

Section 1: Effective Webscraping

Masha Krupenkin¹

April 16, 2019

¹Stanford University

Logistics

Logistics

- ▶ Course website:
https://github.com/justingrimmer/tad_19

Logistics

- ▶ Course website:
`https://github.com/justingrimmer/tad_19`
- ▶ Please sign up for section on Canvas!

Logistics

- ▶ Course website:
`https://github.com/justingrimmer/tad_19`
- ▶ Please sign up for section on Canvas!
- ▶ HW1 due on 4/22!

Scraping

Scraping

- ▶ Urllib package = easy to use for simple scraping tasks

Scraping

- ▶ Urllib package = easy to use for simple scraping tasks
- ▶ BeautifulSoup parses html

Scraping

- ▶ Urllib package = easy to use for simple scraping tasks
- ▶ BeautifulSoup parses html
- ▶ Many scraping tasks are not simple!

Difficult Scraping Tasks

Difficult Scraping Tasks

1. Scraping complicated webpages

Difficult Scraping Tasks

1. Scraping complicated webpages
2. Scraping a large number of webpages

Complex Webpages w/ Selenium

Scraping Complicated Webpages

The image shows a screenshot of a Twitter homepage viewed in a web browser. The browser's address bar displays 'https://twitter.com'. The Twitter interface includes a top navigation bar with links for Home, Moments, Notifications, and Messages. The main content area features a user profile for Masha Krupenkin (@MashaKrupenkin) with 34 tweets, 289 followers, and 6 following. Below the profile is a section for 'Trends for you' listing topics like #TuesdayThoughts, #DontSpoilTheEndgame, #ConcertFail, and Columbine. The central feed shows tweets from Dan Sheehan, Nicole Cliffe, and shauna, along with a 'LIVE Breaking News' alert about President Trump. On the right, there is a promotional banner for Twitter's new features and a 'Who to follow' section with profiles like Philip Bump and Brian Stelter. The Windows taskbar at the bottom shows the search bar and various application icons.

Naive approach...

*untitled-1

```
1 import urllib.request
2 from bs4 import BeautifulSoup
3
4 with urllib.request.urlopen("https://twitter.com/") as response:
5     html = response.read()
6
7 soup = BeautifulSoup(html, "html.parser")
8 print(soup)|
```

Twitter is too smart for this...

Twitter is too smart for this...

- ▶ Urllib package will not give you tweets!

Twitter is too smart for this...

- ▶ Urllib package will not give you tweets!
- ▶ Infinite scrolling poses additional challenge

Twitter is too smart for this...

- ▶ Urllib package will not give you tweets!
- ▶ Infinite scrolling poses additional challenge
- ▶ Ethics: scrape at your own discretion!

Selenium Demo

To the Python!

Selenium Install

`https://selenium-
python.readthedocs.io/installation.html`

Excellent Selenium Tutorial

<https://medium.com/@dawranliou/twitter-scraper-tutorial-with-python-requests-beautifulsoup-and-selenium-part-2-b38d849b07fe>

Selenium Pros & Cons

Selenium Pros & Cons

- ▶ **Pro:** Can scrape webpages that hide content

Selenium Pros & Cons

- ▶ **Pro:** Can scrape webpages that hide content
- ▶ **Pro:** Can fill in forms, searches, etc (NO CAPCHA)

Selenium Pros & Cons

- ▶ **Pro:** Can scrape webpages that hide content
- ▶ **Pro:** Can fill in forms, searches, etc (NO CAPCHA)
- ▶ **Con:** Slow and inefficient

Selenium Pros & Cons

- ▶ **Pro:** Can scrape webpages that hide content
- ▶ **Pro:** Can fill in forms, searches, etc (NO CAPCHA)
- ▶ **Con:** Slow and inefficient
- ▶ **Con:** Potential to miss data

Selenium: Final Verdict

Great for scraping (relatively) small number of highly complex pages

On the other end of the spectrum...

Fast Scraping with Scrapy

Mayor Pete's Website

Pete For America


https://peteforamerica.com

U-SQL Home - Roper Center Google Trends

HOME MEET PETE EVENTS

20 PETE 20

STORE DONATE ENGIES



A FRESH START FOR AMERICA

JOIN TEAM PETE

<input type="text" value="Email address*"/>	<input type="text" value="ZIP code*"/>
<input type="text" value="Cell phone (optional)"/>	<input type="button" value="SUBMIT"/>

Type here to search

03:18 PM 4/16/2019

Scrapy Demo

Command Prompt

Microsoft Windows [Version 10.0.17134.706]

(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\mkurp>cd C:\Users\mkurp\Dropbox\Misc_files\Docs_Grad_School\Teaching\TAD19

C:\Users\mkurp\Dropbox\Misc_files\Docs_Grad_School\Teaching\TAD19>scrapy startproject mayorpete_

Scrapy Demo

Command Prompt

```
Microsoft Windows [Version 10.0.17134.706]  
(c) 2018 Microsoft Corporation. All rights reserved.  
  
C:\Users\mkurp>cd C:\Users\mkurp\Dropbox\Misc_files\Docs_Grad_School\Teaching\TAD19  
  
C:\Users\mkurp\Dropbox\Misc_files\Docs_Grad_School\Teaching\TAD19>cd mayorpete  
  
C:\Users\mkurp\Dropbox\Misc_files\Docs_Grad_School\Teaching\TAD19\mayorpete>scrapy crawl pete
```

Scrapy Spiders

To the Python!

Scrapy Install Guide

`https://docs.scrapy.org/en/latest/intro/install.html`

In-depth Scrapy Tutorial

`https://docs.scrapy.org/en/latest/intro/tutorial.html`

Scrapy Pros & Cons

Scrapy Pros & Cons

- ▶ **Pro:** Incredibly fast Python scraper!

Scrapy Pros & Cons

- ▶ **Pro:** Incredibly fast Python scraper!
- ▶ **Pro:** Can crawl full websites easily

Scrapy Pros & Cons

- ▶ **Pro:** Incredibly fast Python scraper!
- ▶ **Pro:** Can crawl full websites easily
- ▶ **Con:** Too efficient - use `time.sleep()`!

Scrapy Pros & Cons

- ▶ **Pro:** Incredibly fast Python scraper!
- ▶ **Pro:** Can crawl full websites easily
- ▶ **Con:** Too efficient - use `time.sleep()`!
- ▶ **Con:** Cannot scrape complex pages

Scrapy: Final Verdict

Great for bulk scraping

Which scraper should I use?

Which scraper should I use?

- ▶ Choose the correct scraper for the problem

Which scraper should I use?

- ▶ Choose the correct scraper for the problem
- ▶ Urllib good for beginners

Which scraper should I use?

- ▶ Choose the correct scraper for the problem
- ▶ Urllib good for beginners
- ▶ Speed vs complexity tradeoff