Clean.py

import json

import re

INPUT\_FILE = "books\_raw.jsonl"

OUTPUT\_FILE = "books\_clean.jsonl"

EXCHANGE\_RATE = 30000 # 1£ = 30,000 VNĐ

def clean\_price(price\_raw):

if price\_raw is None:

return None

price\_str = str(price\_raw).replace("Â£", "£").strip()

numeric\_part = re.sub(r"[^\d.]", "", price\_str)

try:

gbp\_value = float(numeric\_part) if numeric\_part else None

if gbp\_value is not None:

vnd\_value = int(gbp\_value \* EXCHANGE\_RATE)

return vnd\_value

else:

return None

except ValueError:

return None

def clean\_rating(rating\_raw):

rating\_map = {

"One": 1.0,

"Two": 2.0,

"Three": 3.0,

"Four": 4.0,

"Five": 5.0

}

if isinstance(rating\_raw, str):

return rating\_map.get(rating\_raw.strip(), 0.0)

try:

return float(rating\_raw)

except:

return 0.0

def clean\_book(book):

title = book.get("title")

price\_raw = book.get("price")

rating\_raw = book.get("rating")

url = book.get("url")

if not title or not price\_raw or not url:

return None

price\_cleaned = clean\_price(price\_raw)

rating\_cleaned = clean\_rating(rating\_raw)

if price\_cleaned is None or rating\_cleaned == 0.0:

return None # Bỏ sách không có giá hoặc rating = 0

cleaned = dict(book)

cleaned["price"] = price\_cleaned

cleaned["rating"] = rating\_cleaned

cleaned["title"] = title.strip()

return cleaned

def load\_jsonl(filepath):

data = []

with open(filepath, 'r', encoding='utf-8') as f:

for line in f:

try:

data.append(json.loads(line))

except json.JSONDecodeError:

print("Bo qua dong loi: ", line)

return data

def save\_jsonl(data, filepath):

with open(filepath, 'w', encoding='utf-8') as f:

for item in data:

f.write(json.dumps(item, ensure\_ascii=False) + "\n")

def clean\_and\_deduplicate():

raw\_data = load\_jsonl(INPUT\_FILE)

seen\_titles = set()

cleaned = []

for book in raw\_data:

cleaned\_book = clean\_book(book)

if cleaned\_book is None:

continue

title\_lower = cleaned\_book["title"].lower()

if title\_lower in seen\_titles:

print("Bo trung lap: ", cleaned\_book['title'])

continue

seen\_titles.add(title\_lower)

cleaned.append(cleaned\_book)

save\_jsonl(cleaned, OUTPUT\_FILE)

print("\nDa clean va luu", len(cleaned), "sach vao", OUTPUT\_FILE)

if \_\_name\_\_ == "\_\_main\_\_":

clean\_and\_deduplicate()

* Config.py

MYSQL\_CONFIG = {

"host": "localhost",

"user": "crawler",

"password": "123456",

"database": "crawl\_data\_uk"

}

* Crawl.py

import requests

from bs4 import BeautifulSoup

import json

BASE\_URL = "http://books.toscrape.com"

OUTPUT\_FILE = "books\_raw.jsonl"

BOOKS\_TO\_CRAWL = 1000

BOOKS\_PER\_PAGE = 20

RATING\_MAP = {

"One": 1.0,

"Two": 2.0,

"Three": 3.0,

"Four": 4.0,

"Five": 5.0

}

def get\_page(url):

response = requests.get(url)

response.raise\_for\_status()

return BeautifulSoup(response.text, "html.parser")

def extract\_book\_info(book\_soup):

rating\_str = book\_soup.p['class'][1]

return {

"title": book\_soup.h3.a['title'],

"price": book\_soup.find('p', class\_='price\_color').text,

"rating": RATING\_MAP.get(rating\_str, 0.0), # Mặc định là 0.0 nếu không có rating hợp lệ

"url": BASE\_URL + "/catalogue/" + book\_soup.h3.a['href']

}

def crawl\_books():

books = []

total\_pages = (BOOKS\_TO\_CRAWL + BOOKS\_PER\_PAGE - 1) // BOOKS\_PER\_PAGE

for page in range(1, total\_pages + 1):

print(f"Crawling page {page}...")

url = f"{BASE\_URL}/catalogue/page-{page}.html"

soup = get\_page(url)

articles = soup.find\_all('article', class\_='product\_pod')

for a in articles:

if len(books) < BOOKS\_TO\_CRAWL:

books.append(extract\_book\_info(a))

else:

break

return books

def save\_to\_jsonl(data, filename):

with open(filename, 'w', encoding='utf-8') as f:

for item in data:

f.write(json.dumps(item, ensure\_ascii=False) + '\n')

if \_\_name\_\_ == "\_\_main\_\_":

books = crawl\_books()

save\_to\_jsonl(books, OUTPUT\_FILE)

print(f"Crawl xong {len(books)} sách. Đã lưu vào {OUTPUT\_FILE}")

* save\_to\_mysql.py

import pymysql

import json

from config import MYSQL\_CONFIG # database: crawl\_data\_uk

def create\_books\_uk\_table():

conn = pymysql.connect(

host=MYSQL\_CONFIG["host"],

user=MYSQL\_CONFIG["user"],

password=MYSQL\_CONFIG["password"],

database=MYSQL\_CONFIG["database"],

charset='utf8mb4'

)

cursor = conn.cursor()

cursor.execute("""

CREATE TABLE IF NOT EXISTS books\_uk (

id INT AUTO\_INCREMENT PRIMARY KEY,

title VARCHAR(255) UNIQUE,

price INT,

rating FLOAT,

url TEXT

)

""")

conn.commit()

cursor.close()

conn.close()

print("Đã tạo bảng books\_uk")

def save\_to\_books\_uk(data):

conn = pymysql.connect(

host=MYSQL\_CONFIG["host"],

user=MYSQL\_CONFIG["user"],

password=MYSQL\_CONFIG["password"],

database=MYSQL\_CONFIG["database"],

charset='utf8mb4'

)

cursor = conn.cursor()

insert\_query = """

INSERT INTO books\_uk (title, price, rating, url)

VALUES (%s, %s, %s, %s)

ON DUPLICATE KEY UPDATE

price = VALUES(price),

rating = VALUES(rating),

url = VALUES(url)

"""

for item in data:

cursor.execute(insert\_query, (

item.get("title"),

item.get("price"),

item.get("rating"),

item.get("url")

))

conn.commit()

cursor.close()

conn.close()

print(f"Đã lưu {len(data)} bản ghi vào bảng books\_uk")

if \_\_name\_\_ == "\_\_main\_\_":

create\_books\_uk\_table()

with open("books\_clean.jsonl", "r", encoding="utf-8") as f:

data = [json.loads(line) for line in f]

save\_to\_books\_uk(data)