Filetype Identification

Deliverable 1

Problem Statement:

Identify relevant data sources from where a file type information can be extracted based on filename or file extension. List at least 5 relevant sources and explain the rationale on why they should be used.

Data Sources

1. **FileProInfo:**

The FileProInfo is a website focused on collecting information about file types & extensions and their associated software programs, it provides users with any information related to the file types and their associated applications mainly on Windows, macOS, Linux, Web, iOS, and Android.

- "Web Scraping" is done for the data extraction of the file extension like file Category(Data Files, Video Files,...), file Format(Binary, Text,...), Developer(Microsoft, Autodesk,...), Mime Type(application\/xslt+xml,....), Programming Language(java, python,...).

Source: https://fileproinfo.com/

2.. **FileInfo**:

FileInfo.com contains a searchable database of over 10,000 file extensions with detailed information about the associated file types. You can look up information about unknown file types and find programs that open the files.

- "Web Scraping" is done for the data extraction of the file extension like file Category, Format, Description, Programming Language, Program Support.

Source: https://fileinfo.com/filetypes/

3. **Apache Tika:**

Apache Tika is a library that is used for document type Detection and content extraction from various file formats.

This XML file defines the valid mime types used by Tika. The mime type data within this file is based on information from various sources like Apache Nutch, Apache HTTP Server, the file(1) command, etc.

- The data extraction is done using XML parsing.

Source: https://github.com/apache/tika/blob/master/tika-core/src/main/resources/org/apache/tika/mime/tika-mimetypes.xml

4. File.org:

It is basically a website which provides the service of the various features of the file from its extension. It has a huge database of file extensions (file types) with detailed descriptions.

- "Web Scraping" is done for the data extraction of the file extension like file Description, Various File Openers and viewers.

Source: https://file.org/

5. Connect:

It is basically a website which provides the page of technical terminology, which could be a source for the file extension, which is lesser known. The scenarios where we need to find some rare data, then in that case this can be helpful.

Source: https://www.consp.com/it-information-technology-terminology-dictionary

6. Webopedia:

It is a website which provides a dictionary-like definition for the word, so it is very suitable to fetch data related to extensions. It provides the information by categorizing the information level in various sectors such as computers, IT management, design, applications etc., which could be a plus point if we target information particularly.

Source: https://www.webopedia.com/

7. VidyaGyaan:

It is a website which provides the details of extension and short description about it. It consists of all the extension details on a single page, so when there are time constraints, this can be a very helpful data source.

Source: http://www.vidyagyaan.com/computer-knowledge/list-of-computer-file-extensions-and-their-meaning/

8. Linux Commands:

When examining a Linux file, the contents are only half the story. Every file and directory also has attributes that describe its owner, size, access permissions, and other information. Example: stat, file, lsattr etc

Source:
https://www.oreilly.com/library/view/linux-pocket-guide/9780596806347/ch01s13.html

Apart from listed websites, there are lot of similar websites like File-Extension.org, DotWhat etc are also similar to FileInfo website and provides file detail related to extension.

Deliverable 2:

1. Web scraping the data from FileProInfo.com, Fileinfo.com using java and storing it in *fileproinfo.json* and *fileinfo.json* respectively.

2. And the third data source is generated by extracting *tika.xml* using a java parser and storing it in *tikasource.json*.

3. The user input is a *.csv* file that contains all the input file names. And all the file names are stored in a *list*.

4. The data extracted from the json files is stored in three ConcurrentHashMaps *tikaMap, fileInfoMap, fileProInfoMap* using three threads (multithreading).

5. The input list is approximately divided into four chucks which are worked upon simultaneously using four Threads (multithreading).

6. The threads used to lookup information corresponding to the input will only execute when the maps are completely written to.

7. And finally all the output records are written to their respective output.csv files which are *FileInfoSourceOutput.csv, FileProSourceOutput.csv, TikaSourceOutput.csv*.

Steps to run the program:

1. The code uses three referenced libraries namely *json-simple-1.1.1.jar, jsoup-1.14.1.jar, opencsv-4.1.jar,* these three libraries which will have to be manually added to the project according to the IDE is being used.
2. To execute the code the MainClass.java file has to be run.
3. The input files are being called from the input folder, and the output .csv files will be stored in the output folder. And these output files(*FileInfoSourceOutput.csv, FileProSourceOutput.csv, TikaSourceOutput.csv*.) are being created corresponding to the three data sources.