HTML Primer

Provide a universal language for packaging the layout of content in a document.

[Online HTML Editor]

<tagtype attributes> content </tagtype>

My name is **Derek**

<p style=“font-size: 32px”>My name is <b>Derek</b></p>

<html> surrounds a whole page

<head> contain page meta data

<body> actual content

<div> section

<h1>, <h2>, <h3>, <h4> headers

<p> paragraph

<a> link <a href=”http://www.mcgill.ca”>McGill</a>

<ul> & <ol> lists … <li> items in a list

<img>

<b> & <i> bold/italic

<table> table

Graphical user interface, application

Description automatically generated

A picture containing diagram

Description automatically generated

Web Scraping

0. Decide on the target sits and content

1. download page content

2. extract content e.g. article title, article body, links to other target content…

Downloading

Access the website and download one page of content.

Factors…

• Pretend to be a web browser

• Set User-Agent

• Act like a human

• Avoid exceeding rate limits

• Deal with dynamic pages (sometimes)

Content extraction

Find and extract the desired content from inside the webpage

Site Traversal

• Remember what you’ve visited: Some content we need might be other links we need to gather. Don’t grab the same content multiple times

• Be careful of exponential growth: If every page you visit gives you 2 new pages to collect… your scraper will never finish. (solution: set limits/ sample)

Key tip

• test your scripts at a SMALL scale.

• Add logic to keep the number of pages collected small

• Add print statements to keep track of loops that might have gone crazy

Dev tool

1. Identify the content you want to capture

2. “Inspect” (Dev Tools)

3. Determine the identifying “landmark” elements you can use to programmatically narrow in on the element.

4. Determine the programmatic rules that will allow you to extract what you want.

5. Manually test out the landmarks and rules on a couple other pages

Beautiful soup

Find an element (x.find, x.find\_all)

Grab an element’s parent (x.parent)

Grab an element’s sibling (x.next\_sibling, x.siblings)

wget <url>

Text

Description automatically generated

Text

Description automatically generated