

Data Analysis with Java

데이터 분석 프로그래밍 01

Objective of Course

Handling Big data with Java

- Inspecting, cleansing, transforming, modeling data

Discovering Useful Information

- Using methodologies to analyze big data
- Discovering useful and meaningful information from large quantity of data

The Ethics of Data Scraping

- Studying data scraping ethics

The final result of this course is ...

to build up a server which gets and stores the web data using DBMS

Objective of Today's Class

Distinguishing Data and Information

- ▶ Difference between data and information
- ▶ Types of data
- ▶ Types of information

The Ethics of Data Scraping

- ▶ Studying data scraping ethics

Data Crawling

- ▶ Collect structured data with openCSV and Apache POI

Data and Information

Data

- ▶ Raw and unorganized facts
- ▶ e.g. each student's age

Information

- ▶ Organized, structured or presented set of data
- ▶ e.g. the average age of a class

Q1 : Let's share examples of data and information

Types of Data

Structured Data

- ▶ Data stored in fixed fields or columns such as RDBMS
- ▶ e.g. name, age, gender and etc.

Unstructured Data

- ▶ Data which can't be analyzed or categorized as it is
- ▶ e.g. text, video and audio files

**Q2 : Get some structured data on the web
and point out unstructured data**

Types of Information

Understanding Information Sources

- There are four types of information


Type	Description	Example
Factual	Information that deals with facts	Government Resources, Encyclopedias
Analytical	Interpretation of factual information	Library databases, Academic books
Subjective	Information from only one point of view	Websites, Blogs, Social media
Objective	Information that is understood from multiple view points	Books, Journal articles

Crawling Structured Data(Cont'd)

Download openCSV

► <https://sourceforge.net/projects/opencsv/>

Home / Browse / Development / Data Formats / opencsv




opencsv

Brought to you by: [aruckerjones](#), [sconway](#)

★★★★★ 42 Reviews

Downloads: 493 This Week

Last Update: 2021-04-22

 Download

Get Updates

Share This



Summary	Files	Reviews	Support	Wiki	Tickets ▾	News	Git ▾
---------	-------	---------	---------	------	-----------	------	-------


A Simple CSV Parser for Java under a commercial-friendly Apache 2.0 license

Features


- CSV Parsing

Project Activity



 [Scott Conway](#) posted [a comment](#) on [ticket #92](#) 1 week ago

You can find some more information about the CSVBind annotations at <http://opencsv.sourceforge.net/#annotations> and in the unit tests.

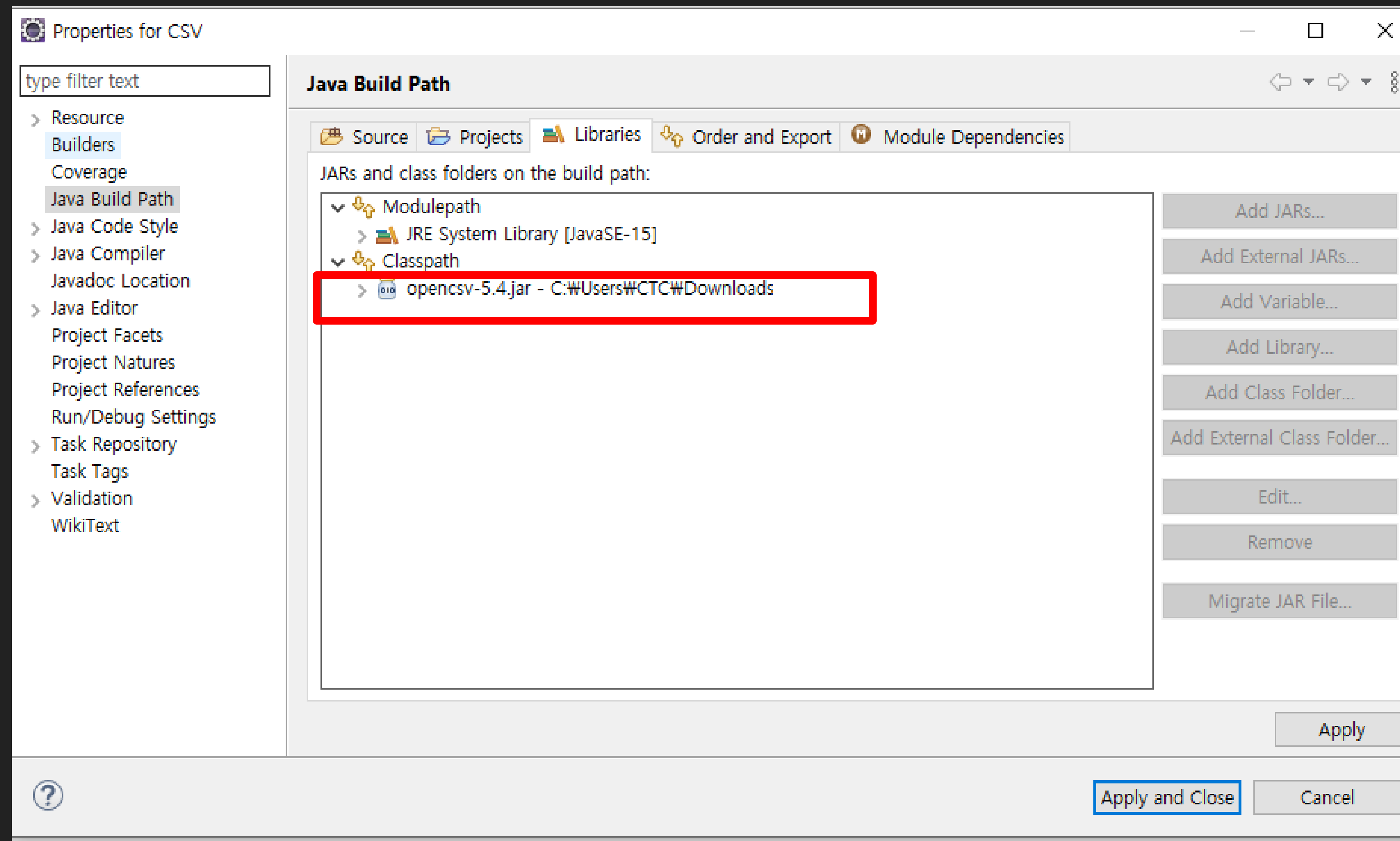
 [Andrew Rucker Jones](#) posted [a comment](#) on [ticket #92](#) 1 week ago

You were never supposed to be using those classes. I can't imagine how you would. Simply annotate your beans with the CsvBind* annotations. If you need more speci...

Crawling Structured Data(Cont'd)

Add the library to the project

- Opencsv.jar



Crawling Structured Data(Cont'd)

READ and WRITE CSV files

- Compose the code below

```
19 public static void main(String[] args) {  
20     String readFileName = "C:/Users/CTC/Desktop/rawData3.csv";  
21     String writeFileName = "C:/Users/CTC/Desktop/rawData4.csv";  
22  
23     CSVReader csvReader;  
24     try {  
25         csvReader = new CSVReader(new InputStreamReader(new FileInputStream(readFileName), "CP949"));  
26         String[] nextLine;  
27         while ((nextLine = csvReader.readNext()) != null) {  
28             System.out.println(nextLine.length + " : " + String.join("|", nextLine));  
29         }  
30     } catch (FileNotFoundException e) {  
31         e.printStackTrace();  
32     } catch (Exception e) {  
33         e.printStackTrace();  
34     }  
35  
36     try {  
37         CSVWriter cw = new CSVWriter(new FileWriter(writeFileName));  
38         String[] data = {"abc", "def", "ghi"};  
39         cw.writeNext(data);  
40         cw.close();  
41     } catch (IOException e) {  
42         e.printStackTrace();  
43     }  
44 }  
45 }
```

INPUT

WRITE

Crawling Structured Data

Exception handling

- You may encounter an exception as below.

```
Exception in thread "main" java.lang.NoClassDefFoundError: org/apache/commons/lang3/ObjectUtils
    at com.opencsv.CSVParser.<init>(CSVParser.java:99)
    at com.opencsv.CSVReader.<init>(CSVReader.java:99)
    at openCSV.Main.main(Main.java:25)
Caused by: java.lang.ClassNotFoundException: org.apache.commons.lang3.ObjectUtils
    at java.base/jdk.internal.loader.BuiltinClassLoader.loadClass(BuiltinClassLoader.java:606)
    at java.base/jdk.internal.loader.ClassLoaders$AppClassLoader.loadClass(ClassLoaders.java:168)
    at java.base/java.lang.ClassLoader.loadClass(ClassLoader.java:522)
    ... 3 more
```

Q3 : Find a way to resolve this problem

Crawling Structured Data II(Cont'd)

Download Apache POI libraries and add them to the project

- <http://poi.apache.org/download.html>



The screenshot shows the Apache POI website. At the top, there's a navigation bar with links to Home, Help, Component APIs, and Getting Involved. A search bar is also present. The main content area is titled "Apache POI - Download Release Artifacts". It includes a section for "Available Downloads" with links to the latest stable release (Apache POI 5.0.0) and archives of all prior releases. Below this, it states that Apache POI releases are available under the Apache License, Version 2.0, and provides instructions on how to verify the integrity of the files. A prominent blue banner announces the "20 January 2021 - POI 5.0.0 available". The text below the banner details the new features and bug fixes in the 5.0.0 release, mentions the release notes and change log, and lists the pre-built binary deployment packages and their checksums. The page also includes a "Binary Distribution" section with links to the binary packages and their checksums, and a "Source Distribution" section with links to the source packages and their checksums. On the left side, there's a sidebar with a menu for Overview (Home, Download, Changelog, Javadocs, Text Extraction, Encryption support, Case Studies, Related projects, Legal) and Apache Wide. There are also logos for "SUPPORT APACHE" and "POWERED BY APACHE" on the left.

Apache Software Foundation > Apache POI >

THE APACHE® SOFTWARE FOUNDATION

APACHE POI

Home Help Component APIs Getting Involved

Search the site with google Search

Last Published: 01/20/2021 08:43:3

Apache POI - Download Release Artifacts

Available Downloads

This page provides instructions on how to download and verify the Apache POI release artifacts. There are different versions available depending on how stable your code should be.

- [The latest stable release is Apache POI 5.0.0](#)
- [Archives of all prior releases](#)

Apache POI releases are available under the [Apache License, Version 2.0](#). See the NOTICE file contained in each release artifact for applicable copyright attribution notices.

To ensure that you have downloaded the true release you should [verify the integrity](#) of the files using the signatures and checksums available from this page.

20 January 2021 - POI 5.0.0 available

The Apache POI team is pleased to announce the release of 5.0.0. Featured are a handful of new areas of functionality and numerous bug fixes.

A summary of changes is available in the [Release Notes](#). A full list of changes is available in the [change log](#). People interested should also follow the [dev list](#) to track progress.

The POI source release as well as the pre-built binary deployment packages are listed below. Pre-built versions of all [POI components](#) are available in the central Maven repository under Group ID "org.apache.poi" and Version "5.0.0".

Binary Distribution

- [poi-bin-5.0.0-20210120.tar.gz](#) (56.01 MB, [signature \(.asc\)](#), checksum: [SHA-256](#), [SHA-512](#))
- [poi-bin-5.0.0-20210120.zip](#) (66.36 MB, [signature \(.asc\)](#), checksum: [SHA-256](#), [SHA-512](#))

Source Distribution

- [poi-src-5.0.0-20210120.tar.gz](#) (106.27 MB, [signature \(.asc\)](#), checksum: [SHA-256](#), [SHA-512](#))
- [poi-src-5.0.0-20210120.zip](#) (110.53 MB, [signature \(.asc\)](#), checksum: [SHA-256](#), [SHA-512](#))

Crawling Structured Data II

READ an XLSX file

```
16 try {
17     String file = "C:/Users/CTC/Desktop/dataraw.xlsx";
18     FileInputStream fis = new FileInputStream(file);
19     XSSFWorkbook workbook = new XSSFWorkbook(fis);
20     XSSFSheet sheet = workbook.getSheet("Sheet5");
21
22     for (int row = 1; row < sheet.getPhysicalNumberOfRows(); row++) {
23         XSSFRow rows = sheet.getRow(row);
24         if (rows != null) {
25             String value = "";
26             int cells = rows.getPhysicalNumberOfCells();
27             for (int column = 0; column <= cells; column++) {
28                 XSSFCell cell = rows.getCell(column);
29                 if (cell != null)
30                     switch (cell.getCellType()){
31                         //case FORMULA:
32                         //    value = cell.getCellFormula();
33                         //    break;
34                         case NUMERIC:
35                             value = cell.getNumericCellValue() + "";
36                             break;
37                         case STRING:
38                             value = cell.getStringCellValue() + "";
39                             break;
40                         case BLANK:
41                             value = cell.getBooleanCellValue() + "";
42                             break;
43                         case ERROR:
44                             value = cell.getErrorCellValue() + "";
45                             break;
46                         default:
47                             break;
48                     }
49                 System.out.print(value + " ");
50             }
51         }
52         System.out.println();
53     }
54 } catch (FileNotFoundException e) {
55     e.printStackTrace();
56 } catch (IOException e) {
57     e.printStackTrace();
58 }
59 }
```

Q4 : You may encounter a problem with executing the program, find a way to resolve it

Q5 : Try writing an XLSX file

Q6 : Try the same things above with an XLS file

P1

Practice for descriptive statistics

- ▶ Visit "data.go.kr"
- ▶ Get at least 10 of CSV files for one theme
- ▶ Merge those files into one file or Get all the rows at once
- ▶ Do descriptive statistics for each column
- ▶ Draw graphs in excel

P2

Practice for descriptive statistics

- ▶ Visit "data.go.kr"
- ▶ Get at least 10 of XLS or XLSX files for one theme
- ▶ Merge those files into one file or Get all the rows at once
- ▶ Do descriptive statistics for each column
- ▶ Draw graphs in excel

Ethics of Web Data Scraping

Web Data Scraping

- ▶ A mechanism to make a computer visit a website automatically and collect some data in the process.
- ▶ Data Scraping can have positive effects if done the right way
- ▶ **However, read the terms of use for the website first!**
- ▶ **Moreover, do not use this data in the wrong way!**

Keywords

What we need to understand are ...

- ▶ ML, DL, the difference between ML and DL
- ▶ Three categories of ML : SL, UL, RL
- ▶ Classification
- ▶ Regression
 - Linear Regression
 - Logistic Regression