UK Data Service

Web Scraping with Google Chrome

Anran Zhao

Research Associate at UK Data Service

Email: anran.zhao@manchester.ac.uk

Economic and Social Research Council

BY-NC-SA

Created by [Organisation], UK Data Service. Copyright © [year] [Institution].

Outline

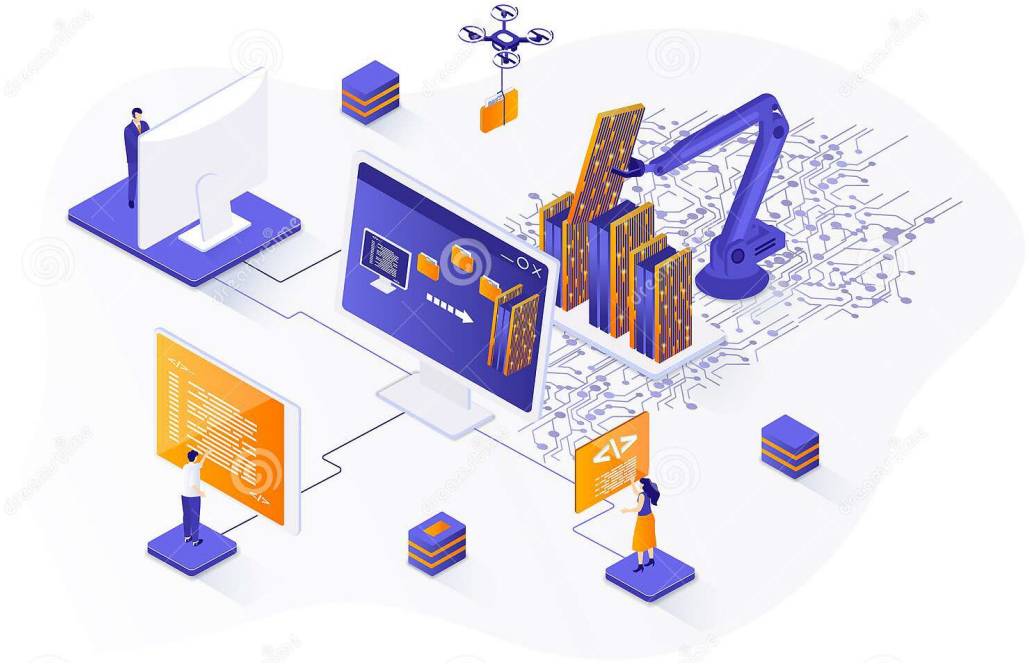
- Introduction (what, why, challenges, tools, etc.)
- Demonstration
  - Scraping one feature
  - Scraping multiple features
  - Scraping links and images
  - Handling ‘next’ pagination
- Q&A

What Is Web Scraping?

- Web scraping is the process of gathering information from the Internet.
  - Copy and paste
  - Automated tools

Why Scrape the Web?

- Web scraping is the process of gathering information from the Internet.
  - Copy and paste
  - Automated tools

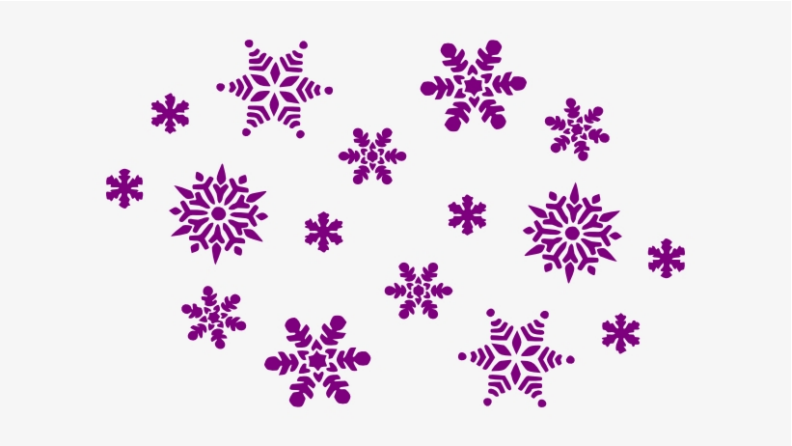


Challenges of Web Scraping

- **Variety**
- Durability

Challenges of Web Scraping

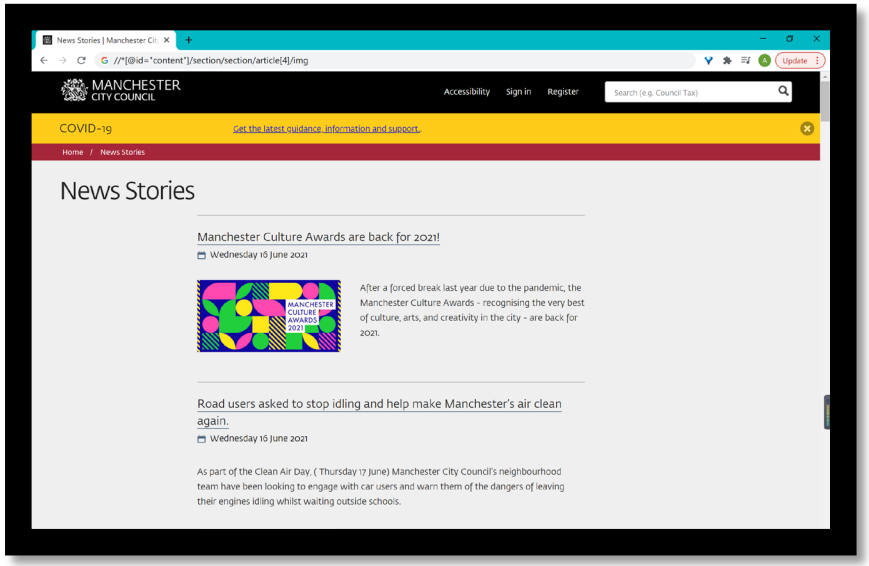
- **Variety**
- Durability



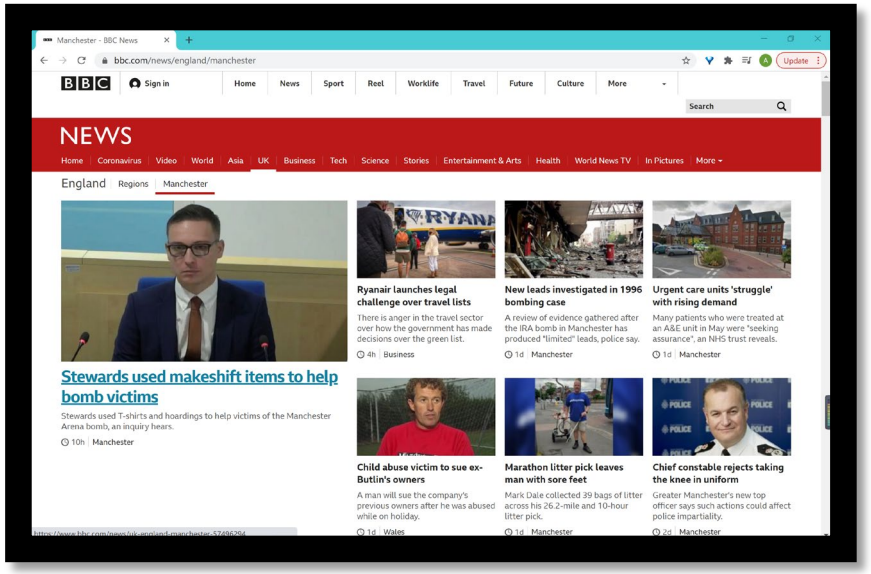
EVERY PAGE IS SPECIAL

Challenges of Web Scraping

- **Variety**
- Durability

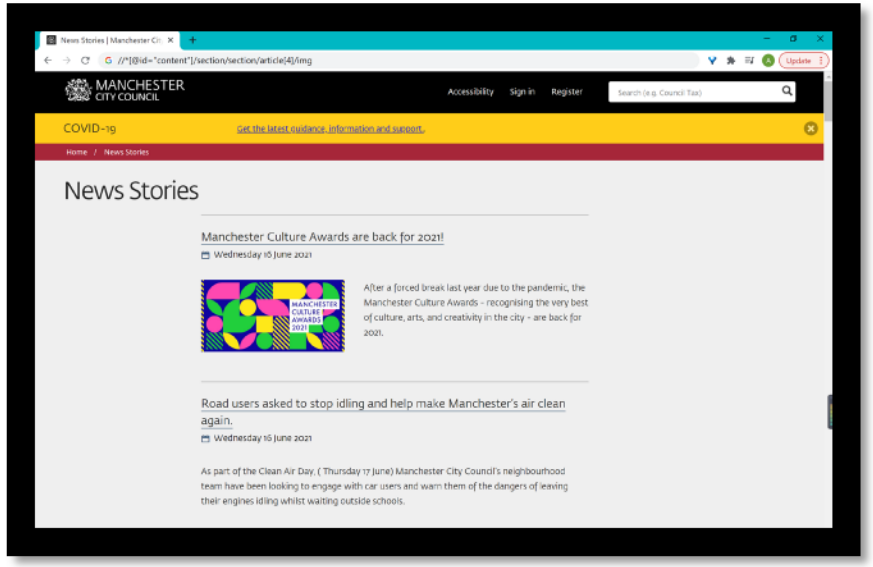


≠



Challenges of Web Scraping

- Variety
- **Durability**



+ □

Challenges of Web Scraping

- ❄Variety
- □ Durability

Common Ways and Tools

- Programming packages: Scrapy, Beautiful Soup, etc.
- API: Facebook, Twitter, etc.
- Software: *ParseHub*, *Dexi*, *Octoparse*, etc.

Resources on Web Scraper

- Tutorial
- Documentation
- Discussion forum

Scraping MCR City Council Website

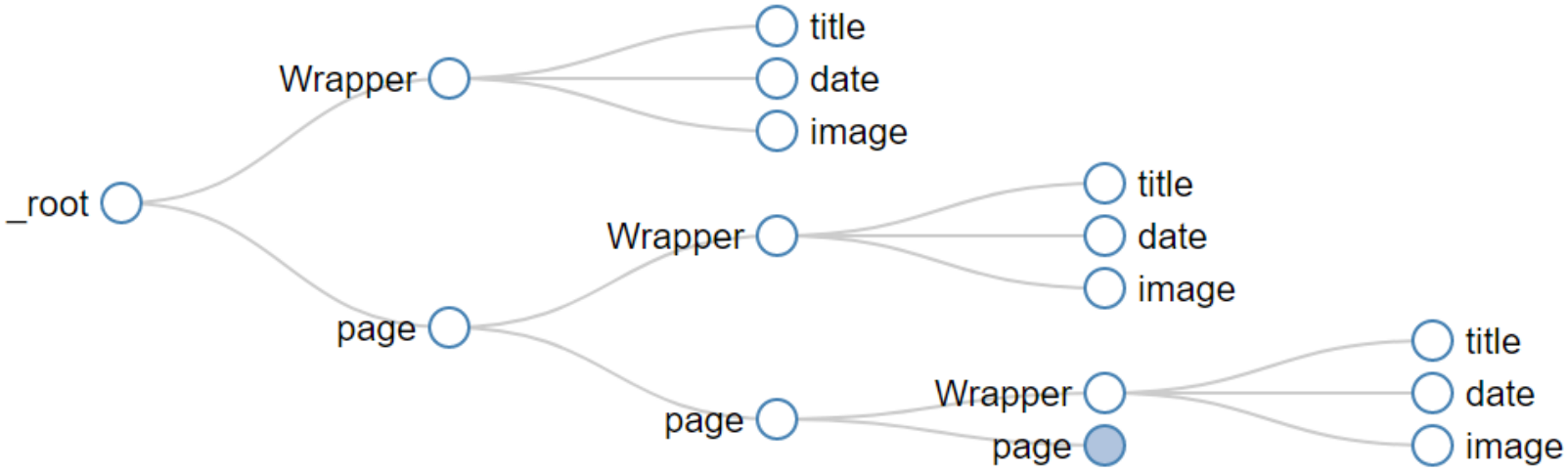
- In this tutorial, you'll build a web scraper that fetches news catalogue from the 'New Stories' page on Manchester City Council site. Your web scraper will pick out the relevant pieces of information and store them in csv file.
- You can scrape any site on the Internet that you can look at, but the difficulty of doing so depends on the site. This tutorial offers you an introduction to web scraping to help you understand the overall process. Then, you can apply this same process for almost every website you'll want to scrape.

Practice

Prerequisite: Chrome Browser

Pagination

- Variety manners of loading more data depending on the website design, E.g.:
- Simple number pagination
- scroll down to load more
- click ‘more’ or ‘next’
- same URL vs multiple URLs


_root
Wrapper
title
date
image
page
Wrapper
title
date
image
page
Wrapper
title
date
image
page

UK Data Service

Thank you

Created by UK Data Archive, UK Data Service. Copyright © 2021 University of Essex.

Economic and Social Research Council

BY-NC-SA