Filetype Identification

(Akhil Rautela,

Ashutosh Mahapatra,

Ayush Kumar)

Deliverable 1:

Identify relevant data sources from where a filetype information 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.

- Relevant data sources from where a filetype information (as described above) can be extracted based on filename or file extension are as follows:

- FileInfo
- Filext
- Wikipedia
- Reviver soft
- Apache Tika

1.FileInfo:

FileInfo.com contains a searchable database of over 10,000 file extensions with detailed information about the associated file types. We can look up information about unknown file types and find programs that open the files. Since 2005, the FileInfo.com 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.

In order to get information about a file we just need to enter it's extension in the searchbox and fileinfo will look for this extension in it's database and will output the result.

We can fetch the results from this source using web scrapping.

Link: https://fileinfo.com/

Search over 10,000 file extensions and software programs.

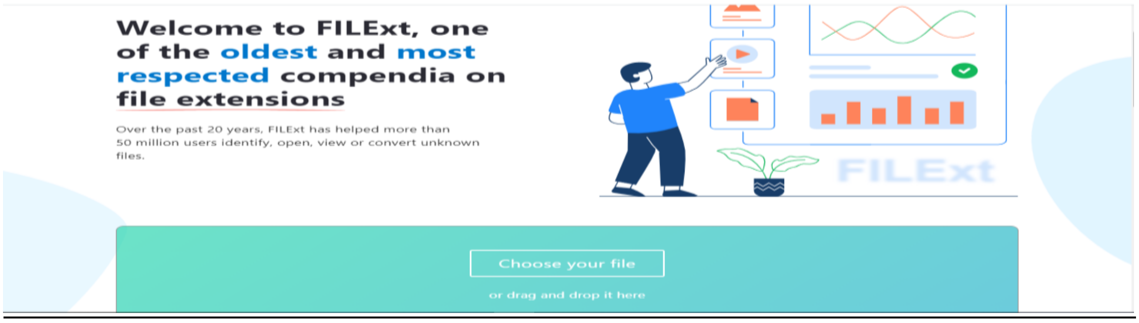
2.Filext:

FILExt is one of the oldest and most respected collections of file formats and file extensions. Over the past 20 years, more than 50 million users have found the right information and tools to open any file on their computer or smartphone. Our knowledge gathered during this period is regularly reviewed and updated. Tom Simondi first provided this information in 2000 as a free online resource for the Internet community.

In order to get information about a file, simply drag and drop a file onto the FILExt website. FILExt will then analyze the file type and immediately preview the file online. File preview recognizes thousands of file types and it's database contains detailed information about almost every file extension there is. If we want to find out about a specific file extension, we can use the search box at the top right of the page.

We can fetch the results from this source using web scrapping.

Link: https://filext.com/



Welcome to FILExt, one of the oldest and most respected compendia on file extensions

Over the past 20 years, FILExt has helped more than 50 million users identify, open, view or convert unknown files.

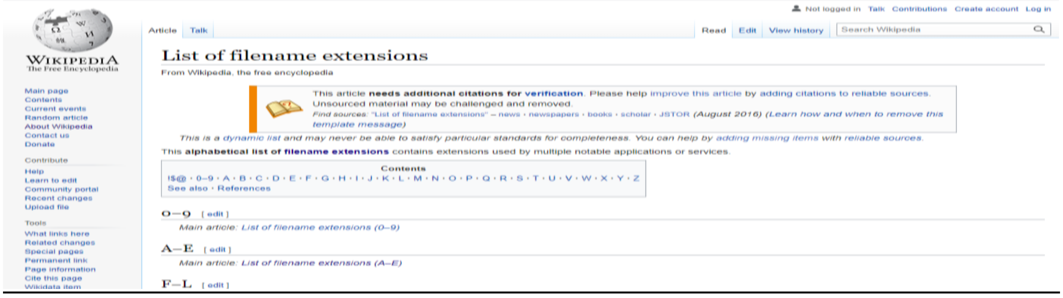
3.Wikipedia:

Wikipedia is an online free content encyclopedia project helping to create a world in which everyone can freely share in the sum of all knowledge. It is supported by the Wikimedia Foundation and based on a model of freely editable content.

To get to know about any kind of file or it's extension ,we simply need to search with it's name or extension or can also look for it in the list of filename extension and can get the information.

We can fetch any kind of information about the file or it's type using the api.

Link: https://en.wikipedia.org/wiki/List_of_filename_extensions



List of filename extensions

From Wikipedia, the free encyclopedia

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed.
Find sources: "List of filename extensions" – news · newspapers · books · scholar · JSTOR (August 2016) (Learn how and when to remove this template message)

This is a dynamic list and may never be able to satisfy particular standards for completeness. You can help by adding missing items with reliable sources.

This **alphabetical list of filename extensions** contains extensions used by multiple notable applications or services.

Contents
!$@ · 0–9 · A · B · C · D · E · F · G · H · I · J · K · L · M · N · O · P · Q · R · S · T · U · V · W · X · Y · Z
See also · References

0–9 [edit]

Main article: List of filename extensions (0–9)

A–E [edit]

Main article: List of filename extensions (A–E)

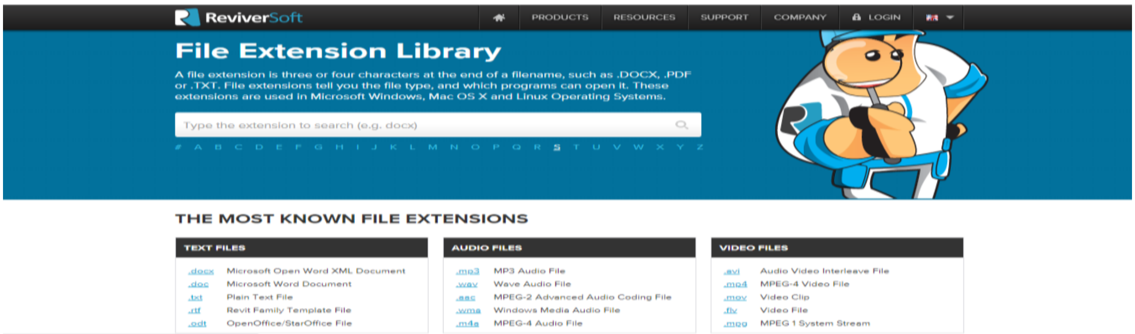
F–L [edit]

4.Reviver Soft:

ReviverSoft provides Award-winning Software and Helpful Tips to make your PC run like NEW again.This website has large collection of file extension library from where we we can search to find out about file extensions and it’s types and can also get information about it’s associated applications. But it’s worth noting that this website may not be updated frequently so it doesn’t have information about all the types of files.

We can fetch the results from this source using web scrapping.

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


ReviverSoft
PRODUCTS
RESOURCES
SUPPORT
COMPANY
LOGIN
Type the extension to search (e.g. docx)
# A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

File Extension Library

A file extension is three or four characters at the end of a filename, such as .DOCX, .PDF or .TXT. File extensions tell you the file type, and which programs can open it. These extensions are used in Microsoft Windows, Mac OS X and Linux Operating Systems.

THE MOST KNOWN FILE EXTENSIONS

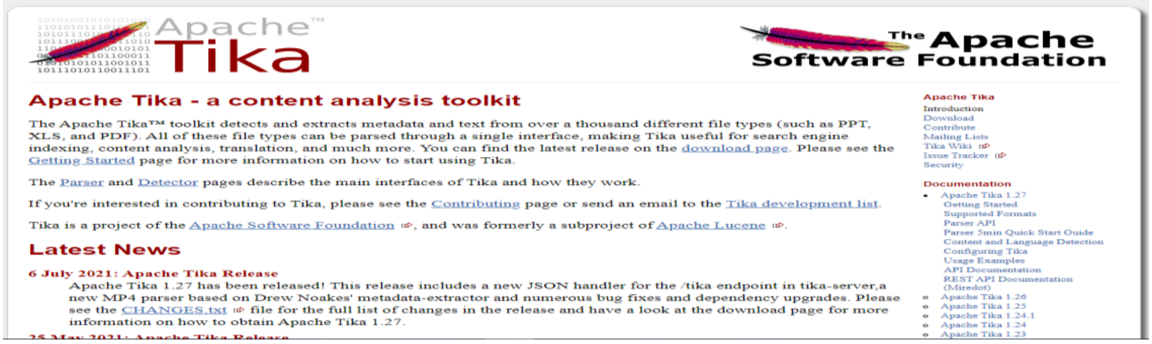
TEXT FILES		AUDIO FILES		VIDEO FILES	
.docx	Microsoft Open Word XML Document	.mp3	MP3 Audio File	.avi	Audio Video Interleave File
.doc	Microsoft Word Document	.wav	Wave Audio File	.mp4	MPEG-4 Video File
.txt	Plain Text File	.aac	MPEG-2 Advanced Audio Coding File	.mov	Video Clip
.rtf	Revit Family Template File	.wma	Windows Media Audio File	.flv	Video File
.odt	OpenOffice/StarOffice File	.m4a	MPEG-4 Audio File	.mpg	MPEG 1 System Stream

5.ApacheTika:

The Apache Tika™ toolkit detects and extracts metadata and text from over a thousand different file types (such as PPT, XLS, and PDF). All of these file types can be parsed through a single interface, making Tika useful for search engine indexing, content analysis, translation, and much more.

Tika provides a single generic api to parse different file formats

Link: https://tika.apache.org/



Apache Tika - a content analysis toolkit

The Apache Tika™ toolkit detects and extracts metadata and text from over a thousand different file types (such as PPT, XLS, and PDF). All of these file types can be parsed through a single interface, making Tika useful for search engine indexing, content analysis, translation, and much more. You can find the latest release on the download page. Please see the Getting Started page for more information on how to start using Tika.

The Parser and Detector pages describe the main interfaces of Tika and how they work.

If you're interested in contributing to Tika, please see the Contributing page or send an email to the Tika development list.

Tika is a project of the Apache Software Foundation, and was formerly a subproject of Apache Lucene.

Latest News

6 July 2021: Apache Tika Release

Apache Tika 1.27 has been released! This release includes a new JSON handler for the /tika endpoint in tika-server,a new MP4 parser based on Drew Noakes' metadata-extractor and numerous bug fixes and dependency upgrades. Please see the CHANGES.txt file for the full list of changes in the release and have a look at the download page for more information on how to obtain Apache Tika 1.27.

25 May 2021: Apache Tika Release