Hướng Dẫn Cào Dữ Liệu Từ Trang VNEXPRESS

I. Cài thư viện cần thiết

- Cài đặt Requests: pip install requests
- Cài đặt Pilow: pip install Pillow
- Cài đặt beautifulsoup4 : pip install beautifulsoup4

II. Cào dữ liệu lấy tin tức về thể thao

1. Lấy dữ liệu

- Dữ liệu lấy sẽ là dữ liệu thể thao của trang vnexpress vì vậy chúng ta sẽ vào trang : https://vnexpress.net/the-thao

- Tiếp theo ta sẽ lấy dữ liệu từ trang này qua Beautiful Soup. Thư viện này sẽ giúp chúng ta phân tách dữ liệu HTML hay XML thành dữ liệu cây, giúp cho dữ liệu gọn gàng hơn :

```
import requests
from bs4 import BeautifulSoup

response = requests.get("https://vnexpress.net/the-thao")
soup = BeautifulSoup(response.content, "html.parser")
print(soup)
```

- Kết quả chúng ta sẽ được trang html.

2. Phân tích dữ liệu

- Tiếp theo chúng ta sẽ phân tích dữ liệu cần lấy. Để xem chi tiết dữ liệu của một bài báo, chúng ta mở https://vnexpress.net/the-thao và ta nhấn F12.
- Mục tiêu chúng ta là tìm ra link của bài báo để trích xuất dữ liệu của từng bài báo. Sau khi phân tích ta thấy rằng link của bài báo nằm trong thẻ <a> và thẻ này nằm trong thẻ h3 và có class là “tilte_news”. Chúng ta sẽ tiến hành lấy link như sau:

```
titles = soup.findAll('h3', class_='title-news')
print(titles)
links = [link.find('a').attrs["href"] for link in titles]
print(links)
```

- Kết quả chúng ta sẽ được danh sách các link của bài báo

```
['https://vnexpress.net/ciro-immobile-va-chiec-giay-vang-kem-hao-nhoang-4139575.html', 'https://vnexpress.net/lampard-chang-ai-nho-doi-a-quan-cup-fa-4139548.html', 'https://vnexpress.net/pirlo-tu-choi-ngoai-hang-anh-de-lam-hlv-u23-juventus-4139649.html']
```

3. Lấy dữ liệu chi tiết từng bài

- Sau khi ta đã có link, chúng ta sẽ tiến hành lấy dữ liệu.

```
for link in links:
    news = requests.get(link)
    print(news)
    soup = BeautifulSoup(news.content, "html.parser")
    print(soup)
    title = soup.find("h1", class_="title-detail")
    print("Tiêu đề: " +title.text)
    abstract = soup.find("p", class_="description")
    print("Mô tả: " +abstract.text)
    body = soup.find("p", class_="Normal")
    print("Nội dung: " +body.text)
```

- Bước đầu tiên chúng ta dùng dòng lặp for duyệt qua từng link, sau đó với mỗi link chúng ta sẽ dùng Beautiful Soup như trên và tiếp tục vào link của bài báo để phân tích. Ví dụ như : tiêu đề của bài báo nằm trong thẻ “h1” và nằm

trong class “title-detail”. Ở đây chúng ta sẽ lấy tiêu đề, mô tả, nội dung.

- Cuối cùng chúng ta sẽ lưu dữ liệu cào được vào file json. Chúng ta sẽ tận dụng thuộc tính __dict__ để xử lý dữ liệu như sau (code hoàn chỉnh để cào dữ liệu):

```
import requests
from bs4 import BeautifulSoup
import json
import pandas as pd


response = requests.get("https://vnexpress.net/the-thao")
soup = BeautifulSoup(response.content, "html.parser")
#print(soup)


titles = soup.findAll('h3', class_='title-news')
#print(titles)


links = [link.find('a').attrs["href"] for link in titles]
#print(links)


class Insert (object):


    def __init__(self, tilte, abstract, body):
        self.tilte = tilte
        self.abstract = abstract
        self.body= body
```

```
last_j=[]
for link in links:
    news = requests.get(link)
    #print(news)
    soup = BeautifulSoup(news.content, "html.parser")
    #print(soup)
    tilte = soup.find("h1", class_="title-detail")
    #print("Tiêu đề: " +tilte.text)
    abstract = soup.find("p", class_="description")
    #print("Mô tả: " +abstract.text)
    body = soup.find("p", class_="Normal")
    #print("Nội dung: " +body.text)
    #dùng thuộc tính dict
    insert = Insert(tilte.text,abstract.text,body.text)
    last_j.append(insert.__dict__)

# Lưu dữ liệu vào file
with open('data.json', 'w', encoding='utf-8') as file:
    json.dump(last_j, file)
```

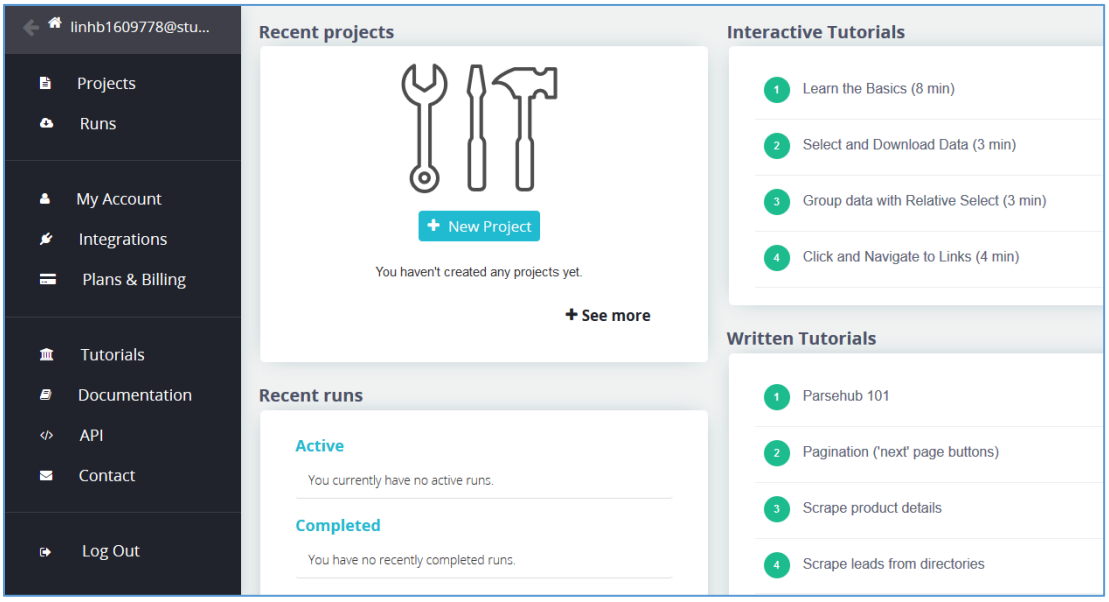
III. Cào dữ liệu với ParseHub

1. Giới thiệu ParseHub

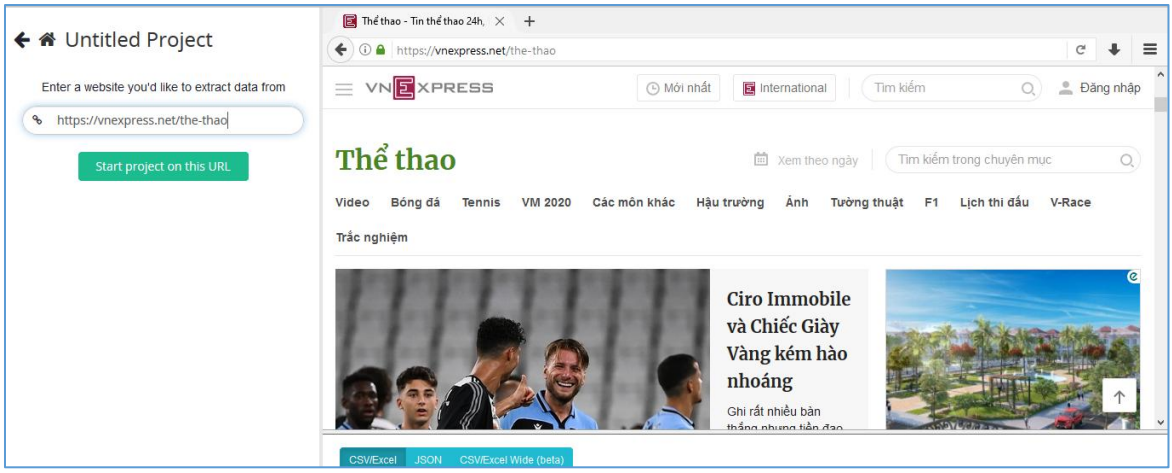
- Phần hướng dẫn bên trên là cách cào dữ liệu "truyền thống". Công cụ sau sẽ hỗ trợ chúng ta việc cào dữ liệu dễ dàng hơn.

2. Chi tiết cào với ParseHub

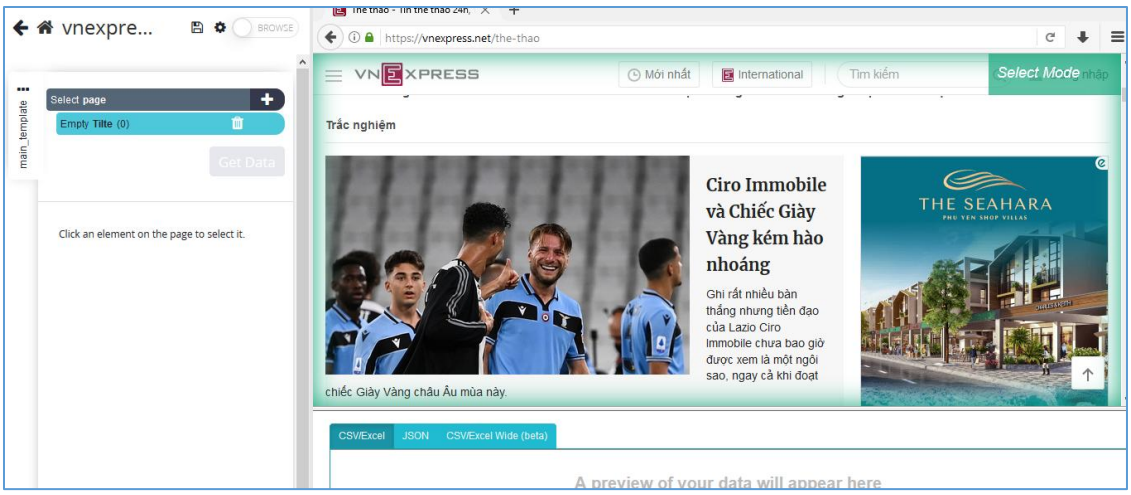
- Tải ParseHub : https://www.parsehub.com/
- Đầu tiên để sử dụng được công cụ này, chúng ta cần đăng kí tài khoản qua link: https://www.parsehub.com/
- Như vậy chúng ta sẽ có giao diện của ParseHub như sau:



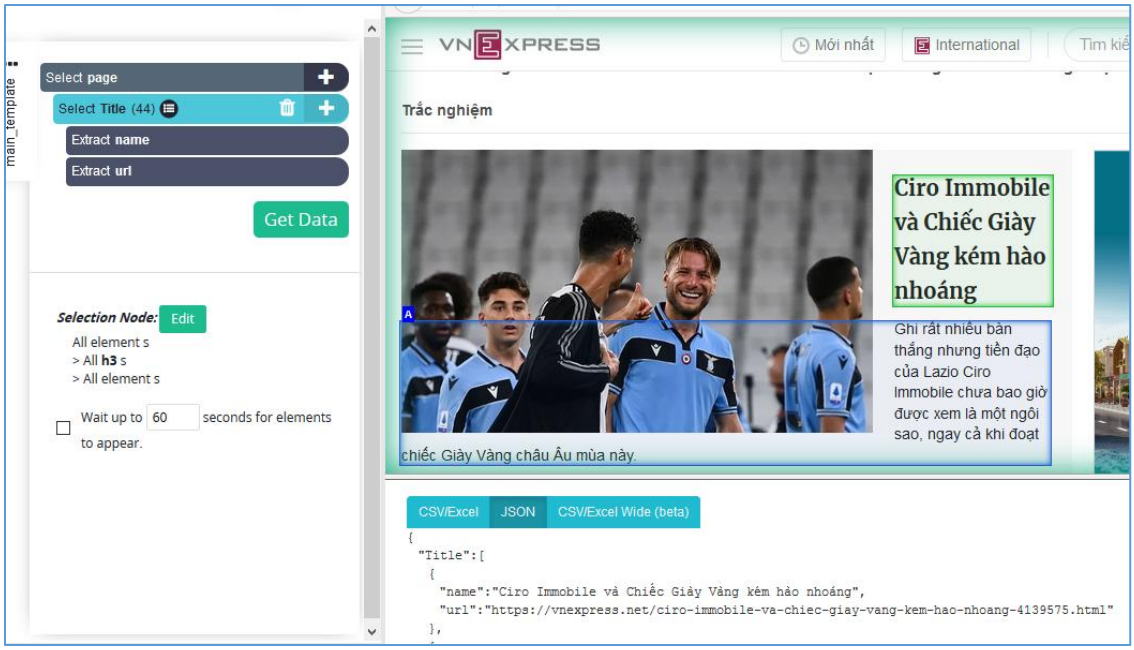
- Chúng ta sẽ tiến hành tạo Projects mới để tiến hành cào dữ liệu.
- Chọn link trang web mà bạn muốn chọn, ở đây chúng ta sẽ lấy trang web: https://vnexpress.net/the-thao



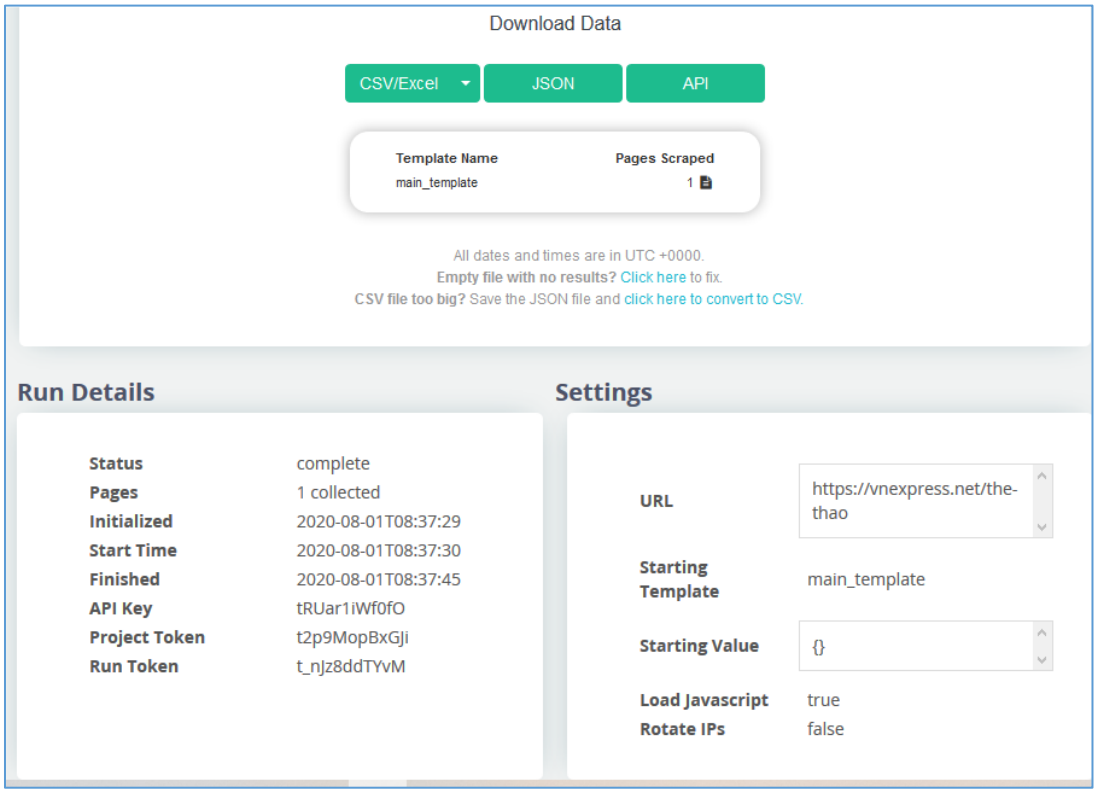
- Chúng ta sẽ điền link trang web vào phần dưới của mục "Enter a website you'd like to extract data from" và sau đó ta sẽ nhấn "Start project on this URL".



- Ở dưới phần Select page chính là thẻ để bạn lấy thông tin ví dụ như gọi thẻ này là Title, thẻ này sẽ lấy các phần tên và link của tiêu đề bài viết. Cách lấy như sau: bạn chỉ cần chọn vào tiêu đề đó và kéo xuống dưới tiếp tục chọn tiêu đề thứ hai, công cụ này sẽ giúp bạn lấy toàn bộ tiêu đề còn lại.
- Kết quả ta được như sau:



- Kết quả trả về sẽ là json hoặc là csv tùy theo lựa chọn của bạn. Cuối cùng bạn chỉ cần nhấn Get Data để lấy dữ liệu.
- Sau đó nhấn thêm RUN.



- Và cuối cùng bạn sẽ tiến hành tải dữ liệu về bằng các định dạng như csv hay json hoặc là kết nối cấu hình API để lấy dữ liệu trực tiếp.
- Để hiểu rõ hơn về phần này các bạn có thể truy cập vào: https://help.parsehub.com/hc/en-us.

HẾT