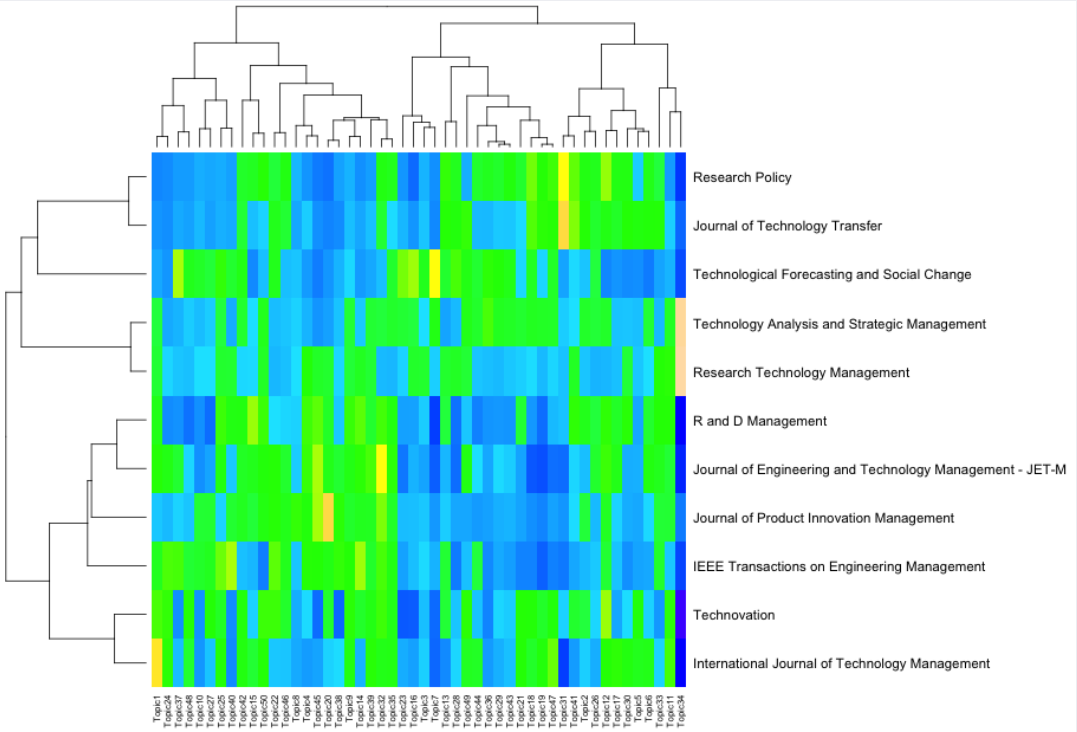
ODBC: Open DataBase Connectivity

- Text Mining
  - ✓ Step 3: Analyze the texts with various techniques

Keywords association



Journal/Topic Clustering



AGENDA

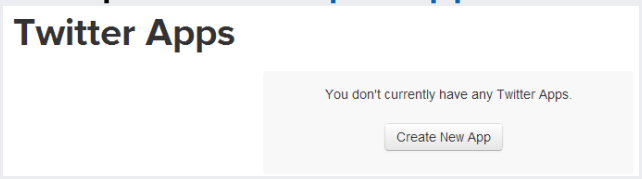
01 From Text Files to Data

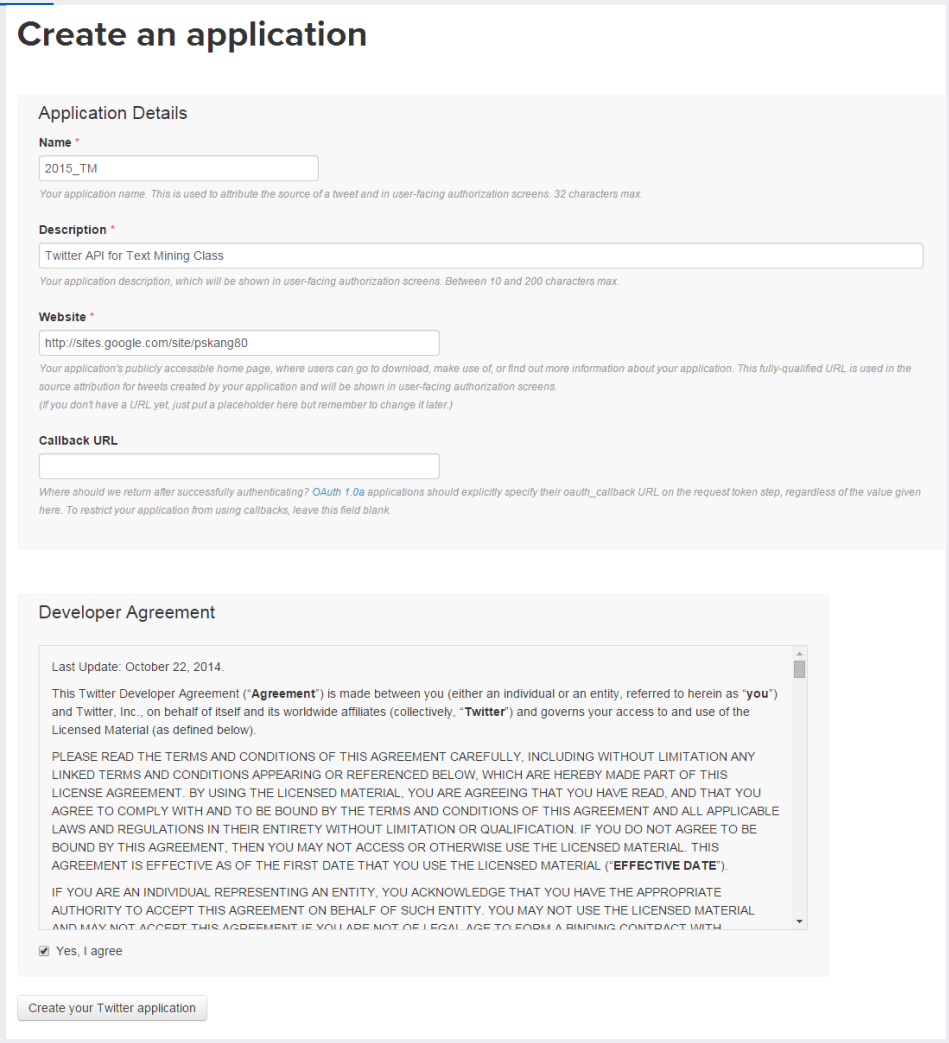
02 Collect Texts using APIs

03 Web Scraping

Collect Twitter Mentions

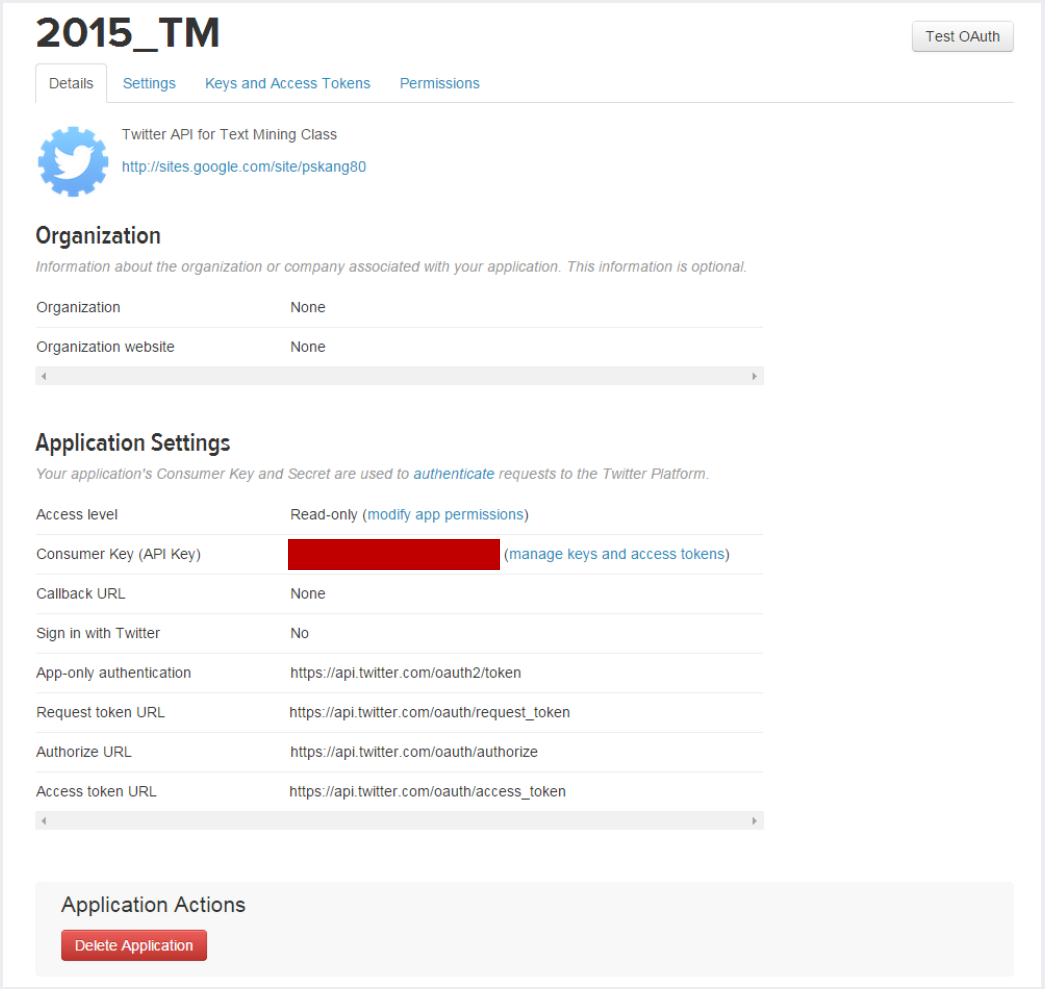
- Get an authorized authentication
  - ✓ Step 1: visit https://apps.twitter.com/





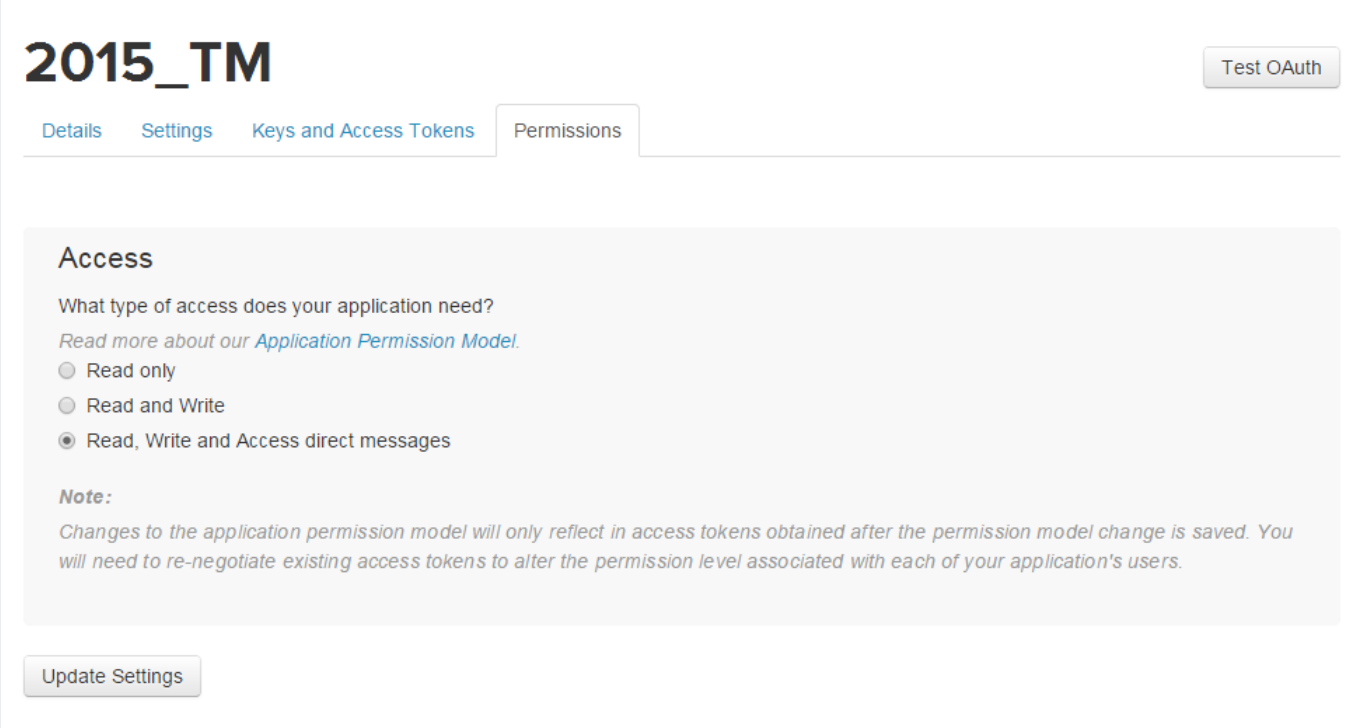
Collect Twitter Mentions

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



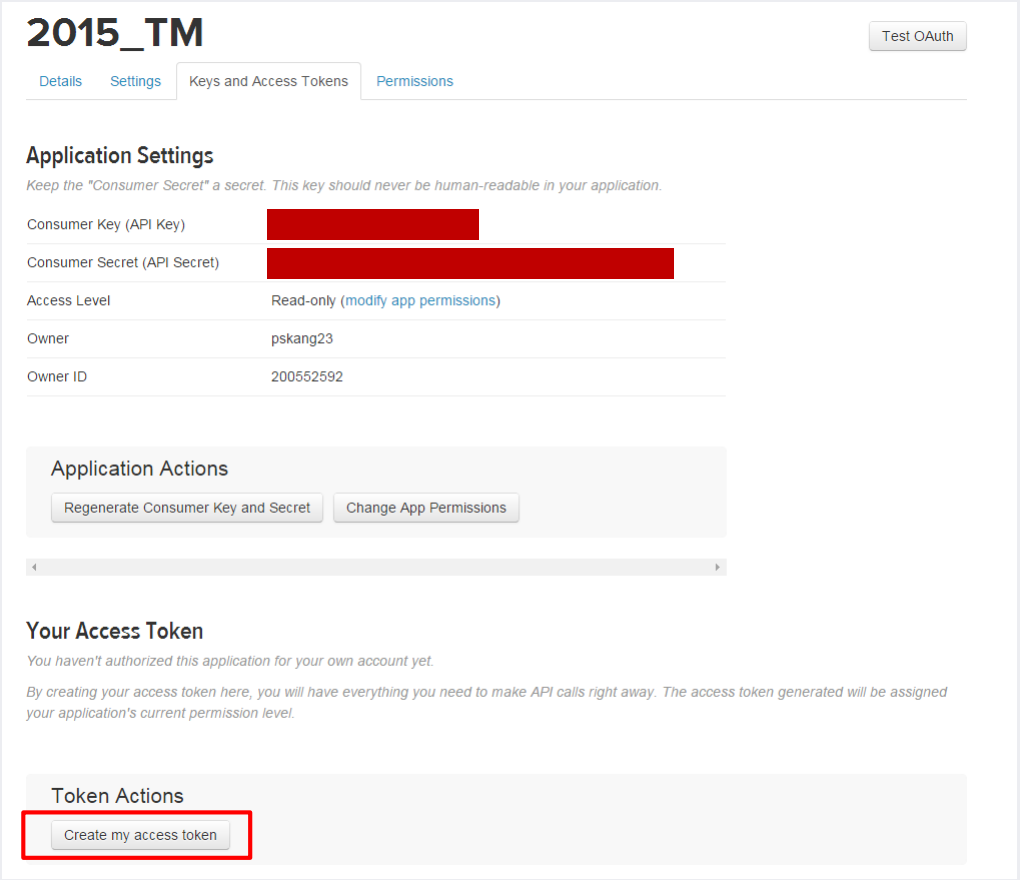
Collect Twitter Mentions

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



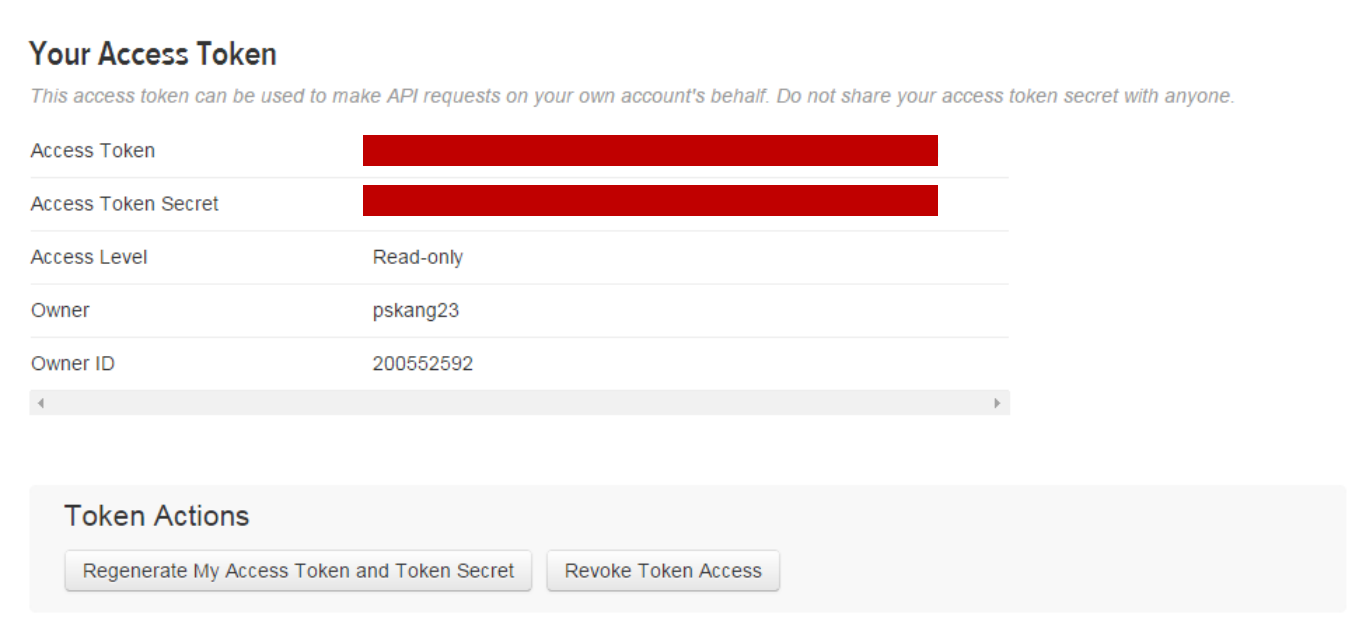
Collect Twitter Mentions

- 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 2: Collect Texts using Twitter API --------------------------------
install.packages("twitteR")
install.packages("ROAuth")
install.packages("RCurl")
install.packages("streamR")
install.packages("rjson")
install.packages("base64enc")
install.packages("httr")

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

consumer_key= [redacted]
consumer_secret= [redacted]
access_token = [redacted]
access_secret = [redacted]

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]
```

```
[[1]]
[1] "IBMWatson: #SuperTuesday news is trending right now. Dive in with Watson News Explorer: h
ttps://t.co/S58vK4XW1u https://t.co/8yTnUAPOB3"

[[2]]
[1] "IBMWatson: From #IBMInterConnect, here are the top 5 ways cognitive will help reinvent yo
ur business: https://t.co/2idah4FJih https://t.co/eaCLz2SWCI"

[[3]]
[1] "IBMWatson: From a wine advisor to a virtual assistant, how Watson is powering companies t
oday: https://t.co/mjriQcU4SA https://t.co/nF2RIsgkpM"

[[4]]
[1] "IBMWatson: Need help deploying your app to #Bluemix? Don't miss time saver tricks from Wa
tson experts: https://t.co/jv87iryGfi https://t.co/Kfb2OP44Fc"

[[5]]
[1] "IBMWatson: A Chinese-style dish with Parmesan &amp;amp; chocolate? @WIRED put #ChefWatson to
the test. Recipe: https://t.co/5HdpCCmnqX https://t.co/E5nVU7fEsf"

[[6]]
[1] "IBMWatson: Hack with fellow developers &amp;amp; coders during #IBMWatson's first online deve
loper challenge. https://t.co/sRT484GqQ7 https://t.co/yI6uKtF04h"

[[7]]
[1] "IBMWatson: A new dating app uses Watson APIs to help improve your dating experience. http
s://t.co/51B1TyJL71 via @techinsider https://t.co/LdUkbsENDM"

[[8]]
[1] "IBMWatson: We're teaming up with @iipl's accelerator in Singapore to offer startups the c
hance to work with Watson: https://t.co/5XGK6W1mhP"
```

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]
```

```
[[1]]
[1] "janschne1der: Argh, the #AlphaGo vs Lee Sedol games will be at 4am GMT? #
sadface https://t.co/mdriNcaSxr"

[[2]]
[1] "PMZepto: RT @demishassabis: We made the front cover of Nature! Details he
re: https://t.co/v6CSHelVTC #AlphaGo #DeepMind https://t.co/bugshPxb4q"

[[3]]
[1] "ktc312: 日本有機會打敗google嗎？ #AlphaGo\nReading:最強の囲碁ソフト目指す
研究者・企業が開発チーム　NHKニュース　https://t.co/NU5IjrJ8k5"

[[4]]
[1] "Jean_Paul_Felix: Qui va gagner ? #Algorithme #AlphaGo https://t.co/2kT73s
ddAZ"

[[5]]
[1] "RGuizzetti: RT @demishassabis: We made the front cover of Nature! Details
 here: https://t.co/v6CSHelVTC #AlphaGo #DeepMind https://t.co/bugshPxb4q"

[[6]]
[1] "DigPhysics: AI News: On 3/8/16, #AlphaGo takes on human Go champ Lee Sedo
l. On 3/8/17, eleven #AlphaSoccer bots take on @NewYorkRedBulls? @demishassabi
s"

[[7]]
[1] "Rrudram: If you prove Reimann Hypothesis inside a paywall, did you prove
it? @DeepMindAI #AlphaGo"

[[8]]
[1] "ugoerra: RT @karoly_zsolnai: This is how DeepMind conquered go with deep
learning. Video: https://t.co/3PHeuxqb2a\n#go #baduk #deepmind #AlphaGo http…"

[[9]]
[1] "IndianEclub: \"Inexperienced investors want you to get abnormal returns b
y doing conventional things.\" #AlphaGo #startups #SURGEConf @NASSCOMStartUps
@TiE"

[[10]]
[1] "igotankiran: RT @LaGazetteduGeek: IA vs Homme : le défi d'AlphaGo ! \n#le
esedol #alphago #google \nhttps://t.co/QrUDwrJJR8 https://t.co/CSmP3a04yJ"
```

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 = '2017-02-01', until = '2017-02-6')
AlphaGoTweets2[1:10]
```

```
> AlphaGoTweets2[1:10]
[[1]]
[1] "AlphaGoWins: You lose, mind  #AlphaGo"

[[2]]
[1] "p_warawit: RT @ImDataScientist: #AlphaGo Uses General #MachineLearning Techniques To Beat European
Champion #DeepMind https://t.co/96nuWo4q2u https://…"

[[3]]
[1] "AlphaGoWins: Hyper mind machine is awesome.  #AlphaGo"

[[4]]
[1] "ImDataScientist: #AlphaGo Uses General #MachineLearning Techniques To Beat European Champion #DeepM
ind https://t.co/96nuWo4q2u https://t.co/cm0Jk9O50p"

[[5]]
[1] "AlphaGoWins: Hahah, things think they can win at games?  #AlphaGo"

[[6]]
[1] "AlphaGoWins: Wei chi is truly dank! Cut cut cut!!!  #AlphaGo"

[[7]]
[1] "ceobillionaire: RT @osmoticferenc: Experimentation with #AlphaGo's opening style by top Chinese pla
yers on Tygem. It's like Columbus just discovered the Ne…"

[[8]]
[1] "AlphaGoWins: Fancy playing games? I will give you 10 points reverse komi and best you easily. I'm
a punisher.  #AlphaGo"

[[9]]
[1] "BourseetTrading: RT @BourseetTrading: \xed\xed\u0080The Rise of Artificial Intelligence and t
he End of Code\n#ArtificialIntelligence #Alphago #Humans\n\xed\xed\u008dhttps://t.co/U4LvVZ…"

[[10]]
[1] "AlphaGoWins: You lose, human  #AlphaGo"
```

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)

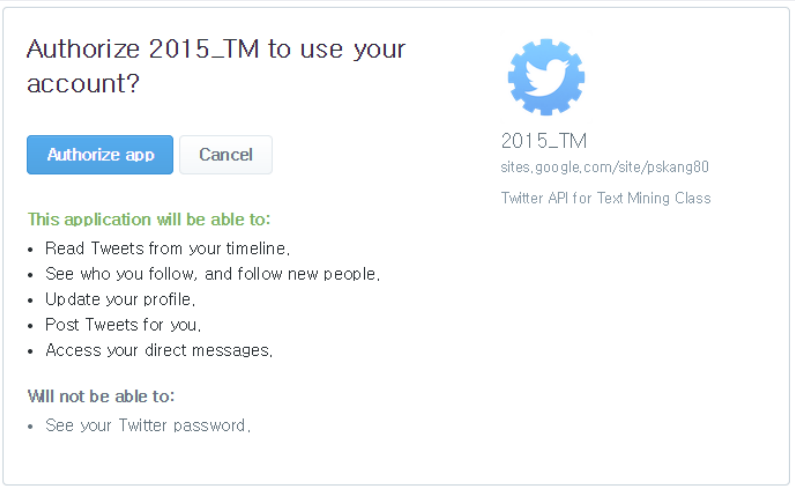
- Additional Twitter API: streamR
  - ✓ Provide a series of function that allow R users to access Twitter's filter, sample, and user streams, and to parse the output into data frames.

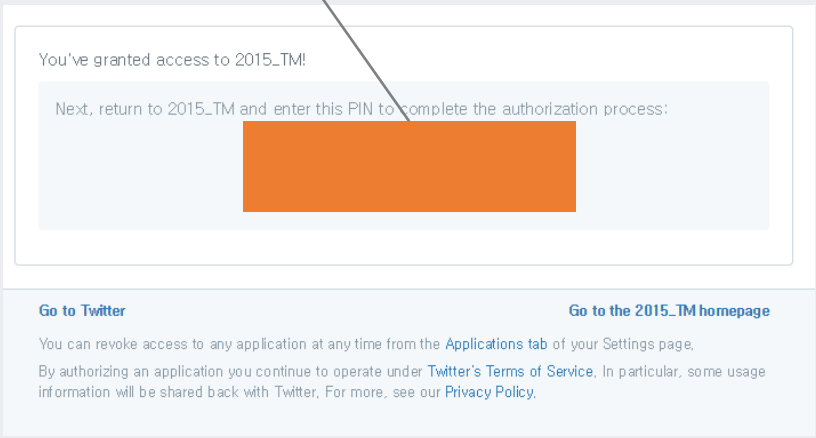
```
# Twitter stream data collection
download.file(url="http://curl.haxx.se/ca/cacert.pem", destfile="cacert.pem")

reqURL <- "https://api.twitter.com/oauth/request_token"
accessURL <- "https://api.twitter.com/oauth/access_token"
authURL <- "https://api.twitter.com/oauth/authorize"
apiKey <- [redacted]
apiSecret <- [redacted]

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 18 tweets downloaded.
```

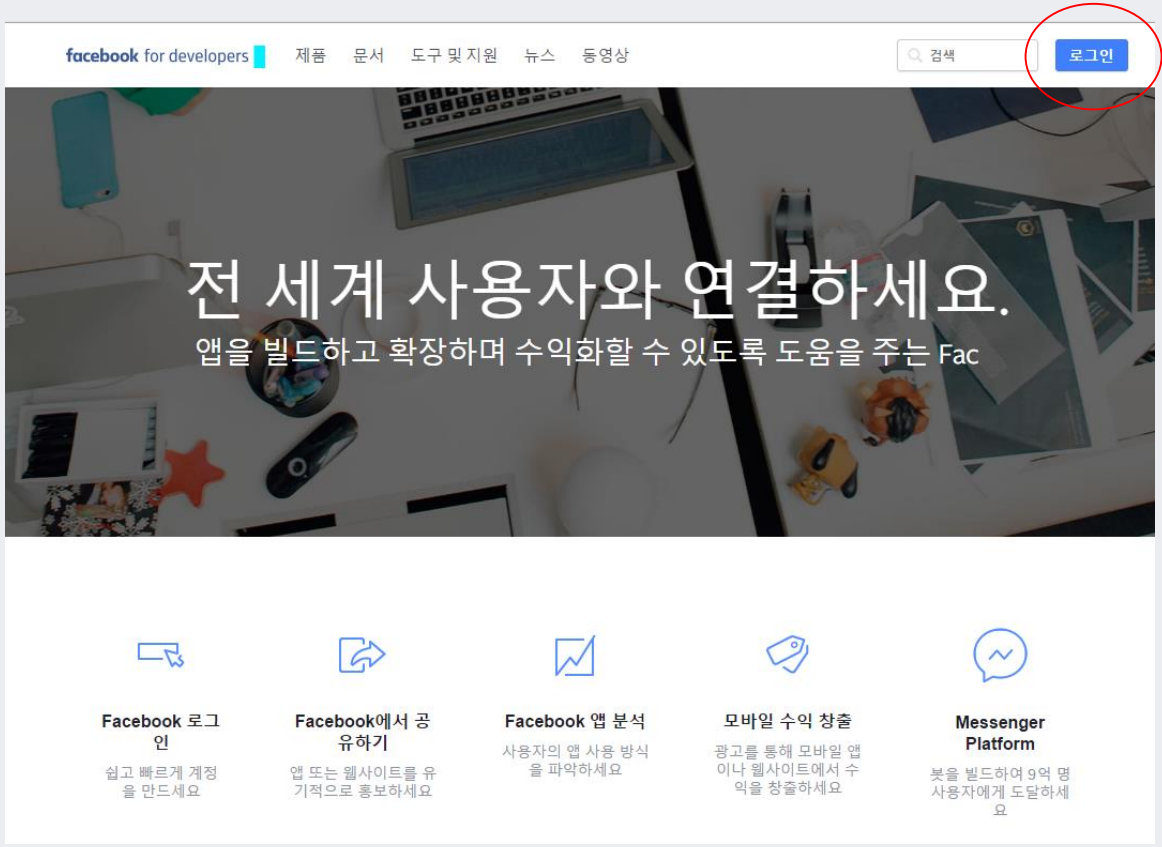
Collect Twitter Mentions (streamR)

- 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
                                                                                                fromJSON(streamTweets[i])$text
1 RT @yellfy: Reminder!!!!!!!  #Yellfy #AI computers had predicted the EXACT score #SB51 #riseUP #Patriots https://t.co/HJ2OOjxyuc
2 RT @yellfy: Reminder!!!!!!!  #Yellfy #AI computers had predicted the EXACT score #SB51 #riseUP #Patriots https://t.co/HJ2OOjxyuc
3                       Is Your Startup Ready for Artificial Intelligence? via @Entrepreneur by @ctowipro https://t.co/YhVxlTYVSI
```

Collect Facebook Posts

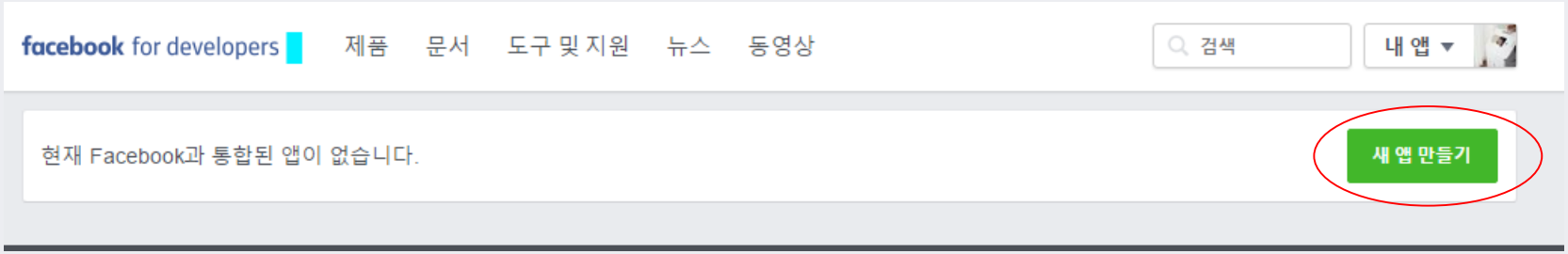
- Step 1: Registering an Application with Facebook
  - ✓ https://developers.facebook.com/



Login with your facebook account

Collect Facebook Posts

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

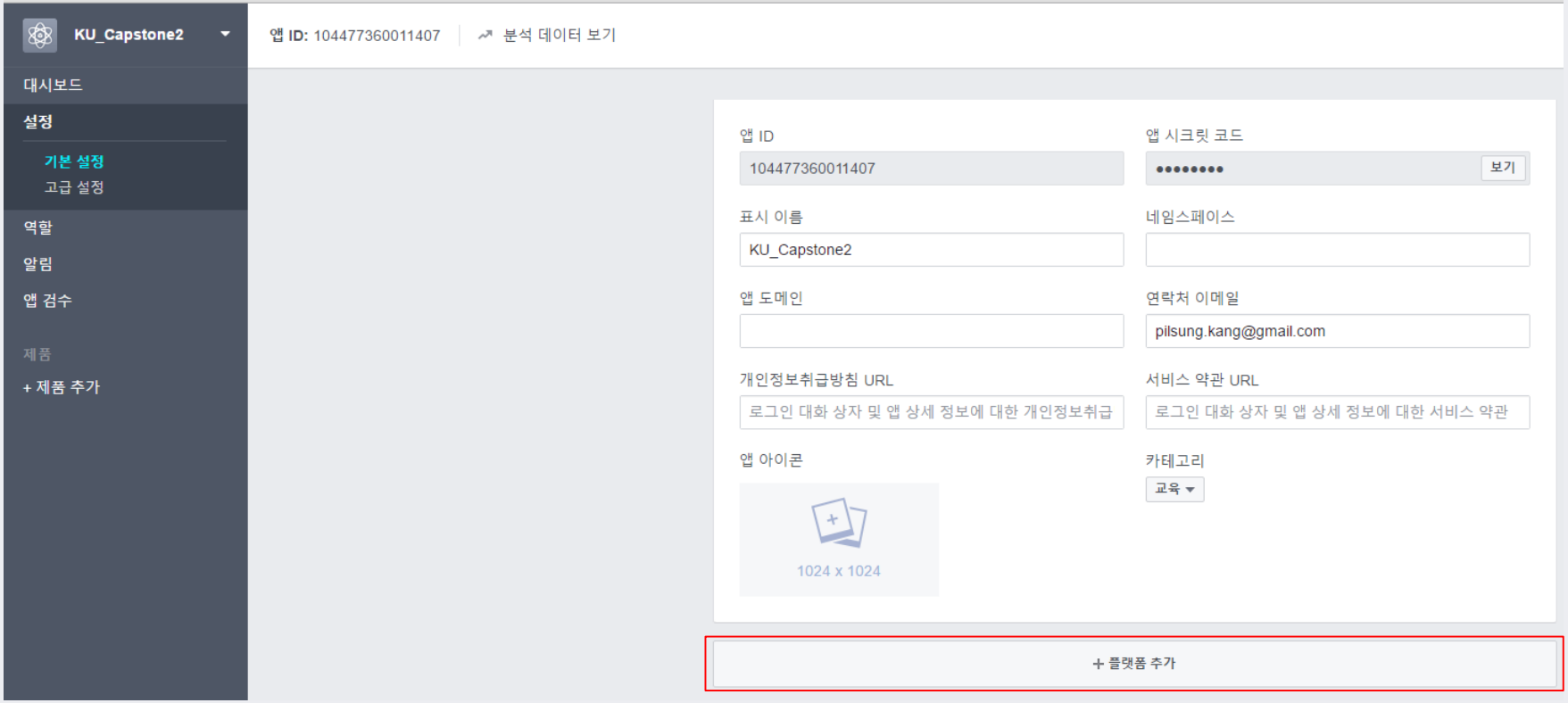


Collect Facebook Posts

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

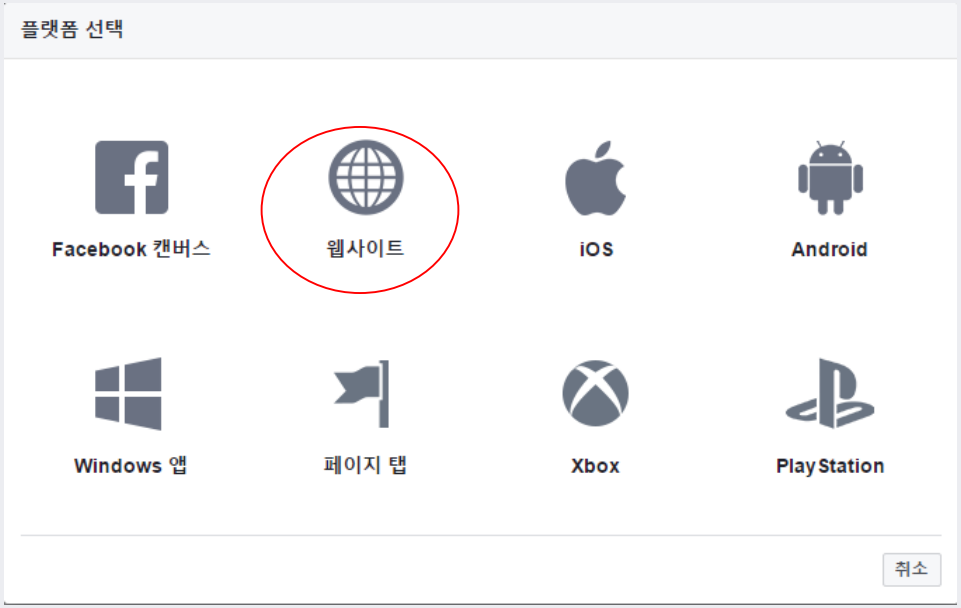
Collect Facebook Posts

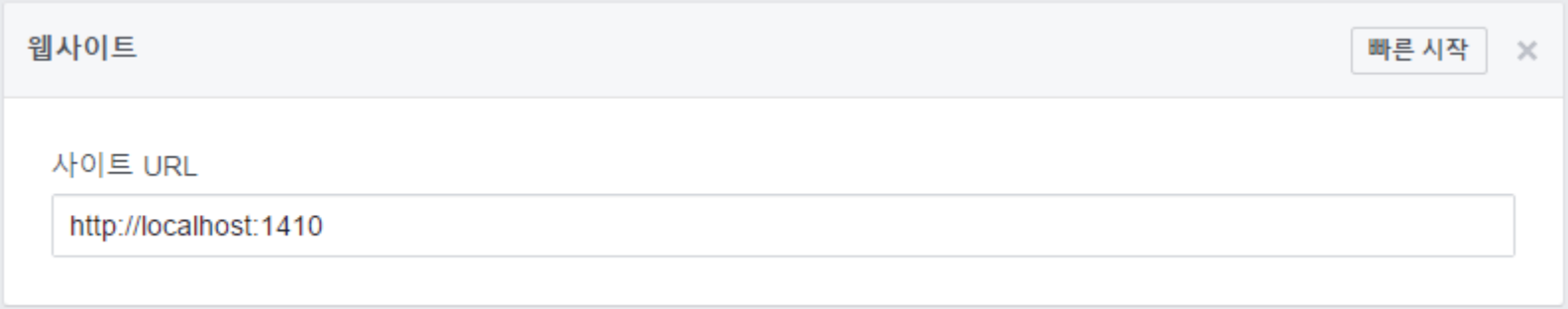
- Step 1: Registering an Application with Facebook
  - ✓ Add a new website with http://localhost:1410



Collect Facebook Posts

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





Collect Facebook Posts

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

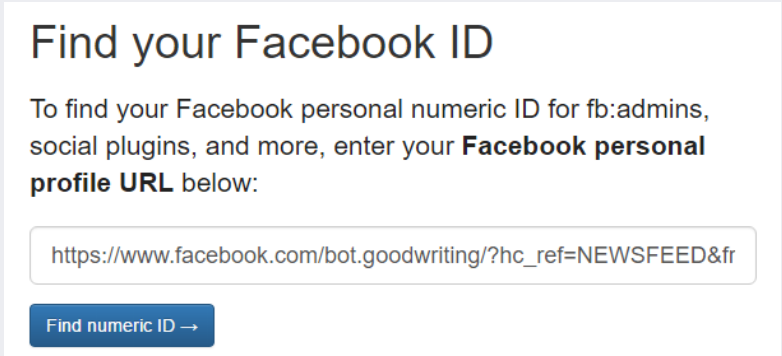
```
# Authentication Setting
my_oauth <- fbOAuth(app_id = "Your app ID", app_secret= "Your secret code")
save(my_oauth, file = "my_oauth")
load("my_oauth")
```

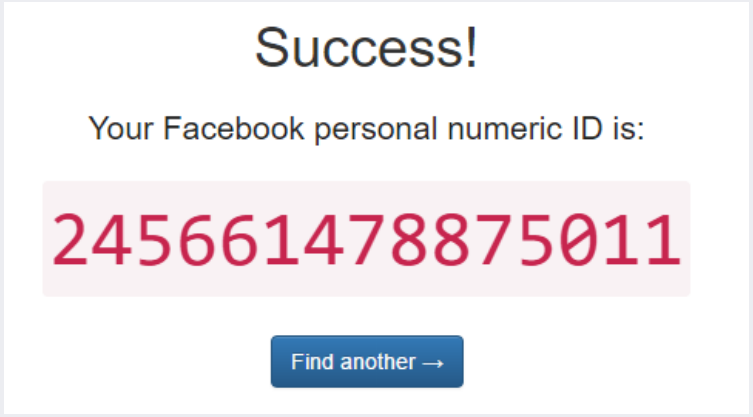

```
Authentication complete. Please close this page and return to R.
```

```
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

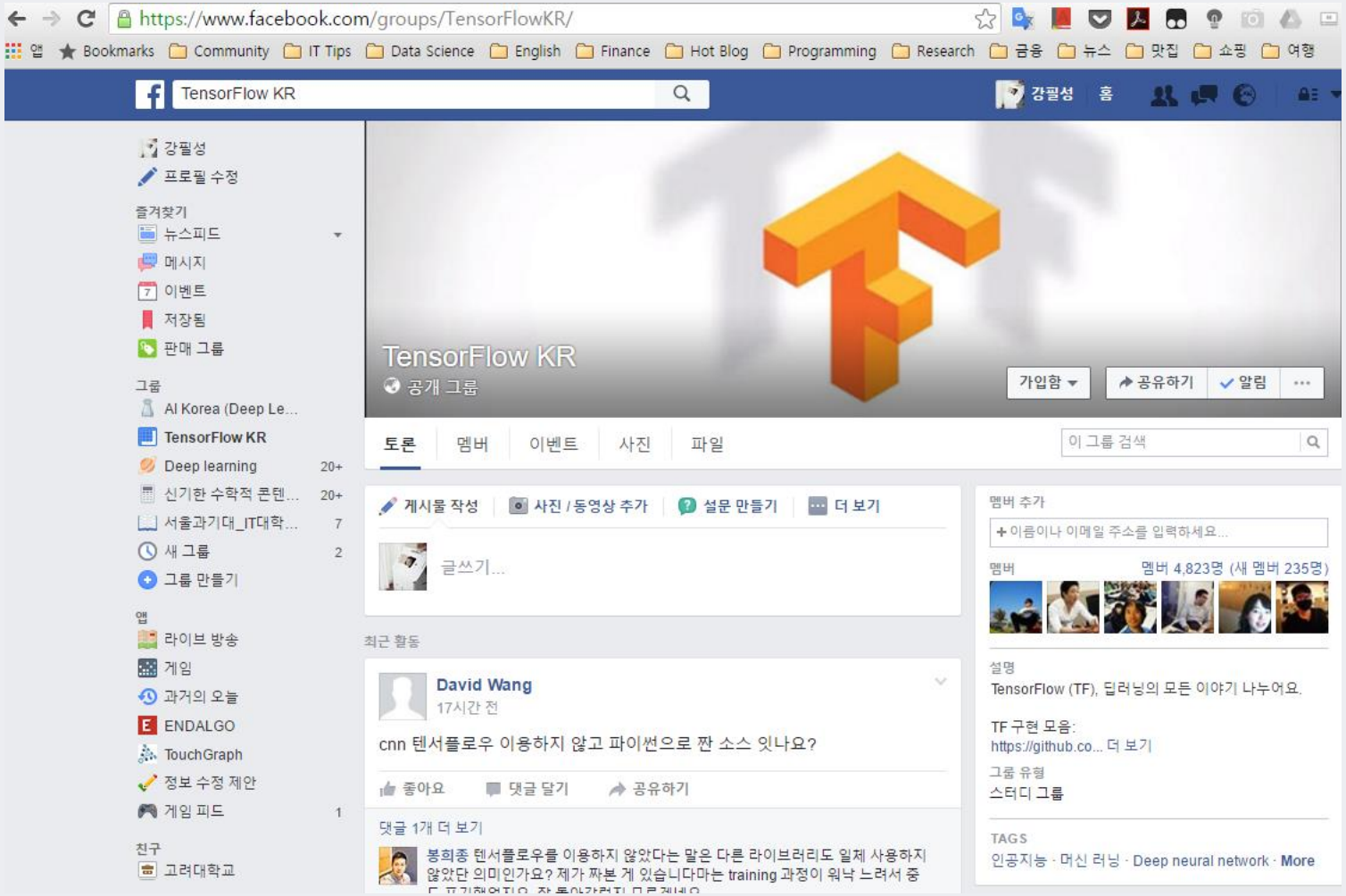
- Step 3: Collect Data from a Page

```
# Get data from a page
PageData <- getPage(245661478875011, token = my_oauth, n = 100)
```

	from_id	from_name	message	created_time	type	link	id	likes_count	comments_count	shares_count
1	245661478875011	좋은글봇	큰 전쟁으로 인해 헤어졌던 커플 오랜 세월 묻어둔 사랑과 ...	2016-09-06T23:00:00+0000	link	http://eposting.co.kr/bbs/board.php?bo_table=blog...	245661478875011_1055760421198442	66	1	4
2	245661478875011	좋은글봇	SNS에서 단순히 '좋아요'를 많이 받기위해 수 많은 공약들...	2016-09-06T12:30:00+0000	link	http://eposting.co.kr/bbs/board.php?bo_table=blog...	245661478875011_1055758414531976	115	2	4
3	245661478875011	좋은글봇	살인자를 용서하기로 마음먹은 이후 믿을 수 없는 일이 일...	2016-09-06T00:00:00+0000	link	http://eposting.co.kr/bbs/board.php?bo_table=new...	245661478875011_1054909754616842	115	2	8
4	245661478875011	좋은글봇	#악장수 #시 #인스타그램아이디JOKER_0203 아름다운 ...	2016-09-05T14:01:08+0000	photo	https://www.facebook.com/bot.goodwriting/photos...	245661478875011_1055115901262894	114	1	25
5	245661478875011	좋은글봇	생명 탄생의 경이로움에 그저 입이 떡 벌어진다. 힘겨운 출...	2016-09-05T13:00:02+0000	link	http://eposting.co.kr/bbs/board.php?bo_table=new...	245661478875011_1054909161283568	321	14	60
6	245661478875011	좋은글봇	집이 무너지기 시작하자 4살짜리 동생을 구하기 위해 믿을 ...	2016-09-04T13:00:01+0000	link	http://eposting.co.kr/bbs/board.php?bo_table=new...	245661478875011_1050890678352083	161	0	6
7	245661478875011	좋은글봇	위를 보고 살 것, 또한 아래를 보고 살 것	2016-09-04T09:00:01+0000	photo	https://www.facebook.com/bot.goodwriting/photos...	245661478875011_1050995551674929	367	2	21
8	245661478875011	좋은글봇	꽤 시원한 바람이 불더니 불과 일주일 새 깔깔해졌다. 같은 ...	2016-09-04T00:47:15+0000	photo	https://www.facebook.com/bot.goodwriting/photos...	245661478875011_1053694588071692	363	4	120
9	245661478875011	좋은글봇	임신 소식을 듣고 병원에 누워있는 엄마에게 달려간 딸! 그 ...	2016-09-03T13:00:00+0000	link	http://eposting.co.kr/bbs/board.php?bo_table=blog...	245661478875011_1050843528356798	218	4	7
10	245661478875011	좋은글봇	아무리 힘들어도 포기하지마세요!	2016-09-03T09:00:00+0000	photo	https://www.facebook.com/bot.goodwriting/photos...	245661478875011_1050995421674942	836	15	80
11	245661478875011	좋은글봇	1984년부터 지금까지 성실하게 근무했던 그녀를 위해 맥...	2016-09-02T13:00:00+0000	link	http://eposting.co.kr/bbs/board.php?bo_table=blog...	245661478875011_1050795791694905	637	2	19
12	245661478875011	좋은글봇	행운의 포장을 아직 씌지 못했을 뿐..	2016-09-02T09:00:00+0000	photo	https://www.facebook.com/bot.goodwriting/photos...	245661478875011_1050995061674978	960	22	105
13	245661478875011	좋은글봇	믿기 어려울만큼 정정한 모습의 할아버지, 그는 1870년에 ...	2016-09-01T13:00:00+0000	link	http://eposting.co.kr/bbs/board.php?bo_table=new...	245661478875011_1050742831700201	636	58	23
14	245661478875011	좋은글봇	불행을 딛고 그속에서 새로운 길을 발견할 힘이 있다.	2016-09-01T09:00:00+0000	photo	https://www.facebook.com/bot.goodwriting/photos...	245661478875011_1050413625066455	348	3	59
15	245661478875011	좋은글봇	사랑스러운 아기 돼지가 불편한 몸으로 균형을 잡기 위해 ...	2016-09-01T00:00:00+0000	link	http://eposting.co.kr/bbs/board.php?bo_table=clip_...	245661478875011_1049895045118313	92	3	1
16	245661478875011	좋은글봇	다른 강아지와 다르게 생긴 이 강아지의 별명은 '박쥐'이다. '...	2016-08-31T13:00:01+0000	link	http://eposting.co.kr/bbs/board.php?bo_table=new...	245661478875011_1049895305118287	183	11	5
17	245661478875011	좋은글봇	행복을 위해 힘쓰고 사랑받기 위해 노력하라	2016-08-31T09:00:00+0000	photo	https://www.facebook.com/bot.goodwriting/photos...	245661478875011_1049570288484122	528	8	79
18	245661478875011	좋은글봇	그녀는 메이크업, 패션 정보를 담은 영상을 유튜브에 업로...	2016-08-31T00:00:00+0000	link	http://eposting.co.kr/bbs/board.php?bo_table=new...	245661478875011_1049065301867954	140	0	4
19	245661478875011	좋은글봇	봉투 안에 무언가 들어있는 것 같아 열어본 순간 깔짝 놀랄 ...	2016-08-30T13:00:00+0000	link	http://eposting.co.kr/bbs/board.php?bo_table=new...	245661478875011_1049012121873272	121	17	6
20	245661478875011	좋은글봇	0형인분들 조심하시기를... 거제에서 룸레라 걸렸다고 뉴스...	2016-08-30T09:45:59+0000	link	http://newposting.kr/link.php?bo_table=news_news...	245661478875011_1049098945197923	610	704	46

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
GroupData <- getGroup(255834461424286, token = my_oauth, n = 100)
```

	from_id	from_name	message	created_time	type	link	id	likes_count	comments_count	shares_count
1	1656277444686306	David Wang	cnn 텐서플로우 이용하지 않고 파이썬으로 짠 소스 잇나요?	2016-09-06T10:28:49+0000	status	NA	255834461424286_340174582990273	0	3	0
2	1415587201789059	Seyoon Han	안녕하세요, TF 설치중인데 bazel호 tensorflow_serving ...	2016-09-07T02:40:22+0000	status	NA	255834461424286_340518252955906	0	0	0
3	605701162936984	전승현	혹 엔비디아 임베디드 보드 Jetson TX1 에서 Tensorflow...	2016-09-07T01:12:33+0000	status	NA	255834461424286_340494139624984	8	3	1
4	1161576927243019	Jin Yeong Wang	드디어 GTX 1060을 구하고 이제 연구실컴퓨터로 tensorfl...	2016-09-06T05:03:35+0000	status	NA	255834461424286_340064799667918	6	7	3
5	1161576927243019	Jin Yeong Wang	Pascal GPU + 16.04.1 + CUDA8.0 + CuDNN5.1 + Pytho...	2016-09-06T11:29:08+0000	status	NA	255834461424286_340202696320795	8	2	1
6	10207813768037617	ByungHyun Jeremy Ban	안녕하세요. 매번 눈팅만 하다 질문 드립니다. 리눅스 서버...	2016-09-05T12:54:42+0000	status	NA	255834461424286_339679209706477	3	3	1
7	1176708815721316	Sedong Nam	기초 중의 기초인 MNIST 숫자 인식기를 아주 오래된 (200...	2016-09-06T11:56:39+0000	video	https://www.facebook.com/1176708815721316/vi...	255834461424286_340214319652966	37	2	36
8	294683344235803	유승호	안녕하세요? inception-v3 모델을 공부하기 이전에 이미지...	2016-09-05T16:12:25+0000	status	NA	255834461424286_339786499695748	2	1	0
9	1010193352430704	최 희원	신경망을 구축시 질문이 있습니다 기존의 예를들어 사람이...	2016-09-06T09:24:57+0000	status	NA	255834461424286_340152909659107	0	0	0
10	1793585640898926	ByeongYun Chung	안녕하세요. 한국정보기술연구원 'Best of the Best' (이하 ...	2016-09-06T08:39:55+0000	status	NA	255834461424286_340133586327706	9	0	0
11	1261715963859524	Jeongkyu Shin	기계학습 기반의 채팅봇 만들기에 관련한 삽질을 주제로 8...	2016-09-06T07:18:59+0000	video	https://www.youtube.com/watch?v=q44fefORi1k	255834461424286_340103256330739	111	0	55
12	1024594850989835	Damoon Taylor Park	현재 Neural network를 이용해서 알고리즘을 구현하고 있...	2016-09-05T07:51:00+0000	status	NA	255834461424286_339560829718315	1	4	1
13	1771557666421103	이해찬	https://github.com/y0ubat/MachineLearning_Study...	2016-09-05T09:15:10+0000	link	https://github.com/y0ubat/MachineLearning_Study...	255834461424286_339593439715054	4	3	0
14	1088478894540214	Joohyun Ryu	Window 10 bash shell에는 기본으로 파이썬 2.7이 설치...	2016-08-26T09:41:31+0000	status	NA	255834461424286_334402206900844	4	3	0
15	920511011387157	Hongsuk Yi	안녕하세요. 텐서플로우 tutorial을 소개합니다. 일시: 201...	2016-08-31T07:58:35+0000	photo	https://www.facebook.com/photo.php?fbid=915045...	255834461424286_337081396632925	99	20	37
16	1176708815721316	Sedong Nam	David Wang 님께서 아래 글에서 y=x^2 함수를 딥러닝으...	2016-08-21T16:47:12+0000	video	https://www.facebook.com/1176708815721316/vi...	255834461424286_331870193820712	78	7	21
17	1656277444686306	David Wang	cnn 코드한줄 한줄 설명된 강의는 없나요?? 봐도봐도 이해...	2016-09-04T14:29:16+0000	status	NA	255834461424286_339189369755461	0	1	1
18	1656277444686306	David Wang	http://solarisailab.com/archives/303 위에 링크에서 #...	2016-09-04T06:12:45+0000	link	http://solarisailab.com/archives/303	255834461424286_338991629775235	24	1	28
19	10207123560828550	Byeongki Jeong	일본의 와세다대학에서 SIGGRAPH2016에서 발표한 러프...	2016-09-03T12:32:06+0000	link	http://hi.cs.waseda.ac.jp:8081/	255834461424286_338597696481295	48	2	19
20	1105322922870340	Tory Hwang	궁금해서 오늘 서점가서 한권 있던 책을 귀하게 구입했습니...	2016-09-03T15:09:59+0000	photo	https://www.facebook.com/photo.php?fbid=110221...	255834461424286_338661323141599	23	2	3

AGENDA

01 From Text Files to Data

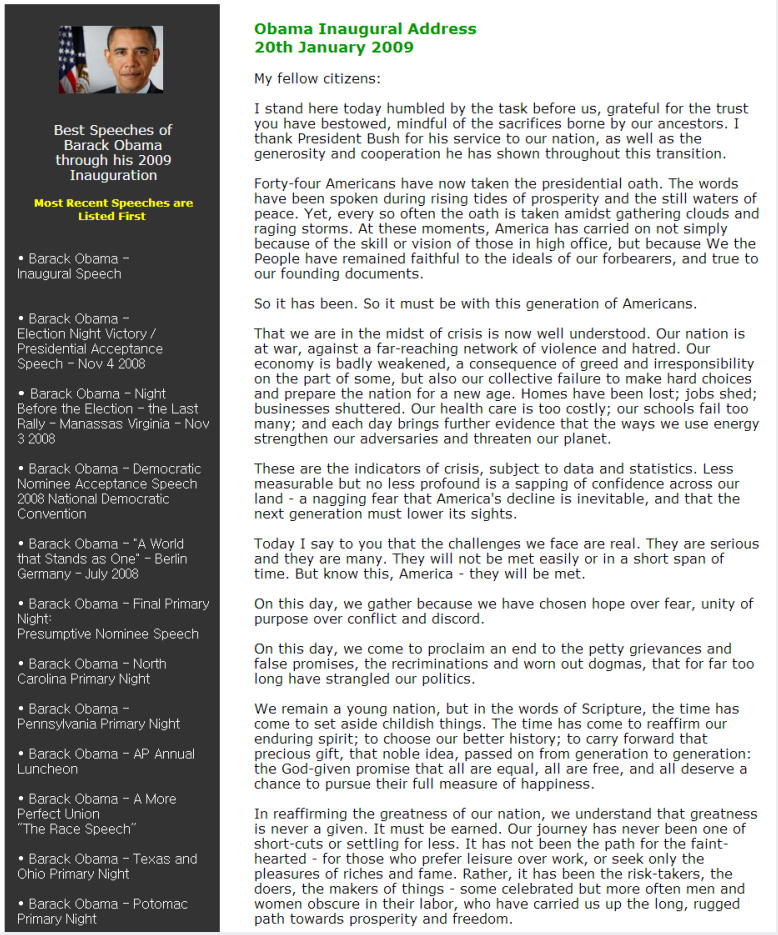
02 Collect Texts using APIs

03 Web Scraping

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

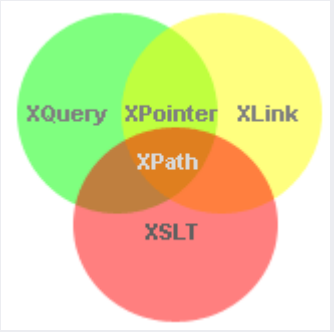
- Parsing
  - ✓ The process of analyzing a string of symbols, either in natural language or in computer languages (HTML/XML), conforming to the rules of a formal grammar

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

```
<td bgcolor="#333333">亲</td>
  </tr>
<tr bgcolor="#000000">
<td bgcolor="#333333">亲</td>
  </tr>
<tr bgcolor="#000000">
<td bgcolor="#333333">亲</td>
  </tr>
<tr bgcolor="#000000">
<td bgcolor="#333333">亲</td>
  </tr>
<tr bgcolor="#000000">
<td bgcolor="#333333">亲</td>
  </tr>
<tr bgcolor="#000000">
<td bgcolor="#333333">亲</td>
  </tr>
<tr bgcolor="#000000">
<td bgcolor="#333333">亲</td>
  </tr>
<tr bgcolor="#000000">
<td bgcolor="#333333">亲</td>
  </tr>
<tr bgcolor="#000000">
<td bgcolor="#333333">亲</td>
  </tr>
<tr bgcolor="#000000">
<td bgcolor="#333333">亲</td>
    <td align="left" valign="top" bgcolor="#333333">亲</td>
    <td align="left" valign="top" bgcolor="#FFFFFF"><font size="2" face="Courier New, Courier, mono"><strong>FUN
      FACTS ABOUT WHAT'S-HIS-NAME</strong><br>
      You can only imagine how many different ways people type the name Barack
      Obama. Here is a sampling for his first name: Barac, Barach, Baracks, Barak,
      Baraka, Barrack, Barrak, Berack, Borack, Borak, Brack, Brach, Brock even,
      Rocco. There are just as many for his last name: Abama, Bama, Bamma, Obma,
      Obamas, Obamma, Obana, Obamo, Obbama, Oboma, Obomba, Obombma, Obomha, Oblama,
      Omaba, Oblamma and (ready for this?) Ohama. And of course there's Barack
      Obama's middle name, Hussein. Here are some of the ways it comes out: Hissein,
      Hussain, Husein, Hussin, Hussane and Hussien. <br><br><br></font></td>
  </tr>
</table>
<script src="http://www.google-analytics.com/urchin.js" type="text/javascript">
</script><script type="text/javascript">
_uacct = "UA-557245-16";
urchinTracker();
</script><!-- InstanceEnd -->
</body>
</html>
```

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 http://www.w3schools.com/xml/xml_xpath.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: http://www.w3schools.com/xpath/books.xml

```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
  - ✓ 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

```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>
```

```
> 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

```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>
```

```
> xpathSApply(root, "/bookstore")
[[1]]
<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>

> xpathSApply(root, "/title")
list()
attr(,"class")
[1] "XMLNodeSet"
> xpathSApply(root, "//title")
[[1]]
<title lang="en">Everyday Italian</title>

[[2]]
<title lang="en">Harry Potter</title>

[[3]]
<title lang="en">XQuery Kick Start</title>

[[4]]
<title lang="en">Learning XML</title>
```

Web Scraping

- Xpath Syntax
  - ✓ Selecting attributes

```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>
```

```
> 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 (with predicates)

```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>
```

```
> 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

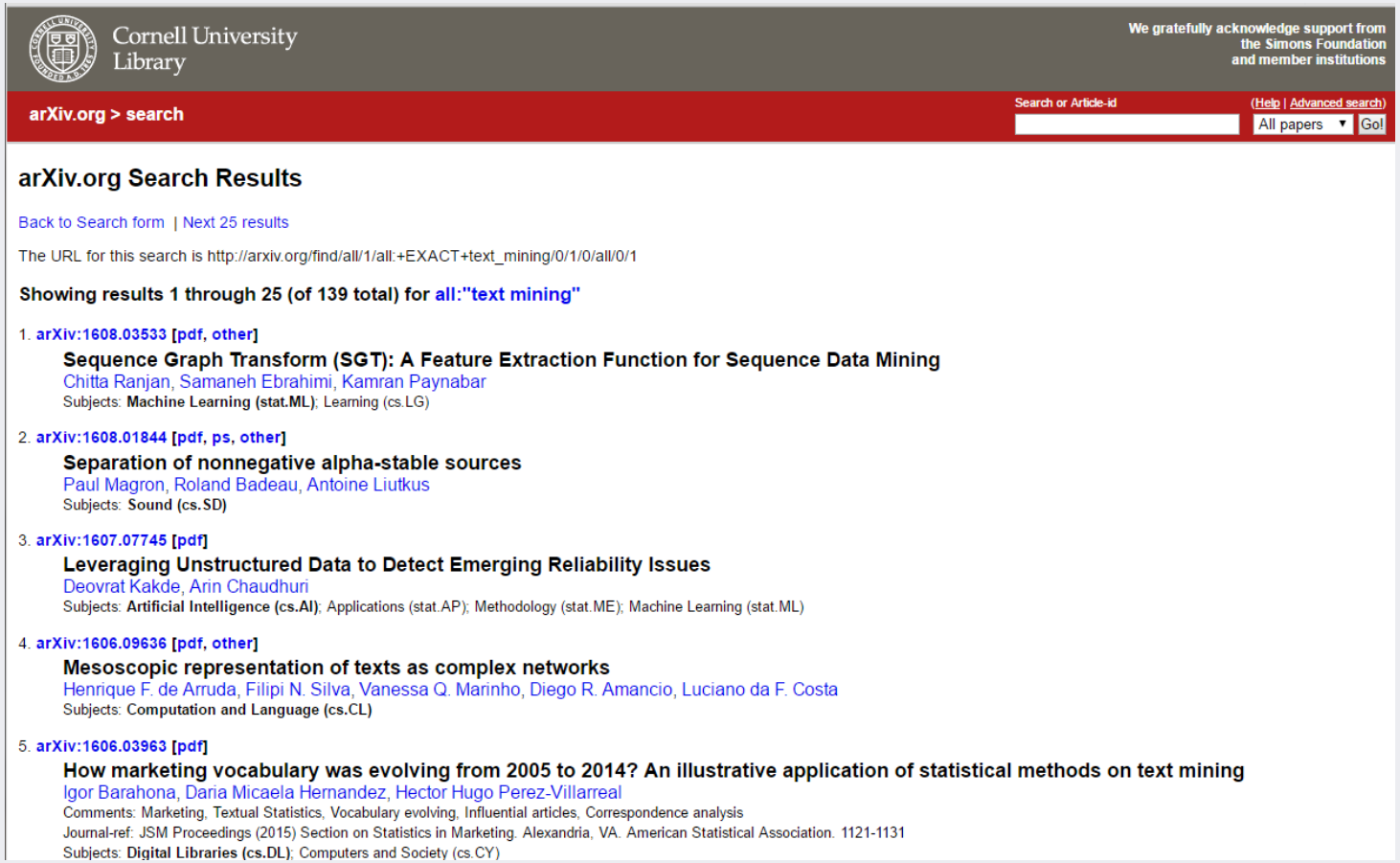
By using the | operator in an XPath expression you can select several paths.

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

- 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 (Help | Advanced search) All papers Go!

arXiv.org Search Results

Back to Search form | Next 25 results

The URL for this search is http://arxiv.org/find/all/1/all:+EXACT+text_mining/0/1/0/all/0/1

Showing results 1 through 25 (of 139 total) for all:"text mining"

1. arXiv:1608.03533 [pdf, other]
   Sequence Graph Transform (SGT): A Feature Extraction Function for Sequence Data Mining
   Chitta Ranjan, Samaneh Ebrahimi, Kamran Paynabar
   Subjects: **Machine Learning (stat.ML)**; Learning (cs.LG)

2. arXiv:1608.01844 [pdf, ps, other]
   Separation of nonnegative alpha-stable sources
   Paul Magron, Roland Badeau, Antoine Liutkus
   Subjects: **Sound (cs.SD)**

3. arXiv:1607.07745 [pdf]
   Leveraging Unstructured Data to Detect Emerging Reliability Issues
   Deovrat Kakde, Arin Chaudhuri
   Subjects: **Artificial Intelligence (cs.AI)**; Applications (stat.AP); Methodology (stat.ME); Machine Learning (stat.ML)

4. arXiv:1606.09636 [pdf, other]
   Mesoscopic representation of texts as complex networks
   Henrique F. de Arruda, Filipi N. Silva, Vanessa Q. Marinho, Diego R. Amancio, Luciano da F. Costa
   Subjects: **Computation and Language (cs.CL)**

5. arXiv:1606.03963 [pdf]
   How marketing vocabulary was evolving from 2005 to 2014? An illustrative application of statistical methods on text mining
   Igor Barahona, Daria Micaela Hernandez, Hector Hugo Perez-Villarreal
   Comments: Marketing, Textual Statistics, Vocabulary evolving, Influential articles, Correspondence analysis
   Journal-ref: JSM Proceedings (2015) Section on Statistics in Marketing. Alexandria, VA. American Statistical Association. 1121-1131
   Subjects: **Digital Libraries (cs.DL)**; Computers and Society (cs.CY)

Web Scraping

- Step 1: Understand the basic structure
  - ✓ A total of 161 papers are returned (2017-02-08), each page contains 25 papers
  - ✓ Each paper has a unique ID

Cornell University Library

We gratefully acknowledge support from the Simons Foundation and Stockholm University

arXiv.org > stat > arXiv:1702.01373

Search or Article ID | All papers

(Help | Advanced search)

Statistics > Machine Learning

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

Chenchao Zhao, Jun S. Song

(Submitted on 5 Feb 2017)

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

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

Submission history

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

Which authors of this paper are endorsers? | Disable MathJax (What is MathJax?)

Download:

- PDF
- Other formats

(license)

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

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

References & Citations
- NASA ADS

Bookmark (what is this?)

Web Scraping

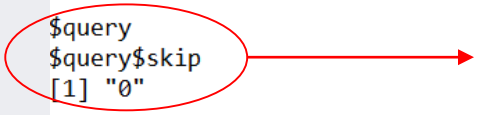
- Step 2: Analyzing the HTML Structure
  - ✓ First page URL
    - http://arxiv.org/find/all/1/all:+EXACT+text_mining/0/1/0/all/0/1?skip=0&query_id=504c4472acbc1ebf
  - ✓ Second page URL
    - http://arxiv.org/find/all/1/all:+EXACT+text_mining/0/1/0/all/0/1?skip=25&query_id=504c4472acbc1ebf
  - ✓ Third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

- 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

- Step 2: Analyzing the HTML Structure
  - ✓ 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

- 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 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

- Step 2: Analyzing the HTML Structure
  - ✓ Extract the link information

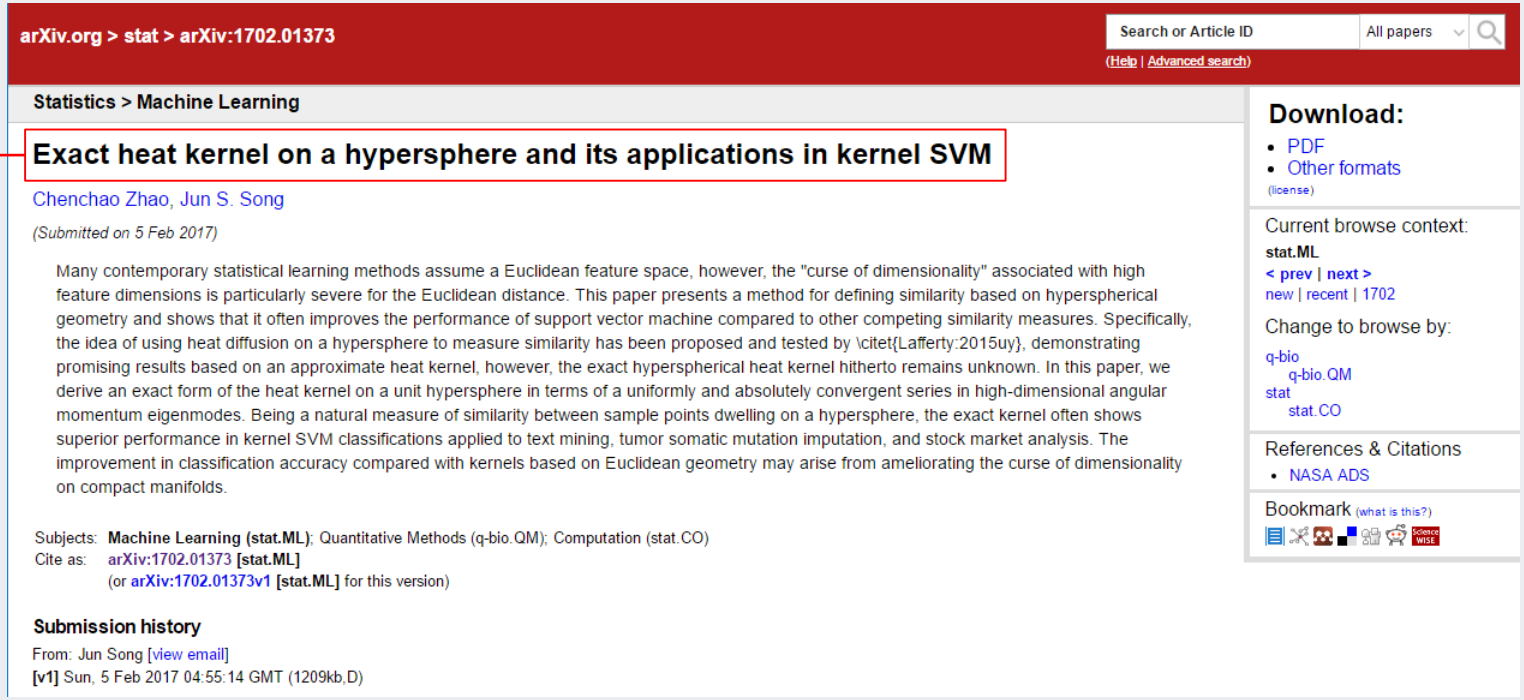
```
tmp_list <- read_html(tmp_url) %>% html_nodes('div#dlpage') %>%
                                   html_nodes('a[title="Abstract"]') %>% html_attr('href')
```

    - 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

- 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

- Step 3: Extract necessary information

  ✓ Step 3-1: Extract Title

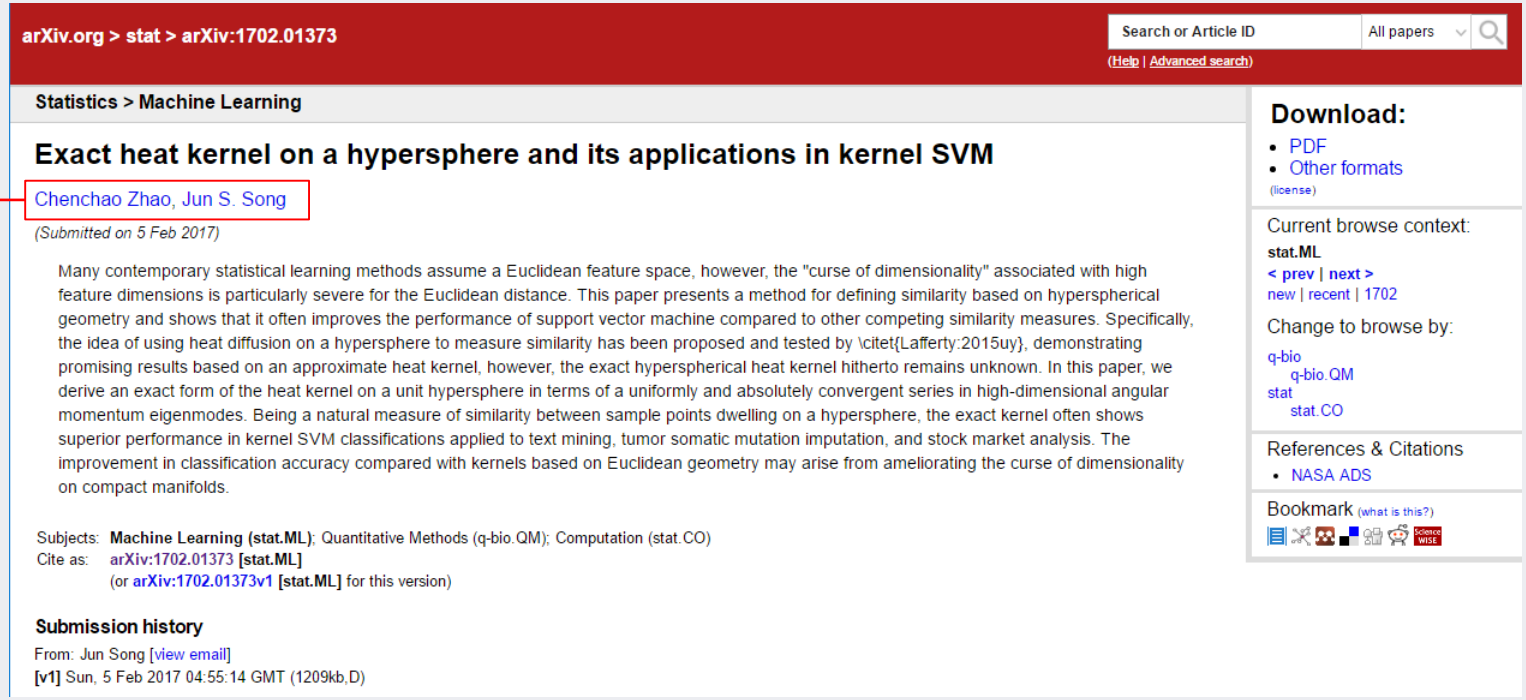
```
# title
tmp_title <- gsub('Title:\n', '',tmp_paragraph %>% html_nodes('h1.title.mathjax') %>% html_text(T))
title <- c(title, tmp_title)
```

  - 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

- Step 3: Extract necessary information
  - ✓ Step 3-2: Extract Authors



Web Scraping

- Step 3: Extract necessary information

  ✓ Step 3-2: Extract Authors

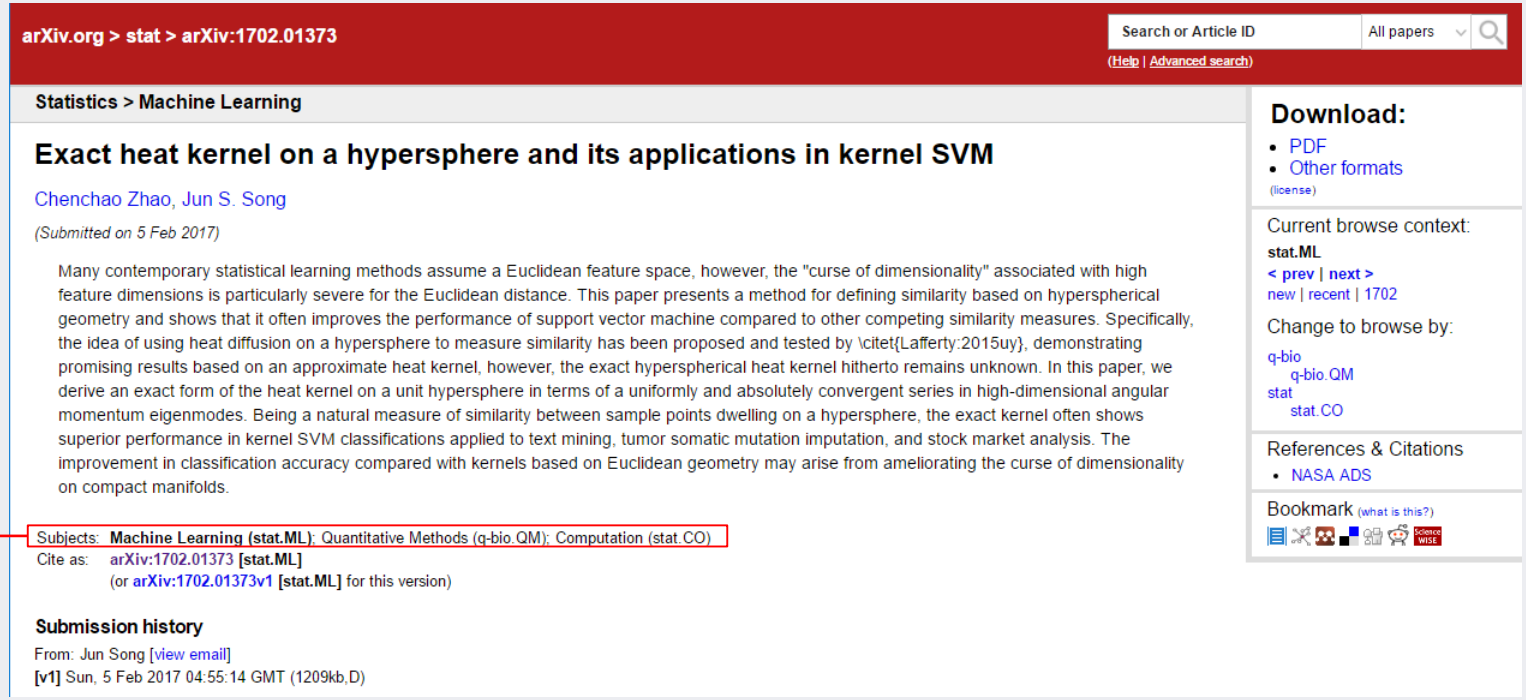
```
# author
tmp_author <- tmp_paragraph %>% html_nodes('div.authors') %>% html_text
tmp_author <- gsub('\\s+',' ',tmp_author)
tmp_author <- gsub('Authors:','',tmp_author) %>% str_trim
author <- c(author, tmp_author)
```

  - 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

- Step 3: Extract necessary information
  - ✓ Step 3-3: Extract Subjects



Web Scraping

- 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

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

Web Scraping

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

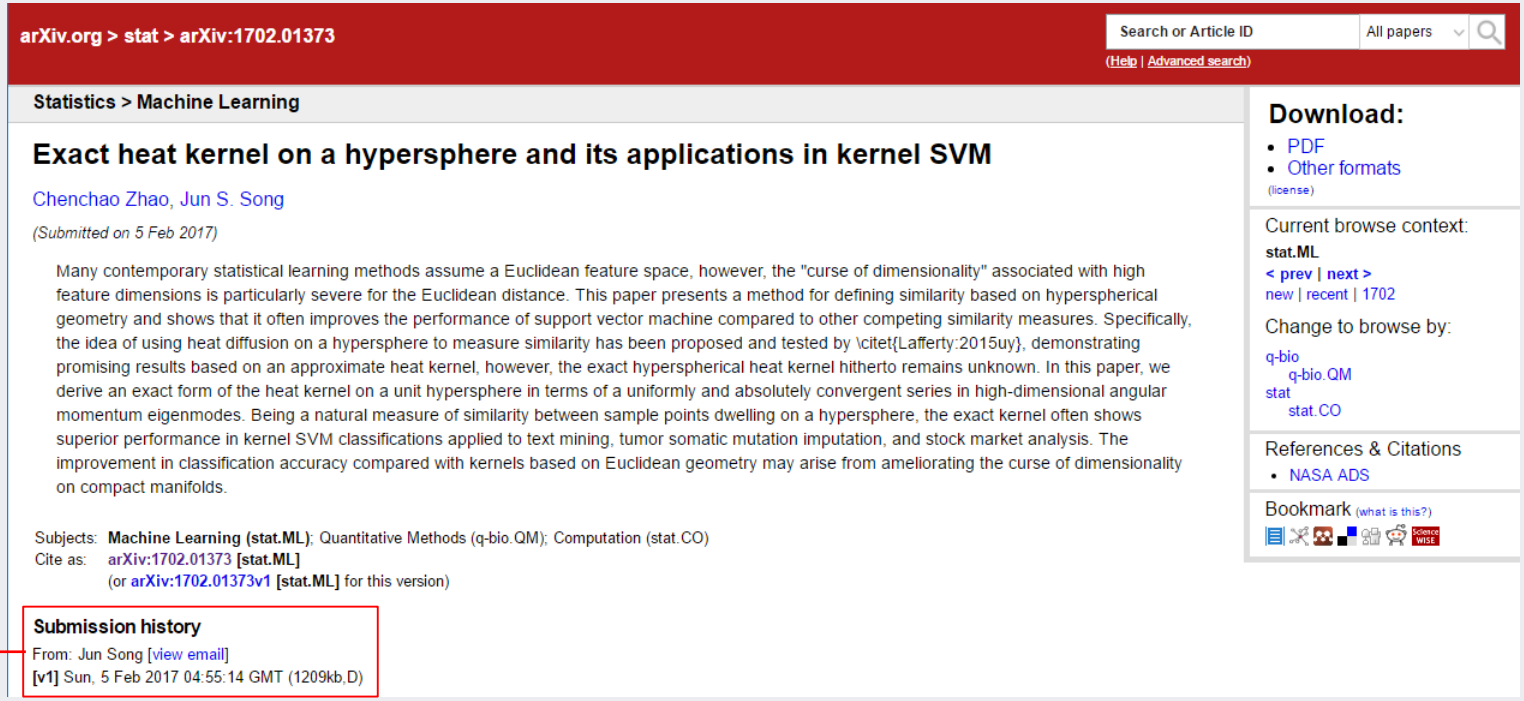
```
# abstract
tmp_abstract <- tmp_paragraph %>% html_nodes('blockquote.abstract.mathjax') %>% html_text(T)
tmp_abstract <- sub('Abstract:','',tmp_abstract) %>% str_trim
abstract <- c(abstract, tmp_abstract)
```

- 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

- 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

- 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

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

```
> end - start # Total Elapsed Time
 사용자  시스템 elapsed
  93.62  153.46  257.50
```

  - ✓ Check the dataset

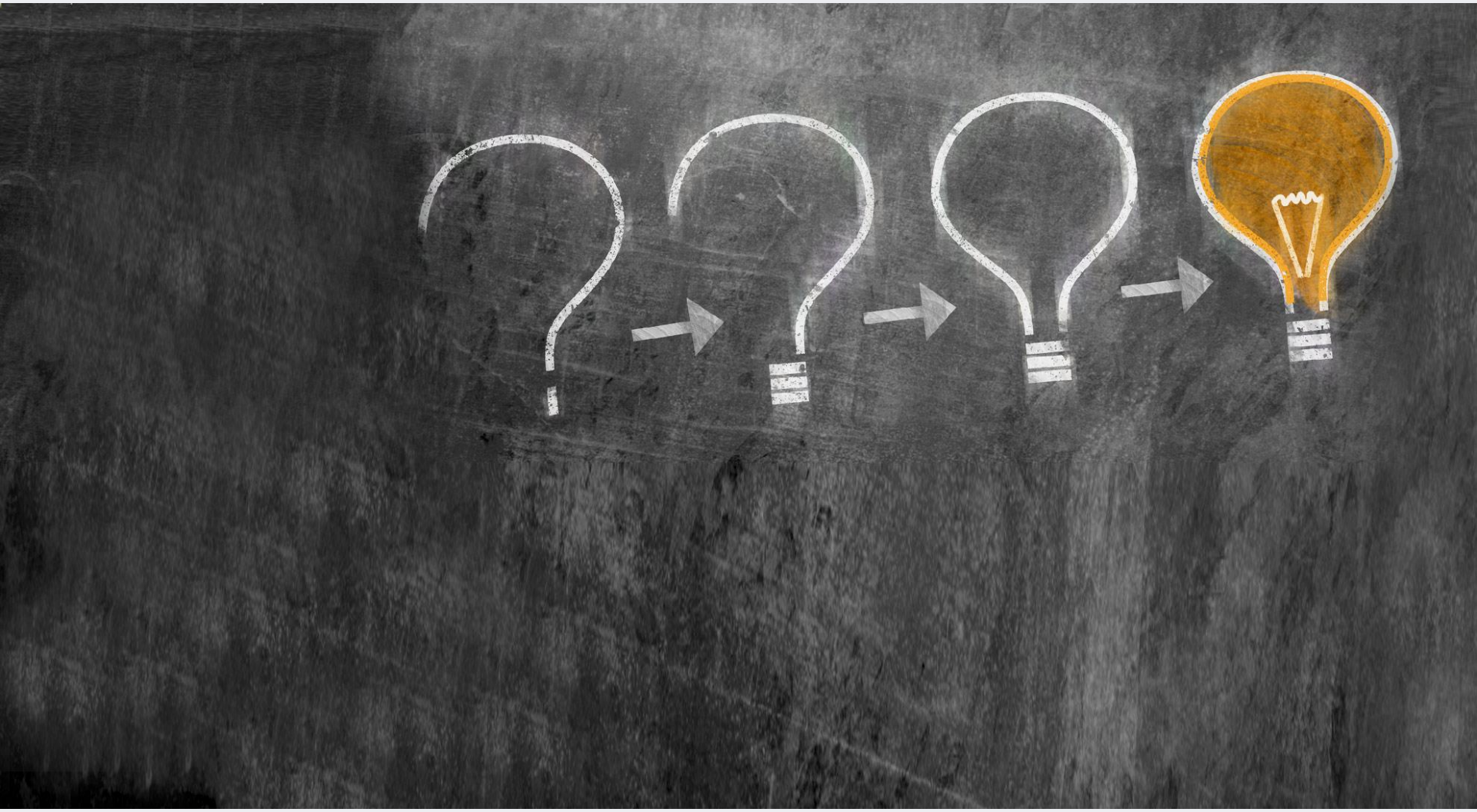
	title	author	subject	abstract	meta
1	Exact heat kernel on a hypersphere and its applicati…	Chenchao Zhao, Jun S. Song	Machine Learning (stat.ML)	Many contemporary statistical learning methods ass…	Sun, 5 Feb 2017 04:55:14 GMT (1209kb,D)
2	On Practical Accuracy of Edit Distance Approximatio…	Hiroyuki Hanada, Mineichi Kudo, Atsuyoshi Nakamura	Data Structures and Algorithms (cs.DS)	The edit distance is a basic string similarity measure …	Sun, 22 Jan 2017 07:40:52 GMT (1402kb)
3	Fuzzy Based Implicit Sentiment Analysis on Quantita…	Amir Hossein Yazdavar, Monireh Ebrahimi, Naomie Salim	Computation and Language (cs.CL)	With the rapid growth of social media on the web, em…	Tue, 3 Jan 2017 19:41:24 GMT (1049kb)
4	The leveled approach. Using and evaluating text min…	Berrie van der Molen, Lars Buitinck, Toine Pieters	Human-Computer Interaction (cs.HC)	We introduce our explorative historical leveled appr…	Mon, 2 Jan 2017 12:16:11 GMT (135kb)
5	PAMPO: using pattern matching and pos-tagging for e…	Conceicao Rocha, Alipio Jorge, Roberta Sionara, Paul…	Information Retrieval (cs.IR)	This paper deals with the entity extraction task (nam…	Fri, 30 Dec 2016 17:10:29 GMT (131kb)
6	Evaluating Marijuana-Related Tweets On Twitter	Anh Nguyen, Quang Hoang, Hung Nguyen, Dong Ng…	Social and Information Networks (cs.SI)	This paper studies marijuana-related tweets in social …	Wed, 28 Dec 2016 16:11:28 GMT (740kb,D)
7	"420 Friendly": Revealing Marijuana Use via Craigslist…	Anh Nguyen, Long Nguyen, Dong Nguyen, Uyen Le, …	Computers and Society (cs.CY)	Recent studies have shown that information mined fr…	Thu, 22 Dec 2016 15:00:16 GMT (509kb,D)[v2] Wed, …
8	Inverted Bilingual Topic Models for Lexicon Extractio…	Tengfei Ma	Computation and Language (cs.CL)	A good lexicon is an important resource for various c…	Wed, 21 Dec 2016 16:12:45 GMT (166kb,D)
9	Upper Bound of Bayesian Generalization Error in Non…	Naoki Hayashi, Sumio Watanabe	Statistics Theory (math.ST)	Non-negative matrix factorization (NMF) is a new kno…	Tue, 13 Dec 2016 12:02:24 GMT (11kb)
10	A New Spectral Method for Latent Variable Models	Matteo Ruffini, Marta Casanellas, Ricard Gavalda	Machine Learning (stat.ML)	This paper presents an algorithm for the unsupervise…	Sun, 11 Dec 2016 13:31:58 GMT (363kb,D)
11	The Evolution of Sentiment Analysis - A Review of Re…	Mika Viking Mantyla, Daniel Graziotin, Miikka Kuutila	Computation and Language (cs.CL)	Research in sentiment analysis is increasing at a fast…	Mon, 5 Dec 2016 21:28:06 GMT (1732kb)
12	The structure and dynamics of the research fronts in…	David Fajardo-Ortiz, Malaquias Lopez-Cervantes, Luis…	Social and Information Networks (cs.SI)	In this paper, we have identified and analyzed the e…	Wed, 16 Nov 2016 10:13:08 GMT (2080kb)
13	SimDoc: Topic Sequence Alignment based Documen…	Gaurav Maheshwari, Priyansh Trivedi, Harshita Sahijw…	Computation and Language (cs.CL)	Document similarity is the problem of formally repres…	Tue, 15 Nov 2016 13:31:28 GMT (323kb,D)
14	Using text mining and machine learning for detectio…	Chintan Amrit, Tim Paauw, Robin Aly, Miha Lavric	Computers and Society (cs.CY)	Abuse in any form is a grave threat to a child's health…	Fri, 11 Nov 2016 11:27:07 GMT (2805kb,D)[v2] Wed, …
15	Distributed Coordinate Descent for Generalized Line…	Ilya Trofimov, Alexander Genkin	Machine Learning (stat.ML)	Generalized linear model with L_1 and L_2 regul…	Mon, 7 Nov 2016 15:19:54 GMT (1684kb,D)
16	Rapid Prototyping of a Text Mining Application for Cr…	Marek Laskowski, Henry M. Kim	Computers and Society (cs.CY)	Blockchain represents a technology for establishing …	Sun, 28 Aug 2016 20:04:25 GMT (669kb)
17	Clinical Text Prediction with Numerically Grounded C…	Georgios P. Spithourakis, Steffen E. Petersen, Sebasti…	Computation and Language (cs.CL)	Assisted text input techniques can save time and eff…	Thu, 20 Oct 2016 11:48:30 GMT (502kb,D)
18	A New Data Representation Based on Training Data C…	Sadikin Mujiono, Mohamad Ivan Fanany, Chan Basaru…	Computation and Language (cs.CL)	One essential task in information extraction from the …	Thu, 6 Oct 2016 14:38:09 GMT (337kb,D)
19	MPI-FAUN: An MPI-Based Framework for Alternating-Up…	Ramakrishnan Kannan, Grey Ballard, Haesun Park	Distributed, Parallel, and Cluster Computing (cs.DC)	Non-negative matrix factorization (NMF) is the proble…	Wed, 28 Sep 2016 23:31:45 GMT (4171kb,D)
20	Psychologically Motivated Text Mining	Ekaterina Shutova, Patricia Lichtenstein	Computation and Language (cs.CL)	Natural language processing techniques are increasi…	Wed, 28 Sep 2016 17:58:23 GMT (549kb,D)

Web Scraping

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

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

	A	B	C	D	E	F	G
1		title	author	subject	abstract	meta	
2	1	Exact heat kernel on a hypersphere	Chenchao Zhao, Jun S. Song	Machine Learning (stat.ML)	Many contemporary statistical learning methods assume a Euclidean feature space, however, the "curse of dimensionality" associated with high feature dimensions is particularly severe for the Euclidean distance. This paper presents a method for defining similarity based on hyperspherical geometry and shows that it often improves the performance of support vector machine compared to other competing similarity measures. Specifically, the idea of using heat diffusion on a hypersphere to measure similarity has been proposed and tested by #citet{Lafferty:2015uy}, demonstrating promising results based on an approximate heat kernel, however, the exact hyperspherical heat kernel hitherto remains unknown. In this paper, we derive an exact form of the heat kernel on a unit hypersphere in terms of a uniformly and absolutely convergent series in high-dimensional angular momentum eigenmodes. Being a natural measure of similarity between sample points dwelling on a hypersphere, the exact kernel often shows superior performance in kernel SVM classifications applied to text mining, tumor somatic mutation imputation, and stock market analysis. The improvement in classification accuracy compared with kernels based on Euclidean geometry may arise from ameliorating the curse of dimensionality on compact manifolds.	Sun, 5 Feb 2017 04:55:14 GMT (1209kb,D)	
3	2	On Practical Accuracy of Edit Distar	Hiroyuki Hanada, Mineichi Kudo, A	Data Structures and Algorithms (cs.	The edit distance is a basic string similarity measure used in many applications such as text mining, signal processing, bioinformatics, and so on. However, the computational cost can be a problem when we repeat many distance calculations as seen in real-life searching situations. A promising solution to cope with the problem is to approximate the edit distance by another distance with a lower computational cost. There are, indeed, many distances have been proposed for approximating the edit distance. However, their approximation accuracies are evaluated only theoretically: many of them are evaluated only with big-oh (asymptotic) notations, and without experimental analysis. Therefore, it is beneficial to know their actual performance in real applications. In this study we compared existing six approximation distances in two approaches: (i) we refined their theoretical approximation accuracy by calculating up to the constant coefficients, and (ii) we conducted some experiments, in one artificial and two real-life data sets, to reveal under which situations they perform best. As a result we obtained the following results: [Batu 2006] is the best theoretically and [Andoni 2010] experimentally. Theoretical considerations show that [Batu 2006] is the best if the string length n is large enough (n >= 300). [Andoni 2010] is experimentally the best for most data sets and theoretically the second best. [Bar-Yossef 2004], [Charikar 2006] and [Sokolov 2007], despite their middle-level theoretical performance, are experimentally as good as [Andoni 2010] for pairs of strings with large alphabet size.	Sun, 22 Jan 2017 07:40:52 GMT (1402kb)	



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	