**BÁO CÁO CRAWLER & DOCKER**

**Môn: Nhập môn dữ liệu lớn**

Tên: Trần Bình Minh

MSSV: 22678911

Link source code: [Big-Data-Subject/(Personal)W1-Crawler\_&\_Docker at main · tbm077861/Big-Data-Subject](https://github.com/tbm077861/Big-Data-Subject/tree/main/(Personal)W1-Crawler_%26_Docker)

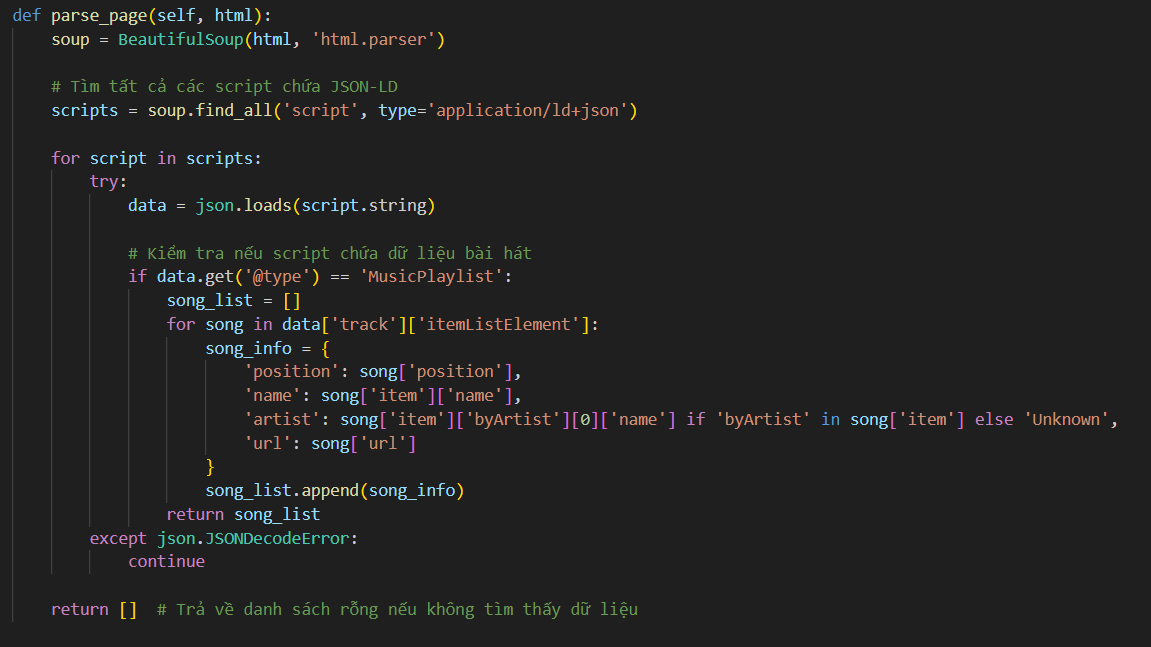
# **1. Crawl và lưu trữ dữ liệu**

## Trang Zing MP3:

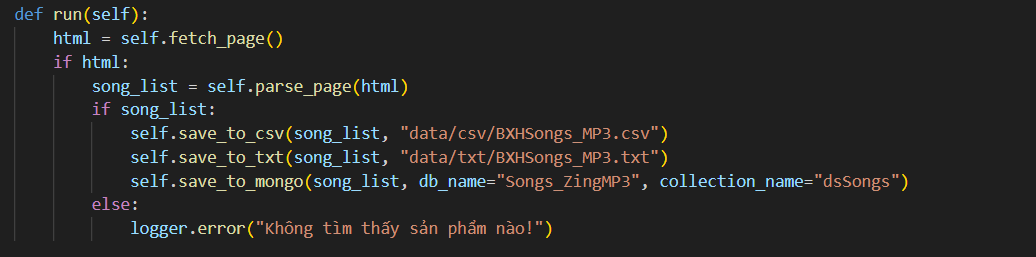
* Mục tiêu: Thu thập các bài hát nằm trong BXH âm nhạc của Zing MP3 gồm các thông tin về thứ hạng(position),Tên bài hát (name),Tác giả (artist),Liên kết (url)

### Crawl dữ liệu:

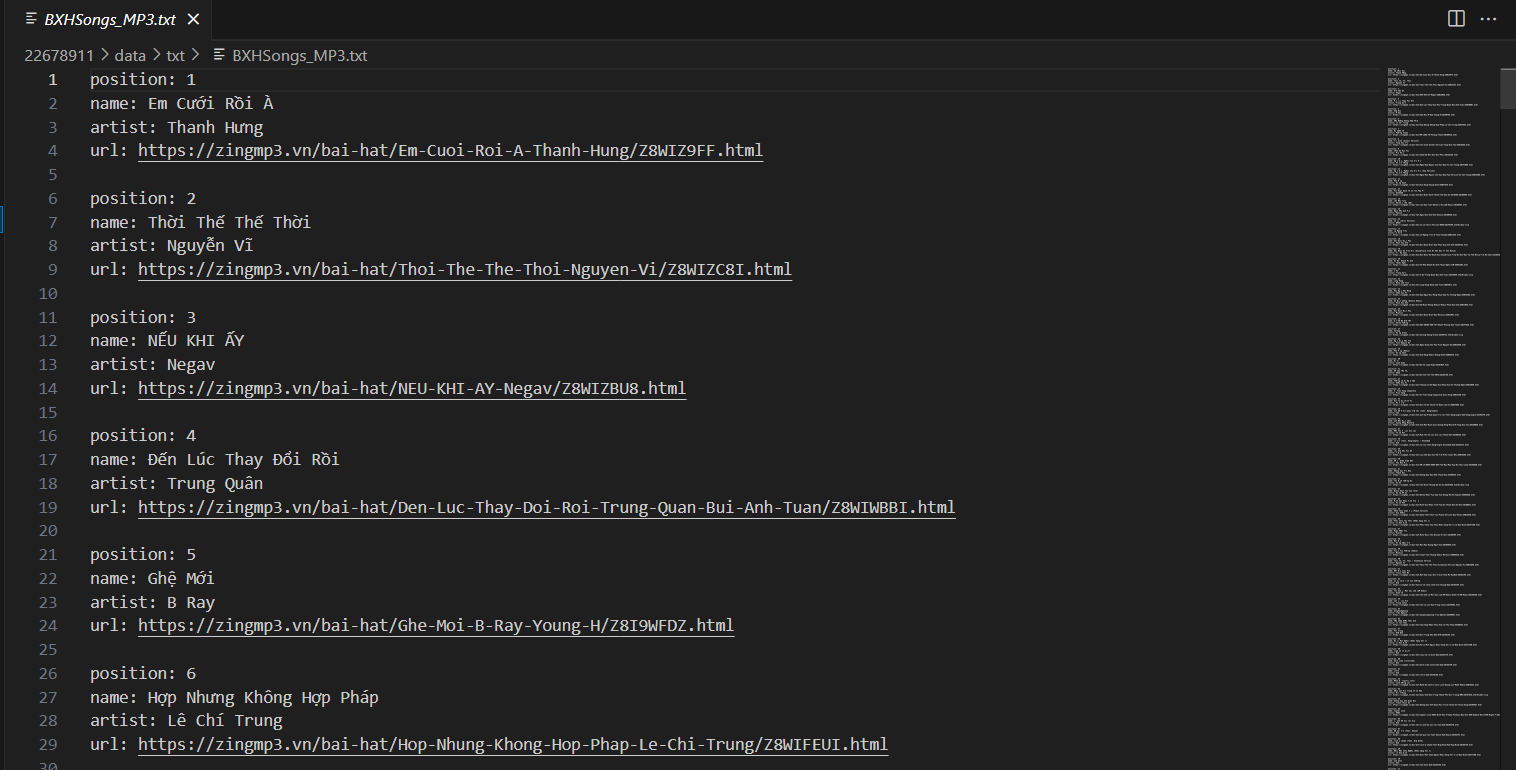
* + Phân tích HTML và lấy thông tin sản phẩm

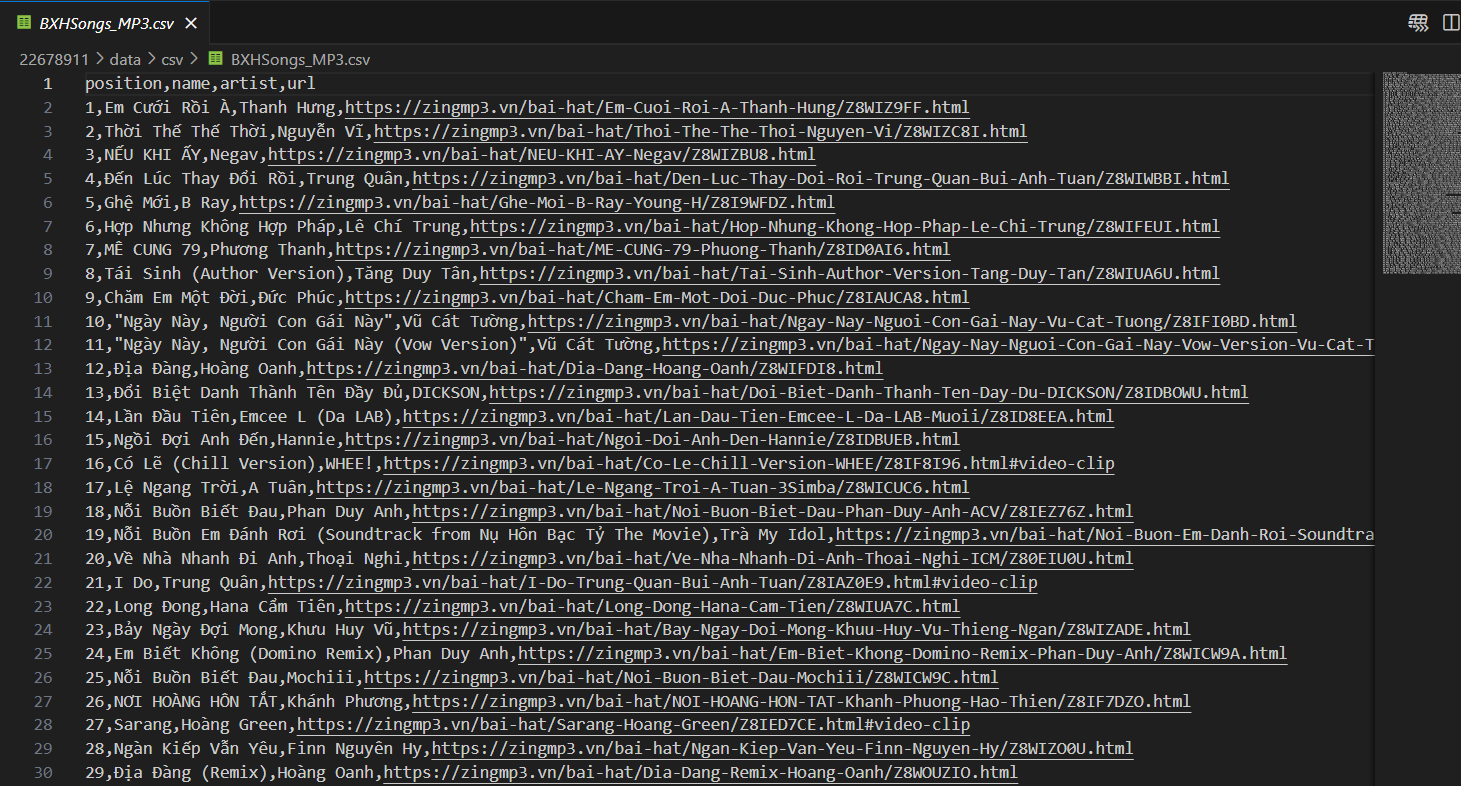


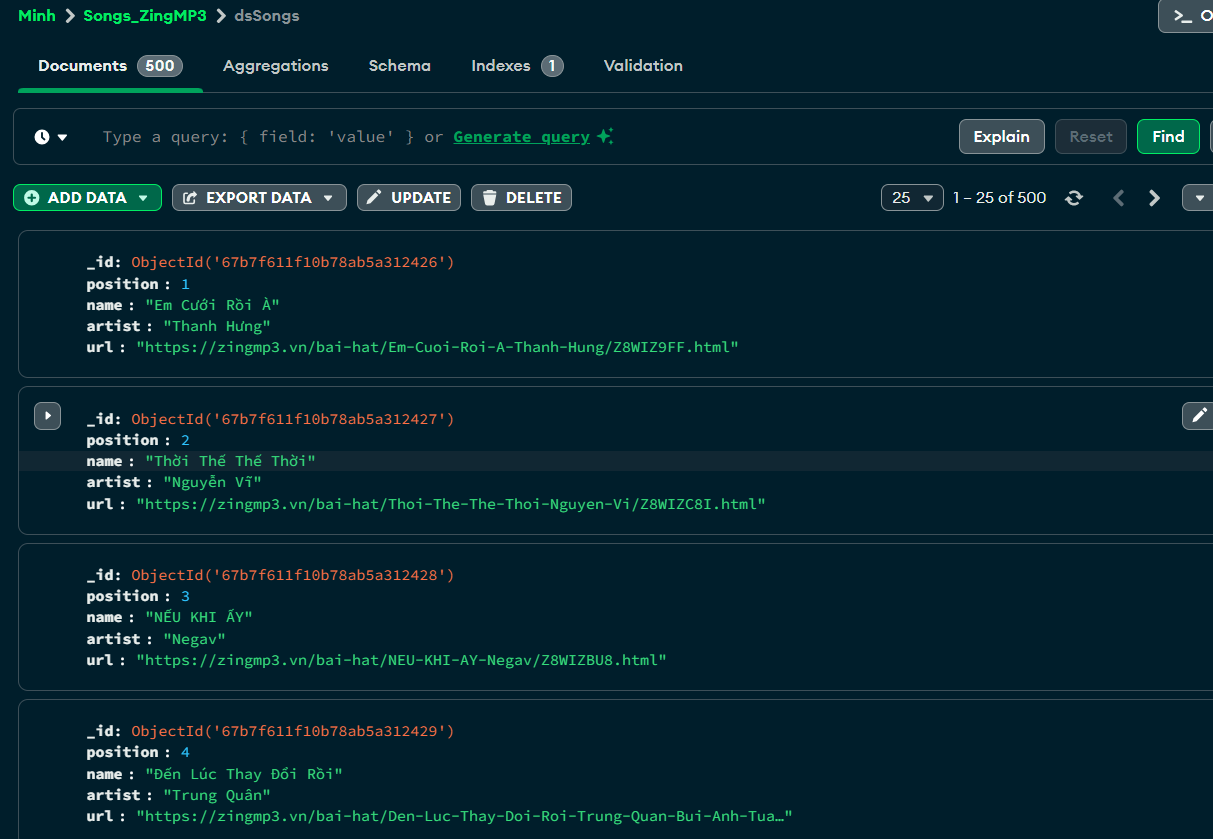
* + Chạy và lưu thông tin vào mongoDB, csv và txt.



### Lưu trữ vào mongoDB, csv, txt.







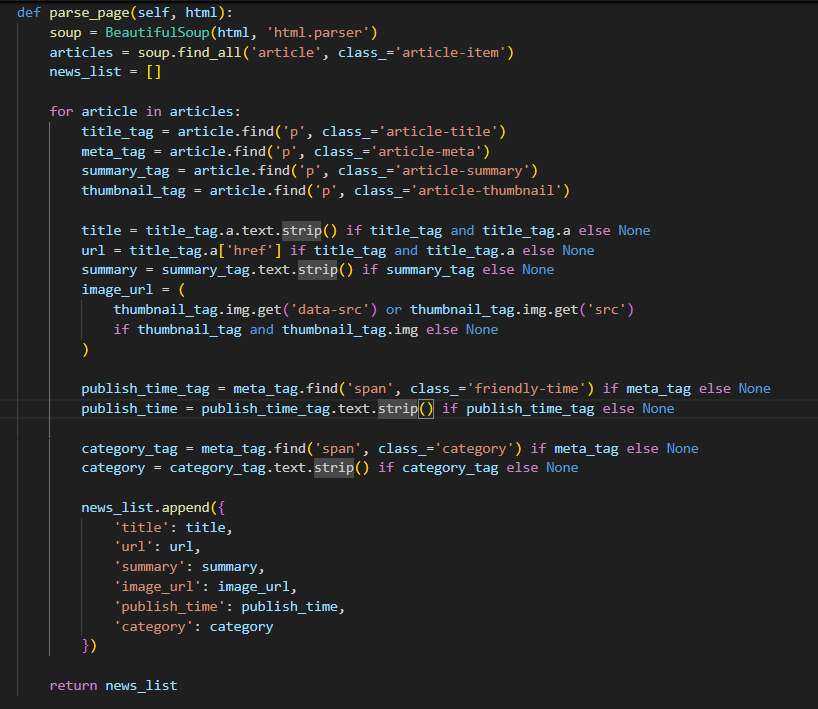
B. Trang Zing News

## Trang Zing News

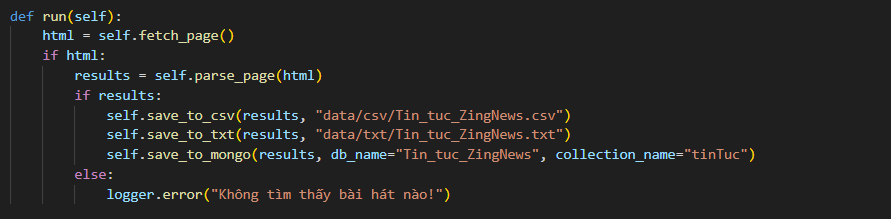
* Mục tiêu: Thu thập các bài báo mới nhất về chủ đề Công Nghệ gồm các trường thông tin Tiêu đề, tóm tắt, liên kết, link hình ảnh, thời gian đăng, thể loại

### Crawl dữ liệu:

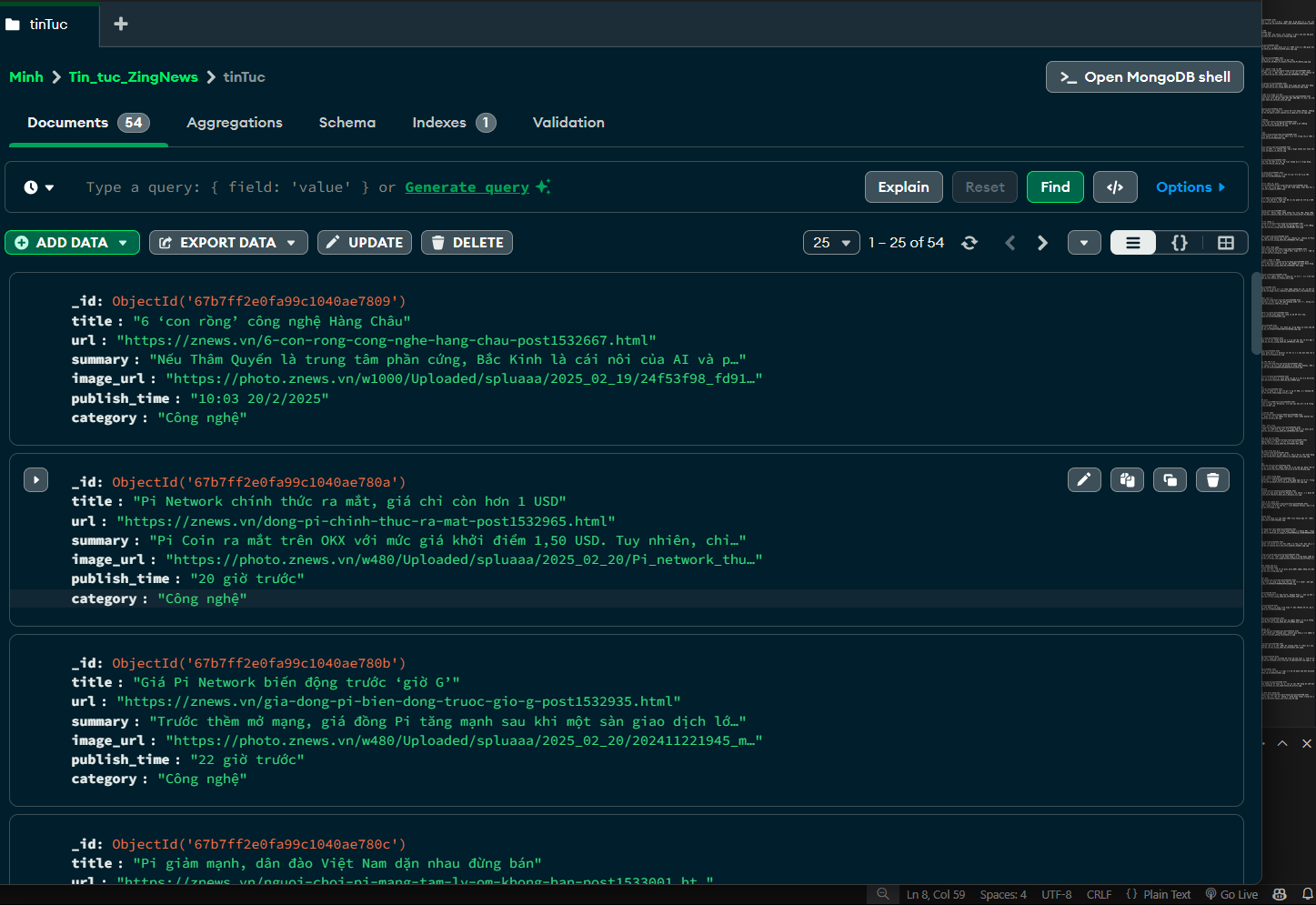
* 1. Phân tích HTML và lấy thông tin bài hát

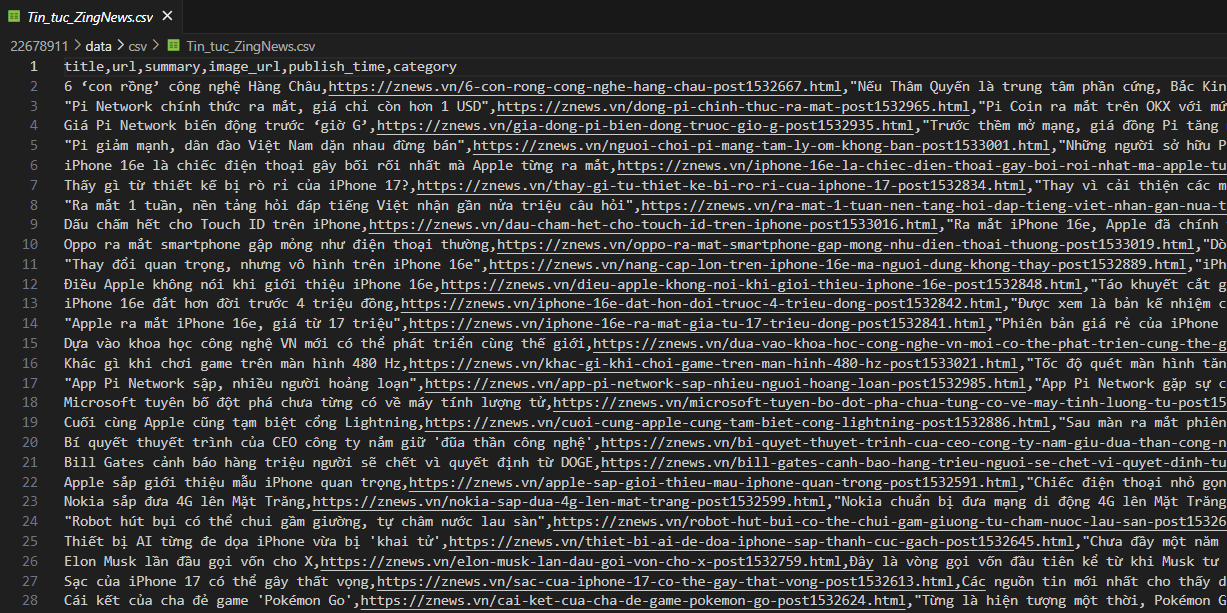


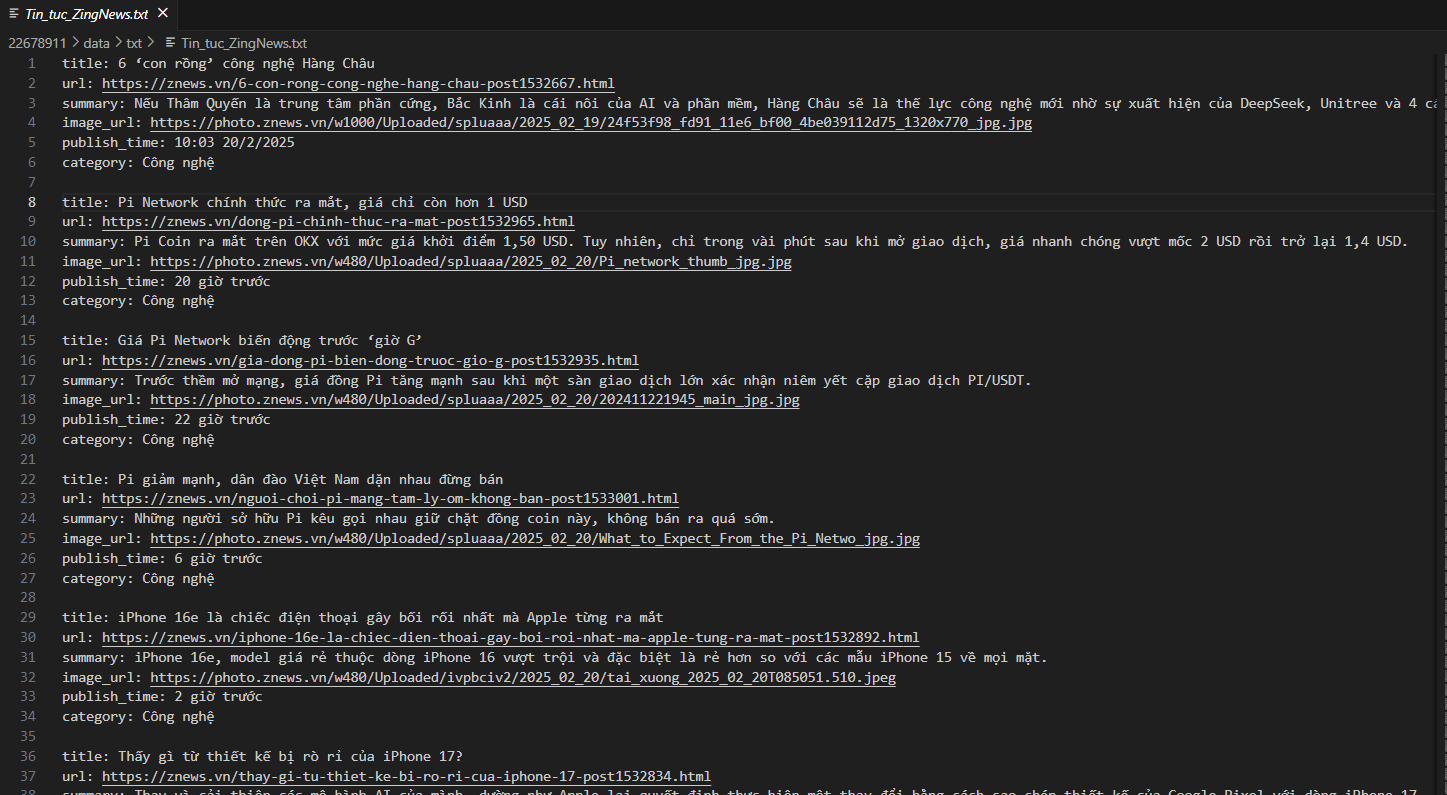
* 1. Chạy và lưu thông tin vào mongoDB, csv và txt.



### Lưu trữ vào mongoDB, csv, txt

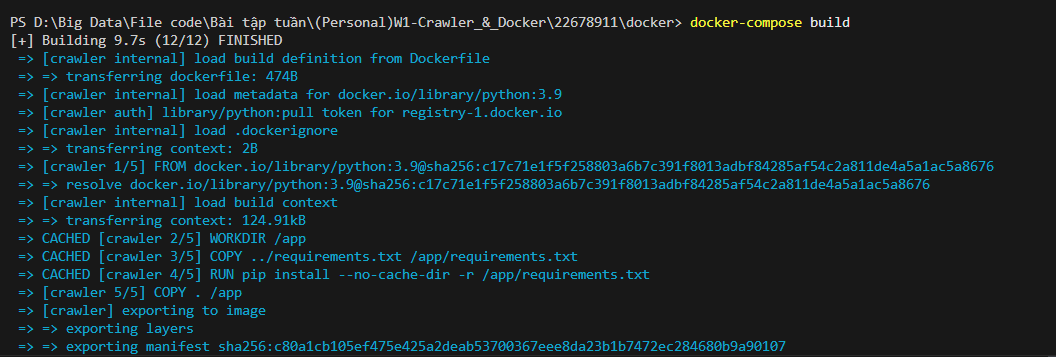




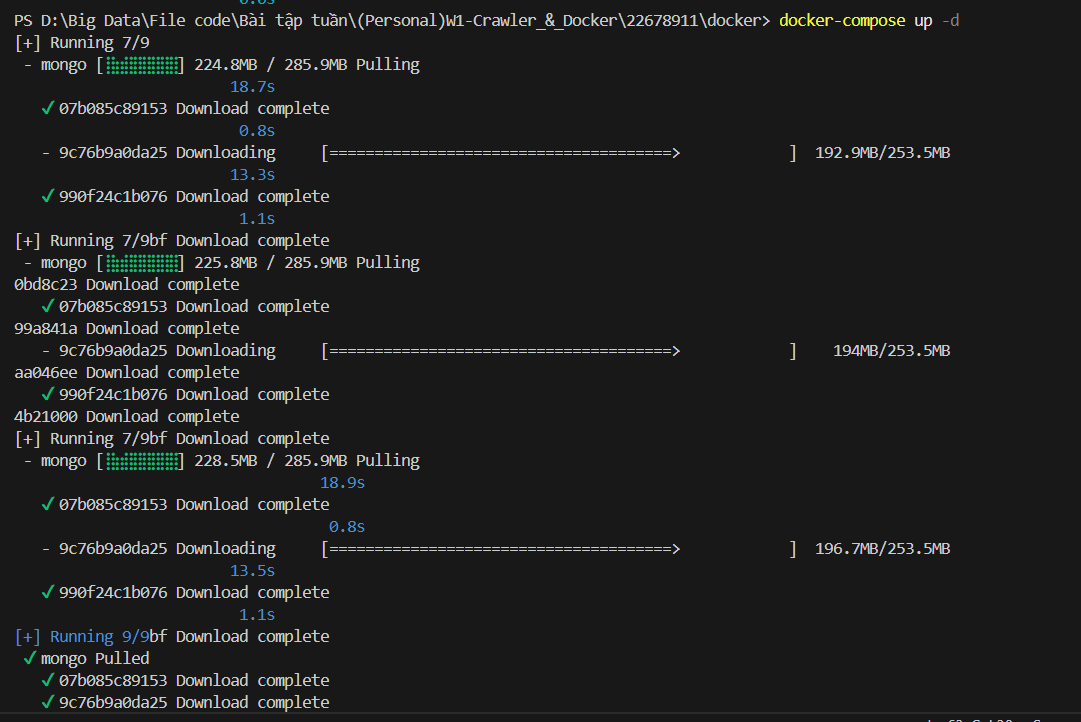


# **2. Đóng gói và triển khai bằng Docker**

* Xây dựng Docker Image:  **docker-compose build**



* Chạy docker container: **docker-compose up -d**



* Kiểm trả container hoạt động: **docker ps**

