CRAWLER & DOCKER

I. Yêu cầu:

Xây dựng hệ thống thu thập dữ liệu tự động từ các trang thương mại điện tử sử dụng Docker và Scrapy, BeautifuSoup

- Crawl dữ liệu sản phẩm từ 2 trang trở lên
- Thu thập các thông tin cần thiết của trang
- Lưu trữ vào mysql, postgre, mongodb, csv, txt
- Sử dụng docker để đóng gói và triển khai

II. Quy định nộp bài:

- T01_MSSV_Hoten.docx
  Trong đó: (1) Link source code trên github, (2) Tài liệu báo cáo kết quả
- **Các trường hợp sao chép: 0 điểm cả 2**

III. Template project crawler

- **Crawler_project được thay thế thành MSSV của sinh viên**

```
crawler_project/
├── docker/
│   ├── Dockerfile
│   └── docker-compose.yml
├── src/
│   ├── crawlers/
│   │   ├── __init__.py
│   │   ├── base_crawler.py      # Class cơ sở cho crawler
│   │   ├── site1_crawler.py     # Crawler cho trang 1
│   │   └── site2_crawler.py     # Crawler cho trang 2
│   ├── database/
│   │   ├── __init__.py
│   │   ├── mysql_handler.py
│   │   ├── postgres_handler.py
│   │   ├── mongo_handler.py
│   │   ├── csv_handler.py
│   │   └── txt_handler.py
│   ├── utils/
│   │   ├── __init__.py
│   │   ├── logger.py
│   │   └── helpers.py
│   └── main.py
├── data/
│   ├── csv/
│   └── txt/
├── config/
│   └── settings.yml
├── requirements.txt
└── README.md
```

IV. Danh sách gợi ý

1. Các trang tin tức: • VnExpress • Báo Thanh Niên • Tuổi Trẻ Online • Vietnamnet • Zing News • VOV News • Dân Trí	**2. Các trang thời tiết:** • nchmf.gov.vn • thoitiet.vn • weather.com (phần Vietnam)	**3. Các trang phim/nhạc:** • phimmoi • HDViet • Nhaccuatui • ZingMP3 • Keeng
4. Trang review/đánh giá: • Foody.vn • Tripadviser (phần Vietnam) • Google Reviews • Tiki Reviews	**5. Trang việc làm:** • VietnamWorks • CareerBuilder • TopCV • ITViec • 24h.com.vn (mục việc làm)	**6. Trang bất động sản:** • Batdongsan.com.vn • Homedy • Propzy • Mogi.vn
7. Trang thể thao: • Bóng đá 24h • Thể Thao 247 • VnExpress Sports • Bongdaplus	**8. Trang giá cả/sản phẩm:** • websosanh.vn • Priceza • Dienmayxanh.com • fptshop.com.vn	

Với mỗi loại trang này, sinh viên có thể thu thập các thông tin như:

- Tin tức: tiêu đề, nội dung, thời gian, chuyên mục
- Việc làm: vị trí, công ty, mức lương, yêu cầu
- Bất động sản: địa chỉ, giá, diện tích, thông tin chi tiết
- Thể thao: kết quả trận đấu, thông tin cầu thủ/đội bóng