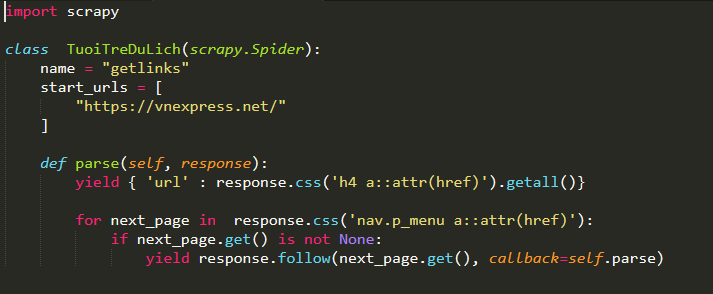
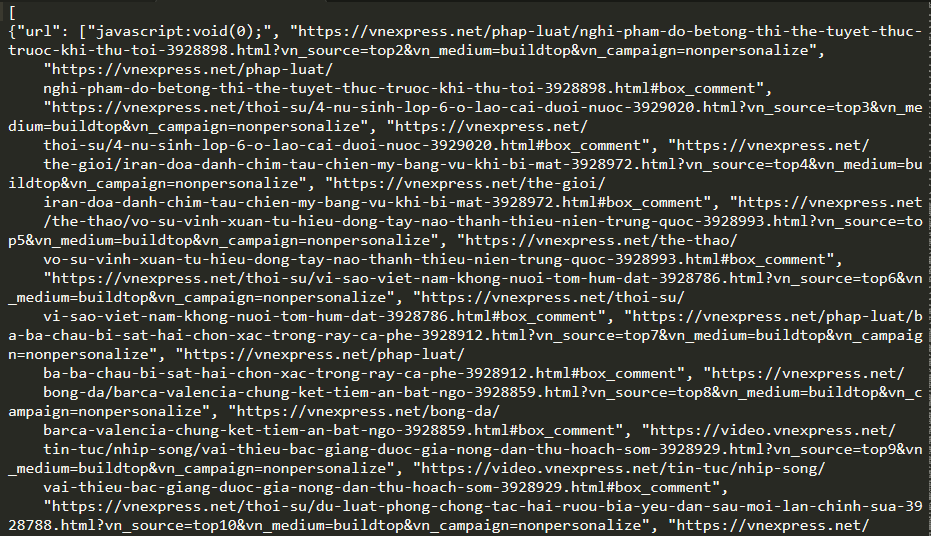
**Dùng Scrapy để crawl dữ liệu từ web:**

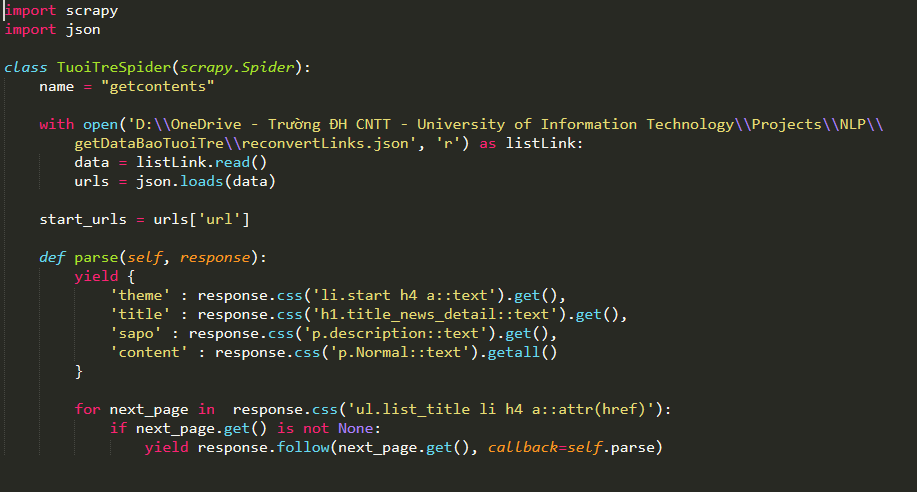
Bước 1: lấy tất cả các link của trang vnexpress



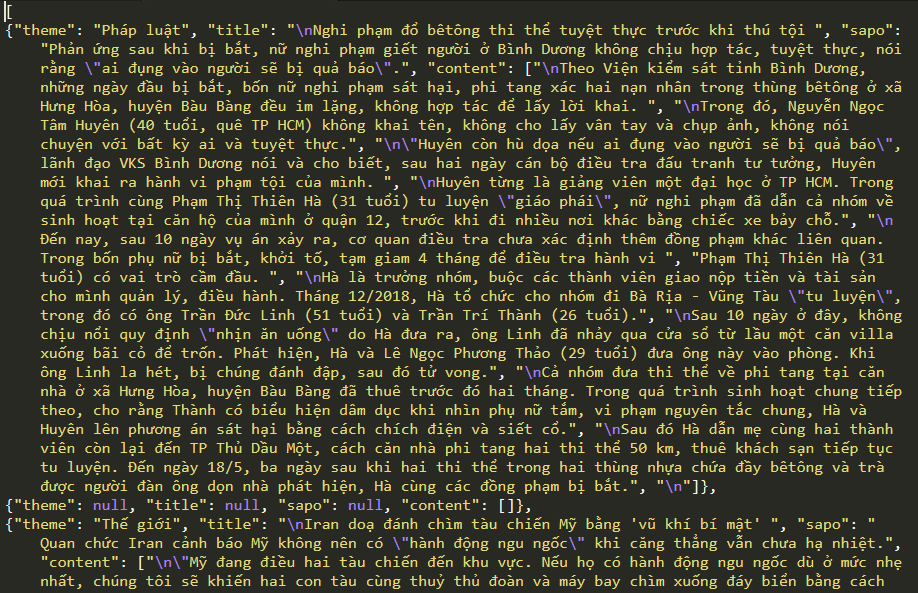
Kết quả thu được:



Bước 2: Lấy content của các bài báo từ các link đã lấy được ở bước 1



Kết quả thu được:



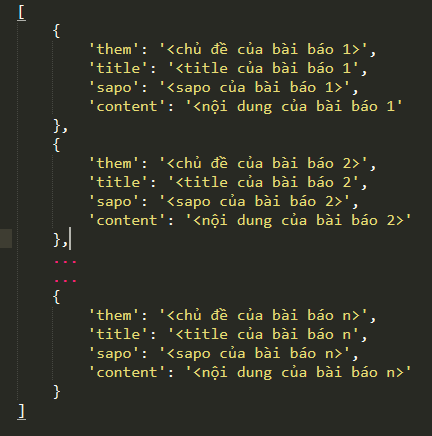
**Hàm xử lý:**



Bước 1: Load data đã lấy về:



file contentsBaoVnexpress.json sẽ có dạng:

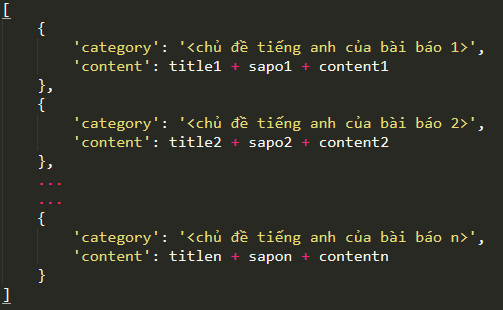


Bước 2: Tạo 1 mảng transfer chưa các object, mỗi object gồm các key:

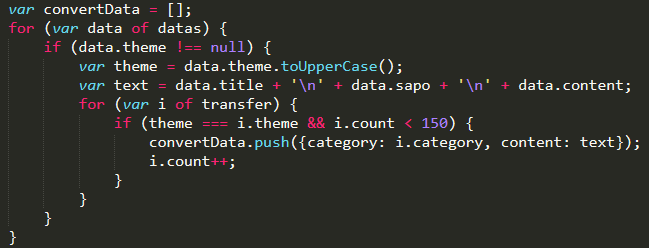
* theme: chủ để tiếng Việt của bài báo
* category: chủ đề tiếng Anh tương ứng
* count: đếm số lượng bài báo của chủ đề



Bước 3: Convert lại data thành dạng:



Bằng đoạn sau:



* Khai báo mảng convertData để lưu data sau khi convert: convertData = [];
* Dùng vòng for lặp các phần từ của biến datas ở **bước 1:**
* Nếu theme của bài báo đang xét khác null thì:

1. Viết hoa theme của bài báo đang xét
2. Khai báo biến text = title bài báo + sapo bài báo + content bài báo
3. Dùng vòng for để kiểm tra bài báo đang xét thuộc chủ đề nào trong mảng transfer ở bước 2 và push bài báo đang xét vào mảng convertData đã khai báo ở trên với từng chủ đề tương ứng, sau đó cho count tăng lên để kiểm soát số lượng bài báo cần lấy của mỗi chủ đề.

Bước 4: Lưu mảng convertData vào file convertDataBaoVnexpress:

