

EE4216 Lab 2 – Web Scraping

You are required to submit the Maven project that contains the solution for Checkpoint 1 to 3, and a lab report to explain the design of your programs and results of each checkpoint with supplementary information such as screenshots and source code as appropriate.

Objectives:

- Learn to scrape data from website using Java.

Modern web applications are data-driven. Data published with standardized formats such as JSON and XML can be parsed and understood by machines easily. However, information published on websites has no common structure and is difficult for computers to process. Such information has tremendous hidden value and remains as the biggest portion of the internet.

Jsoup is a Java based library to work with HTML based content. It provides a very convenient API to extract and manipulate data, using the best of DOM, CSS, and jQuery-like methods. It implements the HTML5 specification, and parses HTML to the same DOM as modern browsers do. This lab will take you through simple and practical methods available in Jsoup library.

- Jsoup Documentation - <https://jsoup.org/cookbook/>

Task 1 – Installing Jsoup Library using Maven

Apache Maven is a build automation tool for Java project. It can download a project's dependency, manage its build process, and generate documentation and report using a single configuration file called Project Object Model (pom.xml).

1. Jsoup runs on Java 7 and up. Start Netbeans and create a Maven project for “Java Application”. Fill in the project name and group ID as follows:

New Java Application

Steps

1. Choose Project
2. Name and Location

Name and Location

Project Name: HelloSoup

Project Location: e:/netbeans/ee4216/2019208/lab2_web_scraping

Project Folder: /ee4216/2019208/lab2_web_scraping/HelloSoup

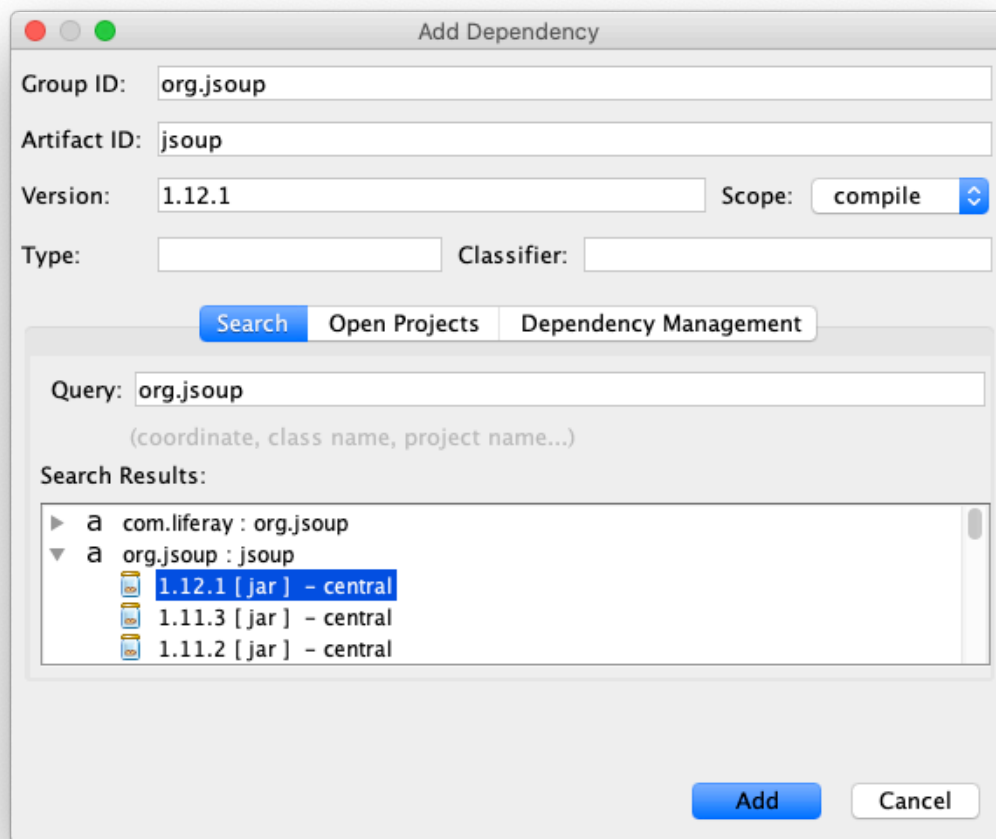
Artifact Id: HelloSoup

Group Id: ee4216.lab2

Version: 1.0-SNAPSHOT

Package: ee4216.lab2.hellosoup (Optional)

2. Create a main class to print out “Hello World!” for testing. Maven will download some files for its own building system when you run your application in the first time.
3. Dependencies can be added to Maven project by editing the pom.xml directly or searching the artifact from the ‘Add Dependency’ window in Netbeans. Let’s enter “org.jsoup” in the Query field and select the latest version (1.12.1) of Jsoup library. We will use it to parse some data from a website in the next task.



Task 2 – Parsing Data from a Website

1. Like a smart browser, Jsoup will make every attempt to create a clean parse from the HTML you provide, regardless of whether the HTML is well-formed or not. The `connect(String url)` method below is used to create a new `Connection` object, and its `get()` method is used to fetch and parses a HTML file using HTTP GET. If an error occurs whilst fetching the URL, it will throw an `IOException`, which you should handle appropriately. The code below prints the page title and HTML source code of CityU's people directory.

```
25 |  
26 |  
27 | Document doc = Jsoup  
28 |     .connect("https://www.cityu.edu.hk/directories/people/academic")  
29 |     .get();  
30 | System.out.println(doc.title());  
31 | System.out.println(doc);
```

2. In Jsoup, a *Document* consists of *Elements*, *TextNodes* and a couple of other miscellaneous nodes such as *Comments* and *Attributes*. Similarly, an *Element* contains a list of children *Nodes*, and has one parent *Element*, which represents a tree structure of the DOM.

Jsoup elements support a CSS like selector syntax to find matching elements, that allows very powerful and robust queries. The following query can select all the English names from the people directory.

```
25 |  
26 | Document doc = Jsoup  
27 |     .connect("https://www.cityu.edu.hk/directories/people/academic")  
28 |     .get();  
29 | Elements names = doc.select("div.name > div.en");  
30 | for(Element name : names) {  
31 |     System.out.println(name.text());  
32 | }  
33 |
```

Checkpoint 1:

Write a program to print out the English name and email address of the staff in the people directory (only the first page), one staff per each line. Note that some staff may have no email address. Explain your code briefly.

```
--- exec-maven-plugin:1.2.1:exec (default-cli) @ HelloSoup ---  
Dr. ABDRABOU Amgad Abdelmohsen Mohammed, mabdrabou2@cityu.edu.hk  
Dr. ABOLUDE Akintayo Temiloluwa, atabolude2@cityu.edu.hk  
Dr. AHN So-Yeon, soyahn@cityu.edu.hk  
Dr. AHN Yoojung, yoojung.ahn@cityu.edu.hk  
Mr. AHUJA Navin GopalDas,  
Ms. AKTER Mastura,  
Mr. ALAM Md Kowsar, mkalam3@cityu.edu.hk  
Dr. ALI Zulfiqar, zulfiqali3@cityu.edu.hk  
Dr. Azizjon ALIMOV, aaalimov@cityu.edu.hk  
Prof. ALLEN Richard William, rwallen@cityu.edu.hk  
Dr. AN Kyoung Jin Alicia, alicia.kjan@cityu.edu.hk  
Mr. AN Pengda, pengdaan2@cityu.edu.hk  
Miss AN Yao,  
Mr. ASIM Muhammad,
```

3. When data are split into multiple pages, it is necessary to follow the pagination links and iterate over the pages. The example code below demonstrates how to print out the link of the next page button. If you want to get an absolute URL, there is an attribute key prefix **abs:** that will cause the attribute value to be resolved against the document's base URI. This is useful when the link you obtained from the DOM is a relative URL.



Checkpoint 2:

Write another program to print out the name and email of all staff in the people directory by iterating over all the pages sequentially. Show the number of people printed at the end and explain your code briefly.

```
Dr. ZHU Zonglong, zonglzhucityu.edu.hk
Dr. ZHUANG Li,
Dr. ZHUANG Xiaosheng, xzhuang7cityu.edu.hk
Dr. ZIA Abdul Wasy, awzia3cityu.edu.hk
Dr. ZOU Heng, hengzou2cityu.edu.hk
Prof. ZUKERMAN Moshe, m.zucityu.edu.hk
Dr. ZUO Zhiya, zhiyazuocityu.edu.hk
Dr. Inez ZWETSL00T, i.m.zwetsl00tcityu.edu.hk
Total number of people: 1650
```

BUILD SUCCESS

Total time: 19.412 s
Finished at: 2020-10-11T00:12:35+08:00

Checkpoint 3:

The performance bottleneck of this web scraper is on the connection time needed for fetching each web document page by page orderly. Redesign your scraper with concurrency to speed up the process. Demonstrate your new program and explain your design.

```
Miss ZHANG Yan, yzhang2324cityu.edu.hk
Miss ZHANG Yizhen,
Mr. ZHANG Yongxiang,
Prof. ZHANG Yuanting, yt.zhangcityu.edu.hk
Dr. ZHANG Yue, yzhang887cityu.edu.hk
Miss ZHANG Yue,
Mr. ZHANG Yue, yzhang742cityu.edu.hk
Miss ZHANG Yuxuan,
Dr. ZHANG Zhedong, zzhan26cityu.edu.hk
Total number of people: 1650
```

BUILD SUCCESS

Total time: 3.462 s
Finished at: 2020-10-11T00:20:05+08:00

END OF THE LAB