a.k.a.

Web scraping in Python

Tomáš Bedřich, @tbedrich

Links

is.gd/web_scraping_in_python_en (this presentation)

github.com/tomasbedrich/web-scraping-in-python

python-requests.org (get, post, Session, …)

beautiful-soup-4.readthedocs.org (find, find_all)

Enviroment setup

python3

pip install requests beautifulsoup4

DevTools (Chrome / Firefox / Safari)

Outline

DevTools

Getting the data

LEVEL 0: Numbers API

LEVEL 1: NASA astronomy pictures

LEVEL 2: IMDB top actors

LEVEL 3: Color scheme uploader

DevTools

Getting the data

CSS selectors - simple and readable

XPath - another syntax

regular expressions - parsing JS code

LEVEL 0

Try *requests* on Numbers API

numbersapi.com

try plain text and json (*?json*)

properly use *params*

try offline

LEVEL 1

Download some images from NASA Astronomy Picture of the Day archive

apod.nasa.gov

invalid HTML

stream=True when getting image

then iterate over response

LEVEL 2

Parse actors for top movies on IMDB

imdb.com/chart/top

mobile version?

LEVEL 3

Create personal color uploader for Coolors.co

coolors.co

python_scraper, secure_password

inspect ajax calls

use *requests.Session*