NVCK training 6 - NVCK scrapy crawl data

Tuesday, December 28, 2021 2:51 PM

Xin chào. Ở bài đầu tiên NVCK training 1 tôi đã giới thiệu các project,
Ở bài này tôi sẽ viết rõ ra những gì tôi muốn làm và làm như thế nào

- [ ] Dữ liệu được lưu trữ tại google drive
- [ ] Cào dữ liệu từ trang truyenfull, tangthuvien về ổ đĩa dưới dang raw, html
- [ ] Từ html chuyển sang file với cấu trúc để xử lý sau này
- [ ] Từ file có cấu trúc lưu vào database

Ở bài sau tôi sẽ nói về chuyển từ raw sang structured text file

Dữ liệu được lưu trữ tại google drive

Nếu lưu vào google drive thì thoải mái bao nhiêu cũng được, vì tôi có mail edu.vn, hiện tại dữ liệu của tôi đã lên đến 75TB và vẫn có thể lưu tiếp thoải mái

Nhưng thi thoảng truy cập lấy file thì hơi đơ đơ, vì phải chờ file download từ google drive về mới có thể đọc được.

Tuy nhiên như thế đã là rất tốt

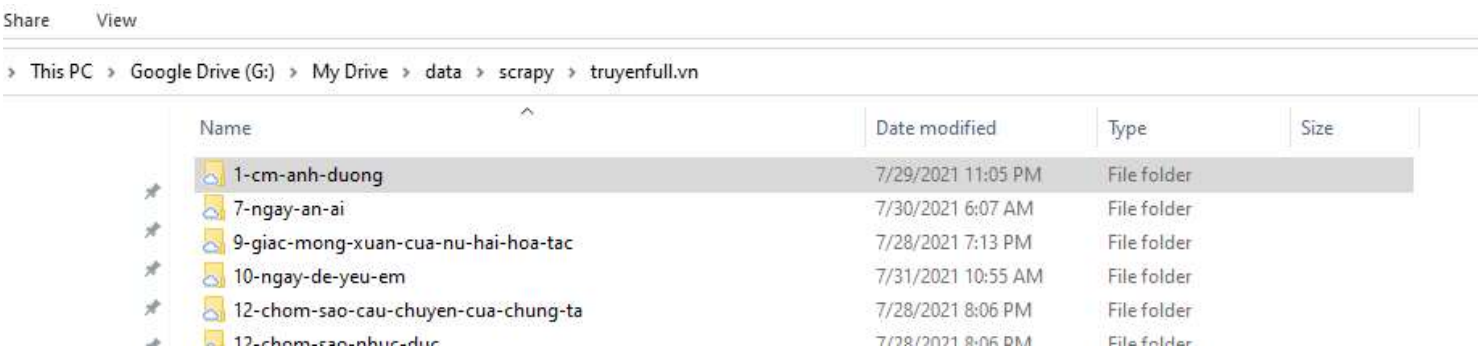
Tôi định lưu trữ các file raw trên gg drive hết,
File có cấu trúc và database thì có thể lưu trên máy vì chúng ta sử dụng thường xuyên và cần nhanh

Scrapy (NVCK backend)

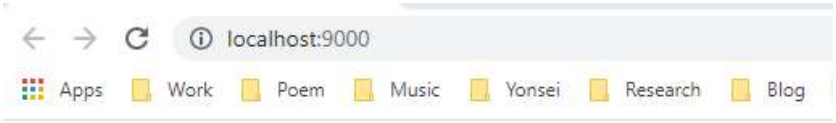
Scrapy dùng để cào dữ liệu website, chúng ta có thể hình dung ta chỉ định trang web

Homepage.com

Trong homepage.com lại có link homepage.com/a homepage.com/b …
Thì scrapy sẽ tự động follow đường dẫn đó và ta có thể lưu trang html lại dạng raw
Tôi sẽ demo ngay bây giờ



Tiến độ (NVCK server)



Raw

Truyện full

Số lượng truyện: 3004
Số lượng chương truyện: 252315

Truyện Tàng thư viện

Số lượng truyện: 99
Số lượng chương truyện: 6126

Processed

Truyện full

Số lượng truyện: 2888
Số lượng chương truyện: 204315

Truyện Tàng thư viện

Số lượng truyện: 55
Số lượng chương truyện:
Đọc online, tải truyện, tìm kiếm mọi thứ, cập nhật từ nhiều nguồn.