

Scraping Media from the Web with Python

INTRODUCTION



Allen O'Neill

ENGINEER

@databytzai



Overview



Differences in scraping web media

Types of web media

Working with large files

Different technologies available



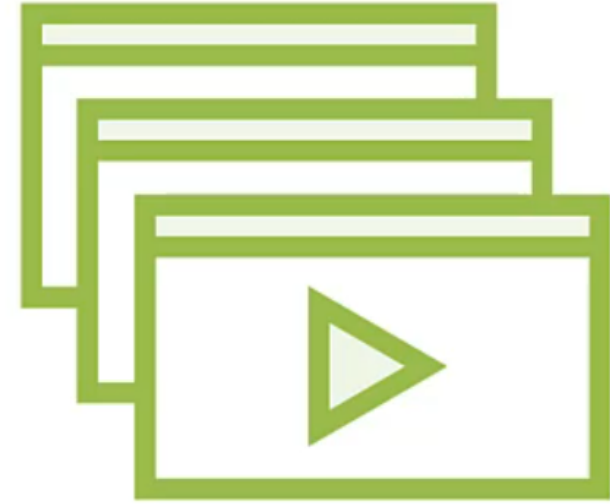
What Is Web Media?



Media



Web Media



Digital Media

Key Differences



Not human readable



Not embedded



Requires processing

Media Grouping



Media Grouping



Media Grouping



DVI



XLR



VGA

Large Media Files



Large File Implications



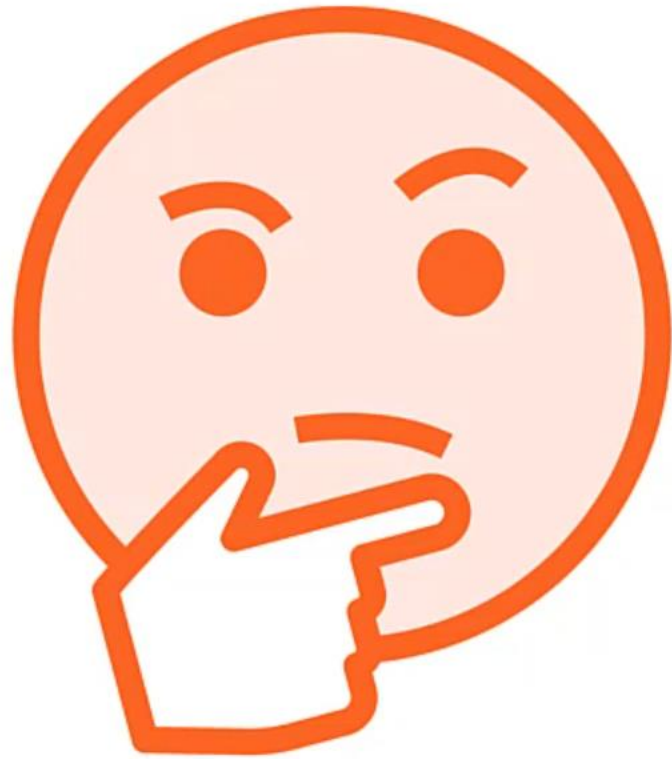
Bandwidth



Storage



Resources



Cost of data:

- In motion
- At rest
- Processing

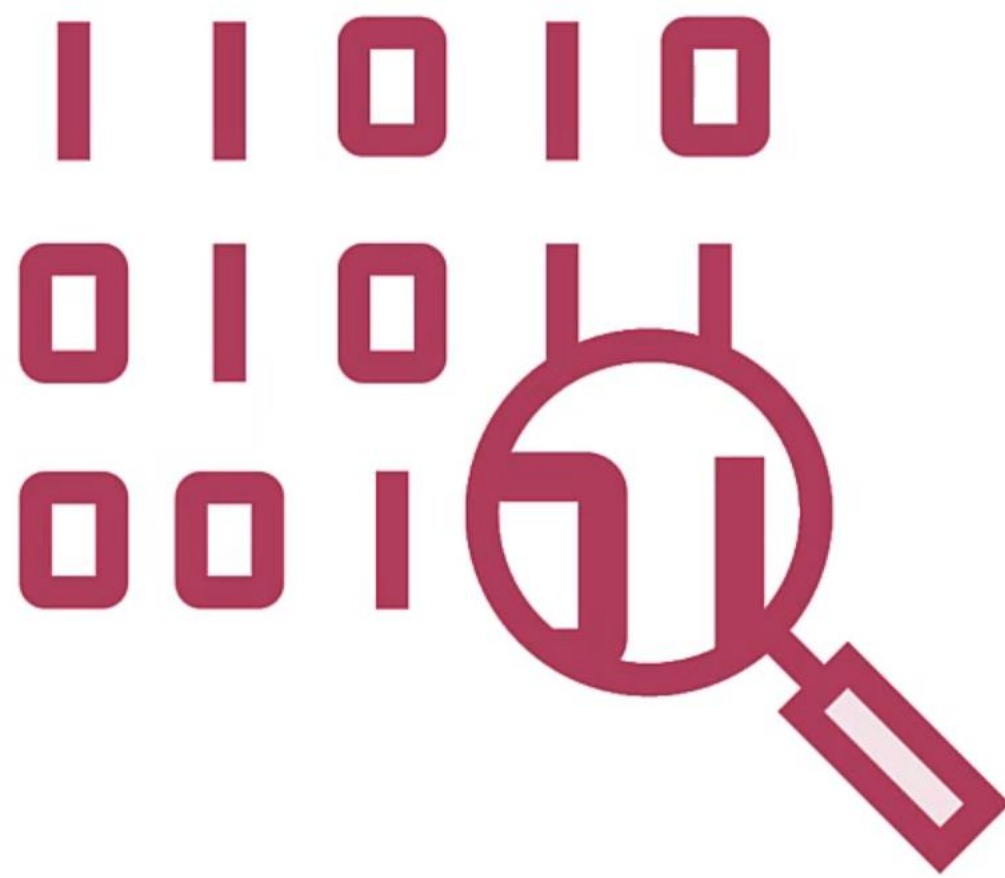
Plan for errors:

- Connection lost
- Partial files
- Corrupt files

Accept-Ranges

The Accept-Ranges response HTTP header is a marker used by the server to advertise its support of partial requests.

How to Auto-resume



HEAD

The HTTP HEAD method requests the headers that would be returned if the HEAD request's URL was instead requested with the HTTP GET method.

This is useful for large file downloads.



Head

File-size

Accept-range

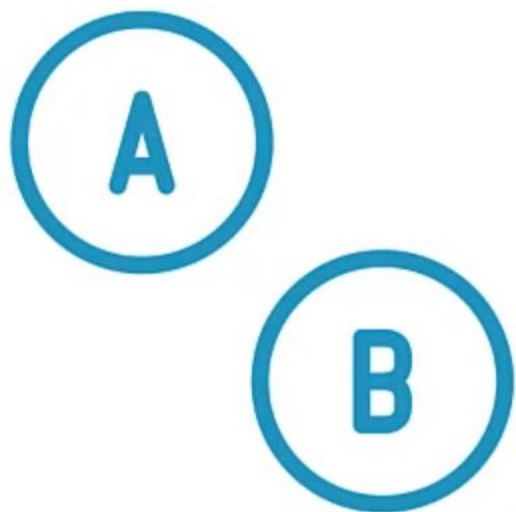
Web Media Tools



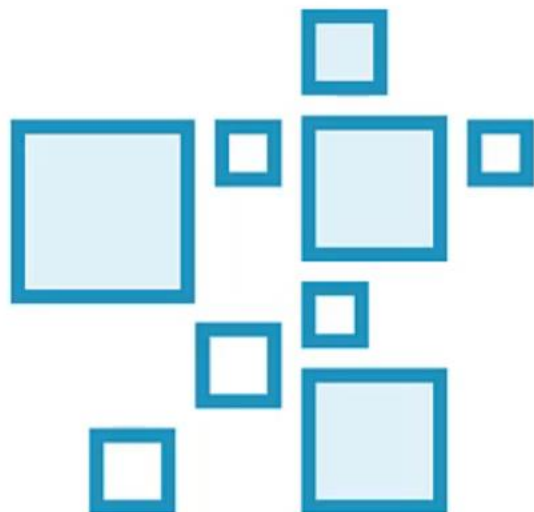
Tools for Web Media



Tools for Web Media



Type



Size



Method

Tools and Methods for Web Media

HTML is readable

Web media is binary

Tools depend on media type

Types of Media



IMAGE: JPG, JPEG, JPE, JFIF



VIDEO: MPE, MPEG, MPG, MP4



AUDIO: WAV, MP3, FLAC, AIFF



DOC: DOC, DOCX, PDF, RTF

Web Media Characteristics

File size

Server resume support

Protected media

Data encoding type