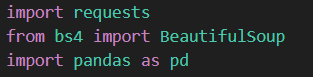
**BÁO CÁO BÀI TẬP CRAWL DỮ LIỆU TỪ NHÀ XUẤT BẢN SPRINGER**

Những phương thức crawl dữ liệu phổ biến nhất hiện nay là:

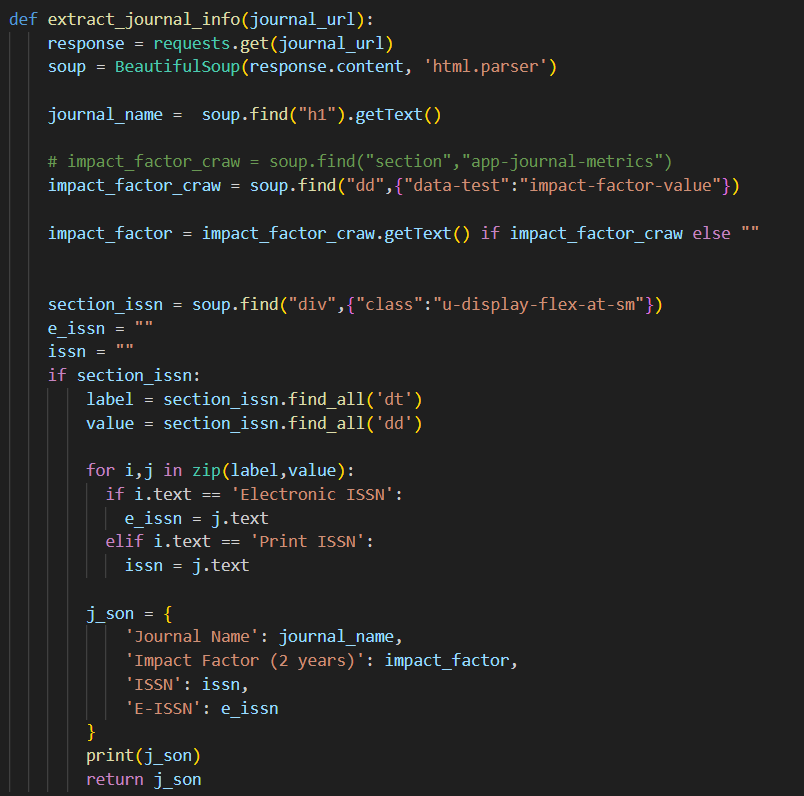
* Sử dụng thư viện hoặc FrameWorks (Scrapy, BeautifulSoup, Requests, Puppeteer của Node.js
* Sử dụng API
* Sử dụng phần mềm thứ ba (Octoparse, Import.io, Apify)

Trong bài tập này em sử dụng thư viện BeautifulSoup của Python để crawl dữ liệu từ nhà xuất bản Springer. Quy trình làm bài tập của em như sau:

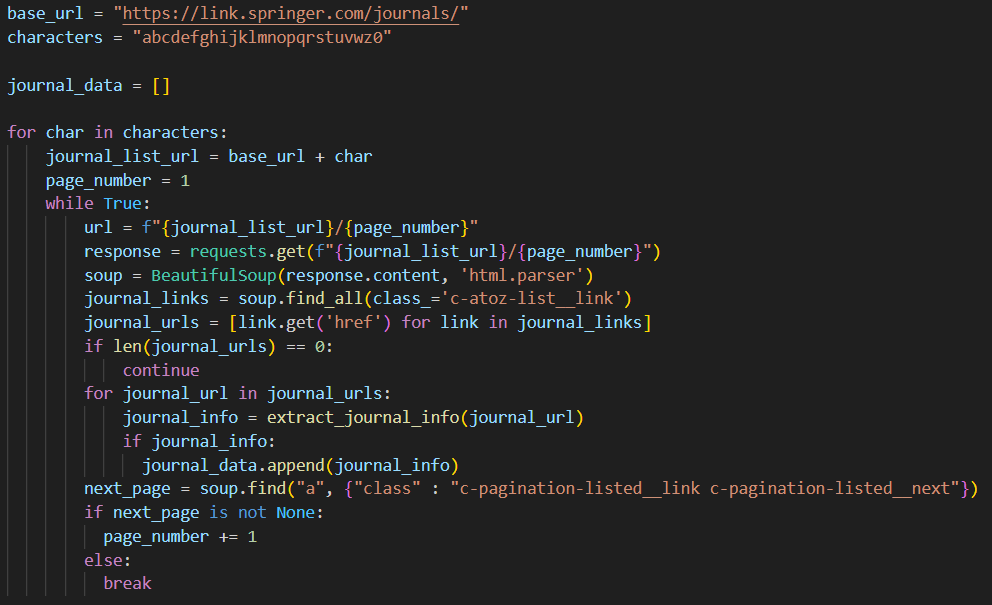
* Đầu tiên em sẽ cài đặt thư viện BeautifulSoup bằng câu lệnh “pip install beautifulsoup4”
* Sau đó ta cài đặt requests để yêu cầu HTTP cho trang web và nhận được content của web “pip install requests”
* Sau khi đã cài đặt hết thì em bắt đầu import



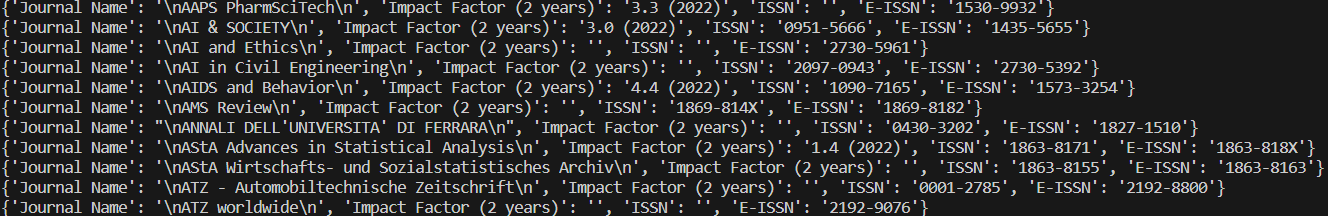
* Em tạo một function để có thể crawl dữ liệu của từng journal trong Springer và ta lấy html của tên của journal, impact factor (2 years), ISSN và e-ISSN. Nếu như trong journal ko có phần nào thì ta để trống phần đó



* Do Spinger có nhiều Journals bắt đầu bằng nhiều chữ cái khác nhau và có nhiều trang nên ta cần phải thực hiện một vòng lặp để có thể crawl dữ liệu từ những Journal đó. Ở đoạn code bên dưới vì ko có journal nào trong chữ “X, Y” và khi em chạy code thì nó lại dừng ở đó nên em đã loại 2 chữ này ra



* Và đây sẽ là kết quả của code



* Sau đó tao cho chúng vào hết 1 Dataframe và xuất nó ra excel



* Có những journal có URL khác với URL ban đầu nên những cái đó thì em phải thêm vào thủ công