**JSoup**

**학습목표**

JSoup을 사용하여 게시판 웹페이지 내용을 가져온다.

가져온 웹페이지 내용에서 원하는 부분만 추출한다.

**목차**

[1. JSoup 라이브러리 2](#_Toc469339144)

[1) 다운로드 2](#_Toc469339145)

[2) 설치 2](#_Toc469339146)

[2. 게시글 목록 가져오기 4](#_Toc469339147)

[1) Example01 4](#_Toc469339148)

[2) Example02 6](#_Toc469339149)

[3) Example03 7](#_Toc469339150)

[4) Example04 8](#_Toc469339151)

[3. 게시글 내용 가져오기 9](#_Toc469339152)

[1) Example05 9](#_Toc469339153)

[2) Example06 10](#_Toc469339154)

[3) Example07a 11](#_Toc469339155)

[4) Example07b 13](#_Toc469339156)

[5) Example07c 14](#_Toc469339157)

[4. 게시글 전체 목록 가져오기 15](#_Toc469339158)

[1) Example08 15](#_Toc469339159)

[2) Example09 17](#_Toc469339160)

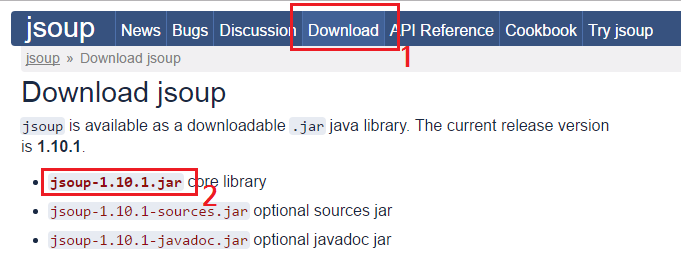
[3) Example10 20](#_Toc469339161)

[4) Example11 21](#_Toc469339162)

# JSoup 라이브러리

## 다운로드

https://jsoup.org/



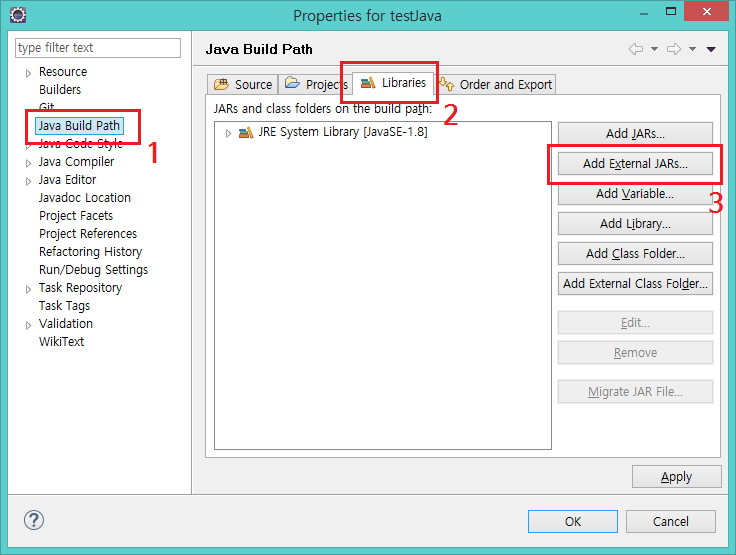
jsoup-1.10.1.jar 파일을 다운로드 하자.

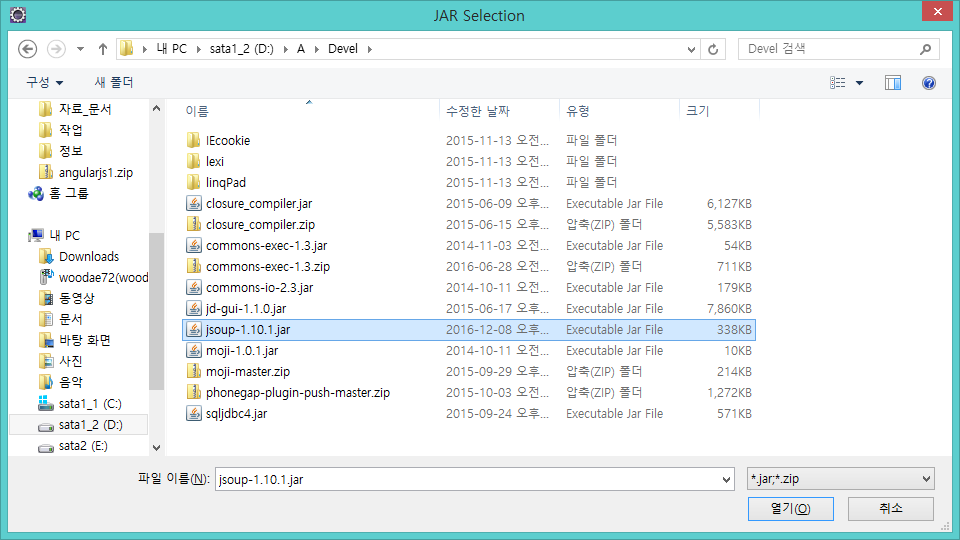
## 설치

jar 파일들을 뫃아 놓을 폴더를 하나 만들고

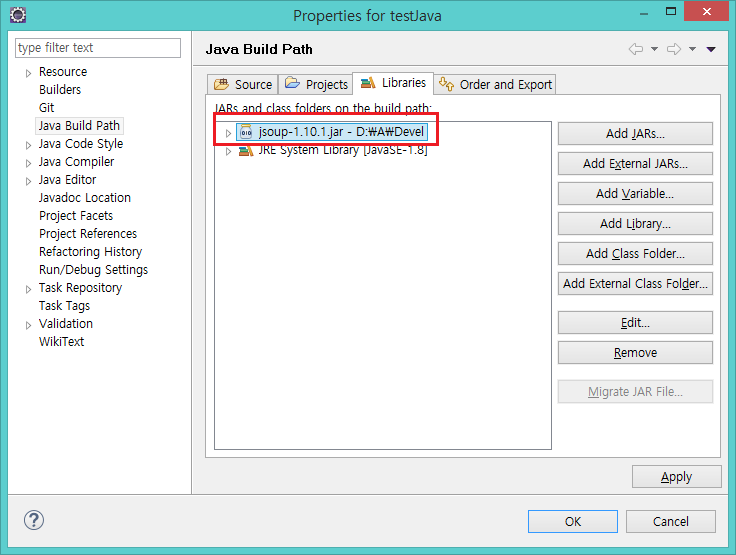
다운로드한 jsoup-1.10.1.jar 파일을 그 폴더에 넣어 두자.

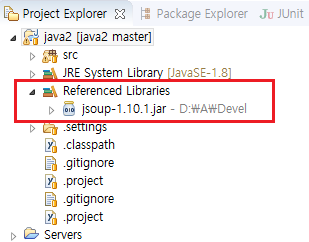
eclipse 메뉴: Project - Properties





jsoup-1.10.1.jar 파일 선택





# 게시글 목록 가져오기

## Example01

"http://www.skhu.ac.kr/board/boardlist.aspx?curpage=1&bsid=10004&searchBun=51"

위 게시글 목록 페이지에서, 게시글 제목은 다음과 같은 HTML 태그이다.

|  |
| --- |
| <td class="left15">  <span id="ctl00\_ContentPlaceHolder1\_ctl00\_rptList2\_ctl00\_lblTitle">  <a href='boardread.aspx?idx=27461&amp;curpage=1&amp;bsid=10004&amp;searchBun=51'>  [교무처] 8학기 이상 재학생 중 취업자에 대한 공...  </a>  </span>  </td> |

Example01.java

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19 | package jsoup1;  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 Example01 {    public static void main(String[] args) throws IOException {  String listUrl = "http://www.skhu.ac.kr/board/boardlist.aspx?curpage=1&bsid=10004&searchBun=51";  Document document = Jsoup.connect(listUrl).get();  Elements tagList = document.select("a");  for (Element tag : tagList )  System.out.println(tag.text());  }  } |

String listUrl = "http://www.skhu.ac.kr/board/boardlist.aspx?curpage=1&bsid=10004&searchBun=51";

게시글 목록 페이지의 URL 이다.

Document document = Jsoup.connect(listUrl).get();

웹서버에 접속해서, 게시글 목록 페이지의 내용을 받아와서,

웹페이지의 내용을 담은 Document 객체를 리턴한다.

Elements tagList = document.select("a");

웹페이지의 내용에 들어있는 a 태그들 전체를 Elements 객체에 담아서 리턴한다.

Elements 객체는 Element 객체들의 목록이 들어있는 객체이다.

즉 a 태그 한 개의 내용이 Element 객체에 들어있고,

Element 객체 여러 개가 Elements 객체에 들어있다.

for (Element tag : tagList )

tagList에 들어있는 Element 객체를 한 개씩 tag 변수로 꺼내서 for 루프를 반복한다.

tagList에 들어있는 Element 객체에는 a 태그 한 개의 내용이 들어있다.

System.out.println(tag.text());

tag 변수가 참조하는 Element 객체에 들어있는 태그의 내용 중에서

**<태그>text</태그>** text 부분만 화면에 출력한다.

Element 클래스의 text() 메소드는, **<태그>text</태그>** text 부분을 리턴한다.

실행결과 출력

|  |
| --- |
| HOME  느티아래  알림마당  대학  대학원  [교무처] 8학기 이상 재학생 중 취업자에 대한 공...  [졸업 자격증 제출] 영어, 일어, 중어, 경영 졸...  [중어중국학과] 2016-2학기 졸업시험 시행 안내  [교무처] 2016년 겨울계절학기 폐강 안내  [교무처]2016년 겨울계절학기 폐강 위험강좌 안내...  [사회복지학과] 재시험 일정 공지 (필독)  [사회복지학과] 졸업시험 결과 확인  [교무처] 군복무기간 중 군e-러닝 취득학점 인정안...  [사회복지학과] 졸업시험 장소 공지  (수정-시간연장)학점포기 일정 안내(2016학년도 ...  2017학년도 1학기 재입학생 모집 공고  [교무처] 2016학년도 겨울계절학기 수강신청 안내...  타 대학 학점교류는 학점교류 게시판을 참고하세요  [사회복지학과] 사회복지학과 졸업시험 (필독)  [학부]2016-2학기 '학기 중 강의 의견 수렴'...  [학부] 2016학년도 2학기 수강과목 철회 안내  2  3  4  5  6  7  8  9  10  11  12  13  14  15 |

게시글 목록 페이지에서 게시글 제목만 <a> 태그인 것이 아니고

페이지 윗 부분의 메뉴와, 페이지 아랫 부분의 페이지 번호 목록도 <a> 태그이다.

게시글 제목에 해당하는 <a> 태그들만 가져오려면 Example01 코드를 수정해야 한다.

## Example02

Example02.java

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19 | package jsoup1;  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 Example02 {    public static void main(String[] args) throws IOException {  String listUrl = "http://www.skhu.ac.kr/board/boardlist.aspx?curpage=1&bsid=10004&searchBun=51";  Document document = Jsoup.connect(listUrl).get();  Elements tagList = document.select("td span a");  for (Element tag : tagList )  System.out.println(tag.text());  }  } |

Elements tagList = document.select("a");

document 웹페이지에 들어있는 모든 <a> 태그들 목록을 리턴한다.

Elements tagList = document.select("td span a");

document 웹페이지에 들어있는 <a> 태그들 중에서

<td> 태그 아래의 <span> 태그 아래의 <a> 태그들의 목록만 리턴한다.

"td span a" 부분은 CSS 실렉터(selector) 문법이다.

CSS 실렉터 문법은 고급웹프로그래밍I에서 자세히 배운다.

위 게시글 목록 페이지에서, 게시글 제목은 다음과 같은 HTML 태그이다.

|  |
| --- |
| <td class="left15">  <span id="ctl00\_ContentPlaceHolder1\_ctl00\_rptList2\_ctl00\_lblTitle">  <a href='boardread.aspx?idx=27461&amp;curpage=1&amp;bsid=10004&amp;searchBun=51'>  [교무처] 8학기 이상 재학생 중 취업자에 대한 공...  </a>  </span>  </td> |

실행결과 출력

|  |
| --- |
| [교무처] 8학기 이상 재학생 중 취업자에 대한 공...  [졸업 자격증 제출] 영어, 일어, 중어, 경영 졸...  [중어중국학과] 2016-2학기 졸업시험 시행 안내  [교무처] 2016년 겨울계절학기 폐강 안내  [교무처]2016년 겨울계절학기 폐강 위험강좌 안내...  [사회복지학과] 재시험 일정 공지 (필독)  [사회복지학과] 졸업시험 결과 확인  [교무처] 군복무기간 중 군e-러닝 취득학점 인정안...  [사회복지학과] 졸업시험 장소 공지  (수정-시간연장)학점포기 일정 안내(2016학년도 ...  2017학년도 1학기 재입학생 모집 공고  [교무처] 2016학년도 겨울계절학기 수강신청 안내...  타 대학 학점교류는 학점교류 게시판을 참고하세요  [사회복지학과] 사회복지학과 졸업시험 (필독)  [학부]2016-2학기 '학기 중 강의 의견 수렴'...  [학부] 2016학년도 2학기 수강과목 철회 안내 |

응용)

Elements tagList = document.select("div span");

document 웹페이지에 들어있는 <span> 태그들 중에서

<div> 태그 아래의 <span> 태그들의 목록만 리턴한다.

## Example03

Example03.java

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19 | package jsoup1;  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 Example03 {    public static void main(String[] args) throws IOException {  String listUrl = "http://www.skhu.ac.kr/board/boardlist.aspx?curpage=1&bsid=10004&searchBun=51";  Document document = Jsoup.connect(listUrl).get();  Elements tagList = document.select("td span a");  for (Element tag : tagList )  System.out.printf("%s %s\n", tag.attr("href"), tag.text());  }  } |

Element 클래스의 text() 메소드는, **<태그>text</태그>** text 부분을 리턴한다.

Element 클래스의 href("애트리뷰트명") 메소드는,

**<태그 애트리뷰트명="애트리뷰트값">text</태그>** 애트리뷰트값 부분을 리턴한다.

|  |
| --- |
| <td class="left15">  <span id="ctl00\_ContentPlaceHolder1\_ctl00\_rptList2\_ctl00\_lblTitle">  <a href='boardread.aspx?idx=27461&amp;curpage=1&amp;bsid=10004&amp;searchBun=51'>  [교무처] 8학기 이상 재학생 중 취업자에 대한 공...  </a>  </span>  </td> |

tag.attr("href") 메소드는 노란색으로 칠한 부분을 리턴한다.

tag.text() 메소드는 하늘색으로 칠한 부분을 리턴한다.

위 태그에서 td 태그의 class 애트리뷰트 값은 "left15" 이다.

span 태그의 id 애트리뷰트 값은 "ctl00\_ContentPlaceHolder1\_ctl00\_rptList2\_ctl00\_lblTitle" 이다.

실행결과 출력

|  |
| --- |
| boardread.aspx?idx=27461&curpage=1&bsid=10004&searchBun=51 [교무처] 8학기 이상 재학생 중 취업자에 대한 공...  boardread.aspx?idx=27677&curpage=1&bsid=10004&searchBun=51 [졸업 자격증 제출] 영어, 일어, 중어, 경영 졸...  boardread.aspx?idx=27673&curpage=1&bsid=10004&searchBun=51 [중어중국학과] 2016-2학기 졸업시험 시행 안내  boardread.aspx?idx=27612&curpage=1&bsid=10004&searchBun=51 [교무처] 2016년 겨울계절학기 폐강 안내  boardread.aspx?idx=27606&curpage=1&bsid=10004&searchBun=51 [교무처]2016년 겨울계절학기 폐강 위험강좌 안내...  boardread.aspx?idx=27599&curpage=1&bsid=10004&searchBun=51 [사회복지학과] 재시험 일정 공지 (필독)  boardread.aspx?idx=27590&curpage=1&bsid=10004&searchBun=51 [사회복지학과] 졸업시험 결과 확인  boardread.aspx?idx=27536&curpage=1&bsid=10004&searchBun=51 [교무처] 군복무기간 중 군e-러닝 취득학점 인정안...  boardread.aspx?idx=27516&curpage=1&bsid=10004&searchBun=51 [사회복지학과] 졸업시험 장소 공지  boardread.aspx?idx=27496&curpage=1&bsid=10004&searchBun=51 (수정-시간연장)학점포기 일정 안내(2016학년도 ...  boardread.aspx?idx=27462&curpage=1&bsid=10004&searchBun=51 2017학년도 1학기 재입학생 모집 공고  boardread.aspx?idx=27457&curpage=1&bsid=10004&searchBun=51 [교무처] 2016학년도 겨울계절학기 수강신청 안내...  boardread.aspx?idx=27404&curpage=1&bsid=10004&searchBun=51 타 대학 학점교류는 학점교류 게시판을 참고하세요  boardread.aspx?idx=27228&curpage=1&bsid=10004&searchBun=51 [사회복지학과] 사회복지학과 졸업시험 (필독)  boardread.aspx?idx=27218&curpage=1&bsid=10004&searchBun=51 [학부]2016-2학기 '학기 중 강의 의견 수렴'...  boardread.aspx?idx=27066&curpage=1&bsid=10004&searchBun=51 [학부] 2016학년도 2학기 수강과목 철회 안내 |

## Example04

Example04.java

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29 | package jsoup1;  import java.io.IOException;  import java.util.regex.Matcher;  import java.util.regex.Pattern;  import org.jsoup.Jsoup;  import org.jsoup.nodes.Document;  import org.jsoup.nodes.Element;  import org.jsoup.select.Elements;  public class Example04 {    public static void main(String[] args) throws IOException {  String listUrl = "http://www.skhu.ac.kr/board/boardlist.aspx?curpage=1&bsid=10004&searchBun=51";  Document document = Jsoup.connect(listUrl).get();  Elements tagList = document.select("td span a");    String regex = "boardread.aspx\\?idx=([0-9]+)&";  Pattern pattern = Pattern.compile(regex);    for (Element tag : tagList ) {  String s = tag.attr("href");  Matcher matcher = pattern.matcher(s);  String idx = matcher.find() ? matcher.group(1) : null;  System.out.printf("%s %s\n", idx, tag.text());  }  }  } |

"boardread.aspx\\?idx=([0-9]+)&"

위 정규식은 아래에서 노란색 부분과 일치한다.

<a href='boardread.aspx?idx=27461&amp;curpage=1&amp;bsid=10004&amp;searchBun=51'>

"boardread.aspx\\?idx=([0-9]+)&"

위 정규식에서 노란색 괄호 부분은 아래에서 노란색 부분과 일치한다.

<a href='boardread.aspx?idx=27461&amp;curpage=1&amp;bsid=10004&amp;searchBun=51'>

String idx = matcher.find() ? matcher.group(1) : null;

정규식에서 첫 번째 괄호에 일치하는 부분만 구한다.

실행결과 출력

|  |
| --- |
| 27461 [교무처] 8학기 이상 재학생 중 취업자에 대한 공...  27677 [졸업 자격증 제출] 영어, 일어, 중어, 경영 졸...  27673 [중어중국학과] 2016-2학기 졸업시험 시행 안내  27612 [교무처] 2016년 겨울계절학기 폐강 안내  27606 [교무처]2016년 겨울계절학기 폐강 위험강좌 안내...  27599 [사회복지학과] 재시험 일정 공지 (필독)  27590 [사회복지학과] 졸업시험 결과 확인  27536 [교무처] 군복무기간 중 군e-러닝 취득학점 인정안...  27516 [사회복지학과] 졸업시험 장소 공지  27496 (수정-시간연장)학점포기 일정 안내(2016학년도 ...  27462 2017학년도 1학기 재입학생 모집 공고  27457 [교무처] 2016학년도 겨울계절학기 수강신청 안내...  27404 타 대학 학점교류는 학점교류 게시판을 참고하세요  27228 [사회복지학과] 사회복지학과 졸업시험 (필독)  27218 [학부]2016-2학기 '학기 중 강의 의견 수렴'...  27066 [학부] 2016학년도 2학기 수강과목 철회 안내 |

# 게시글 내용 가져오기

## Example05

Example05.java

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22 | package jsoup1;  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 Example05 {    public static void main(String[] args) throws IOException {  String articleUrl =  "http://www.skhu.ac.kr/board/boardread.aspx?idx=27461&curpage=1&bsid=10004&searchBun=51";  Document document = Jsoup.connect(articleUrl).get();  Elements tagList = document.select("span#ctl00\_ContentPlaceHolder1\_ctl00\_lbContents");    for (Element tag : tagList )  System.out.println(tag.text());  }  } |

"http://www.skhu.ac.kr/board/boardread.aspx?idx=27461&curpage=1&bsid=10004&searchBun=51";

위 URL은 게시글 아이디(idx)가 27461인 게시글의 본문 내용을 보여주는 웹페이지의 URL 이다.

위 게시글 내용 웹페이지에서 게시글 내용은 다음과 같은 HTML 태그이다.

|  |
| --- |
| <td colspan="4" class="board\_view\_con">  <span id="ctl00\_ContentPlaceHolder1\_ctl00\_lbContents">  ... 게시글 내용 ...  </span>  </td> |

Elements tagList = document.select("span#ctl00\_ContentPlaceHolder1\_ctl00\_lbContents");

웹페이지 내용에 들어있는 span 태그들 중에서

id="ctl00\_ContentPlaceHolder1\_ctl00\_lbContents" 애트리뷰트를 갖는 span 태그들의 목록만 리턴한다.

Element 클래스의 text() 메소드는, **<태그>text</태그>** text 부분을 리턴한다.

그런데 text 부분에서 HTML 태그들을 모두 제거하고 리턴한다.

즉 text 부분에서

text.replaceAll("</?[^>]+>", "")

위 코드를 실행하여 태그 부분을 제거하고 남은 문자열을 리턴한다.

## Example06

Example06.java

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21 | package jsoup1;  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 Example06 {    public static void main(String[] args) throws IOException {  String articleUrl =  "http://www.skhu.ac.kr/board/boardread.aspx?idx=27461&curpage=1&bsid=10004&searchBun=51";  Document document = Jsoup.connect(articleUrl).get();  Elements tagList = document.select("span#ctl00\_ContentPlaceHolder1\_ctl00\_lbContents");    for (Element tag : tagList )  System.out.println(tag.html());  }  } |

Element 클래스의 html() 메소드는, **<태그>text</태그>** text 부분을 있는 그대로 리턴한다.

## Example07a

Example07a.java

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35 | package jsoup1;  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 Example07a {    public static void main(String[] args) throws IOException {  String articleUrl =  "http://www.skhu.ac.kr/board/boardread.aspx?idx=27461&curpage=1&bsid=10004&searchBun=51";  Document document = Jsoup.connect(articleUrl).get();    Elements tags = document.select("span#ctl00\_ContentPlaceHolder1\_ctl00\_lbTitle");  Element tag = tags.get(0);  String title = tag.text();    tags = document.select("span#ctl00\_ContentPlaceHolder1\_ctl00\_lbDate");  tag = tags.get(0);  String date = tag.text();  tags = document.select("span#ctl00\_ContentPlaceHolder1\_ctl00\_lbWriter");  tag = tags.get(0);  String author = tag.text();  tags = document.select("span#ctl00\_ContentPlaceHolder1\_ctl00\_lbContents");  tag = tags.get(0);  String body = tag.text();    System.out.printf("%s\n %s\n %s\n %s\n", title, date, author, body);  }  } |

게시글 내용 보기 웹페이지에서 게시글 제목 부분의 HTML 태그는 다음과 같다.

|  |
| --- |
| <tr>  <th scope="row" abbr="제목">  <img src="/common/images/txt/txt\_board\_list\_th\_title.gif" alt="제목" />  </th>  <td colspan="3">  <span id="ctl00\_ContentPlaceHolder1\_ctl00\_lbTitle">  [교무처] 8학기 이상 재학생 중 취업자에 대한 공결 안내  </span>  </td>  </tr> |

게시글 내용 보기 웹페이지에서 게시글 작성일, 작성자 부분의 HTML 태그는 다음과 같다.

|  |
| --- |
| <tr>  <th scope="row">  <img src="/common/images/txt/txt\_board\_list\_th\_date.gif" alt="작성일" />  </th>  <td>  <span id="ctl00\_ContentPlaceHolder1\_ctl00\_lbDate">2016-11-09</span>  </td>  <th scope="row">  <img src="/common/images/txt/txt\_board\_list\_th\_write.gif" alt="작성자" />  </th>  <td>  <span id="ctl00\_ContentPlaceHolder1\_ctl00\_lbWriter">교무처(교무팀)</span>  </td>  </tr> |

게시글 내용 보기 웹페이지에서 게시글 내용은 다음과 같은 HTML 태그이다.

|  |
| --- |
| <td colspan="4" class="board\_view\_con">  <span id="ctl00\_ContentPlaceHolder1\_ctl00\_lbContents">  ... 게시글 내용 ...  </span>  </td> |

Elements tags = document.select("span#ctl00\_ContentPlaceHolder1\_ctl00\_lbTitle");

현재 document의 내용 중에서

<span id="ctl00\_ContentPlaceHolder1\_ctl00\_lbTitle"> 형태의 span 태그의 목록을 리턴한다.

Element tag = tags.get(0);

태그 목록에서 첫째 태그 객체를 리턴한다. (index == 0)

"태그#아이디"

이 형태의 CSS 실렉터는 <태그 id="아이디"> 형태의 태그를 가르킨다.

## Example07b

Example07b.java

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30 | package jsoup1;  import java.io.IOException;  import org.jsoup.Jsoup;  import org.jsoup.nodes.Document;  import org.jsoup.select.Elements;  public class Example07b {    public static void main(String[] args) throws IOException {  String articleUrl =  "http://www.skhu.ac.kr/board/boardread.aspx?idx=27461&curpage=1&bsid=10004&searchBun=51";  Document document = Jsoup.connect(articleUrl).get();    Elements tags = document.select("span#ctl00\_ContentPlaceHolder1\_ctl00\_lbTitle");  String title = tags.get(0).text();    tags = document.select("span#ctl00\_ContentPlaceHolder1\_ctl00\_lbDate");  String date = tags.get(0).text();  tags = document.select("span#ctl00\_ContentPlaceHolder1\_ctl00\_lbWriter");  String author = tags.get(0).text();  tags = document.select("span#ctl00\_ContentPlaceHolder1\_ctl00\_lbContents");  String body = tags.get(0).text();    System.out.printf("%s\n %s\n %s\n %s\n", title, date, author, body);  }  } |

Elements tags = document.select("span#ctl00\_ContentPlaceHolder1\_ctl00\_lbTitle");

Element tag = tags.get(0);

String title = tag.text();

위 코드와 아래 코드는 동일하다.

Elements tags = document.select("span#ctl00\_ContentPlaceHolder1\_ctl00\_lbTitle");

String title = tags.get(0).text();

tags.get(0).text()

먼저 tags 객체의 get 메소드를 호출한다.

그 다음 get 메소드가 리턴하는 객체의 text 메소드를 호출한다.

tags 객체는 Elements 클래스 객체이고

Elements 클래스의 get 메소드는 Element 객체를 리턴한다.

text 메소드는 Element 클래스의 메소드이다.

## Example07c

Example07c.java

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22 | package jsoup1;  import java.io.IOException;  import org.jsoup.Jsoup;  import org.jsoup.nodes.Document;  public class Example07c {    public static void main(String[] args) throws IOException {  String articleUrl =  "http://www.skhu.ac.kr/board/boardread.aspx?idx=27461&curpage=1&bsid=10004&searchBun=51";  Document document = Jsoup.connect(articleUrl).get();    String title = document.select("span#ctl00\_ContentPlaceHolder1\_ctl00\_lbTitle").get(0).text();  String date = document.select("span#ctl00\_ContentPlaceHolder1\_ctl00\_lbDate").get(0).text();  String author = document.select("span#ctl00\_ContentPlaceHolder1\_ctl00\_lbWriter").get(0).text();  String body = document.select("span#ctl00\_ContentPlaceHolder1\_ctl00\_lbContents").get(0).text();    System.out.printf("%s\n %s\n %s\n %s\n", title, date, author, body);  }  } |

Elements tags = document.select("span#ctl00\_ContentPlaceHolder1\_ctl00\_lbTitle");

String title = tags.get(0).text();

위 코드와 아래 코드는 동일하다.

String title = document.select("span#ctl00\_ContentPlaceHolder1\_ctl00\_lbTitle").get(0).text();

document.select("span#ctl00\_ContentPlaceHolder1\_ctl00\_lbTitle").get(0).text();

먼저 document 객체의 select 메소드를 호출한다.

그 다음 select 메소드가 리턴하는 객체의 get 메소드를 호출한다.

그 다음 get 메소드가 리턴하는 객체의 text 메소드를 호출한다.

document 객체는 Document 클래스 객체이고

Document 클래스의 select 메소드는 Elements 객체를 리턴한다.

Elements 클래스의 get 메소드는 Element 객체를 리턴한다.

text 메소드는 Element 클래스의 메소드이다.

# 게시글 전체 목록 가져오기

## Example08

Example08.java

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27 | package jsoup1;  import java.io.IOException;  import java.util.regex.Matcher;  import java.util.regex.Pattern;  import org.jsoup.Jsoup;  import org.jsoup.nodes.Document;  import org.jsoup.nodes.Element;  public class Example08 {    public static void main(String[] args) throws IOException {  String listUrl = "http://www.skhu.ac.kr/board/boardlist.aspx?curpage=1&bsid=10004&searchBun=51";  Document document = Jsoup.connect(listUrl).get();  Element aTag = document.select("a.nextL").get(0);    String regex = "boardlist.aspx\\?curpage=([0-9]+)&";  Pattern pattern = Pattern.compile(regex);    Matcher matcher = pattern.matcher(aTag.attr("href"));  String s = matcher.find() ? matcher.group(1) : "0";  int lastPage = Integer.valueOf(s);    System.out.println(lastPage);  }  } |

"http://www.skhu.ac.kr/board/boardlist.aspx?curpage=1&bsid=10004&searchBun=51"

위 URL은 게시글 목록 페이지의 URL 이다.

게시글 목록에서 게시판 마지막 페이지로 가려면 아래의 링크를 클릭하면 된다.



게시판 마지막 페이지로 가는 링크의 HTML 태그는 다음과 같다.

|  |
| --- |
| <a href='boardlist.aspx?curpage=44&amp;bsid=10004&amp;searchBun=51' class="nextL" title="맨뒤로">  <img src="/common/images/common/btn/btn\_nextL.gif" alt="맨뒤로" />  </a> |

Element aTag = document.select("a.nextL").get(0);

현재 document의 내용 중에서

<a class="nextL"> 형태의 a 태그의 목록을 리턴한다.

"태그.클래스"

이 형태의 CSS 실렉터는 <태그 class="클래스"> 형태의 태그를 가르킨다.

"boardlist.aspx\\?curpage=([0-9]+)&";

위 정규식은 아래에서 노란색 부분과 일치한다.

<a href='boardlist.aspx?curpage=44&amp;bsid=10004&amp;searchBun=51' class="nextL" title="맨뒤로">

"boardlist.aspx\\?curpage=([0-9]+)&";

위 정규식에서 노란색 괄호 부분은 아래에서 노란색 부분과 일치한다.

<a href='boardlist.aspx?curpage=44&amp;bsid=10004&amp;searchBun=51' class="nextL" title="맨뒤로">

이 부분은 게시글 목록에서 마지막 페이지의 번호이다.

String idx = matcher.find() ? matcher.group(1) : null;

정규식에서 첫 번째 괄호에 일치하는 부분만 구한다.

int lastPage = Integer.valueOf(s);

문자열을 int 값으로 변환한다.

실행결과 출력

|  |
| --- |
| 44 |

## Example09

Example09.java

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33 | package jsoup1;  import java.io.IOException;  import java.util.regex.Matcher;  import java.util.regex.Pattern;  import org.jsoup.Jsoup;  import org.jsoup.nodes.Document;  import org.jsoup.nodes.Element;  import org.jsoup.select.Elements;  public class Example09 {    public static void main(String[] args) throws IOException {  String listUrl = "http://www.skhu.ac.kr/board/boardlist.aspx?curpage=1&bsid=10004&searchBun=51";  Document document = Jsoup.connect(listUrl).get();  Elements tagList = document.select("table.board\_list tbody tr");  String regex = "boardread.aspx\\?idx=([0-9]+)&";  Pattern pattern = Pattern.compile(regex);    for (Element trTag : tagList ) {  Element aTag = trTag.select("td span a").get(0);  Matcher matcher = pattern.matcher(aTag.attr("href"));  String idx = matcher.find() ? matcher.group(1) : null;  String title = aTag.text();  String author = trTag.select("td:nth-child(4)").text();  String date = trTag.select("td:nth-child(5)").text();  String count = trTag.select("td:nth-child(6)").text();    System.out.printf("%s %s %s %s %s\n", idx, title, author, date, count);  }  }  } |

String listUrl = "http://www.skhu.ac.kr/board/boardlist.aspx?curpage=1&bsid=10004&searchBun=51";

게시글 목록 페이지 URL 이다.

Document document = Jsoup.connect(listUrl).get();

게시글 목록 페이지 내용을 웹서버에서 받아온다

Elements tagList = document.select("table.board\_list tbody tr");

document 내용 중에서

<table class="board\_list"> 태그 아래의

<tbody> 태그 아래의 <tr> 태그 목록을 리턴한다.

<thead> 태그 이래의 <tr> 태그는 포함되지 않는다.

<table class="board\_list"> 태그가 아닌 다른 <table> 태그 아래의 <tr> 태그도 포함되지 않는다.

for (Element trTag : tagList ) {

tagList 목록에 들어있는 tr 태그를 하나씩 trTag 변수로 꺼내서 for 루프를 반복한다.

Element aTag = trTag.select("td span a").get(0);

trTag가 참조하는 tr 태그의 아래에서

td 태그 아래의 span 태그 아래의 a 태그를 구한다.

trTag 객체는 Element 클래스 객체이다.

Element 클래스의 select 메소드는 Elements 객체를 리턴한다.

Element 클래스의 select 메소드는 자기 자신의 아래에서 태그들을 찾아서 리턴한다.

Document 클래스의 select 메소드는 문서 전체에서 태그들을 찾아서 리턴한다.

String author = trTag.select("td:nth-child(4)").text();

trTag가 참조하는 <tr> 태그 아래에서 4 번째 자식인 <td> 태그를 찾아서 태그 사이 내용을 리턴한다.

String date = trTag.select("td:nth-child(5)").text();

trTag가 참조하는 <tr> 태그 아래에서 5 번째 자식인 <td> 태그를 찾아서 태그 사이 내용을 리턴한다.

"태그:nth-child(번호)"

이 형태의 CSS 실렉터는 **번호** 번째의 자식인 **태그**를 찾는다.

게시글 목록 페이지의 게시글 목록 부분 HTML 태그는 다음과 같다.

|  |
| --- |
| <table summary="...생략..." class="board\_list">  ... 생략 ...  <tbody>  <tr>  <td>공지</td>  <td class="left15">  <span id="ctl00\_ContentPlaceHolder1\_ctl00\_rptList2\_ctl00\_lblTitle">  <a href='boardread.aspx?idx=27461&amp;curpage=1&amp;bsid=10004&amp;searchBun=51'>  [교무처] 8학기 이상 재학생 중 취업자에 대한 공...  </a>  </span></td>  <td>... 생략 ...</td>  <td>교무처(교무팀...</td>  <td>2016-11-09</td>  <td>2056</td>  </tr>  <tr>  <td><span id="ctl00\_ContentPlaceHolder1\_ctl00\_rptList\_ctl00\_lblNum">647</span></td>  <td class="left15">  <span id="ctl00\_ContentPlaceHolder1\_ctl00\_rptList\_ctl00\_lblTitle">  <a href='boardread.aspx?idx=27612&amp;curpage=1&amp;bsid=10004&amp;searchBun=51'>  [교무처] 2016년 겨울계절학기 폐강 안내  </a>  </span>  </td>  <td></td>  <td>교무처(교무팀...</td>  <td>2016-12-02</td>  <td>465</td>  </tr>  <tr>  <td><span id="ctl00\_ContentPlaceHolder1\_ctl00\_rptList\_ctl01\_lblNum">646</span></td>  <td class="left15">  <span id="ctl00\_ContentPlaceHolder1\_ctl00\_rptList\_ctl01\_lblTitle">  <a href='boardread.aspx?idx=27606&amp;curpage=1&amp;bsid=10004&amp;searchBun=51'>  [교무처]2016년 겨울계절학기 폐강 위험강좌 안내...  </a>  </span>  </td>  <td></td>  <td>교무처(교무팀...</td>  <td>2016-12-02</td>  <td>282</td>  </tr>  ... 생략 ...  </tbody>  </table> |

<tr> 태그 아래에 6 개의 <td> 태그가 있다.

제목은 2 번째 <td> 태그이다.

작성자는 4 번째, 작성일은 5 번째, 조횟수는 6 번째 <td> 태그이다.

실행결과 출력

|  |
| --- |
| 27461 [교무처] 8학기 이상 재학생 중 취업자에 대한 공... 교무처(교무팀... 2016-11-09 5765  27677 [졸업 자격증 제출] 영어, 일어, 중어, 경영 졸... 교무처(학과지... 2016-12-12 87  27673 [중어중국학과] 2016-2학기 졸업시험 시행 안내 교무처(학과지... 2016-12-09 131  27612 [교무처] 2016년 겨울계절학기 폐강 안내 교무처(교무팀... 2016-12-02 615  27606 [교무처]2016년 겨울계절학기 폐강 위험강좌 안내... 교무처(교무팀... 2016-12-02 401  27599 [사회복지학과] 재시험 일정 공지 (필독) 교무처(학과지... 2016-12-01 265  27590 [사회복지학과] 졸업시험 결과 확인 교무처(학과지... 2016-11-30 292  27536 [교무처] 군복무기간 중 군e-러닝 취득학점 인정안... 교무처(교무팀... 2016-11-18 383  27516 [사회복지학과] 졸업시험 장소 공지 교무처(학과지... 2016-11-17 393  27496 (수정-시간연장)학점포기 일정 안내(2016학년도 ... 교무처(교무팀... 2016-11-15 655  27462 2017학년도 1학기 재입학생 모집 공고 교무처(교무팀... 2016-11-09 834  27457 [교무처] 2016학년도 겨울계절학기 수강신청 안내... 교무처(교무팀... 2016-11-08 1457  27404 타 대학 학점교류는 학점교류 게시판을 참고하세요 교무처(교무팀... 2016-10-31 671  27228 [사회복지학과] 사회복지학과 졸업시험 (필독) 교무처(학과지... 2016-10-05 887  27218 [학부]2016-2학기 '학기 중 강의 의견 수렴'... 교무처(교무팀... 2016-10-04 613  27066 [학부] 2016학년도 2학기 수강과목 철회 안내 교무처(교무팀... 2016-09-12 1165 |

## Example10

Example10.java

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58 | package jsoup1;  import java.io.IOException;  import java.util.regex.Matcher;  import java.util.regex.Pattern;  import org.jsoup.Jsoup;  import org.jsoup.nodes.Document;  import org.jsoup.nodes.Element;  import org.jsoup.select.Elements;  public class Example10 {  static final String listUrl =  "http://www.skhu.ac.kr/board/boardlist.aspx?curpage=9999&bsid=10004&searchBun=51";    public static void main(String[] args) throws IOException {  String regex = "boardread.aspx\\?idx=([0-9]+)&";  Pattern pattern = Pattern.compile(regex);  // 마지막 페이지 번호를 구한다.  int pageCount = getPageCount();    // 1 페이지 부터 마지막 페이지까지  for (int i=1; i <= pageCount; ++i) {    // i번 페이지의 내용을 서버에서 가져온다.  String url = listUrl.replace("9999", String.valueOf(i));  Document document = Jsoup.connect(url).get();  Elements tagList = document.select("table.board\_list tbody tr");    // 현재 페이지의 게시글 목록의 내용을 출력한다.  for (Element trTag : tagList ) {  Element aTag = trTag.select("td span a").get(0);  Matcher matcher = pattern.matcher(aTag.attr("href"));  String idx = matcher.find() ? matcher.group(1) : null;  String title = aTag.text();  String author = trTag.select("td:nth-child(4)").text();  String date = trTag.select("td:nth-child(5)").text();  String count = trTag.select("td:nth-child(6)").text();    System.out.printf("%s %s %s %s %s\n", idx, title, author, date, count);  }  }  }    static int getPageCount() throws IOException {  String url = listUrl.replace("9999", "1");  Document document = Jsoup.connect(url).get();  Element aTag = document.select("a.nextL").get(0);  String regex = "boardlist.aspx\\?curpage=([0-9]+)&";  Pattern pattern = Pattern.compile(regex);  Matcher matcher = pattern.matcher(aTag.attr("href"));  String s = matcher.find() ? matcher.group(1) : "0";  int pageCount = Integer.valueOf(s);  return pageCount;  }  } |

"공지" 게시글은 모든 게시글 목록 페이지에 반복해서 나타난다.

게시글 목록 페이지의 내용을 가져올 때 마다 "공지" 게시글이 포함되어 있다.

Example10 코드가 출력하는 게시글 목록에 "공지" 게시글이 여러 번 등장한다.

## Example11

Example11.java

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99  100  101 | package jsoup1;  import java.io.IOException;  import java.util.Comparator;  import java.util.TreeSet;  import java.util.regex.Matcher;  import java.util.regex.Pattern;  import org.jsoup.Jsoup;  import org.jsoup.nodes.Document;  import org.jsoup.nodes.Element;  import org.jsoup.select.Elements;  public class Example11 {    static final String listUrl =  "http://www.skhu.ac.kr/board/boardlist.aspx?curpage=999&bsid=10004&searchBun=51";  static final String articleUrl =  "http://www.skhu.ac.kr/board/boardread.aspx?idx=999&bsid=10004&searchBun=51";  static class ArticleComparator implements Comparator<Article> {  @Override  public int compare(Article a1, Article a2) {  return a1.getIdx() - a2.getIdx();  }  }  public static void main(String[] args) throws IOException {  Article[] articles = getArticles();  for (Article article : articles) {  int length = Math.min(30, article.getBody().length());  System.out.printf("%d %s %s %s %s %s\n",  article.getIdx(),  article.getTitle(),  article.getAuthor(),  article.getDate(),  article.getCount(),  article.getBody().substring(0, length));  }  }  // 게시판의 전체 게시글 목록을 리턴한다. (본문 포함)  static Article[] getArticles() throws IOException {  String regex = "boardread.aspx\\?idx=([0-9]+)&";  Pattern pattern = Pattern.compile(regex);  TreeSet<Article> set = new TreeSet<Article>(new ArticleComparator());  int pageCount = getPageCount();    for (int i=1; i <= pageCount; ++i) {  System.out.printf("%d page\n", i);    String url = listUrl.replace("999", String.valueOf(i));  Document document = Jsoup.connect(url).get();  Elements tagList = document.select("table.board\_list tbody tr");    for (Element trTag : tagList ) {  Element aTag = trTag.select("td span a").get(0);  Matcher matcher = pattern.matcher(aTag.attr("href"));  String s = matcher.find() ? matcher.group(1) : null;  int idx = Integer.valueOf(s);  String title = aTag.text();  String author = trTag.select("td:nth-child(4)").text();  String date = trTag.select("td:nth-child(5)").text();  String count = trTag.select("td:nth-child(6)").text();  String body = getBody(idx);    Article article = new Article();  article.setIdx(idx);  article.setTitle(title);  article.setAuthor(author);  article.setDate(date);  article.setCount(count);  article.setBody(body);  set.add(article);  }  }  return set.toArray(new Article[0]);  }    // 마지막 페이지 번호를 리턴한다  static int getPageCount() throws IOException {  String url = listUrl.replace("999", "1");  Document document = Jsoup.connect(url).get();  Element aTag = document.select("a.nextL").get(0);  String regex = "boardlist.aspx\\?curpage=([0-9]+)&";  Pattern pattern = Pattern.compile(regex);  Matcher matcher = pattern.matcher(aTag.attr("href"));  String s = matcher.find() ? matcher.group(1) : "0";  int pageCount = Integer.valueOf(s);  return pageCount;  }    // 게시글 아이디가 idx인 게시글 본문을 리턴한다  static String getBody(int idx) throws IOException {  String url = articleUrl.replace("999", String.valueOf(idx));  Document document = Jsoup.connect(url).get();  String body = document.select("span#ctl00\_ContentPlaceHolder1\_ctl00\_lbContents").get(0).html();  return body;  }  } |

줄24의 getArticle 메소드가 리턴하는 전체 게시글 목록에

"공지" 게시글이 중복되어 들어있지 않도록 구현했다.

TreeSet<Article> set = new TreeSet<Article>(new ArticleComparator());

TreeSet 컬렉션 객체에 게시글(Article)을 add 하면

먼저 그 게시글과 동일한 게시글이 this 컬렉션에 들어있는지 확인한다.

이미 들어있다면 중복해서 add 하지 않는다.

동일한 게시글인지 비교할 때 ArticleComparator 객체의 compare 메소드를 호출하여 비교한다.

객체를 찾을 때, ArrayList 클래스의 contains 메소드는 선형 탐색(linear search)을 한다.

성능 복잡도는 O(N) 이다.

객체를 찾을 때 TreeSet 클래스 이진 탐색(binary search)을 한다.

성능 복잡도는 O(logN) 이다.