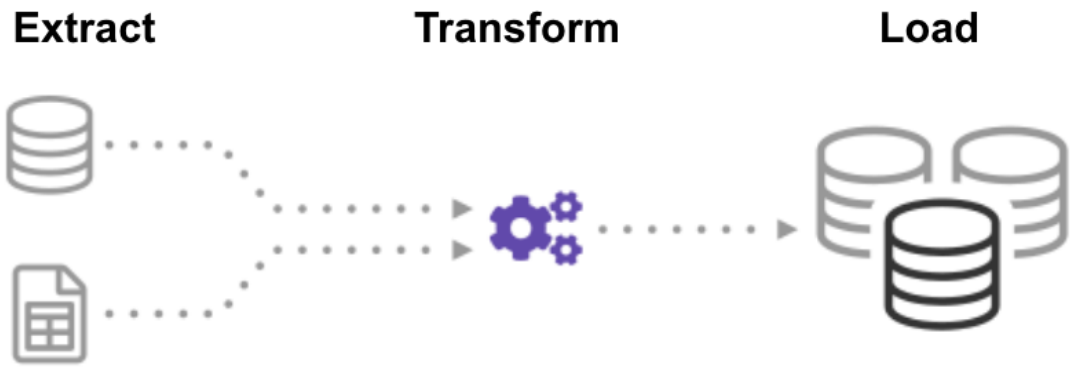
Frontrunner program

Week 1

7-8 February 2022

Extract, Transfer, Load (ETL)

- ETL is a data engineers process to
  - **Extract** data from different sources
    - Sensor data
    - Databases
    - APIs, webpages
  - **Transfer** the data into a usable format
  - **Load** data into the system that end-user can access



ETL

- Here we want to focus on the first step of the ETL pipeline: Extract
  - Web scraping: Collect data from the web

Build web crawler in Python

1. Download the HTML (web content) from a URL
    - Requests
2. Parse HTML to extract information (Link, text, image)

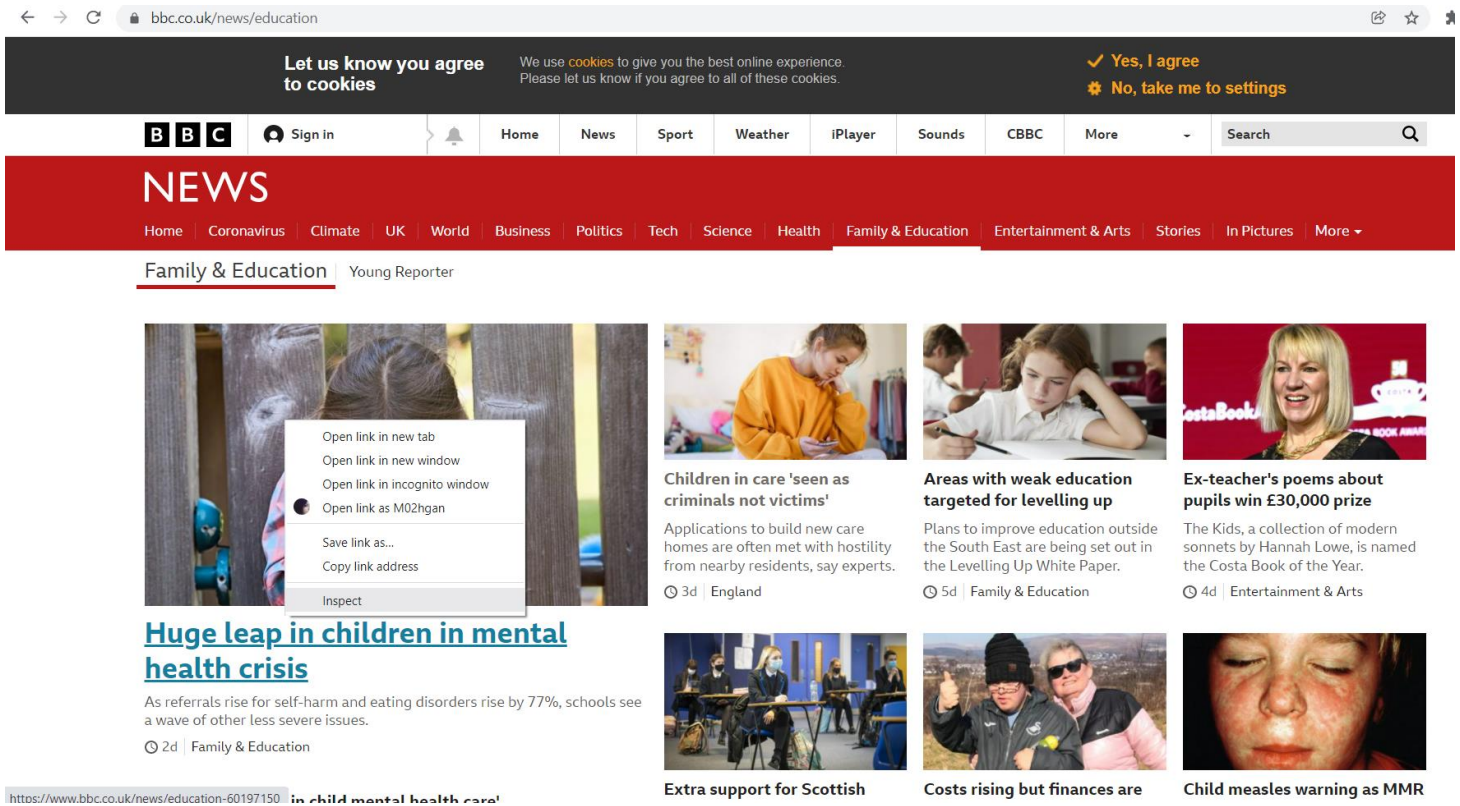
    - Beautiful Soup: is a Python library for pulling data out of HTML and XML files
        - Source: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Parse HTML file

```
# <html>
#  <head>
#   <title>
#    The Dormouse's story
#   </title>
#  </head>
#  <body>
#   <p class="title">
#    <b>
#     The Dormouse's story
#    </b>
#   </p>
#   <p class="story">
#    Once upon a time there were three little sisters; and their names were
#    <a class="sister" href="http://example.com/elsie" id="link1">
#     Elsie
#    </a>
#    ,
#    <a class="sister" href="http://example.com/lacie" id="link2">
#     Lacie
#    </a>
#    and
#    <a class="sister" href="http://example.com/tillie" id="link3">
#     Tillie
#    </a>
#    ; and they lived at the bottom of a well.
#   </p>
#   <p class="story">
#    ...
#   </p>
#  </body>
# </html>
```

Inspect

- Inspect indicate HTML code behind the page
- HTML code indicates the web content
- Web content
  - Textual data
  - Image
  - Link to another webpage



Web inspect example

He's drawn himself "feeling sad because I feel like I'm trapped in the telly", while his grandma tries "to scare Covid away".

Pupils at Seascape Primary School have been taking part in mental health workshops

Isaac's drawing is a task as part of a mental health workshop he's in with other Year 3 and Year 4 pupils at Seascape Primary School in Peterlee, County Durham.

Grace, 8, remembers her parents telling her about the pandemic beginning.

She's drawn herself "curled in the corner" of a prison cell, "because I felt trapped in lockdown".

- **'We've lost faith in child mental health care'**
- **Child mental health not improved since lockdown**

Could a greener asthma inhaler help you breathe better?

Why do energy prices have to go up? And other questions

```
<figure class="ssrcss-wpgbih-StyledFigure e34k3c23">
  <div class="ssrcss-ab5fd8-StyledFigureContainer e34k3c21">
    <span class="ssrcss-1hq4gmv-Placeholder e16icw910">
      <span>
        <img alt="Isaac and Grace" srcset="https://ichef.bbci.co.uk/news/240/cpsprodpb/E665/production/_123118985_ti4igitw.jpg 240w, https://ichef.bbci.co.uk/news/320/cpsprodpb/E665/production/_123118985_ti4igitw.jpg 320w, https://ichef.bbci.co.uk/news/480/cpsprodpb/E665/production/_123118985_ti4igitw.jpg 480w, https://ichef.bbci.co.uk/news/624/cpsprodpb/E665/production/_123118985_ti4igitw.jpg 624w, https://ichef.bbci.co.uk/news/800/cpsprodpb/E665/production/_123118985_ti4igitw.jpg 800w, https://ichef.bbci.co.uk/news/976/cpsprodpb/E665/production/_123118985_ti4igitw.jpg 976w" src="https://ichef.bbci.co.uk/news/976/cpsprodpb/E665/production/_123118985_ti4igitw.jpg" width="976" height="549" loading="lazy" class="ssrcss-1drmwog-Image ee0ct7c0"> flex == $0
        <noscript></noscript>
      </span>
    </span>
  </div>
```

span.ssrcss-1hq4gmv-Placeholder.e16icw910 span img.ssrcss-1drmwog-Image.ee0ct7c0

Styles | Computed | Layout | Event Listeners | DOM Breakpoints | Properties | Accessibility

Filter :hov .cls

```
element.style {
}
.ssrcss-1hq4gmv-Placeholder img {                    <style>
  overflow: hidden;
  position: absolute;
  inset: 0px;
  display: flex;
  -webkit-box-pack: center;
  justify-content: center;
  -webkit-box-align: center;
  align-items: center;
  width: 100%;
  height: 100%;
  object-fit: cover;
}
.ssrcss-1drmwog-Image {                  education-60197150:24
```

Source : https://www.bbc.co.uk/news/education-60197150

First week activity

- Get familiar with HTML file
- Try to parse HTML files using Beautiful Soup
- Try to download and parse any website HTML code using request in python

- It’s nice to:
  - Create a document from what you have learned in the first week
  - Store your code in GitHub, so you can share it with others. It is also good evidence that shows you coding skills
    - No worries if you think you are not good at coding, you can start from today ☺

Useful link

- Beautiful Soup

- Useful Examples :
  - Link
  - Link

- Feel free to send me an email if you have any question: mt19258@essex.ac.uk