**Data Scraping for Asahi and Yomiuri Newspaper Website**

This note summarizes the steps necessary to automatically obtain the number of search results for Japanese Uncertainty Index using Java. The note is organized as follows. Section I introduces the Asahi&Yomiuri newspaper database that students/faculty at the University of Chicago have through the University library website. Section II contains the automation programs written to obtain the number of hits based on the set of key terms for the Japanese Uncertainty Index. Appendix attaches the source code behind. If you already know about the Asahi/Yomiuri database, you can skip Section I and start directly from Section II.

**Section I: How to access Yomiuri and Asahi website.**

1. Go to the website <http://www.lib.uchicago.edu/e/index.html>

2. Select “Databases & Articles” under “SEARCH”

3. Type in “Yomidas Rekishikan”

4. Select “Yomidas Rekishikan (Yomiuri Shinbun)” (If you’re on campus, this will directly give you access. If not you need to type in your Cnet ID and Password)

5. Click on “ログイン”

6. You’ll be directed to [https://database.yomiuri.co.jpHYPERLINK "https://database.yomiuri.co.jp/rekishikan/"/HYPERLINK "https://database.yomiuri.co.jp/rekishikan/"rekishikan/](https://database.yomiuri.co.jp/rekishikan/)

7. Select “平成”.

8. Type in (不透明OR不確実) AND (経済) AND (政策 OR 税 OR 歳出 OR 公共事業費 OR 公共投資 OR 国費 OR 規制 OR 日本銀行 OR 連銀 OR 財政 OR 赤字 OR 負債)” and specify the Year, Month and Date and hit Search.

9. You will obtain the search results. Be careful that the dates of the last month have to have exact matches to obtain the results. (be careful with loop years).

When you change the set of words in the automation code, I do recommend that you first try manually and see if you get a result. If you are typing the search string in a wrong fashion, you will not obtain any result in automation system as well. **You have to be careful that you only use Japanese input mode for Japanese characters but nothing else (e.g. “(“ ,“OR”, “AND”, ”)”). Use English input for everything else.**

The Asahi Website is done in a similar manner with “Kikuzo II Visual for Libraries” in Step 3. And the set of key words used for the Japanese Uncertainty Index is (as of August 6, 2012) is “(不透明+不確実)&(経済)&(政策+税+歳出+公共事業費+公共投資+国費+規制+日本銀行+日銀+財政+赤字+負債)”.

The Asahi Website contains not only newspapers but magazines so pick only “朝日新聞” under “対象紙誌名”.

**Section II:**

Everything you need to run this code is:

* Java 7 (Notice that you need to run on PC not on Mac to run this code as the latest **Java 7 is not supported by Mac OS X**).
* 6 files contained in JPNautoCode.zip
* Asahi.ini
* Asahi.jar
* Yomiuri.ini
* Yomiuri.jar
* start.bat
* Structure.ini

The only two files that RAs need to know to run the codes are Asahi.ini, Yomiuri.ini. The following is the structure of the ini files.

**STEP 1: Download Java 7**

1.

Download Java 7 from the following website.

<http://www.oracle.com/technetwork/java/javase/downloads/index.html>

Or

<http://www.oracle.com/technetwork/java/javase/downloads/jdk7-downloads-1637583.html>

**STEP 2: Change the Asahi.ini and Yomiuri.ini files.**

The structure of the ini files are shown below. (This is from Structure.ini).

<UChicagoLoginName>

<UChicagoLoginPW>

<SearchString>

<Optional:PathToSaveOutputIn Example:/Ippei OR C:/Ippei>

Explanation:

<UChicagoLoginName> is your CNet Login name. (e.g. ishibata)

<UChicagoLoginPW> is your CNet Login Password.

<SearchString> is the set of the key words you use to search in Asahi/Yomiuri database. (e.g. (不透明 OR 不確実）　 AND (経済) AND (政策 OR 税 OR 歳出 OR 公共事業費 OR 公共投資 OR 国費 OR 規制 OR 日本銀行 OR 日銀 OR 財政 OR 赤字 OR 負債) for Yomiuri and (不透明+不確実)&(経済)&(政策+税+歳出+公共事業費+公共投資+国費+規制+日本銀行+日銀+財政+赤字+負債) for Asahi).

Finally, <Optional:PathToSaveOutputIn> specifies where to save the output txt files on your PC. Example:/Ippei OR C:/Ippei>. You need to specify this.

To summarize, the Yomiuri.ini and Asahi.ini should look as follows:

Examples:

Asahi.ini

ishibata

xxx

(不透明+不確実)&(経済)&(政策+税+歳出+公共事業費+公共投資+国費+規制+日本銀行+日銀 +財政+赤字+負債)

C:/Ippei

Yomiuri.ini

ishibata

password

(不透明 OR 不確実) AND (経済) AND (政策 OR 税 OR 歳出 OR 公共事業費 OR 公共投資 OR 国費 OR 規制 OR 日本銀行 OR 日銀 OR 財政 OR 赤字 OR 負債)

C:/Ippei

There are 4 things that RAs need to be careful with.

First, as mentioned in Section I, **MAKE SURE TO USE JAPANESE INPUT ONLY FOR THE SEARCH KEY WORDS but nothing else. Use the US input to type “(“ “ “ “)” “OR” “AND” “+” . They have to be in the ASCII format (check if the lines are in ASCII at** [**http://rishida.net/tools/conversion/**](http://rishida.net/tools/conversion/) **).**

Second, RAs need to change the format of ini file. Do not change the space or anything that is in the ini file. If you change them, you will get an error! This is very important!

Third**, this program must be run from OFF-CAMPUS! It will not work on-campus!**

Fourth**, Macintosh OS X does not support Java 7 as of August, 2012, you need to run on Windows PC.**

**STEP 3: Run the jar files (Asahi.jar and Yomiuri.jar) using the batch mode.**

start.bat file is a batch file that runs the Java codes Asahi.jar and Yomiuri.jar. The text file, start.bat, reads

java -jar Asahi.jar

java -jar Yomiuri.jar

Therefore, this program runs the Asahi.jar and Yomiuri.jar. Alternatively, you can run these files in “cmd” in windows. Open up "cmd" and just type in **java –jar Asahi.jar** etc.

To run this batch file, you just need to put all five files in the same folder and **double click start.bat.** You will obtain the two txt files, Asahi.txt and Yomiuri.txt, in your specified location (e.g. C:/Ippei). The two output text files contain the Year Month and the number of hits obtained as a result of searching for the set of words. The output files would look like follows:

….

20110401:20110431,29

20110501:20110531,24

20110601:20110631,35

20110701:20110731,31

….

Congratulations. You’re done!

**Appendix**

Asahi.jar and Yomiuri.jar files are basically the set of programming files, and all the detailed programming go way beyond the scope of what a beginner in Java can understand. The following is the java source code that is contained in .jar file.

import java.io.File;

import javax.net.ssl.TrustManager;

import javax.net.ssl.X509TrustManager;

import javax.net.ssl.HttpsURLConnection;

import javax.net.ssl.SSLContext;

import jsoup.nodes.Document;

import jsoup.select.Elements;

public class Asahi extends Utilities {

private static final String URL = "[http://proxy.uchicago.edu/login?url=http://database.asahi.com/library2e/";//https://shibboleth2.uchicago.edu/idp/Authn/UserPassword](http://proxy.uchicago.edu/login?url=http://database.asahi.com/library2e/%22;//https://shibboleth2.uchicago.edu/idp/Authn/UserPassword)

private static final String URL2 = "<https://shibboleth2.uchicago.edu/idp/profile/SAML2/POST/SSO>";

private static final String URL3 = "<https://shibboleth2.uchicago.edu/idp/Authn/UserPassword>";

private static final String URL4 = "<https://login.proxy.uchicago.edu/Shibboleth.sso/SAML2/POST>";

private static final String URL5 = "<https://database-asahi-com.proxy.uchicago.edu/library2e/login/login.php>";

private static final String URL6 = "<http://database.asahi.com.proxy.uchicago.edu/library2e/topic/t-list.php>";

private static String OUTPUT\_FOLDER = (isWindows7() ? "D:/Ippei" : (isWindowsXP() ? "C:/Dropbox/J-Java/Ippei" : "/Ippei"));

private static final int SLEEP\_INBETWEEN = 2000; // 2s

private static String USERNAME;

private static String PASSWORD;

private static String SEARCH;

private static boolean stop = true;

public static void main(String[] args) {

// Create a trust manager that does not validate certificate chains

TrustManager[] trustAllCerts = new TrustManager[]{

new X509TrustManager() {

public java.security.cert.X509Certificate[] getAcceptedIssuers() {

return null;

}

public void checkClientTrusted(

java.security.cert.X509Certificate[] certs, String authType) {

}

public void checkServerTrusted(

java.security.cert.X509Certificate[] certs, String authType) {

}

}

};

// Install the all-trusting trust manager

try {

SSLContext sc = SSLContext.getInstance("SSL");

sc.init(null, trustAllCerts, new java.security.SecureRandom());

HttpsURLConnection.setDefaultSSLSocketFactory(sc.getSocketFactory());

} catch (Exception e) {

}

String[] ini = IO.loadString("Asahi.ini").split("\r\n");

if (ini.length > 2) {

stop = false;

USERNAME = ini[0];

PASSWORD = ini[1];

SEARCH = ini[2];

}

if (ini.length > 3)

OUTPUT\_FOLDER = ini[3];

(new File(OUTPUT\_FOLDER)).mkdirs();

Document d;

Elements e;

d = HtmlReaderJsoupEdit.loadDocument(URL, false);//, HtmlReader.Method.POST, "j\_username", "ishibata", "j\_password", "XXX"

if (d != null && !stop) {

// print("%s", d.html());

e = d.select("input");

// System.setProperty("javax.net.ssl.trustStore", "uchicago.edu.jks");

d = HtmlReaderJsoupEdit.loadDocument(URL2, true, HtmlReaderJsoupEdit.Method.POST, "", "RelayState", e.get(0).attr("value"), "SAMLRequest", e.get(1).attr("value"));

if (d != null) {

// print("%s", d.html());

d = HtmlReaderJsoupEdit.loadDocument(URL3, true, HtmlReaderJsoupEdit.Method.POST, "", "j\_username", USERNAME, "j\_password", PASSWORD);

if (d != null) {

// print("%s", d.html());

e = d.select("input");

// System.setProperty("javax.net.ssl.trustStore", "proxy.uchicago.edu.jks");

d = HtmlReaderJsoupEdit.loadDocument(URL4, true, HtmlReaderJsoupEdit.Method.POST, "", "RelayState", e.get(0).attr("value"), "SAMLResponse", e.get(1).attr("value"));

if (d != null) {

// print("%s", d.html());

d = HtmlReaderJsoupEdit.loadDocument(URL5, true, HtmlReaderJsoupEdit.Method.POST, "", "RelayState", e.get(0).attr("value"), "SAMLResponse", e.get(1).attr("value"));

if (d != null) {

// print("%s", d.outputSettings().charset());

// print("%s", d.html());

e = d.select("input");

String loginSID = new String(e.get(0).attr("value"));

for (int y=1984; y<=2013; y++) // 1984

for (int m=1; m<=12; m++) {

String year = String.format("%04d", y);

String month = String.format("%02d", m);

String issueDate = year + month + "01:" + year + month + "31";

print("Looking at %s", issueDate);

d = HtmlReaderJsoupEdit.loadDocument(URL6, true, HtmlReaderJsoupEdit.Method.POST, "text/html;charset=euc-jp",

"rdoSrchMode", "0",

"chkShishi", "N",

"txtWord", SEARCH, // "(不透明+不確実)&(経済)&(政策+税+歳出+公共事業費+公共投資+国費+規制+日本銀行+日銀+財政+赤字+負債)"

"hdnItaiji", "1",

"rdoSrchTerm", "1",

"cmbIDFy", year,

"cmbIDFm", month,

"cmbIDFd", "01",

"cmbIDDm", "FROM",

"cmbIDTy", year,

"cmbIDTm", month,

"cmbIDTd", "31",

"rdoSrchItem", "M\_H",

"txtBunrui", "",

"txtMenmei", "",

"cmbDspNum", "10",

"rdoDspOrder", "NEW",

"loginSID", loginSID,

"srchId", "12",

"srchKind", "zent",

"selectNo", "",

"kiriId", "",

"cond", "MH&2&3,MidashiHonbun,1," + SEARCH + ",ShishiName,1,N,HakkouDate,5," + issueDate,

"ronri", "MH&2&3",

"hdnSrchMode", "0",

"hdnShishi", "N",

"hdnShishiOrg", "N",

"hdnWord", SEARCH,

"hdnKeyword", SEARCH,

"hdnSrchTerm", "1",

"hdnIssueDate", issueDate,

"hdnIDFy", year,

"hdnIDFm", month,

"hdnIDFd", "01",

"hdnIDDm", "FROM",

"hdnIDTy", year,

"hdnIDTm", month,

"hdnIDTd", "31",

"hdnSrchItem", "M\_H",

"hdnBunrui", "",

"hdnKan", "",

"hdnKanOrg", "",

"hdnMenmei", "",

"hdnHochi", "",

"hdnHochiOrg", "",

"hdnIssueS", "",

"hdnIssueSOrg", "",

"hdnDspNum", "10",

"hdnDspOrder", "NEW",

"hdnDispSoukensuOnlyFlag", "0",

"hdnSrchStart", "",

"hdnSrchEnd", "",

"hdnAsahi", "Y",

"hdnAera", "Y",

"hdnWeekly", "Y");

if (d != null) {

e = d.select("span.fontcolor001");

IO.saveString(OUTPUT\_FOLDER + "/results.txt", issueDate + "," + e.get(0).html() + "\r\n", true);

print("%s", e.get(0).html());

}

sleep(SLEEP\_INBETWEEN);

}

}

}

}

}

// d = HtmlReader.loadDocument("http://www.bankruptcydata.com/NGR2000SubscrWelcome.asp", true);

// print("%s", d.html());

}

}

}