Table of Contents

1. Objectives

The main objective of the website is providing a place that users can search and approach up-to-date information about Covid 19 from the reliable online Vietnamese newspapers including Tuoi Tre, Dan Tri and VnExpress news. The information includes the quantity of total confirmed cases, confirmed deaths and confirmed being cured in Vietnam and around the world, which are released officially by the WHO. Besides, users can also approach news related to Covid 19 to get further information about the situation of Corona Pandemic in the country and over the world.

2. Elements

- Data crawler
- Web interface

3. Process

3.1. Data crawler

- Crawler was built with python
- Library: Scrapy
- How can it work:
  - Get HTML response from URL
    - Target to the page that we want to crawl data by using
      def start_requests(self)

  - Save images, captions and links to the news
    - Get images' source:

      for i in range(1, len(response.css("img::attr(src)").extract())):
      source.append(response.css("img::attr(src)").extract()[i])

    - Get captions' source:
      for i in range(1, len(response.css("img::attr(alt)").extract())):
      caption.append(response.css("img::attr(alt)").extract()[i])
    - Get links to the news:
      for i in range(len(response.css("a.news-item__avatar::attr(href)").extract())):
      news_source.append("https://dantri.com.vn" + str(response.css("a.news-item__avatar::attr(href)").extract()[i]))

  - Store crawled data locally (VnExpress for example):
    - for i in range(0, len(source)):
      vnexpress_covid['img'].append({
      "source": source[i],
      "caption": caption[i],
      "news_source": news_source[i] })

```
with open(f"{JSON_directory}image_vnexpress.json", "w") as image_file:
    json.dump(vnexpress_covid, image_file)
```

3.2. Design

- Bootstrap
- Custom CSS style

3.3. How to run the project:

- Clone repo of the project to your device by
- git clone https://github.com/namluu25/covidcrawl

- Copy the project folder to htdocs folder
- Crawl the latest data and copy all .json files into the root of the project folder
- Turn on Xampp
- Access localhost/covidcrawl-main

3.4. How to crawl newest data:

- Open Terminal from the project folder
- Target to covidcrawl/spiders
- Type

python3 covid.py

- Inside spiders folder, crawled data is inside JSON Image file, html page is in HTML File folder

4. Difficulties and limitations: VnExpress has some encoded imagines which are unable to get and decode