Summary

Filetype Identification

Work sample by BlueOptima to identify different sources to identify details of a file type by using file names 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)

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

Solution

Flow of Execution

Deliverable 1 - Identification and Analysis of Data Sources.

⇨ Identify at least 5 different Data sources.
⇨ Expand on the rationale for using the Data source.

Deliverable 2 - Implementation and Outputting of information about the given input file types.

⇨ Extract (Web scraping) data from Fileinfo.com using Chrome Web Extension and store in **FileInfo.json** file.
⇨ Extract (Web scraping) data from FileProInfo using Chrome Web Extension and store in **fileProInfo.json** file.
⇨ Extract (web scraping) data from Howopen source using Chrome Web Extension and store in **Extensions.json** file.
⇨ Create an default input **defaultInput.txt** file and **input.txt** file for passing all the inputs.

⇨ Implement the Java main Program - **fileTypeIdentifier.java**

- •⟩ Store all the input filenames in a list.
- •⟩ Access various data sources (extracted previously in .json files) and load each data source into Hashmaps **extensionsMap**, **fileInfoMap** & **ProInfoMap**.
- •⟩ For each file Extension input, parse it in the HashMaps to search for required data.
- •⟩ Write the information about each file input in **output.txt** file.

Input

The input file is found in the '**inputFileTypes**' directory of the **file_type_identifier**. We have taken filenames with its extension in a text file (**input.txt**) as shown below.

input.txt

```
Main.java
Python.py
program.CPP
tinplated.CCP
Readme.PDF
enlighten.KPL
```

Output

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

output.txt

```
File          : Main.java

Category      : Developer File

Type          : Java Source Code File

Description: A JAVA file is a source code file written in the Java
programming language, which was originally developed by Sun
Microsystems but is now maintained by Oracle.  It uses an object-
oriented approach, where structured data types, called classes, are
used to instantiate objects at runtime.

Programs    : File Viewer Plus, Oracle Java Virtual Machine, Eclipse IDE
for Java Developers, Google Android Studio, Oracle NetBeans, Xinox
JCreator, ES-Computing EditPlus, Microsoft Notepad, gVim, Other text
editor, Apple Xcode, MacroMates TextMate, MacVim, javac, GNU
Emacs, Vi, File Viewer for Android

==================================================
```

Steps to Run the Program

Step 1.

⇨ In **/inputFileTypes/** Create your input file in text(.txt) as given in the above input format or just use the pre built one.

Step 2.

⇨ Execute the main program:
/src/main/java/file_type_identifier/fileTypeIdentifier.java

Step 3.

⇨ Enter the input file name example: **input.txt** in the console or else it will take the default **defaultInput.txt** on return.

Step 4.

⇨ Check the output in **output.txt** file in the main directory.

Developers

- Raj Anand
- Vinay Achari