

Scraping Media from the Web with Python

PROCESSING WEB MEDIA FILES



Allen O'Neill

ENGINEER

@databytzai

Overview/ Summary



Converting between media types

Searching within media

Analysing media to extract text

Working with Images



```
pip install Pillow
```



Python Pillow

- Image manipulation
- Convert format

Converting Images



Resizing Images

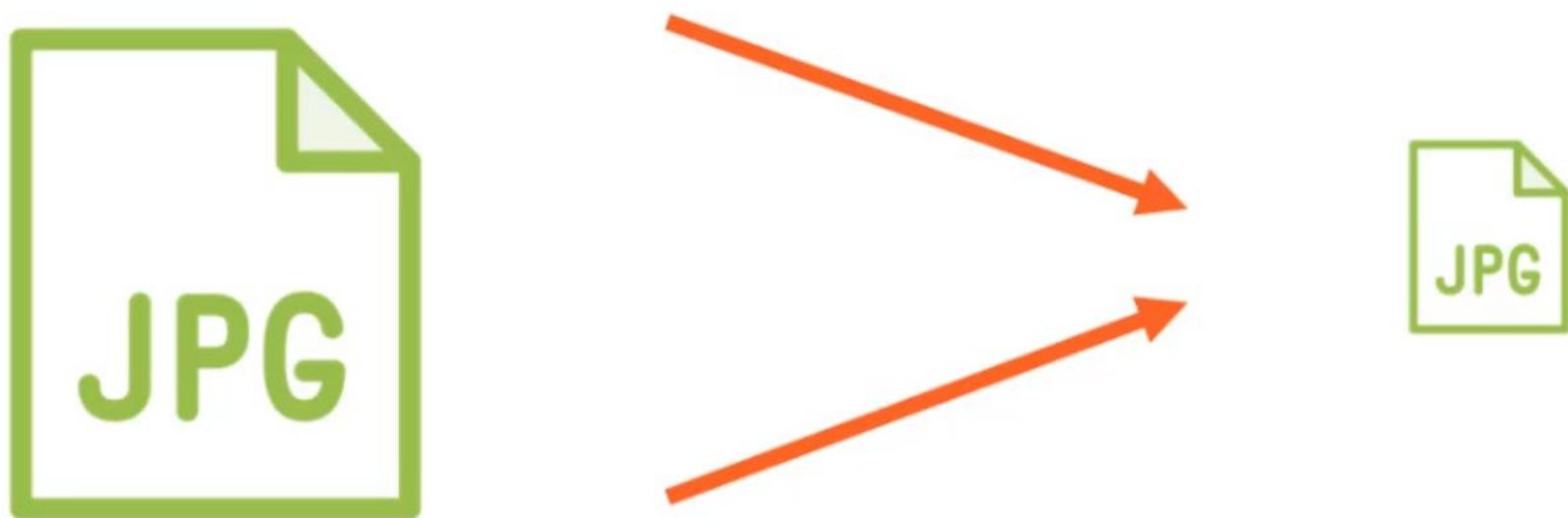


Image EXIF Metadata



- i Copyright
- i Geo-location
- i Encoding
- i Shutter-speed

EXIF data

EXIF = “Exchangeable Image File” – a standard for digital image information. Information stored includes shutter speed, exposure compensation, lenses used and other useful information including GPS location.

Audio processing

Pydub & Matchering

pydub 0.24.1

```
pip install pydub
```



Manipulate audio with an simple and easy high level interface

matchering 2.0.3

```
pip install matchering
```



Audio Matching and Mastering Python Library

Video processing

MoviePy

moviepy 1.0.3

```
pip install moviepy
```



Video editing with Python

Document Media Processing

Document Media Conversion



Adobe Acrobat



Microsoft Word

Extracting Information from Web Media

Data to Information



Raw Format



Process Media



Extract Added Value




Images

- Extracting text information
- Optical character recognition

Tesseract

```
sudo apt-get install tesseract-ocr
```

pytesseract 0.3.7

`pip install pytesseract` 

Python-tesseract is a python wrapper for Google's Tesseract-OCR

Audio to Text



Speech Recognition

SpeechRecognition 3.8.1

```
pip install SpeechRecognition
```



Library for performing speech recognition, with support
for several engines and APIs, online and offline.

Audio from Video



Course Summary



Introduction to web media

Identifying and extracting media links

Downloading web media

Processing different web media