File Type Identification

Worksample by BlueOptima to identify file type using various sources by using files name and file extensions.

Problem Statement:

With the enormous number of languages and file types used for writing logical source or for data purposes, it is very important for a product like BlueOptima to effectively identify and categorize a file into its type. And this has to be done solely based on Extension and Name of the file itself. This work sample requires you to identify different sources that could be used to identify details of a file type like following (but not limited to)

- Short Description (explaining the usage of the file type)
- Category (i.e. Logical Source, Configuration, Data, etc.)
- Language Family (Java, Python, Perl, etc.)
- Programming Paradigm (Procedural, OOP, Dynamic, etc)
- Associated applications

Solution (Execution Flow)

- [x] **Deliverable 1 - Identification and Analysis of Data Sources.**
    - [x] Identify at least 5 different Data sources.
    - [x] Expand on the rationale for using the Data source.
- [x] **Deliverable 2 - Implementation and Presentation of information about the given input file types.**
    - [x] Extract (Web scraping) data from Fileinfo.com using python script and store in **sourceFileInfo.json** file.
    - [x] Extract tika.xml using java parser and store in**sourceTika.json** file.
    - [ ] Extract (web scraping) data from IANA source using python script and store in .json file.
    - [x] Create an input**input.csv** file for passing all the inputs.
    - [x] Implement the main Program -**fileTypeIdentification.java**
        - [x] Store all the input filenames in a list.
        - [x] Access various data sources (extracted previously in .json files) and load each data source in the main memory (hash maps) `fileInfoHM, tikaHM`.
        - [x] For each file Extension input, parse it in the hash maps to search for required data in a priority.
        - [x] Write the information about each file input in `output.txt`

Input

The input file is found in the 'data' directory of the `File-Type-Identification`. We have taken filenames with its extension in a `csv file` as shown below.

> /input/input.csv

```
binarySort.CPP
linkList.cpp
Readme.pdf
fibonacci.XCODEPROJ
about.txt
scrape.py
xmlParser.java
```

Output

The output for the program is written on a text file `output.txt` in the main directory. Given below is a sample output.

> output.txt

```
_____________________________________________________________________________

File: binarySort.CPP
_____________________________________________________________________________

      Category    : Developer File
      Type        : C++ Source Code File
      Description : A CPP file is a source code file written in C++, a popular prc
      Programs    : File Viewer Plus, Microsoft Visual Studio 2017, Microsoft Visu
```

Steps to Run the Program

1. In `/input/` Create your input file in csv as given in the above input format or just use the pre built one.
2. Execute the main program: `/src/fileTypeIdentification/FileTypeIdentification.java`
3. Enter the input file name example: `input1.csv` in the console or else it will take the default `input0.csv` on return.
4. Check the output in `output.txt` file in the main directory.

> `java version "12.0.2" 2019-07-16`
> https://www.oracle.com/technetwork/java/javase/downloads/jdk12-downloads-5295953.html

Developers

- Mohammed Ataaur Rahaman
- Siddharth Singh
- Shivani Bangalore