Python Class 4: Scrape the web!

Michelle Torres

August 14, 2016

Overview

1. Call the website and open it
2. Extract the html code
3. Retrieve information using the names of the nodes, tags, ids, etc.
4. Store it in lists or directly to CSV files

GETTING TO KNOW THE PAGE SOURCE

```
<!DOCTYPE html>
<html>
<head>
<title>Page Title</title>
</head>
<body>

<h1>My First Heading</h1>
<p>My first paragraph.</p>

</body>
</html>
```

Why do we care about another approach?

- Beautiful Soup and Urllib are “crawlers” that operate in the “background”.

Why do we care about another approach?

- Beautiful Soup and Urllib are “crawlers” that operate in the “background”.
- They are incredibly fast...

Why do we care about another approach?

- Beautiful Soup and Urllib are “crawlers” that operate in the “background”.
- They are incredibly fast...
- ... but also easier to detect and block

Why do we care about another approach?

- Beautiful Soup and Urllib are “crawlers” that operate in the “background”.
- They are incredibly fast...
- ... but also easier to detect and block
- Plus, they sometimes a bit restrictive if you don’t know the exact node or thing to extract

SELENIUM

- Selenium is a “remote driver” of your favorite browser

SELENIUM

- Selenium is a “remote driver” of your favorite browser
- Therefore, you can pretty much simulate behavior of a human “surfing the web”

SELENIUM

- Selenium is a “remote driver” of your favorite browser
- Therefore, you can pretty much simulate behavior of a human “surfing the web”
- With the right tricks, the likelihood of tracking and blocking your “bot” decreases.

SELENIUM

- Selenium is a “remote driver” of your favorite browser
- Therefore, you can pretty much simulate behavior of a human “surfing the web”
- With the right tricks, the likelihood of tracking and blocking your “bot” decreases.
- It also offers flexibility in terms of “unknown” items: you can even look by name of buttons in the page

SELENIUM

- Selenium is a “remote driver” of your favorite browser
- Therefore, you can pretty much simulate behavior of a human “surfing the web”
- With the right tricks, the likelihood of tracking and blocking your “bot” decreases.
- It also offers flexibility in terms of “unknown” items: you can even look by name of buttons in the page
- There are some downsides though...

Selenium

- Selenium is a “remote driver” of your favorite browser
- Therefore, you can pretty much simulate behavior of a human “surfing the web”
- With the right tricks, the likelihood of tracking and blocking your “bot” decreases.
- It also offers flexibility in terms of “unknown” items: you can even look by name of buttons in the page
- There are some downsides though...
  - It is slower

SELENIUM

- Selenium is a “remote driver” of your favorite browser
- Therefore, you can pretty much simulate behavior of a human “surfing the web”
- With the right tricks, the likelihood of tracking and blocking your “bot” decreases.
- It also offers flexibility in terms of “unknown” items: you can even look by name of buttons in the page
- There are some downsides though...
  - It is slower
  - It is dependent on your internet connection quality

SELENIUM

- Selenium is a “remote driver” of your favorite browser
- Therefore, you can pretty much simulate behavior of a human “surfing the web”
- With the right tricks, the likelihood of tracking and blocking your “bot” decreases.
- It also offers flexibility in terms of “unknown” items: you can even look by name of buttons in the page
- There are some downsides though...
  - It is slower
  - It is dependent on your internet connection quality
  - Sometimes it is *very* annoying!

Some tricks and advice to become a pro-scraper

- Google Chrome is better to track nodes and page sources

Some tricks and advice to become a pro-scraper

- Google Chrome is better to track nodes and page sources
- Inspect the source and get to know your document/website!

Some tricks and advice to become a pro-scraper

- Google Chrome is better to track nodes and page sources
- Inspect the source and get to know your document/website!
- Use the 'Copy Xpath' command if you're having troubles

Some tricks and advice to become a pro-scraper

- Google Chrome is better to track nodes and page sources
- Inspect the source and get to know your document/website!
- Use the 'Copy Xpath' command if you're having troubles
- If you have a complex website, visit the websites that store java scripts. They give hints on how information is displayed

Some tricks and advice to become a pro-scraper

- Google Chrome is better to track nodes and page sources
- Inspect the source and get to know your document/website!
- Use the 'Copy Xpath' command if you're having troubles
- If you have a complex website, visit the websites that store java scripts. They give hints on how information is displayed
- Use time breaks to avoid being blocked