

# The Economics Crime over Time and Space Theory, Practice and Applications

*(A primer in) Webscraping with Python*

Arpita Ghosh

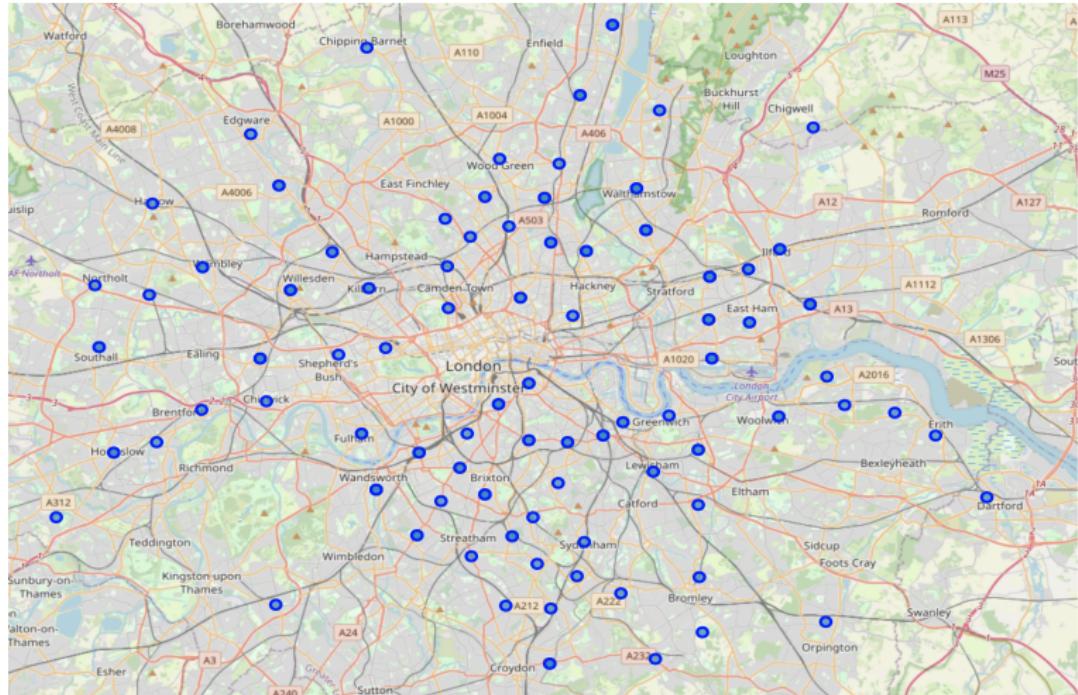
*22<sup>nd</sup> September 2022*

# Why do we care?

- Vast and unused data online - newspapers, blogs, etc
- Quite rich data that can capture outcomes and in cases even treatments
- External validity - opinions, concepts, or even for robustness
- Computing power has become cheap and tools are available to use
- We like having more data!

# At the end of my talk...

we will have mapped the unsolved murders in London (during 2005-2021), getting the data scraped from a website.



# Aim for this session

- ➊ Look at the website murdermap
  - Explore how to understand HTML (a little)
- ➋ We focus on the unsolved murder cases tab
- ➌ Understand how the website is organised and what do we want to obtain
  - Write a script to get the 1st level of data
- ➍ Write a script to actually obtain the raw data of interest

# Assumption/Goal/Declaration

## Goal

We are only interested to extract the date and location of these murders

*There is other info in the websites like name and age, but we ignore them.*

## Assumption

We are not looking to perfect the scraper

*There are ways to make the scraper work more efficiently.*

## Declaration

I do cheat a little at places and don't write a beautiful code

*The aim is to give the knowhow not have a data for research.*

# HTML structure

## Hyper Text Markup Language

Essentially standard language for building web pages; with a series of elements

*Mainly has two parts - HEAD (title of page, HTML version etc) and BODY (everything that the website needs to display)*

*Two type of HTML in scraping websites:*

- Static websites: like Wikipedia or similar - use requests with Scrapy, BeautifulSoup, Selenium etc
- Dynamic websites: newspapers, social media etc. Can be also user specific like need to log in etc. - use Selenium or Splash

**Most websites are dynamic, it's a matter of when it changes.**

HTML structure

# Locators: Xpath and CSS Selector

## Web-element locators

Types of locators: element ID, field name, text, link text, CSS class, **Xpath, CSS selectors**

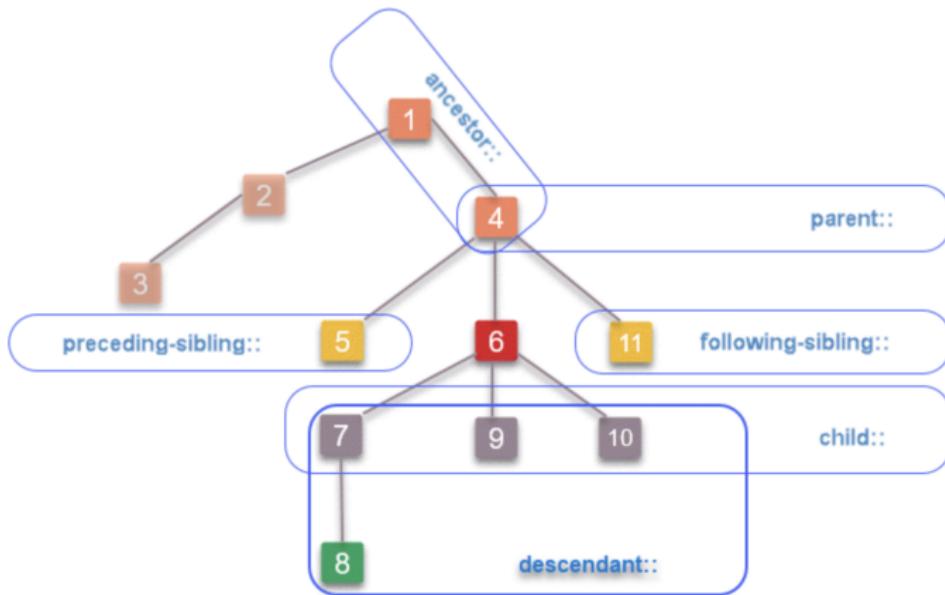
- *XPath*: Used in Selenium to identify web elements uniquely on a web page. Can be absolute or relative. Always better to use relative Xpath (probability of getting it wrong is less).
- *CSS selectors*: Cascading Style Sheets (CSS) is a declarative language to control how websites look in browser - CSS selectors are used to locate HTML elements we want to style. There are different ones like element, id, class selectors. Prefer id as that is unique in an HTML.

*Both of these are most frequently used to locate elements - CSS has better performance and speed than Xpath but Xpath is more versatile (can be used to locate elements up or down). We will not go in more details in this session.*

# Xpath hierarchy

Xpath axes: locating elements relative to known elements

Ancestors, descendant, following, following sibling, preceding, child, parent etc



Source : XPath Expression components. <http://www.iro.umontreal.ca/~lapalme/ForestInsteadOfTheTrees/HTML/ch04s01.html>

Let's look at the website - unsolved tab

On Thursday, 22 June, 1995, Christine McGovern was found strangled in her flat in Walthamstow, east London. The 47 year-old lived alone with her Rottweiler dog at a flat...

I 2 3 4 5 ... 7 8 9 10 Next »

div.wp-block-uagb-post-g  
rid.uagb-post-grid.uagb-p  
ost\_\_image-position-back  
ground.u...

667 x 3442.91

## Unsolved murders in London: 2005

These 13 murders in London in 2005 remain...

## Unsolved Murders in London: 2006

These 12 cases of homicide in London remain...

Elements Console Sources Network Performance Memory > 1

```
<header class="entry-header">
  <h1 class="entry-title"></h1>
</header>
<div class="entry-content">
  ::before
  <div></div>
  <div class="wp-block-uagb-post-grid uagb-post_grid uagb-post__image-position-enabled uagb-block-55474a28 items-uagb-post__columns-2 is-grid uagb-post__columns-tablet-2 uagb-post__columns-mobile-1 uagb-post__equal-height" data-total="13"></div>
<h2></h2>
<div class="wp-block-uagb-post-grid uagb-post_grid uagb-post__image-position-disabled uagb-block-15678afe items-uagb-post__columns-1 is-grid uagb-post__columns-tablet-2 uagb-post__columns-mobile-1 uagb-post__equal-height" data-total="1"></div>
<article class="uagb-post_inner-wrap">
  <h3 class="uagb-post_title uagb-post_text">=</h3>
  <a href="https://www.murdermap.co.uk/unsolved-murders/unsolved-murders-london-2005/" target="_self" rel="bookmark noopener noreferrer"> Unsolved murders in London: 2005</a>
</div>
<div class="uagb-post_text uagb-post-grid-byline"></div>
<div class="uagb-post_text uagb-post_excerpt"> These 13 murders in London in 2005 remain...</div>
</article>
<article class="uagb-post_inner-wrap"></article>
<article class="uagb-post_inner-wrap"></article>
<article class="uagb-post_inner-wrap"></article>
<article class="uagb-post_inner-wrap"></article>
<article class="uagb-post_inner-wrap"></article>
<article class="uagb-post_inner-wrap" data-bbox="113 688 906 704" data-label="Text"><div></div>
```

*Open the murdermap website in Google and scroll down to the cases by year, right click on an element and select inspect.*

# Website structure - two cases: 2005, 2011



Benjamin Onwuka, 24, was shot in the head in Maxilla Walk, Harlesden, on 2 January 2005. He died a short time later in hospital.

Four men were arrested and a £10,000 reward for information was offered but nobody has ever been charged.

Contact Crimestoppers on 0800 555 111.

The screenshot shows two separate sections of a website. The left section is for the murder of Trevor Ellis, featuring a large photograph of him, his name, and a brief description: "Trevor Ellis, 24, was shot in the head in Maxilla Walk, Harlesden, on 2 January 2005. He died a short time later in hospital." Below this is a paragraph about four arrests and a £10,000 reward. The right section is for the murder of Ronnie Khan, featuring a large photograph of him, his name, and a brief description: "Ronnie Khan, 24, was stabbed to death in southeast London, on 12 February 2011. He had spent the afternoon watching football with friends before getting the 9.17pm train from Bickley to Penge East to see his girlfriend. On arrival at the station just after 9.30pm, crossed over the footbridge towards the platform and saw a man carrying a rucksack, Samuel, who was carrying a pink T-mobile carrier bag. Samuel last seen at the bus stop opposite Bailey Place at about 9.30pm. Ten minutes later he was found slumped over a bench in a nearby park. The phone number which was not connected to Samuel or anyone he knew and has never been activated. Seconds later the phone was used to dial 9999 and the call was received at 9.48pm. The police were called and found the victim with Bailey Place. His wallet had been stolen but was found to have £100 in it. It was pronounced dead an hour later. A postmortem was carried out." Both sections include a "View Largest Contentful Paint (LCP) details" button.

The screenshot shows the same two sections of the website as the previous one, but with the browser's developer tools open. The left section's HTML is as follows:

```
<div class="row">
    <div class="col-6">
        <img alt="Photograph of Trevor Ellis" data-bbox="148 268 308 398"/>
        <div>
            <strong>Trevor Ellis</strong>
            <p>Trevor Ellis, 24, was shot in the head in Maxilla Walk, Harlesden, on 2 January 2005. He died a short time later in hospital.</p>
            <p>Four men were arrested and a £10,000 reward for information was offered but nobody has ever been charged.</p>
            <p>Contact Crimestoppers on 0800 555 111.</p>
        </div>
    </div>
    <div class="col-6">
        <img alt="Photograph of Ronnie Khan" data-bbox="528 268 744 398"/>
        <div>
            <strong>Ronnie Khan</strong>
            <p>Ronnie Khan, 24, was stabbed to death in southeast London, on 12 February 2011. He had spent the afternoon watching football with friends before getting the 9.17pm train from Bickley to Penge East to see his girlfriend. On arrival at the station just after 9.30pm, crossed over the footbridge towards the platform and saw a man carrying a rucksack, Samuel, who was carrying a pink T-mobile carrier bag. Samuel last seen at the bus stop opposite Bailey Place at about 9.30pm. Ten minutes later he was found slumped over a bench in a nearby park. The phone number which was not connected to Samuel or anyone he knew and has never been activated. Seconds later the phone was used to dial 9999 and the call was received at 9.48pm. The police were called and found the victim with Bailey Place. His wallet had been stolen but was found to have £100 in it. It was pronounced dead an hour later. A postmortem was carried out.</p>
        </div>
    </div>
</div>
```

The right section's HTML is as follows:

```
<div class="row">
    <div class="col-6">
        <img alt="Photograph of Trevor Ellis" data-bbox="148 268 308 398"/>
        <div>
            <strong>Trevor Ellis</strong>
            <p>Trevor Ellis, 24, was shot in the head in Maxilla Walk, Harlesden, on 2 January 2005. He died a short time later in hospital.</p>
            <p>Four men were arrested and a £10,000 reward for information was offered but nobody has ever been charged.</p>
            <p>Contact Crimestoppers on 0800 555 111.</p>
        </div>
    </div>
    <div class="col-6">
        <img alt="Photograph of Ronnie Khan" data-bbox="528 268 744 398"/>
        <div>
            <strong>Ronnie Khan</strong>
            <p>Ronnie Khan, 24, was stabbed to death in southeast London, on 12 February 2011. He had spent the afternoon watching football with friends before getting the 9.17pm train from Bickley to Penge East to see his girlfriend. On arrival at the station just after 9.30pm, crossed over the footbridge towards the platform and saw a man carrying a rucksack, Samuel, who was carrying a pink T-mobile carrier bag. Samuel last seen at the bus stop opposite Bailey Place at about 9.30pm. Ten minutes later he was found slumped over a bench in a nearby park. The phone number which was not connected to Samuel or anyone he knew and has never been activated. Seconds later the phone was used to dial 9999 and the call was received at 9.48pm. The police were called and found the victim with Bailey Place. His wallet had been stolen but was found to have £100 in it. It was pronounced dead an hour later. A postmortem was carried out.</p>
        </div>
    </div>
</div>
```

*First look at the url links: they differ widely, not just the years.  
Second look at the way the paragraphs are arranged: sometimes with separators after the picture block and sometimes not! Messy website!*

# Some important things

## API vs Scraping

Both gives access to data on websites: if publicly available, use API (Application Programming Interface). But can be expensive, limited, or might not even exist. In that case can build essentially own API by web scraping.

- Ethics: read note of terms of use; create new data, not duplicate it; reasonable queries; providing User Agent string with clear intentions/contact (easier with requests, more complicated in Selenium) etc.
- Use of VPNs/proxy: allows crawling websites with reduced chances of getting blocked
- Imitate human actions: Using random time delays between commands or actions, having user agent etc.

# Extracting URLs

- We need to get the target urls first from the unsolved tab in murdermap.
  - <https://www.murdermap.co.uk/unsolved-murders/unsolved-homicides-2012/> or
  - <https://www.murdermap.co.uk/unsolved-murders/londons-unsolved-murders-2010/> or
  - <https://www.murdermap.co.uk/unsolved-murders/unsolved-murders-london-2005/> etc
- Good to know how to do in stages, as it can be used in scraping other sites, newspapers - like BBC, Wayback Machine etc

# Extracting URLs - the code

murdermap.co.uk/unsolved/

On Thursday, 22 June, 1995, Christine McGovern was found strangled in her flat in Walthamstow, east London. The 47 year-old lived alone with her Rottweiler dog at a flat...

1 2 3 4 5 ... 7 8 9 10 Next »

## Cases by Year

### Unsolved murders in London: 2005

These 13 murders in London in 2005 remain...

### Unsolved Murders in London: 2006

*Write a loop to extract the href attributes from the elements with the given Xpath: 17 years - all urls*

# Extracting Data

- Target: dates and locations of unsolved murders in London.
- Looking at the texts - mainly the first 2 paragraphs are of interest.
- Have the date and location along with name, age, how etc.
- Sometimes the 1st case of a year start after 3 dots following the picture block.
- Sometimes they start without any separator.
- Messy structure  $\Rightarrow$  nested loops and exceptions handling

# Extracting Data - the code

Oktay Erbasi

Daniel Duke

Philip Silvester

Victims in unsolved murder cases in the year 2009

Eighty-one year-old Molly Morgan died after being mugged in the street on January 15, 2009.

Molly was on her way to Kenton Library to hear a lecture on 'Buildings of London' when she was attacked in Streatham Road, Harrow, at 7.40pm.

Her handbag was pulled so violently that she fell and suffered severe head injuries and a broken left arm.

Elements Console Sources Network Performance Memory

... standard.hentry.category-unsolved-murders.tag-260.tag-unsolved-by-year.entry .entry-content p ...

.entry-content > p:nth-child(3)

Console What's New

Highlights from the Chrome 105 update

Step-by-step replay in the Recorder panel

Set a breakpoint and replay a recording step by step in

*Split in types: when there is a separator in the beginning (Type 1) extract 2 'p' sibling texts, if there are separators in the page but not before the 1st case (Type 2) extract the 2 sibling where separators are and the 1st case as well, if there is no separators at all (Type 3) extract paragraph texts as cases.*

# Wrapping up

- Observed how to extract some text data from a messy website.
- Words of caution:
  - must be careful with number of attempts
  - Scraper.py can be improved to count how many paragraphs are there for each case and extract 2 or 1 from it
  - entire code is specific to this case - NOT a generic scraper

```
<html>
```

```
  <head>
```

```
    <title>Page title</title>
```

```
  </head>
```

```
<body>
```

```
    <h1>This is a heading</h1>
```

```
    <p>This is a paragraph.</p>
```

```
    <p>This is another paragraph.</p>
```

```
  </body>
```

```
</html>
```

Back