Java Programming #3: Web Crawler

2015/04/27

Yerim Choi

Java Assignment

- Topics
  - ~~Assignment 1 : Self introduction~~
  - ~~Assignment 2 : Java basics~~
  - ~~Assignment 3 : File I/O~~
  - Assignment 4 : Web crawler
  - Assignment 5 : Machine learning algorithm

Assignment#3 Review

Web crawler

Java Programming #3

Previously…

Jsoup

- http://jsoup.org/

jsoup: Java HTML Parser

`jsoup` is a Java library for working with real-world HTML. It provides a very convenient API for extracting and manipulating data, using the best of DOM, CSS, and jquery-like methods.

`jsoup` implements the WHATWG HTML5 specification, and parses HTML to the same DOM as modern browsers do.

- scrape and parse HTML from a URL, file, or string
- find and extract data, using DOM traversal or CSS selectors
- manipulate the HTML elements, attributes, and text
- clean user-submitted content against a safe white-list, to prevent XSS attacks
- output tidy HTML

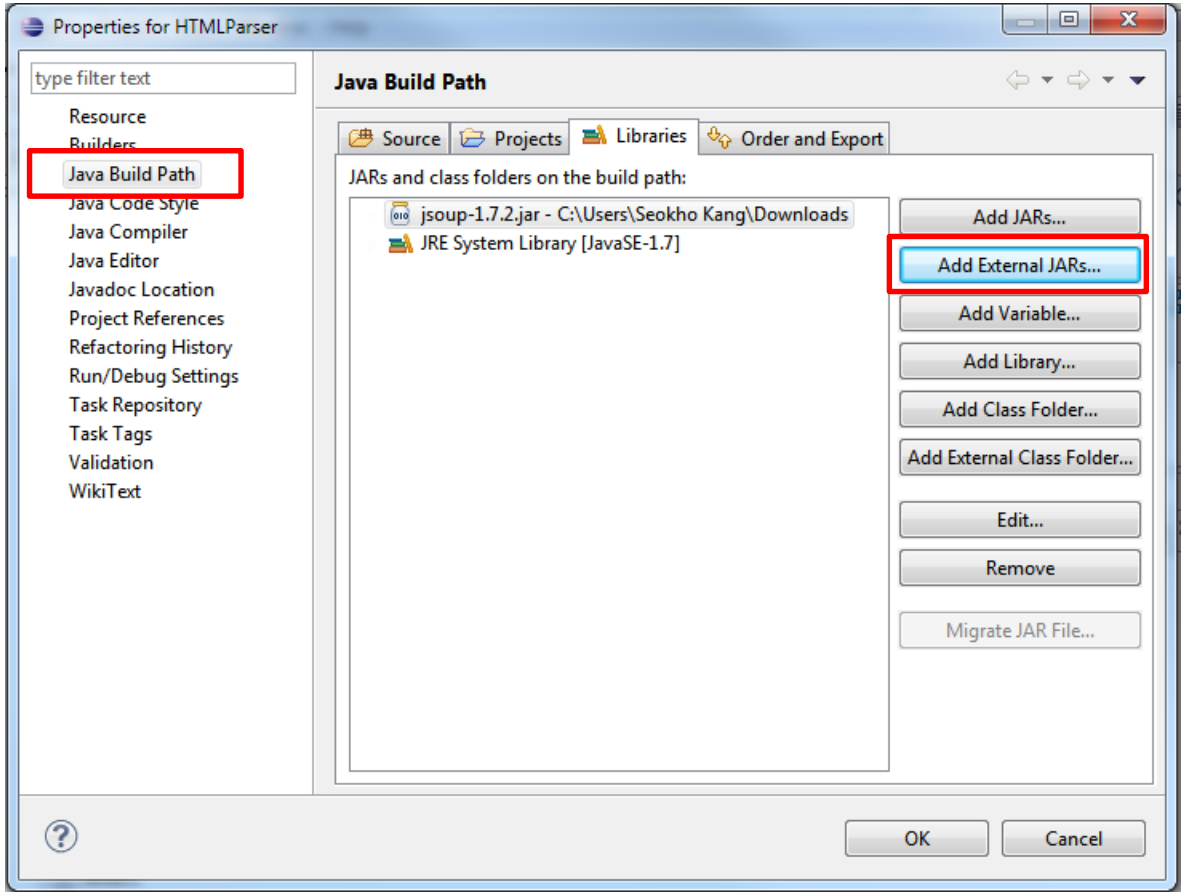
jsoup is designed to deal with all varieties of HTML found in the wild; from pristine and validating, to invalid tag-soup; jsoup will create a sensible parse tree.

Downloading jsoup

1. click
2. download

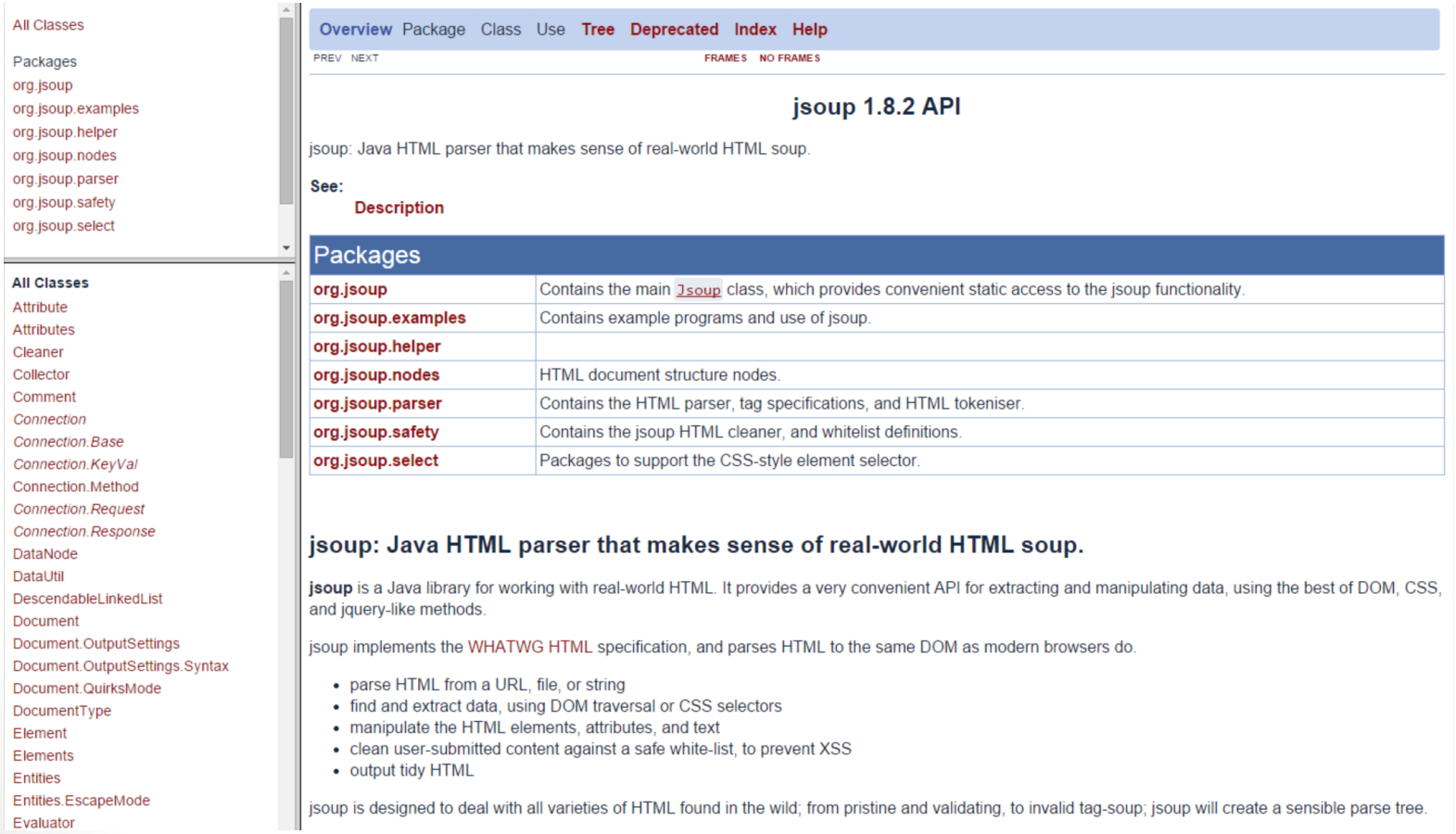
Installing jsoup

- JAVA Project properties → Add External JAR



API docs

- http://jsoup.org/apidocs/



All Classes

Packages
org.jsoup
org.jsoup.examples
org.jsoup.helper
org.jsoup.nodes
org.jsoup.parser
org.jsoup.safety
org.jsoup.select

All Classes
Attribute
Attributes
Cleaner
Collector
Comment
Connection
Connection.Base
Connection.KeyVal
Connection.Method
Connection.Request
Connection.Response
DataNode
DataUtil
DescendableLinkedList
Document
Document.OutputSettings
Document.OutputSettings.Syntax
Document.QuirksMode
DocumentType
Element
Elements
Entities
Entities.EscapeMode
Evaluator

Overview Package Class Use Tree Deprecated Index Help

PREV NEXT FRAMES NO FRAMES

jsoup 1.8.2 API

jsoup: Java HTML parser that makes sense of real-world HTML soup.

See:
Description

Packages	
org.jsoup	Contains the main `Jsoup` class, which provides convenient static access to the jsoup functionality.
org.jsoup.examples	Contains example programs and use of jsoup.
org.jsoup.helper	
org.jsoup.nodes	HTML document structure nodes.
org.jsoup.parser	Contains the HTML parser, tag specifications, and HTML tokeniser.
org.jsoup.safety	Contains the jsoup HTML cleaner, and whitelist definitions.
org.jsoup.select	Packages to support the CSS-style element selector.

jsoup: Java HTML parser that makes sense of real-world HTML soup.

jsoup is a Java library for working with real-world HTML. It provides a very convenient API for extracting and manipulating data, using the best of DOM, CSS, and jquery-like methods.

jsoup implements the WHATWG HTML specification, and parses HTML to the same DOM as modern browsers do.

- parse HTML from a URL, file, or string
- find and extract data, using DOM traversal or CSS selectors
- manipulate the HTML elements, attributes, and text
- clean user-submitted content against a safe white-list, to prevent XSS
- output tidy HTML

jsoup is designed to deal with all varieties of HTML found in the wild; from pristine and validating, to invalid tag-soup; jsoup will create a sensible parse tree.

Selector Syntax

- http://jsoup.org/apidocs/org/jsoup/select/Selector.html

org.jsoup.select

Class Selector

```
java.lang.Object
  └org.jsoup.select.Selector
```

```
public class Selector

      extends
      Object
```

CSS-like element selector, that finds elements matching a query.

Selector syntax

A selector is a chain of simple selectors, separated by combinators. Selectors are case insensitive (including against elements, attributes, and attribute values).

The universal selector (*) is implicit when no element selector is supplied (i.e. `*.header` and `.header` is equivalent).

Pattern	Matches
`*`	any element
`tag`	elements with the given tag name
`ns|E`	elements of type E in the namespace *ns*
`#id`	elements with attribute ID of "id"
`.class`	elements with a class name of "class"
`[attr]`	elements with an attribute named "attr" (with any value)
`[^attrPrefix]`	elements with an attribute name starting with "attrPrefix". Use to find elements with HTML5 datasets
`[attr=val]`	elements with an attribute named "attr", and value equal to "val"
`[attr="val"]`	elements with an attribute named "attr", and value equal to "val"

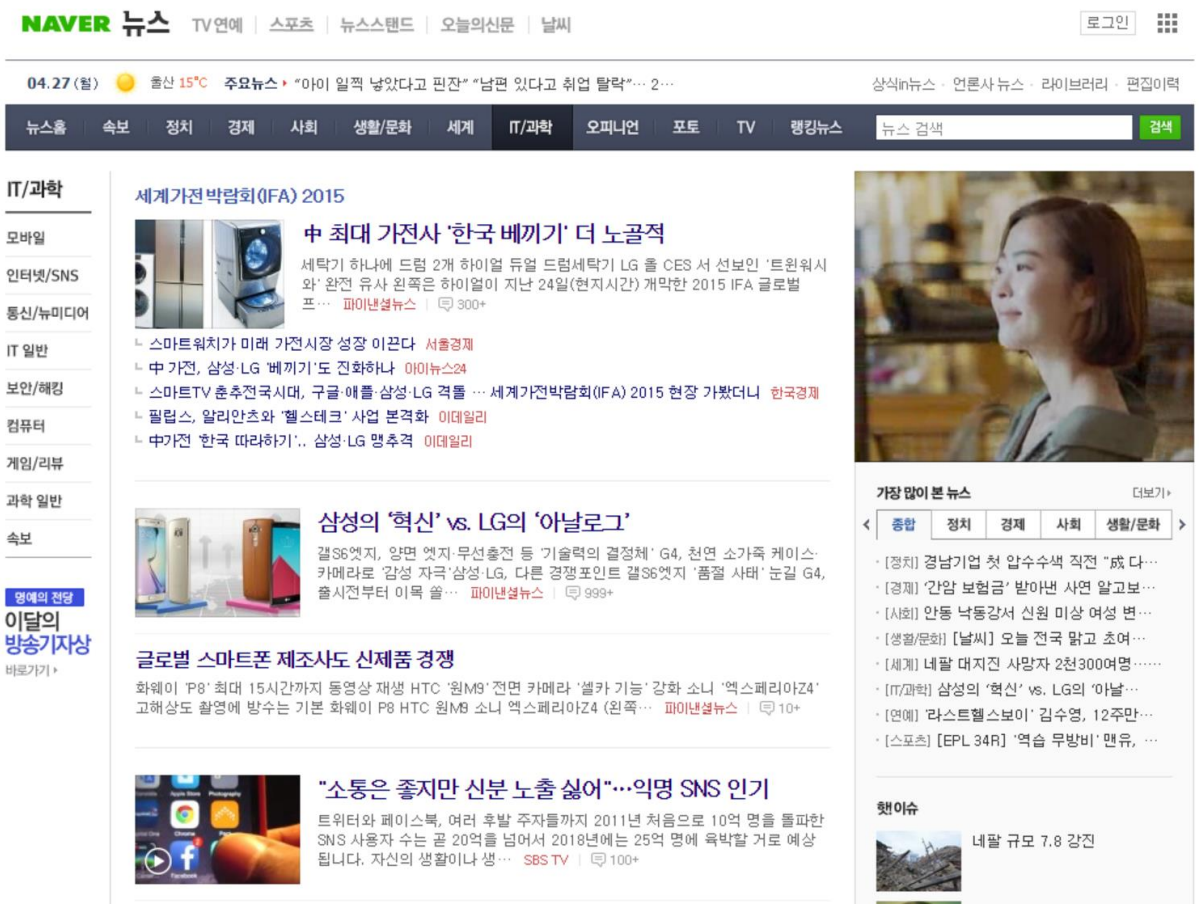
Basic Structure

```
Document doc = Jsoup.connect([URL]).get();      ← [URL]에서 HTML Source를 받아옴
Elements es=doc.select([Syntax]);               ← 특정 Syntax를 만족하는 Element들을 선택

for(Element e: es){                             ← 각 Element들에 대해서
        System.out.println(e.text());           ← Text 출력
        System.out.println(e.attr([attributeName]));  ← 특정 Attribute의 값 출력
}
```

Example 1: Naver News

- 네이버 뉴스 IT/과학 페이지의 기사 제목을 수집하고 싶다면?



http://news.naver.com/main/main.nhn?mode=LSD&mid=shm&sid1=105

Example 1: Naver News

- HTML 구문 분석

```
<div class="section_body" id="section_body">
<ul class="type02_headline" style="margin-bottom:18px;">
<li class="_rcount" data-comment="{gno:'news421,0001386493',nclicks:'arl.cmt'}" ><a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shm
샤오미 지분 매입"  id='949984_26478920'><strong>印 타타, 중국 스마트폰 메이커 샤오미 지분 매입</strong></a><img src="http://imgnews.naver.net/
title="포토"><span class="writing">뉴스1</span></li>
<li class="_rcount" data-comment="{gno:'news001,0007554968',nclicks:'arl.cmt'}" ><a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shm
세이프가드' 공조"  id='949984_26478687'><strong>정부, 中·터키와 '터키 휴대전화 세이프가드' 공조</strong></a><span class="writing">연합뉴스</spa
<li class="_rcount" data-comment="{gno:'news055,0000314034',nclicks:'arl.cmt'}" ><a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shm
지'?…익명 SNS의 그늘"  id='949984_26478306'><strong>포인트 챙기는 고객은 '별거지'?…익명 SNS의 그늘</strong></a><strong class="r_ico r_vod_t>
</li>
<li class="_rcount" data-comment="{gno:'news138,0002028112',nclicks:'arl.cmt'}" ><a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shm
서버 성장 이끄는 원동력" "  id='949984_26478052'><strong> "데스크톱가상화(VDI), 시스코 서버 성장 이끄는 원동력" </strong></a><img src="http:/
height="13" alt="포토" title="포토"><span class="writing">디지털데일리</span></li>
<li class="_rcount" data-comment="{gno:'news022,0002821835',nclicks:'arl.cmt'}" ><a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shm
복" "  id='949984_26478050'><strong> "유전정보 해독 노화·질병 극복" </strong></a><img src="http://imgnews.naver.net/image/news/2009/ico_photo
class="writing">세계일보</span></li>
</ul>
<ul class="type02" style="margin-bottom:18px;">
<li class="_rcount" data-comment="{gno:'news214,0000491328',nclicks:'arl.cmt'}" ><a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shm
현실 시장 잡아라…콘텐츠 개발 치열"  id='949984_26478023'><strong>[뉴스플러스] 오감 즐기는 가상현실 시장 잡아라…콘텐츠 개발 치…</strong></a>
<li class="_rcount" data-comment="{gno:'news029,0002280247',nclicks:'arl.cmt'}" ><a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shm
어떻게 뚫었나 보니"  id='949984_26477505'><strong>러시아, 오바마 이메일 해킹… 어떻게 뚫었나 보니</strong></a><span class="writing">디지털타임
<li class="_rcount" data-comment="{gno:'news029,0002280290',nclicks:'arl.cmt'}" ><a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shm
석 계약했다가…"  id='949984_26477373'><strong>유명 개발자 커뮤니티 믿고 덥석 계약했다
토"><span class="writing">디지털타임스</span></li>
<li class="_rcount" data-comment="{gno:'news029,0002280314',nclicks:'arl.cmt'}" ><a hr
포… 피해 예방하려면"  id='949984_26477133'><strong>'한국형 랜섬웨어' 무차별 유포… 피
<li class="_rcount" data-comment="{gno:'news029,0002280294',nclicks:'arl.cmt'}" ><a hr
증에 숨은 비밀"  id='949984_26477105'><strong>삼성전기 1분기 영업익 300% 급증에 숨은 비
</ul>
<ul class="type02" style="margin-bottom:18px;">
<li class="_rcount" data-comment="{gno:'news001,0007554912',nclicks:'arl.cmt'}" ><a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shm
표…TV사업 적자 기로"  id='949984_26477078'><strong>삼성·LG전자 주중 동시 실적발표…TV사업 적자 기로</strong></a><span class="writing">연합뉴
<li class="_rcount" data-comment="{gno:'news030,0002352517',nclicks:'arl.cmt'}" ><a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shm
반하는 5가지 규정은?"  id='949984_26477076'><strong>구글이 꼽았다! 앱 개발사가 위반하는 5가지 규정은?</strong></a><span class="writing">전자신
<li class="_rcount" data-comment="{gno:'news030,0002352524',nclicks:'arl.cmt'}" ><a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shm
향기업"  id='949984_26477075'><strong>`주홍글씨`에 두번 우는 기술우향기업</strong></a><span class="writing">전자신문</span></li>
```

div tag(class="section_body") 다음에 있는 ul tag 다음 li tag 다음 a tag
➔ [class=mlist2 no_bg] li a

Example 1: Naver News

- Source Code

```
package java3;

import java.io.IOException;

import org.jsoup.Jsoup;
import org.jsoup.nodes.Document;
import org.jsoup.nodes.Element;
import org.jsoup.select.Elements;

public class Example1 {

	public static void main(String[] args) throws IOException {

		Document doc =
		Jsoup.connect("http://news.naver.com/main/main.nhn?mode=LSD&mid=shm&sid1=10
		5").get();
		Elements newsTitle = doc.select("div[class=section_body] ul li a ");

		for(Element e : newsTitle){

			System.out.println(e.text());
			System.out.println(e.attr("href"));
		}
	}
}
```

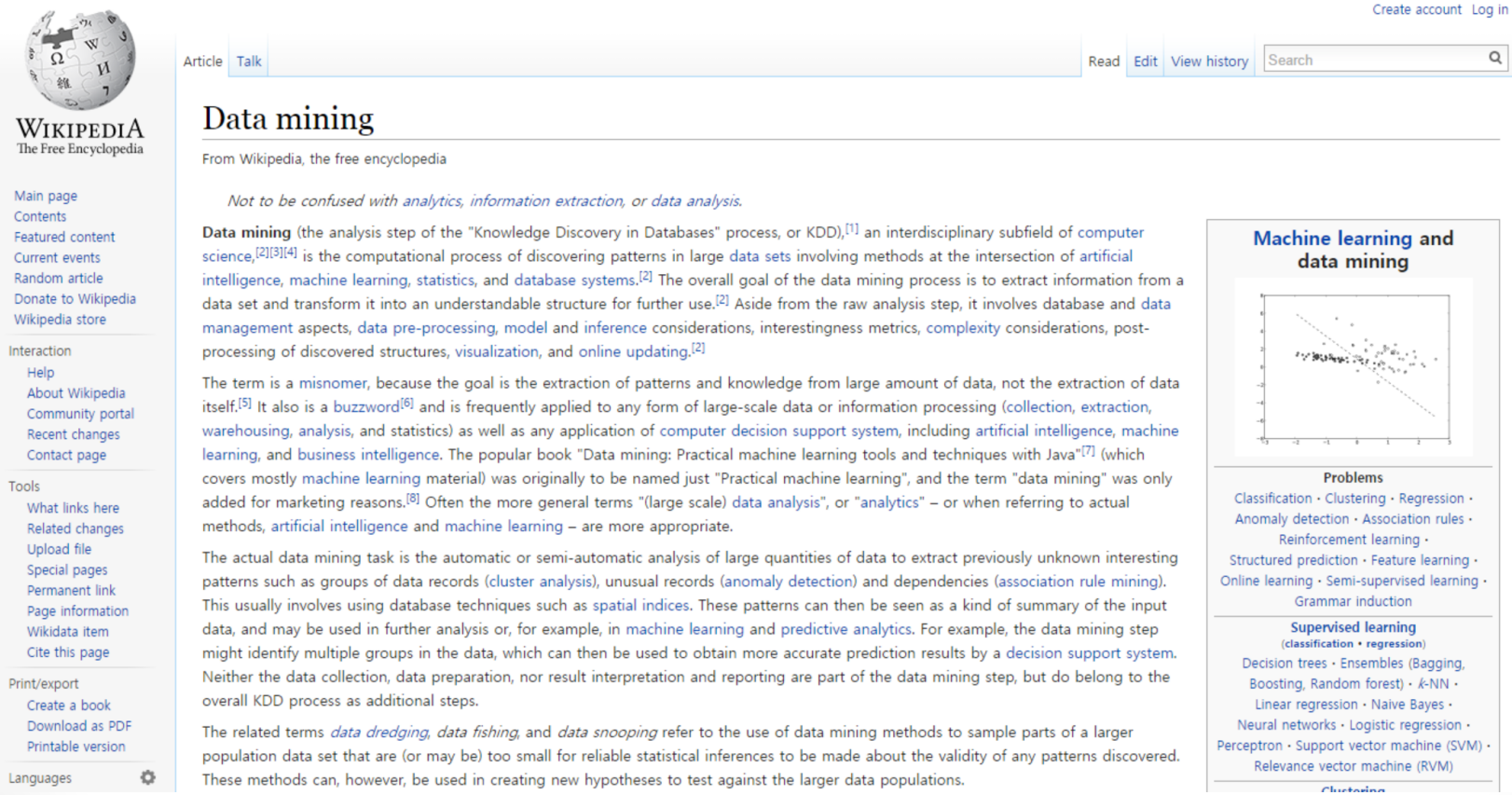
Example 1: Naver News

- Result…

```
印 타타, 중국 스마트폰 메이커 샤오미 지분 매입
http://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=421&aid=0001386493
정부, 中·EU와 '터키 휴대전화 세이프가드' 공조
http://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=001&aid=0007554968
포인트 챙기는 고객은 '별거지'?…익명 SNS의 그늘
http://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=055&aid=0000314034
“데스크톱가상화(VDI), 시스코 서버 성장 이끄는 원동력”
http://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=138&aid=0002028112
“유전정보 해독 노화·질병 극복”
http://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=022&aid=0002821835
[뉴스클러스] 오감 즐기는 가상현실 시장 잡아라…콘텐츠 개발 치…
http://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=214&aid=0000491328
러시아, 오바마 이메일 해킹… 어떻게 됐었나 보니
http://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=029&aid=0002280247
유명 개발자 커뮤니티 믿고 덥석 계약했다가…
http://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=029&aid=0002280290
‘한국형 랜섬웨어’무차별 유포… 피해 예방하려면
http://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=029&aid=0002280314
삼성전기 1분기 영업익 300% 급증에 숨은 비밀
http://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=029&aid=0002280294
삼성·LG전자 주중 동시 실적발표…TV사업 적자 기로
http://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=001&aid=0007554912
구글이 꼽았다! 앱 개발사가 위반하는 5가지 규정은?
http://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=030&aid=0002352517
`주홍글씨`에 두번 우는 기술우량기업
http://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=030&aid=0002352524
[에디슨프로젝트]<상>성과 내는 에디슨 사업
http://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=030&aid=0002352527
```

Example 2: Wikipedia

- 위키피디아에서 관련 키워드를 수집하고 싶다면?



Create account Log in

Article Talk | Read Edit View history | Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Languages

Data mining

From Wikipedia, the free encyclopedia

Not to be confused with analytics, information extraction, or data analysis.

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD),[1] an interdisciplinary subfield of computer science,[2][3][4] is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.[2] The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.[2] Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.[2]

The term is a misnomer, because the goal is the extraction of patterns and knowledge from large amount of data, not the extraction of data itself.[5] It also is a buzzword[6] and is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence, machine learning, and business intelligence. The popular book "Data mining: Practical machine learning tools and techniques with Java"[7] (which covers mostly machine learning material) was originally to be named just "Practical machine learning", and the term "data mining" was only added for marketing reasons.[8] Often the more general terms "(large scale) data analysis", or "analytics" – or when referring to actual methods, artificial intelligence and machine learning – are more appropriate.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps.

The related terms *data dredging*, *data fishing*, and *data snooping* refer to the use of data mining methods to sample parts of a larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations.

Machine learning and data mining

Problems
Classification · Clustering · Regression · Anomaly detection · Association rules · Reinforcement learning · Structured prediction · Feature learning · Online learning · Semi-supervised learning · Grammar induction

Supervised learning
(classification • regression)
Decision trees · Ensembles (Bagging, Boosting, Random forest) · *k*-NN · Linear regression · Naive Bayes · Neural networks · Logistic regression · Perceptron · Support vector machine (SVM) · Relevance vector machine (RVM)

Clustering

http://en.wikipedia.org/wiki/Data_mining

Example 2: Wikipedia

- HTML 구문 분석

<!-- bodycontent -->
<div id="mw-content-text" lang="en" dir="ltr" class="mw-content-ltr"**>** Not to be confused with analytics, information extraction, or data analysis.
<p>Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD),^{[1]} an interdisciplinary subfield of **computer science**,^{[2]}^{[3]}^{[4]} is the computational process of discovering patterns in large **data sets** involving methods at the intersection of **artificial intelligence**, machine learning, statistics, and database systems.^{[2]} The overall goal ….

- **div tag(id=mw-content-title) 다음에 있는 a tag중에서 title attribute가 있는 경우**
- **div[id=mw-content-text] a[title]**

Example 2: Wikipedia

- Source Code

```
package java3;

import java.io.IOException;

import org.jsoup.Jsoup;
import org.jsoup.nodes.Document;
import org.jsoup.nodes.Element;
import org.jsoup.select.Elements;

public class Example2 {

	public static void main(String[] args) throws IOException {

		Document doc =
		Jsoup.connect("http://en.wikipedia.org/wiki/Data_mining").get();
		Elements words = doc.select("div[id=mw-content-text] a[title]");

		String title = doc.title();
		System.out.println(title);

		for(Element e : words){
			System.out.println(e.attr("title"));
		}
	}
}
```

Example 2: Wikipedia

- Result…

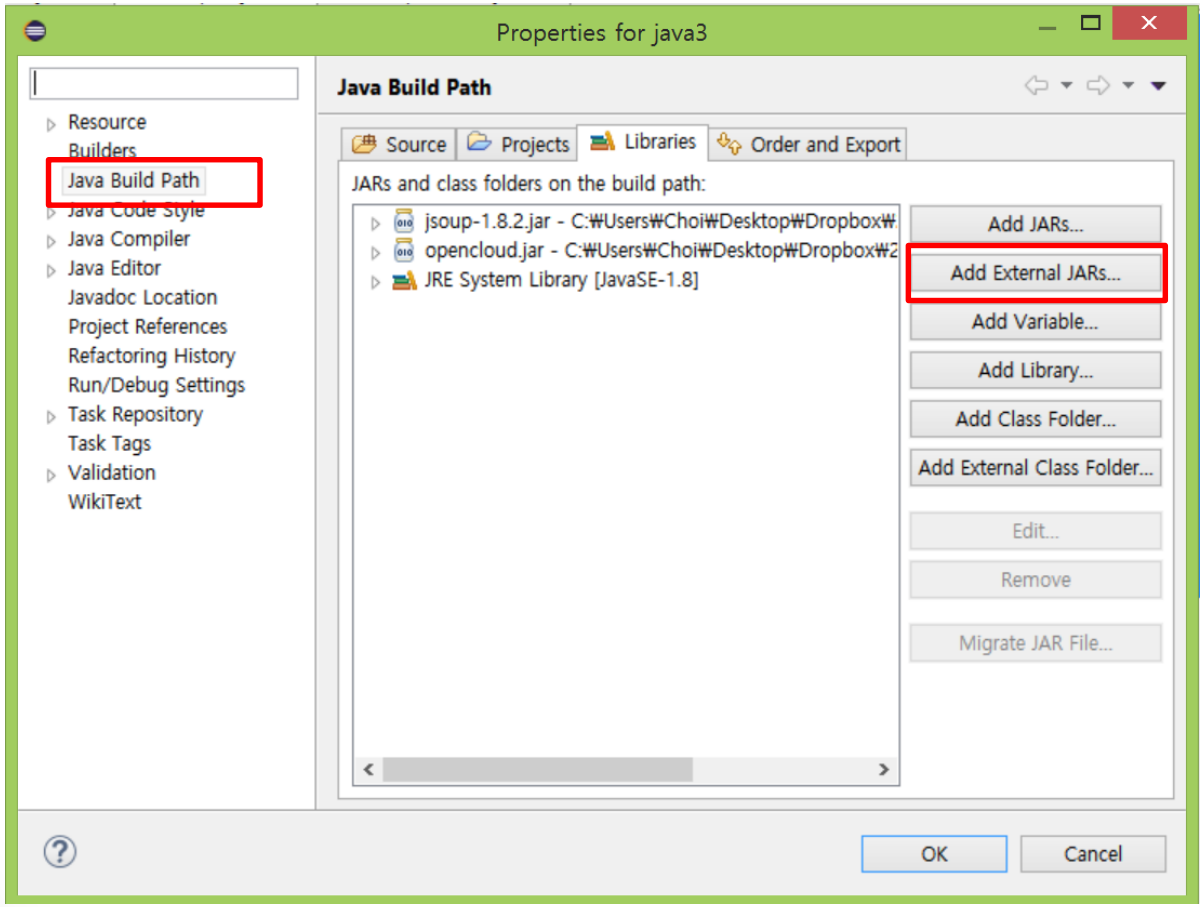
```
Data mining - Wikipedia, the free encyclopedia
Analytics
Information extraction
Data analysis
Computer science
Data set
Artificial intelligence
Machine learning
Statistics
Database system
Data management
Data Pre-processing
Statistical model
Statistical inference
Computational complexity theory
Data visualization
Online algorithm
Buzzword
Data collection
Information extraction
Data warehouse
Data analysis
Decision support system
Artificial intelligence
Machine learning
```

External Library

Java Programming #3

Installing External Library

- JAVA Project properties → Add External JAR



Example 3: Tag Cloud

- Search “tag cloud java library”

http://opencloud.mcavallo.org/

Example 3: Tag Cloud

Quick start

You can create a simple tag cloud following these steps:

1. Create a Cloud object and set its properties. One of the most common properties is the maximum weight value, that defines the range of weight values assigned to tags. It can be set to a convenient value, e.g. the maximum font size. For the minimum weight value can often be kept the default value of zero.

```
Cloud cloud = new Cloud();  // create cloud
cloud.setMaxWeight(38.0);   // max font size
```

2. Populate the tag cloud by creating Tag objects and adding them to the cloud. As said before, the Cloud object by default counts the number of times that a tag has been added, so that more frequent tags will have a higher score.

```
Tag tag = new Tag("Google", "http://www.google.com");   // creates a tag
cloud.addTag(tag);                                      // adds it to the cloud
```

3. Call the tags method of the Cloud class to obtain a list of the tags composing the tag cloud, each with its own weight assigned. Then cycle through the list and write the HTML code.

```
<div>
<% for (Tag tag : cloud.tags()) { %>
<a href="<%= tag.getLink() %>" style="font-size: <%= tag.getWeight() %>px;"%gt;<%= tag.getName() %></a>
<% } %>
</div>
```

In this example the getLink, getWeight and getName are used to compose the HTML link.

Example 3: Tag Cloud

```
package java3;

import javax.swing.*;
import org.mcavallo.opencloud.*;

public class TestOpenCloud {

    private static final String[] WORDS = {"hello","hello","hello","hello","hello","world","world","world","world","world","world",
    "world","world","world","world","world","world","Korea","Korea","Korea"};

    protected void initUI(String[] wrods) {
        JFrame frame = new JFrame(TestOpenCloud.class.getSimpleName());
        frame.setDefaultCloseOperation(JFrame.EXIT_ON_CLOSE);
        JPanel panel = new JPanel();
        Cloud cloud = new Cloud();

        for (String s : WORDS) {
                cloud.addTag(s);
        }
        for (Tag tag : cloud.tags()) {
            final JLabel label = new JLabel(tag.getName());
            label.setOpaque(false);
            label.setFont(label.getFont().deriveFont((float) tag.getWeight() * 10));
            panel.add(label);
        }

        frame.add(panel);
        frame.setSize(800, 600);
        frame.setVisible(true);
    }

    public static void main(String[] args) {

    String[] words = WORDS;
        SwingUtilities.invokeLater(new Runnable() {
            @Override
            public void run() {
                new TestOpenCloud().initUI(words);
            }
        });
    }
}
```

Example 3: Tag Cloud

- Results

Assignment #4

Java Programming #3

Assignment #4 : Web crawler

- Web site를 crawling하여 단어를 수집한 후 이를 이용하여 informative한 tag cloud를 만드시오.
  - 단어 예시 → Naver IT 뉴스기사 제목을 구성하는 단어
  - Informative?
  - Tab cloud 구축은 example 3 참고

Java Assignment

- Source code 30%
  - 작성한 package 폴더
  - Workspace>project>src>id

- Report 70% (under 2 pages)
  - 과제 내용에 대해 정리
    - 코드 설명, 사용한 함수 설명 등 자유롭게

Assignment #4 : Web crawler

- Submission guideline
  - Due 5/18 00:00
  - Mail to iangoozh@gmail.com
  - Title of mail : [Assignment N] 20XX-XXXXX YYY
  - Title of files
    - Package name : id
    - Report : id
    - Package and report의 압축 파일 : [Assignment N] 20XX-XXXXX YYY

End of Document