Crawling SIX ITB with jsoup: Java HTML Parser

Naufal Malik Rabbani - 13514052
Computer Science/Informatics Program
Bandung Institute of Technology (ITB)
Ganeca Street 10th, Bandung City 40132, Indonesia
13514052@std.stei.itb.ac.id

***Abstract*—As students who are active in various activities/organization other than academic, attention should be paid to the colleagues of their academic. One way is to crawling SIX ITB website. From there we can think if there are colleagues who have problems in the academic (needs advocacy), and we can know the schedule of their lectures to easier determine the time for organization gather.**

Keywords—crawling, SIX ITB

I. INTRODUCTION (*HEADING 1*)

Students holding positions in an organization, such as a division chairman or department chairman, are important to know the commitment of each member. Since the student's primary objective is at his academic level, it is advisable for the chairman to see the status of the academic busyness of his members.

Figure 1 SIX ITB website interface

During this time the division chiefs usually ask members of their members' lectures manually. By doing SIX ITB crawling, the chairman can immediately know the activity of its members. That way, if the chairman wants to hold a gathering division, then can look for the most appropriate time.

There are IF 2012 students who have already done SIX ITB crawling but only take name and NIM information only, not along with the schedule of the lecture.

Because the SIX ITB website is a simple website, the crawling process can be done by sending a simple HTTP request and then parse the content in HTTP response.

This discussion is about how to send and receive HTTP connection to SIX ITB, then how to get the content using jsoup: Java HTML Parser.

II. RELATED WORKS

A. *Review of HTTP Connection*

HTTP is based on the client-server architecture model and a stateless request/response protocol that operates by exchanging messages across a reliable TCP/IP connection.

An HTTP "client" is a program (Web browser or any other client) that establishes a connection to a server for the purpose of sending one or more HTTP request messages. An HTTP "server" is a program (generally a web server like Apache Web Server or Internet Information Services IIS, etc.) that accepts connections in order to serve HTTP requests by sending HTTP response messages.

HTTP makes use of the Uniform Resource Identifier (URI) to identify a given resource and to establish a connection. Once the connection is established, **HTTP messages** are passed in a format similar to that used by the Internet mail [RFC5322] and the Multipurpose Internet Mail Extensions (MIME) [RFC2045]. These messages include **requests** from client to server and **responses** from server to client which will have the following format:

HTTP requests and HTTP responses use a generic message format of RFC 822 for transferring the required data. This generic message format consists of the following four items:

- A start-line
- Zero or more header fields followed by CRLF
- An empty line (i.e., a line with nothing preceeding the CRLF) indicating the end of the header fields
- Optionally a message-body

The request **method** indicates the method to be performed on the resource identified by the given **Request-URI**. The method is case-sensitive and should always be mentioned in uppercase.

- The GET method is used to retrieve information from the given server using a given URI. Requests using GET should only retrieve data and should have no other effect on the data.
- A POST request is used to send data to the server, for example, customer information, file upload, etc. using HTML forms.

The Request-URI is a Uniform Resource Identifier and identifies the resource upon which to apply the request. The most common form of Request-URI is that used to identify a resource on an origin server or gateway.

For example, a client wishing to retrieve a resource directly from the origin server would create a TCP connection to port 80 of the host "www.w3.org" and send the following lines:

```
GET /pub/WWW/TheProject.html HTTP/1.1
Host: www.w3.org
```

Examples of Response Message:

```
HTTP/1.1 200 OK
Date: Mon, 27 Jul 2009 12:28:53 GMT
Server: Apache/2.2.14 (Win32)
Last-Modified: Wed, 22 Jul 2009 19:15:56
GMT
Content-Length: 88
Content-Type: text/html
Connection: Closed
<html>
<body>
<h1>Hello, World!</h1>
</body>
</html>
```

III. METHODS

Crawling here uses the Java language with an additional library that is jsoup: Java HTML Parser. This jsoup will be implemented with the library to send HTTP requests i.e. java.net and java.io.

Make sure the device is connected to the local ITB network or using the ITB VPN. After successfully sending a request and getting a response, parse the HTML using jsoup and save the content.

Repeat HTTP request and response process for Daftar Peserta Kelas and Jadwal Kuliah of all faculties and study programs. Then if you want, crawl also for semester, year, and different curriculum by replacing URI.

IV. EXPERIMENT RESULTS

The simple HTTP request response method is able to retrieve the data contained in SIX ITB. The total time required to retrieve all data on the Jadwal Kuliah and Daftar Peserta Kelas is as follows:

Table 1 Total time for crawling

With VPN ITB	With local connection/wi-fi
20 – 30 minutes	3 – 5 minutes

From these results it can be seen that HTTP requests run much faster if the distance of the destination server is not far away (both are in ITB, no need to pass the proxy provider).

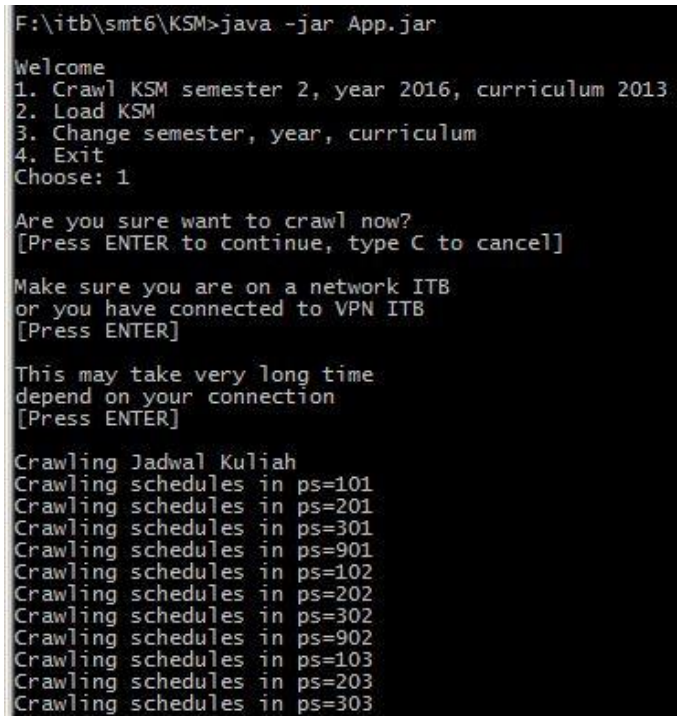

Figure 2 Crawling process

After the data successfully crawled, we can process the data to get more information, such as:

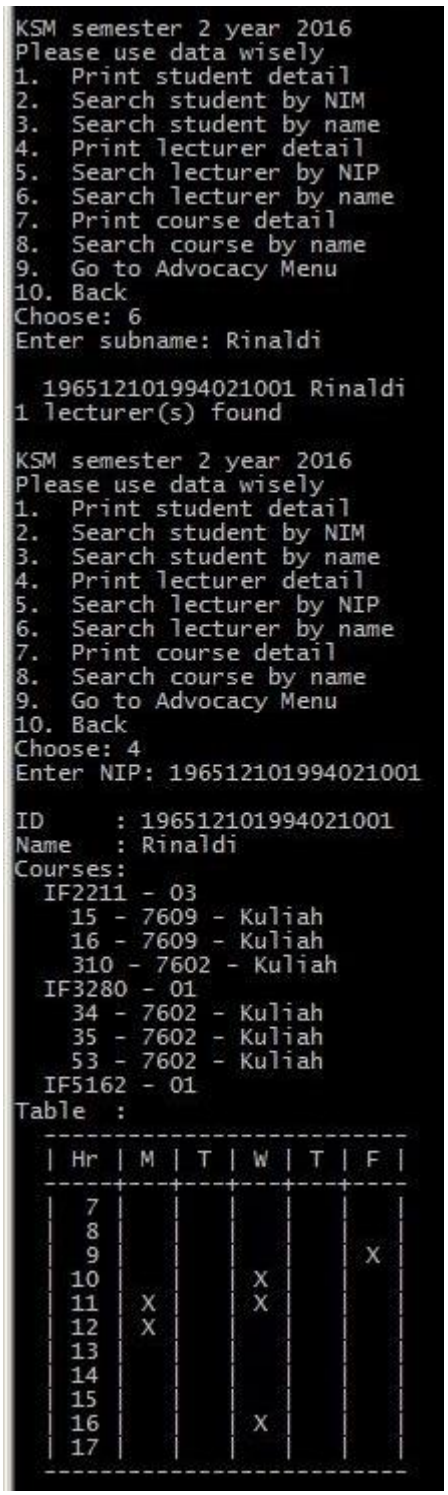

Figure 3 Search and display info about a lecturer

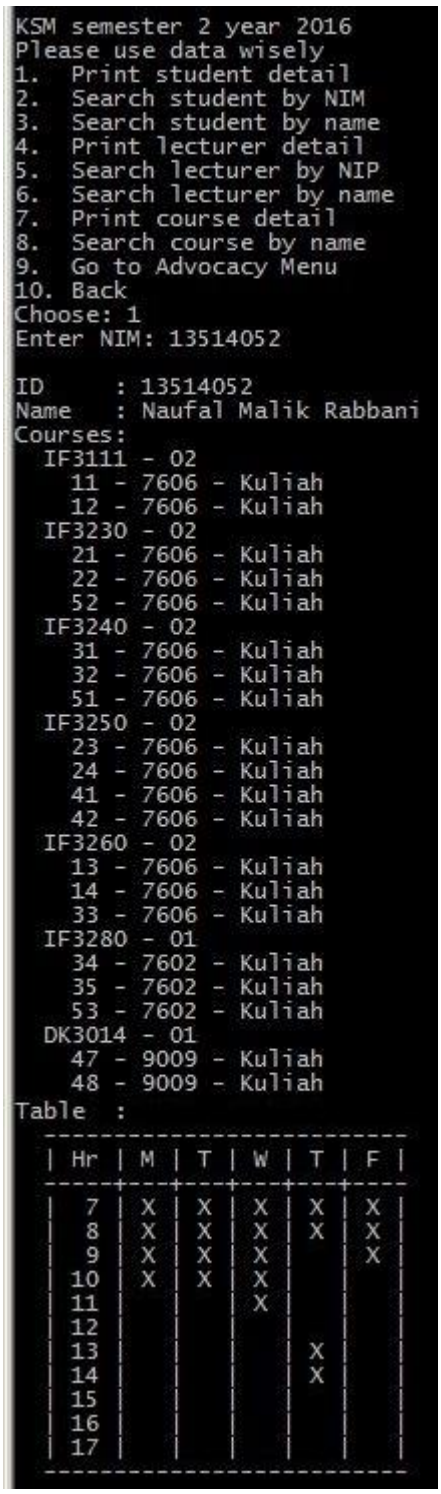

Figure 4 Search and display info about a student

```
KSM semester 2 year 2016
Please use data wisely
1.  Print student detail
2.  Search student by NIM
3.  Search student by name
4.  Print lecturer detail
5.  Search lecturer by NIP
6.  Search lecturer by name
7.  Print course detail
8.  Search course by name
9.  Go to Advocacy Menu
10. Back
Choose: 7
Enter course ID   : IF3280
Enter class number: 01

ID       : IF3280
Name     : Socio-informatika dan Profesionalisme
Class    : 01
SKS      : 3
Schedule:
  34 - 7602 - Kuliah
  35 - 7602 - Kuliah
  53 - 7602 - Kuliah
Lecturer:
  196512101994021001 Rinaldi
Students:
  13513043 Agung Baptiso Sorlawan
  13514001 Joshua Salimin
  13514004 Catherine Pricilla
  13514007 Sri Umay Nur'aini Sholihah
  13514010 Febi Agil Ifdillah
  13514013 Anwar Ramadha
  13514016 Alif Bhaskoro
  13514019 Wiega Sonora
  13514022 Taufic Leonardo Sutejo
  13514025 Ratnadira Widyasari
  13514028 Dharma Kurnia Septialoka
  13514031 Andri Hardono Hutama
  13514034 Evita Chandra
  13514037 Cendhika Imantoro
  13514040 Devin Lukianto
  13514043 Vitra Chandra
  13514046 Albert Logianto
  13514049 Ade Surya Ramadhani
  13514052 Naufal Malik Rabbani
  13514055 Yeksadiningrat Av
  13514058 Jason Jeremy Iman
  13514061 Robert Sebastian Herlim
  13514064 Kharis Isriyanto
  13514070 Dendy Suprihady
  13514073 Muhammad Naufal
  13514079 Ade Yusuf Rahardian
  13514082 Rio Chandra Rajagukguk
  13514085 Christian Anthony S
  13514088 Alfonsus Raditya Arsadjaja
  13514091 Adam Rotal Yuliandaru
  13514094 Kevin Supendi
  13514097 Stefanus Agus Haryono
  13514100 Albertus Kelvin
  13514103 Muhammad Reifiza
  13514106 Hasna Nur Karimah
  13514109 Resa Kemal Saharso
  13515601 Dandu Satyanuraga
37 student(s) found
```

Figure 5 Search and display info about a course/lecture

```
Advocacy Menu semester 2 year 2016
Please use data wisely
1. Search students who needs advocacy
2. Load another KSM
3. Print KSM(s) loaded
4. Back
Choose: 3

  Semester 1 year 2016
  Semester 2 year 2015
  Semester 1 year 2015
  Semester 2 year 2014
  Semester 1 year 2014
  Semester 2 year 2013
  Semester 1 year 2013

Advocacy Menu semester 2 year 2016
Please use data wisely
1. Search students who needs advocacy
2. Load another KSM
3. Print KSM(s) loaded
4. Back
Choose: 1
Enter 3 digits Major ID: 135

  Saving result to Advokasi-20162-135.csv
  Save success
```

Figure 6 Search students who needs advocacy

	A	B	C	D
1	Daftar Mahasiswa Advokasi Program Studi 135			
2	Semester 2 Tahun 2016 (80 mahasiswa)			
3	Matkul yang tercetak berkali-kali menandakan sudah mengulang berkali-kali			
4				
5	Nama	NIM	Kode Matkul	Nama Matkul
6	Thea Olivia	13511001	IF2210	Pemrograman Berorientasi Objek
7	Thea Olivia	13511001	IF2210	Pemrograman Berorientasi Objek
8	Thea Olivia	13511001	IF4091	Tugas Akhir I & Seminar
9	Tito D Kesumo Siregar	13511018	IF4091	Tugas Akhir I & Seminar
10	Tito D Kesumo Siregar	13511018	IF4091	Tugas Akhir I & Seminar
11	Tito D Kesumo Siregar	13511018	IF4091	Tugas Akhir I & Seminar
12	Tito D Kesumo Siregar	13511018	IF4091	Tugas Akhir I & Seminar
13	Lubis Sucipto	13511025	IF4091	Tugas Akhir I & Seminar
14	Lubis Sucipto	13511025	IF4091	Tugas Akhir I & Seminar
15	Lubis Sucipto	13511025	IF4091	Tugas Akhir I & Seminar
16	Lubis Sucipto	13511025	IF4091	Tugas Akhir I & Seminar
17	Lubis Sucipto	13511025	IF4091	Tugas Akhir I & Seminar
18	Akbar Juang Saputra	13511026	IF4040	Pemodelan Data Lanjut
19	Akbar Juang Saputra	13511026	IF4092	Tugas Akhir II
20	Akbar Juang Saputra	13511026	KU2071	Pancasila dan Kewarganegaraan
21	Isabella Julia Putri	13511033	IF3280	Socio-informatika dan Profesionalisme
22	Isabella Julia Putri	13511033	IF4040	Pemodelan Data Lanjut
23	Isabella Julia Putri	13511033	IF4091	Tugas Akhir I & Seminar
24	Isabella Julia Putri	13511033	IF4091	Tugas Akhir I & Seminar
25	Isabella Julia Putri	13511033	IF4091	Tugas Akhir I & Seminar

Figure 7 Result of Informatics students who needs advocacy

V. DISCUSSION

Crawling SIX ITB is very easy by simply sending a request, then parse HTML response using jsoup, then process the data as we want to get certain information.

But because it is too easy like this, very susceptible to our information can be taken by anyone, although in SIX ITB itself there is a menu to see KSM:

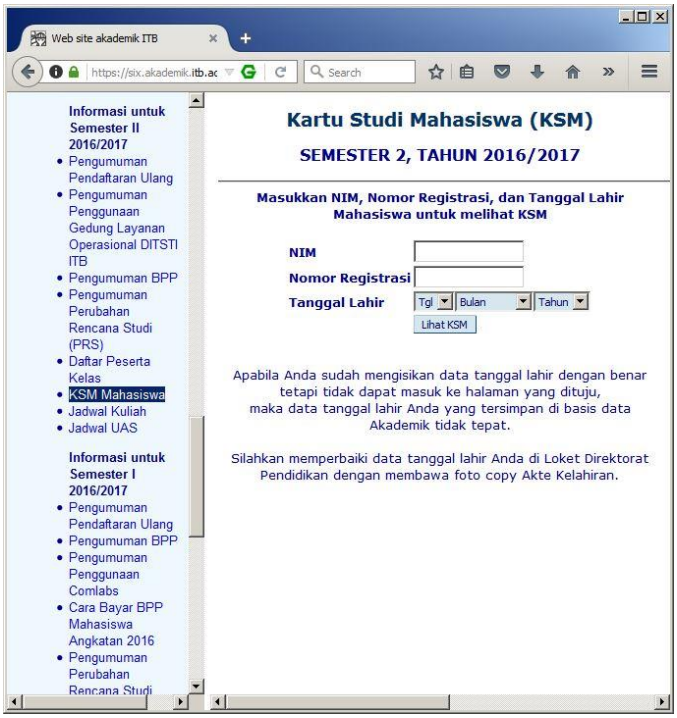

Figure 8 Menu to see KSM on SIX ITB website

It's as if the menu is useless for someone who has succeed to crawl SIX. With SIX crawling and processing the data, he/she can look exactly schedules in someone's KSM.

But it's okay, as long as this information is used well by the division / department chairman, hopefully it can be easier in managing colleagues in the organization or academic.

ACKNOWLEDGMENT

Alhamdulillaah, first of all I would like to thank both my parents, then thanks to IF 2004 who has supported my lecture fund, then thanks to Mr. Rinaldi, Mrs. Dessi, and Mrs. Ayu as lecturer of IF3280 course, then thank you Zaky's IF 2012 NIM Finder as my inspiration to do SIX ITB crawling, and to my friends who contribute indirectly for making this paper.

REFERENCES

[1] Web site akademik ITB: https://six.akademik.itb.ac.id/ accessed on May 5th 2017 3:47 PM

[2] jsoup Java HTML Parser: https://jsoup.org/ accessed on May 5th 2017 4:11 PM

[3] HTTP Tutorial: https://www.tutorialspoint.com/http/index.htm accessed on May 5th 2017 4:04 PM

[4] NIM Finder: https://azaky.github.io/nim-finder/ accessed on May 5th 2017 4:05 PM

[5] NIM Finder repository: https://github.com/azaky/nim-finder accessed on May 5th 2017 4:05 PM

DECLARATION

I hereby declare that the paper I am writing is my own, not an adaptation, or a translation of someone else's paper, and not plagiarism.

Bandung, May 5th 2017

Naufal Malik Rabbani - 13514052