

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



The screenshot shows the jsoup website's navigation bar with links: jsoup, News, Bugs, Discussion, Download, API Reference, Cookbook, and Try jsoup. Below the navigation bar is a breadcrumb trail: jsoup » Download jsoup. The main heading is "Download jsoup". The text states: "jsoup is available as a downloadable .jar java library. The current release version is 1.8.2." Below this is a list of download links:

- **jsoup-1.8.2.jar** core library
- jsoup-1.8.2-sources.jar optional sources jar
- jsoup-1.8.2-javadoc.jar optional javadoc jar

Below the list is the section "What's new" with the text: "See the 1.8.2 release announcement for the latest changes, or the changelog for the full history."

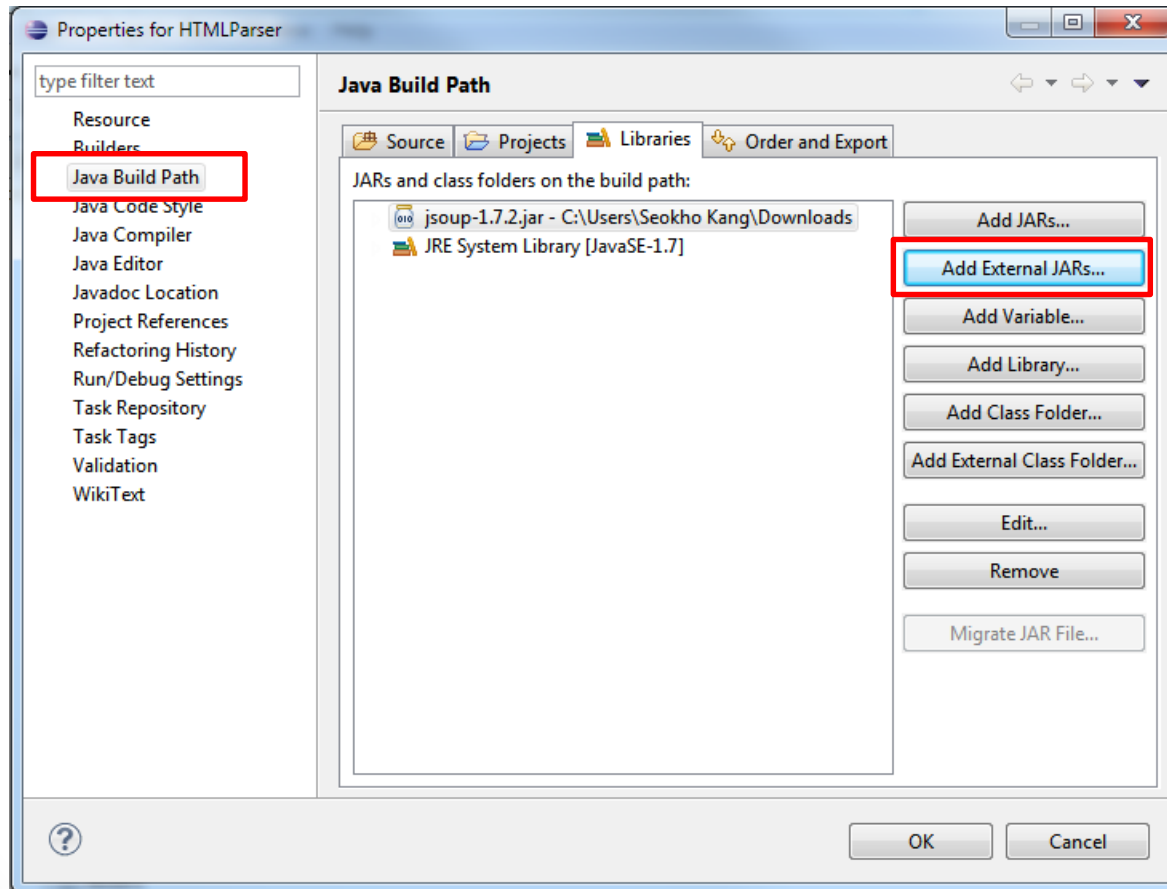
Maven

If you use **Maven** to manage the dependencies in your Java project (and you should!), you do not need to download; just place the following into your POM's `<dependencies>` section:

```
<dependency>
  <!-- jsoup HTML parser library @ http://jsoup.org/ -->
  <groupId>org.jsoup</groupId>
  <artifactId>jsoup</artifactId>
  <version>1.8.2</version>
</dependency>
```

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 <code>Jsoup</code> 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

```
*
tag
ns|E
```

```
#id
```

```
.class
```

```
[attr]
```

```
[^attrPrefix]
```

```
[attr=val]
```

```
[attr="val"]
```

Matches

any element

elements with the given tag name

elements of type E in the namespace ns

elements with attribute ID of "id"

elements with a class name of "class"

elements with an attribute named "attr" (with any value)

elements with an attribute name starting with "attrPrefix". Use to find elements with HTML5 datasets

elements with an attribute named "attr", and value equal to "val"

elements with an attribute named "attr", and value equal to "val"

Basic Structure

```
Document doc = Jsoup.connect([URL]).get();  
Elements es=doc.select([Syntax]);  
  
for(Element e: es){  
    System.out.println(e.text());  
    System.out.println(e.attr([attributeName]));  
}
```

← [URL]에서 HTML Source를 받아옴
← 특정 Syntax를 만족하는 Element들을 선택
← 각 Element들에 대해서
← Text 출력
← 특정 Attribute의 값 출력

Example 1: Naver News

- 네이버 뉴스 IT/과학 페이지의 기사 제목을 수집하고 싶다면?

NAVER 뉴스 TV연예 스포츠 뉴스스탠드 오늘의신문 날씨 로그인

04.27 (월) ☀️ 출산 15°C 주요뉴스 ▶ “아이 일찍 낳았다고 핀잔” “남편 있다고 취업 탈락”... 2... 상식뉴스 · 언론사뉴스 · 라이브러리 · 편집이력

뉴스를 속보 정치 경제 사회 생활/문화 세계 IT/과학 오피니언 포토 TV 행킹뉴스 뉴스 검색 검색

IT/과학

모바일
인터넷/SNS
통신/뉴미디어
IT 일반
보안/해킹
컴퓨터
게임/리뷰
과학 일반
속보

영예의 전당
이달의 방송기자상
바로가기 ▶

세계가전박람회(IFA) 2015

中 최대 가전사 '한국 베끼기' 더 노골적
세탁기 하나에 드럼 2개 하이얼 듀얼 드럼세탁기 LG 올 CES 서 선보인 '트윈워시와' 완전 유사 왼쪽은 하이얼이 지난 24일(현지시간) 개막한 2015 IFA 글로벌 프... [파이낸셜뉴스](#) | 👁 300+

- 스마트워치가 미래 가전시장 성장 이끈다 [서울경제](#)
- 中 가전, 삼성·LG '베끼기'도 진화하나 [아이뉴스24](#)
- 스마트TV 춘추전국시대, 구글·애플·삼성·LG 격돌 ... 세계가전박람회(IFA) 2015 현장 가봤더니 [한국경제](#)
- 필립스, 알리안츠와 '헬스테크' 사업 본격화 [이데일리](#)
- 中가전 '한국 따라하기'.. 삼성·LG 맹추격 [이데일리](#)

삼성의 '혁신' vs. LG의 '아날로그'
갤S6엣지, 양면 엣지·무선충전 등 '기술력의 결정체' G4, 천연 소가죽 케이스·카메라로 감성 자극·삼성·LG, 다른 경쟁포인트 갤럭시·'품질 사태' 눈길 G4, 출시전부터 이목 끌... [파이낸셜뉴스](#) | 👁 999+

글로벌 스마트폰 제조사도 신제품 경쟁
화웨이 'P8' 최대 15시간까지 동영상 재생 HTC '원M9' 전면 카메라 '샷카 기능' 강화 소니 '엑스페리아Z4' 고해상도 촬영에 방수는 기본 화웨이 P8 HTC 원M9 소니 엑스페리아Z4 (왼쪽... [파이낸셜뉴스](#) | 👁 10+

"소통은 좋지만 신분 노출 싫어"...익명 SNS 인기
트위터와 페이스북, 여러 후발 주자들까지 2011년 처음으로 10억 명을 돌파한 SNS 사용자 수는 곧 20억을 넘어서 2018년에는 25억 명에 육박할 거로 예상됩니다. 자신의 생활이나 생... [SBS TV](#) | 👁 100+

가장 많이 본 뉴스 [더보기+](#)

종합 정치 경제 사회 생활/문화

- [정치] 경남기업 첫 압수수색 직전 "咸..."
- [경제] '간암 보험금' 받아낸 사연 알고보...
- [사회] 안동 낙동강서 신원 미상 여성 변...
- [생활/문화] [날씨] 오늘 전국 맑고 초여...
- [세계] 네팔 대지진 사망자 2천300여명
- [IT/과학] 삼성의 '혁신' vs. LG의 '아날...
- [연예] '라스트헬스보이' 김수영, 12주만...
- [스포츠] [EPL 34R] '역습 무방비' 맨유, ...

핫이슈

네달 규모 7.8 강진

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',nclinks:'ar1.cmt'}" >a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shn
샤오미 지문 매입" id='949984_26478920'><strong>샤 타타, 중국 스마트폰 메이커 샤오미 지문 매입</strong></a><span class="writing">뉴스1</span></li>
<li class="_rcount" data-comment="{gno:'news001,0007554968',nclinks:'ar1.cmt'}" >a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shn
세이프가드' 공조" id='949984_26478687'><strong>정부, 中·티와 '터키 휴대전화 세이프가드' 공조</strong></a><span class="writing">연합뉴스</sp;
<li class="_rcount" data-comment="{gno:'news055,0000314034',nclinks:'ar1.cmt'}" >a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shn
지'?...익명 SNS의 그늘" id='949984_26478306'><strong>포인트 행기는 고객은 '별거지'?...익명 SNS의 그늘</strong></a><strong class="r_ico_r_vod_t;
</li>
<li class="_rcount" data-comment="{gno:'news138,0002028112',nclinks:'ar1.cmt'}" >a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shn
서버 성장 이끄는 원동력" id='949984_26478052'><strong>"데스크톱가상화(VDI), 시스템 서버 성장 이끄는 원동력" </strong></a><span class="writing">디지털데일리</span></li>
<li class="_rcount" data-comment="{gno:'news022,0002821835',nclinks:'ar1.cmt'}" >a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shn
복" id='949984_26478050'><strong>"유전정보 해독 노화·질병 극복" </strong></a>세계일보</span></li>
</ul>
<ul class="type02" style="margin-bottom:18px;">
<li class="_rcount" data-comment="{gno:'news214,0000491328',nclinks:'ar1.cmt'}" >a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shn
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',nclinks:'ar1.cmt'}" >a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shn
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',nclinks:'ar1.cmt'}" >a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shn
석 계약했다가..." id='949984_26477373'><strong>유명 개발자 커뮤니티 밀고 업적 계약했다
토"><span class="writing">디지털타임스</span></li>
<li class="_rcount" data-comment="{gno:'news029,0002280314',nclinks:'ar1.cmt'}" >a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shn
포... 피해 예방하려면" id='949984_26477133'><strong>'한국형 랜섬웨어' 무차별 유포... 피
<li class="_rcount" data-comment="{gno:'news029,0002280294',nclinks:'ar1.cmt'}" >a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shn
중에 숨은 비밀" id='949984_26477105'><strong>삼성전기 1분기 영업이익 300% 급증에 숨은 비
</li>
<ul class="type02" style="margin-bottom:18px;">
<li class="_rcount" data-comment="{gno:'news001,0007554912',nclinks:'ar1.cmt'}" >a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shn
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',nclinks:'ar1.cmt'}" >a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shn
반하는 5가지 규정은?" id='949984_26477076'><strong>구글이 뽑았다! 앱 개발사가 위반하는 5가지 규정은?</strong></a><span class="writing">전자신
<li class="_rcount" data-comment="{gno:'news030,0002352524',nclinks:'ar1.cmt'}" >a href="http://news.naver.com/main/read.nhn?mode=LSD&mid=shn
량기업" id='949984_26477075'><strong>주홍글씨'에 두 번 우는 기술우량기업</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=105").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...

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, ㄱ·EU와 '터키 휴대전화 세이프가드' 공조
<http://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=001&aid=0007554968>
포인트 채기는 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>
[뉴스플러스] 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 들었나 보니
<http://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=029&aid=0002280247>
유명 개발자 커뮤니티 믿고 업적 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 영업이익 300% 급증에 숨은 비밀
<http://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=029&aid=0002280294>
삼성·LG전자 주종 동시 실적발표...TV사업 적자 기록
<http://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=001&aid=0007554912>
구글이 끝났다! 앱 개발사가 위반하는 5가지 규정은?
<http://news.naver.com/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=030&aid=0002352517>
‘주홍글씨’에 두 번 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?



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

Article **Talk** Read **Edit** View history

Data mining

From Wikipedia, the free encyclopedia

Not to be confused with [analytics](#), [information extraction](#), or [data analysis](#).

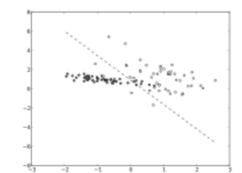
Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD),^[1] an interdisciplinary subfield of [computer science](#),^{[2][3][4]} is the computational process of discovering patterns in large [data sets](#) involving methods at the intersection of [artificial intelligence](#), [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"><div class="dablink">Not to be confused
with <a href="/wiki/Analytics" title="Analytics">analytics</a>, <a href="/wiki/Information_extraction"
title="Information extraction">information extraction</a>, or <a href="/wiki/Data_analysis" title="Data
analysis">data analysis</a>.</div>
<p><b>Data mining</b> (the analysis step of the "Knowledge Discovery in Databases" process, or KDD),<sup
id="cite_ref-Fayyad_1-0" class="reference"><a href="#cite_note-Fayyad-
1"><span></span>1<span></span></span></a></sup> an interdisciplinary subfield of <a
href="/wiki/Computer_science" title="Computer science">computer science</a>,<sup id="cite_ref-acm_2-0"
class="reference"><a href="#cite_note-acm-2"><span></span>2<span></span></span></a></sup><sup id="cite_ref-
britannica_3-0" class="reference"><a href="#cite_note-brittanica-
3"><span></span>3<span></span></span></a></sup><sup id="cite_ref-elements_4-0" class="reference"><a
href="#cite_note-elements-4"><span></span>4<span></span></span></a></sup> is the computational process of
discovering patterns in large <a href="/wiki/Data_set" title="Data set">data sets</a> involving methods at the
intersection of <a href="/wiki/Artificial_intelligence" title="Artificial intelligence">artificial intelligence</a>, <a
href="/wiki/Machine_learning" title="Machine learning">machine learning</a>, <a href="/wiki/Statistics"
title="Statistics">statistics</a>, and <a href="/wiki/Database_system" title="Database system" class="mw-
redirect">database systems</a>.<sup id="cite_ref-acm_2-1" class="reference"><a href="#cite_note-acm-
2"><span></span>2<span></span></span></a></sup> 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...

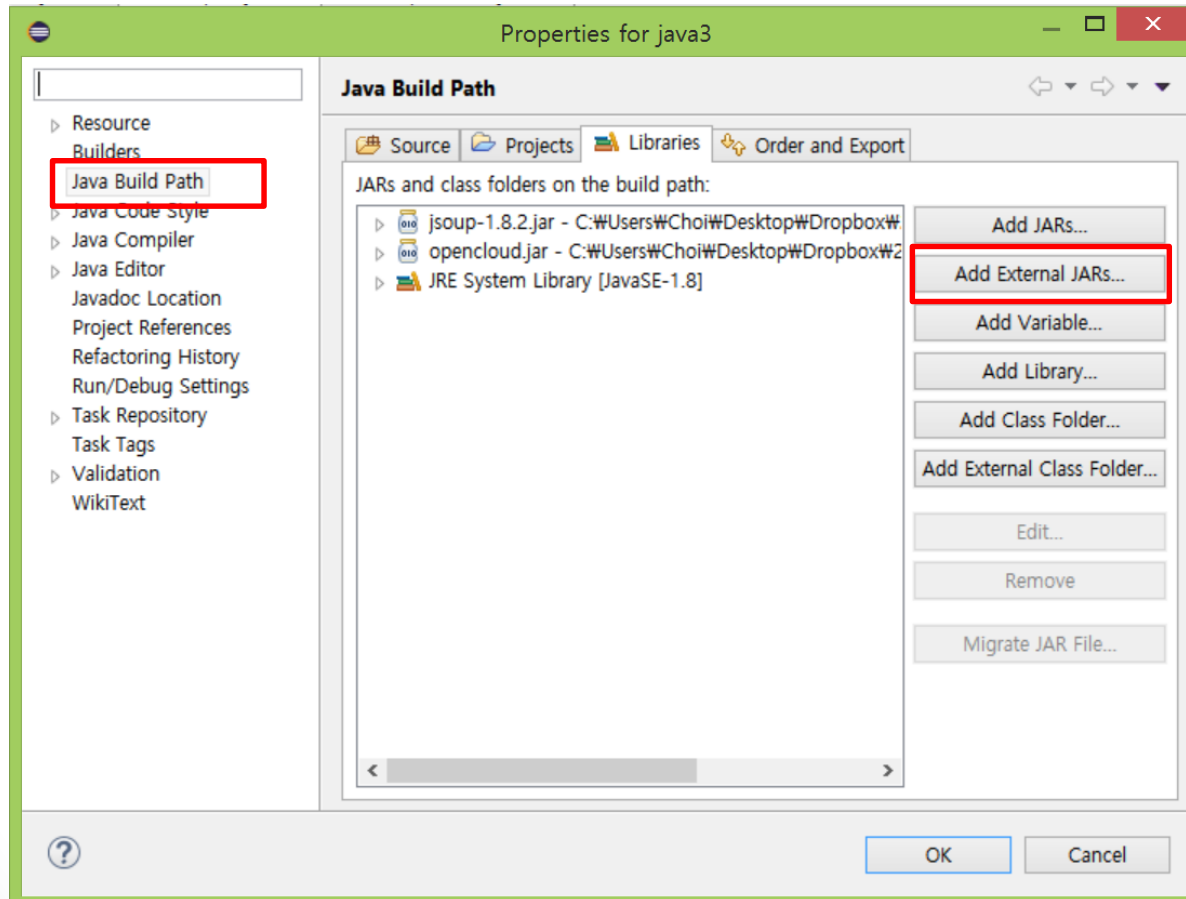
```
Data mining - Wikipedia, the free encyclopedia
Analytics
Information extraction
Data analysis
Computer science
Data set
Artificial intelligence
Machine learning
Statistics
Database system
Data management
Data Pre-processing
Statistical model
Statistical inference
Computational complexity theory
Data 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”

OpenCloud

[Home](#)

[Download](#)

[Getting Started Guide](#)

[Javadoc](#)

[Samples](#)

[Project Page](#)

[Support](#)

OpenCloud is a Java library for generating and managing tag clouds, similar to those found in many websites. The aim of the project is to provide an easy to use library, but versatile enough to be used in various applications.

Basically the user inserts words or word/link pairs into the cloud and sets the parameters that control its structure, such as the number of tags composing the cloud, the weight range, the type of ordering etc. Then the library assigns a weight to each tag based on its frequency (or score) and gives back the resulting list of weighted tags. The appearance of the cloud can be controlled through HTML/CSS instructions inside the JSPs or using the provided formatter class.

The tag cloud is intended to be used also as a knowledge management tool, for example to get a quick visualisation of the main topics of a document. Hence there are methods to extract tags from a generic text and to filter unwanted words according to a dictionary.

Your feedback is very appreciated and will help to improve the library. If you encounters any problem, or want to request a feature or just want to make a comment, feel free to post a message on the [forum](#).

art australia baby beach birthday blue bw california canada canon cat chicago
china christmas city dog england europe family festival flower flowers food
france friends fun germany holiday india italy japan london me mexico
music nature new newyork night nikon nyc paris park party people
portrait sanfrancisco sky snow spain summer sunset taiwan tokyo travel trip uk usa
vacation water wedding

sourceforge

<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;"><%= 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[] words) {
        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);
        }

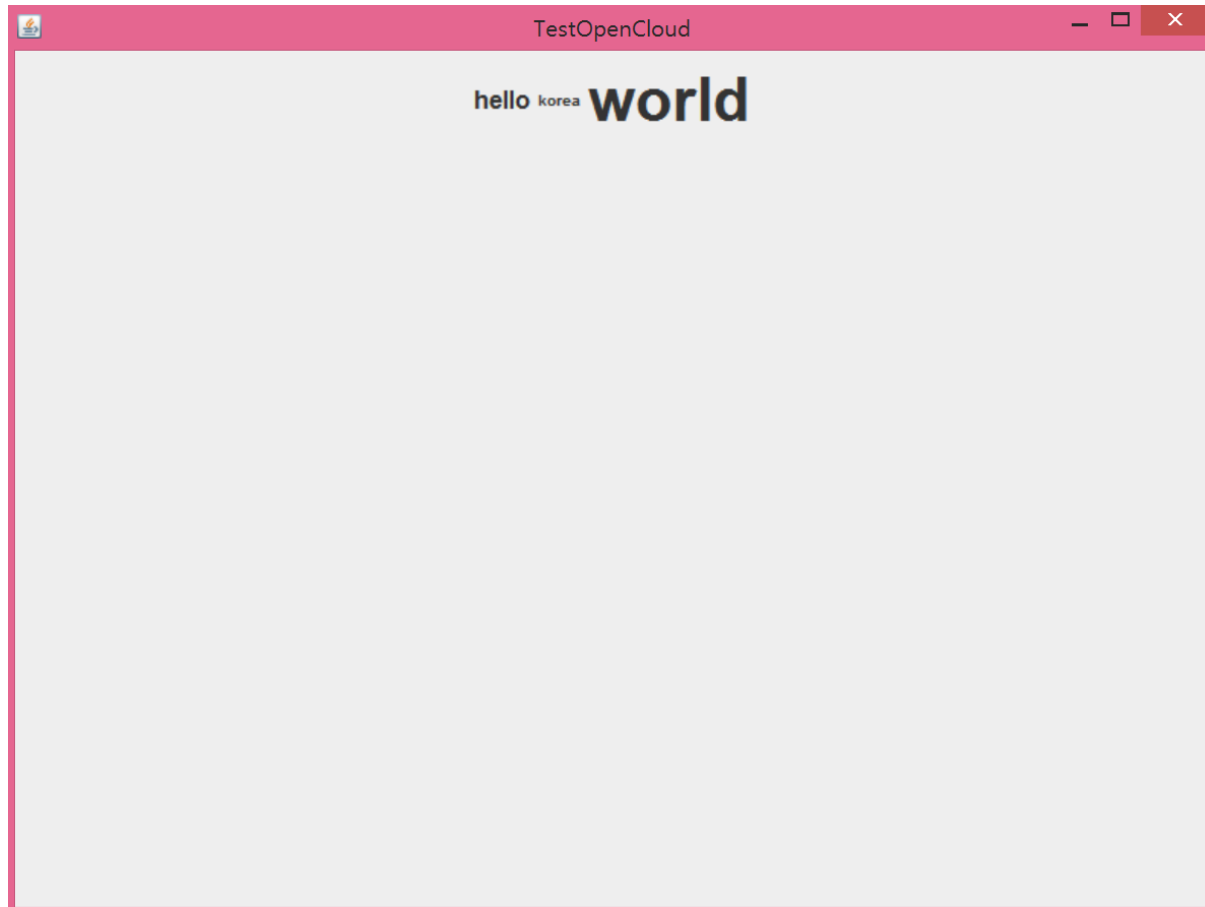
        frame.add(panel);
        frame.setSize(800, 600);
        frame.setVisible(true);
    }

    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