Filetype Identification

Deliverable 1

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. Language Family (Java, Python, Perl, etc.)

Identify relevant data sources from where a filetype information (as described above) can be extracted based on filename or file extension. List at least 5 relevant sources and explain the rationale on why it should be used.

Data Sources:

- Fileinfo
- dotwhat
- reviversoft
- openWith
- file-extensions
- Apache OpenOffice

1. fileinfo:

Link: https://fileinfo.com/

FileInfo is a website found in the year 2005 and it contains a searchable database of over 10,000 file extensions with detailed information about the associated file types. It also gives the information about unknown file types and finds programs that open these kinds of files.

FileInfo team has worked with developers, large and small, to create a central file extensions registry. FileInfo.com has become the authoritative website where developers can submit new file extensions and provide information about file types.
We scraped data from this website to get all the necessary data needed for the project regarding the filetype. Data which we scraped from this website is:

- File Category
- File Type (CPP, Java, Perl, C, Python etc.)
- Description about the extension
- Applications for the respective file

2. dotWhat:

Link: http://dotwhat.net/type/developer-files

dotWhat is a website which started in 2005. It is one of the leading file extension resource and the most detailed database which covers all the various operating system like Windows, Mac, Unix. They also take suggestions in the form of input from their users to make the website better and more resourceful.

Few resources to keep their database up to date and running smooth used by this website are:

1. NetLingo.com
2. ComputerResource.org
3. whatisprocess.com

We scraped data from this website to get all the necessary data needed for the project regarding the filetype. Data which we scraped from this website is:

• File Category
• File Type (CPP, Java, Perl, C, Python etc.)
• Description about the extension
• Applications for the respective file

3. reviversoft:

Link: https://www.reviversoft.com/

ReviverSoft was found to provide trusted resources to help its users repair, optimize and maintain your computer for optimum performance.

The website invests heavily in research and development to deliver the best possible products, professional instructional videos, how-to articles, informative blog posts and other resources like file extensions.

We scraped data from this website to get all the necessary data needed for the project regarding the filetype. Data which we scraped from this website is:

• File Category
• File Type (CPP, Java, Perl, C, Python etc.)
• Description about the extension
• Applications for the respective file

4. openWith:

Link: https://www.openwith.org/categories/source-code

OpenWith started not very long ago in 2019 but is a website with good amount of data and has one purpose that is to provide its users with open source database containing all the different file extensions. It also provides with free programs that can run and create each type of file.

5. File-extensions:

Link: https://www.file-extensions.org/

The File-Extensions.org is a website found in the year 2000, it is focused on collecting of information about file extensions and their associated programs, so that it can provide with any information that is needed by its customer related to different type of files. File-Extensions.org is a huge and regularly updated database offering as much as maximum possible amount of information for thousands of file types and their associated applications mainly on the Microsoft Windows platform, and along with that also for MAC OS and Unix based operating systems. In many cases this website also provides basic instructions on how to open or convert certain filetypes.

It also encourages the site visitors to suggest new file extensions or any other additional information, which can help in improving the website.
Various files which can be scraped from the website are developer files, program executable files, audio files, documents, accounting, financial, tax files and many more.

6. Apache Open Office:

Link: https://www.openoffice.org/

Apache OpenOffice is the leading open-source office software suite. This website can be used to scrape different types of file extensions. The page contains the filename extensions, or suffixes, used in the source code.