데이터융합SW과
김규석 교수

Data Analysis with Java

데이터 분석 프로그래밍01

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

Source : https://en.wikipedia.org/wiki/Bitwise_operation

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

Source : https://en.wikipedia.org/wiki/Bitwise_operation

Types of Information

Understanding Information Sources

- There are four types of information

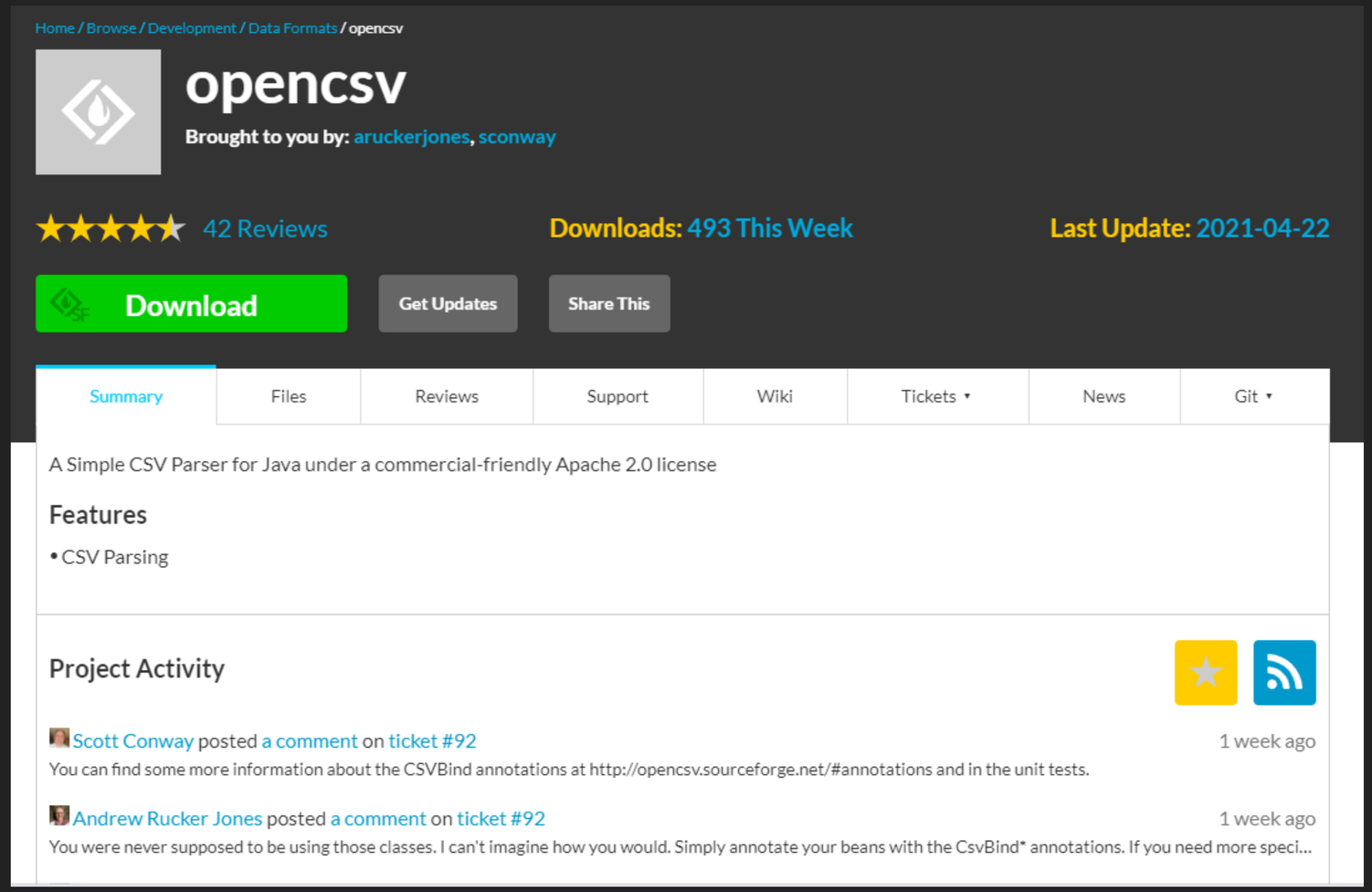
Type	Description	Example
Factual	Information that deals with facts	Government Resources, Encyclopedias
Analytical	Interpretation of factual information	Library databases, Academic books
Subjective	Information from only one point of view	Websites, Blogs, Social media
Objective	Information that is understood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

Source : https://libguides.astate.edu/information/types

Crawling Structured Data(Cont'd)

Download openCSV

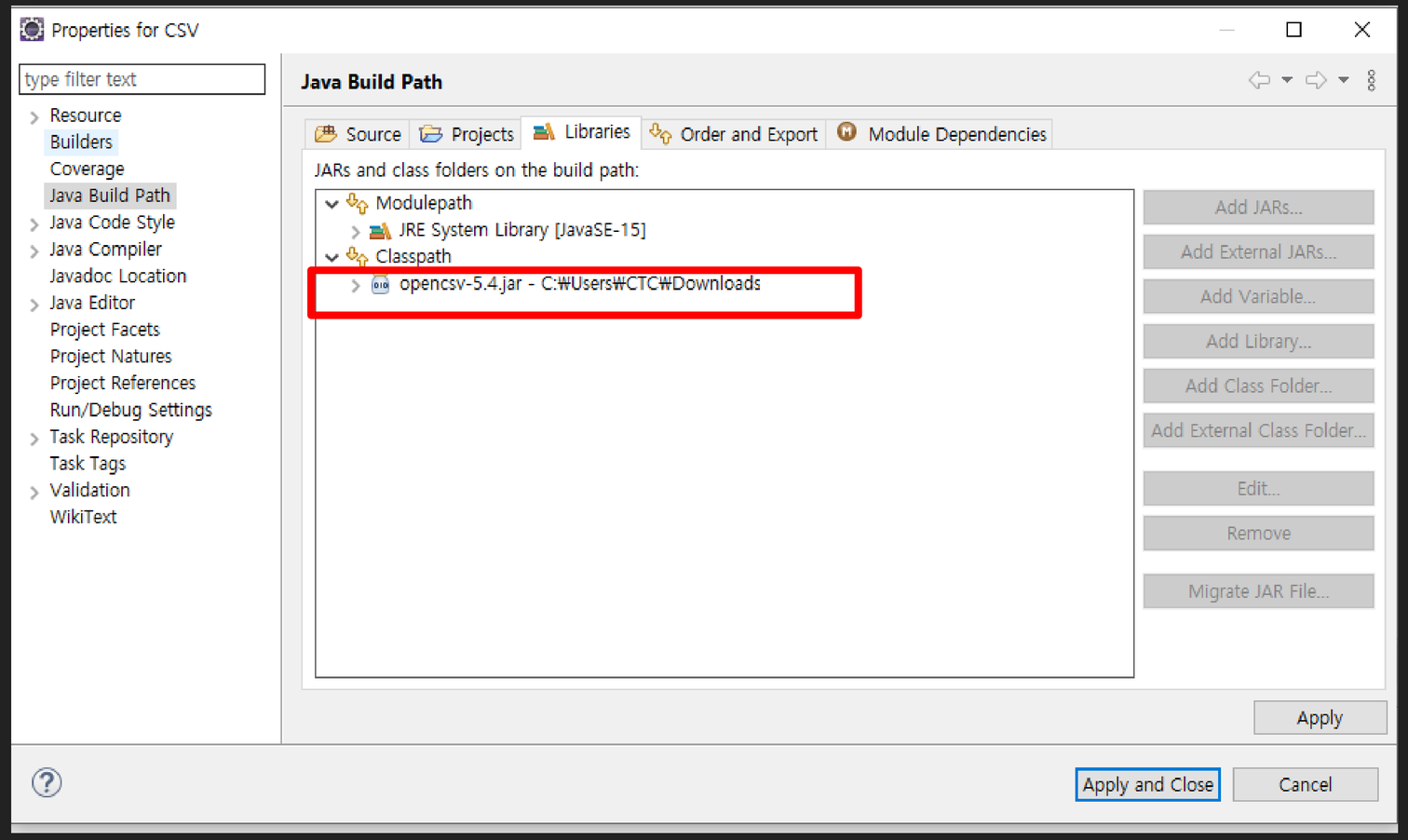
- https://sourceforge.net/projects/opencsv/



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

```java
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.

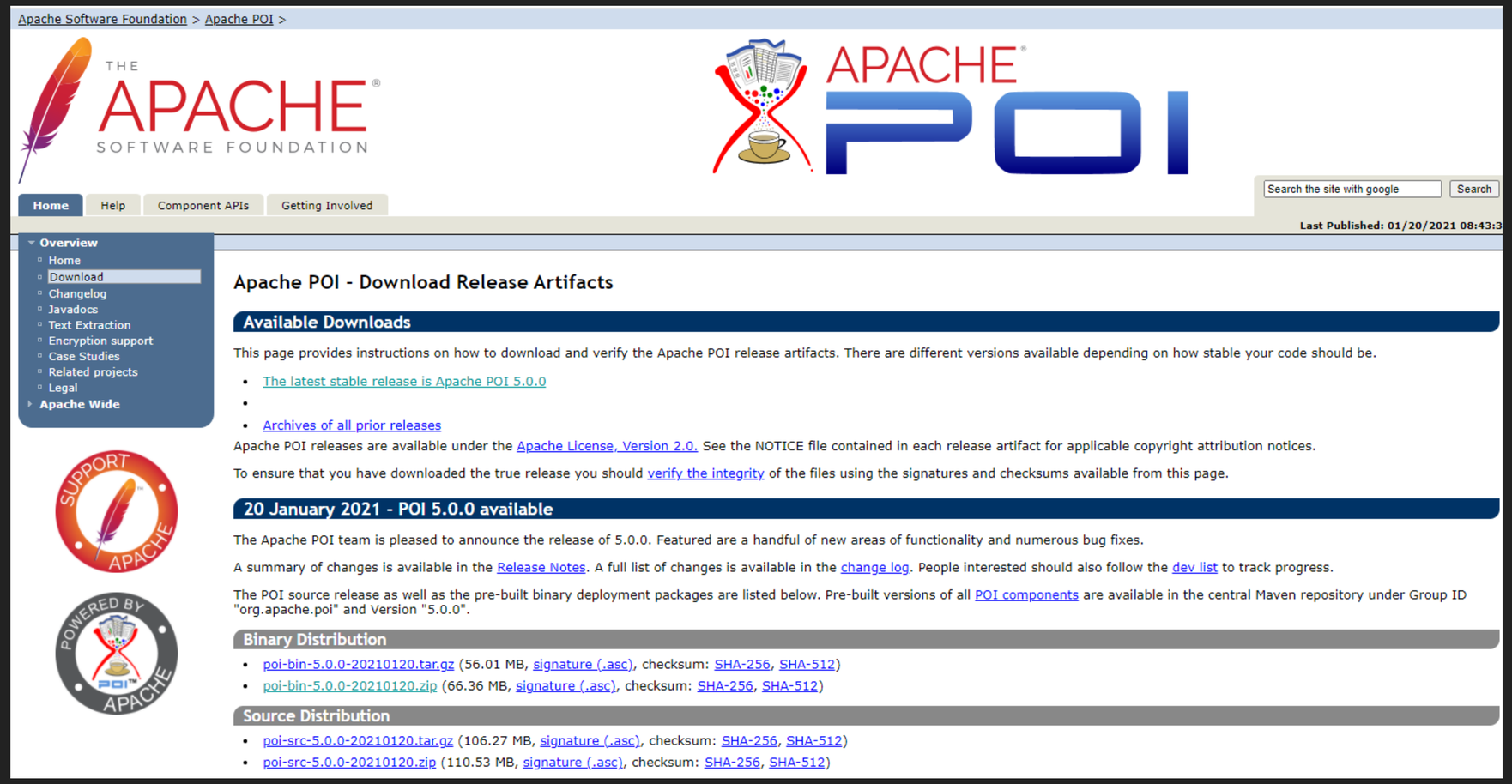
```
Exception in thread "main" java.lang.NoClassDefFoundError: org/apache/commons/lang3/ObjectUtils
        at com.opencsv.CSVParser.<init>(CSVParser.java:99)
        at com.opencsv.CSVReader.<init>(CSVReader.java:99)
        at openCSV.Main.main(Main.java:25)
Caused by: java.lang.ClassNotFoundException: 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



Crawling Structured Data II

READ an XLSX file

```java
        try {
            String file = "C:/Users/CTC/Desktop/dataraw.xlsx";
            FileInputStream fis = new FileInputStream(file);
            XSSFWorkbook workbook = new XSSFWorkbook(fis);
            XSSFSheet sheet = workbook.getSheet("Sheet5");

            for (int row = 1; row < sheet.getPhysicalNumberOfRows(); row++) {
                XSSFRow rows = sheet.getRow(row);
                if (rows != null) {
                    String value = "";
                    int cells = rows.getPhysicalNumberOfCells();
                    for (int column = 0; column <= cells; column++) {
                        XSSFCell cell = rows.getCell(column);
                        if (cell != null)
                            switch (cell.getCellType()){
                                //case FORMULA:
                                    //value = cell.getCellFormula();
                                    //break;
                                case NUMERIC:
                                    value = cell.getNumericCellValue() + "";
                                    break;
                                case STRING:
                                    value = cell.getStringCellValue() + "";
                                    break;
                                case BLANK:
                                    value = cell.getBooleanCellValue() + "";
                                    break;
                                case ERROR:
                                    value = cell.getErrorCellValue() + "";
                                    break;
                                default:
                                    break;
                        }
                        System.out.print(value + " ");
                    }
                }
                System.out.println();
            }
        } catch (FileNotFoundException e) {
            e.printStackTrace();
        } catch (IOException e) {
            e.printStackTrace();
        }
    }
```

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!**

Source : https://www.promptcloud.com/blog/is-data-scraping-ethical/

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