Deliverable 1

Statement:

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.

Identified Data Sources

1. **File Info**
2. **Apache Tika**
3. **Apache OpenOffice**
4. **Internet Assigned Numbers Authority (IANA)**
5. **Windows Registry [HKEY_CLASSES_ROOT]**
6. **filext.com website**
7. **Webopedia Website**
8. **Reviver Soft Website**
9. **Linux file Command [$ file]**

File Info – Website

This website founded in 2005, has been constantly working with developers to create central file extensions registry. This website has become authoritative website where developers can submit new file extensions and provide information about file types. Presently it has more than 10000 file extensions with detailed information about the associated file types.

We can scrape data from file Info Website to get all the necessary data related to a file extension. Data which could be scrapped or crawled from this website are

- File Category (Developer file, Database file … etc)
- File Type / Language Family (Java, Perl, C, … etc)
- Short Description about the file type
- Associated Applications

Link: https://fileinfo.com/browse/

Apache Tika

Apache Tika is a library which is used for document type Detection and content extraction from various file formats. Tika provides a single generic API to parse different file formats. From the link given below, tika-mimetypes.xml is a licensed to the Apache Software Foundation (ASF). This xml file defines 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.

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

Apache OpenOffice

This is an online website having various file extensions with its file Types listed for Source code files and database files. This website could also be scarped to get the file types associated with the file extensions.

Link: https://www.openoffice.org/dev_docs/source/file_extensions.html

Internet Assigned Number Authority – IANA

Founded in 1988 IANA is a non-profit private American corporation that oversees global IP address allocation, root zone management in DNS, autonomous system number allocation, media types for MIME files, and various others. What comes to our interest is that it has information about various MIME file types. This website has registries for application, audio, font, image, text, video, etc to name few file types. We can go to the respective webpage and get files description, its type, Security considerations, published specification, associated applications and others

Link: https://www.iana.org/assignments/media-types/media-types.xhtml

Windows Registry (regedit)

As windows registry is a hierarchical database that stores low-level settings for the windows OS. For example, when a program is installed a new subkey containing setting such as a program's location, its version and how to start the program are all added to the windows registry. So from HKEY_CLASSES_ROOT we can get all the associated extensions available on a particular Machine.

Using Hivex which is a C API to extract contents of Registry hive files. It is designed to be secure against buggy or malicious registry files. The extracted data could also be exported as an XML file.

Link: https://github.com/libguestfs/hivex

FILExt – website

Founded in 2000, being one of the oldest websites to store and update various file extensions has helped more than 50 million users identify, open, view, or convert unknow files. With just the file extension one can know brief summary about the file type, associated applications with it, and know to common problems related the file type.

Link: https://filext.com

Webopedia – Website

Webopedia is an online tech Dictionary for IT professionals, educators and students which provides definitions to words, phrases and abbreviations related to computing and information technology. So this website is one of the source to find some short description about a given file extension. But its worth noting that this website may not be updated frequently and has very less number of file extensions.

Link: https://www.webopedia.com/quick_ref/fileextensionsfull.asp

Reviver Soft – Website

Reviver soft was founded to provide trusted resources to help repair, optimize and maintain our computer for optimum performance. This website has a decent collection of file extension library which could be searched to find out about the file extensions brief description and how to open it (associated applications). Also, one could find a word of warning related to each file extensions.

Link: https://www.reviversoft.com/file-extensions/

Linux file (command)

This file command is a standard program of linux and linux like OS for recognising the various types of data contained in a computer file. This file command on linux machines could be used to get file type information for various file extensions. File command tests each argument in an attempt to classify it. There are three tests performed in this order: filesystem tests, magic test and language tests. The file system test gives us the file type required.

Link: https://linux.die.net/man/1/file