



Curtin University

Using the Web

ISYS5002, School of Marketing and Management

ELECTRONIC WARNING NOTICE FOR COPYRIGHT STATUTORY LICENCES

WARNING

This material has been reproduced and communicated to you by or on behalf of **Curtin University** in accordance with section 113P of the *Copyright Act 1968 (the Act)*

The material in this communication may be subject to copyright under the Act. Any further reproduction or communication of this material by you may be the subject of copyright protection under the Act.

Do not remove this notice.



I acknowledge the traditional custodians of the land on which I work and live, and recognise their continuing connection to land, water and community. I pay respect to elders past, present and emerging.

What is Web Scraping



Image Source: <https://i.imgur.com/6zM7JBq.png>

Why is Web Scraping Used?

- Price Comparison
- Email address gathering
- Social Media Scraping
- Research and Development
- Job Listings

Is Web Scraping Legal?

- Copyright law protects original work
 - Fair dealing
- Data not consider original work
- Read Website Terms of Service
- Check robots.txt

Is Python Good for Web Scraping

- Ease of Use
- Large Collection of Libraries
- Dynamically Typed
- Small code large Task
- Community

How Scrape

- Find URL
- Inspect the Page
- Find the data want to extract
- Write the code
- Run code and extract the data
- Store data in the required format

Libraries

- Selenium – automate browser activities
- BeautifulSoup – parsing HTML and XML documents
- Requests – make HTTP request simpler

Anatomy of URL

- Scheme

https://developer.mozilla.org

- Domain

https://developer.mozilla.org/en-US/docs/Learn/

- Path

https://developer.mozilla.org/en-US/docs/Learn/

- Parameters

https://developer.mozilla.org/en-US/search?q=URL

HTML Tags

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

<h1>This is a Heading</h1>
<p>This is a paragraph.</p>

</body>
</html>
```