**LAB 4**

# Thông tin chung

* Thực hành chương 3: Crawling website, .
* Sinh viên lưu file với tên theo định dạng sau: Lab4\_MSSV.ipynb,
* Sinh viên nộp lên Elearning của lớp học tại buổi thực hành 4.
* Deadline: 11h30, ngày 07/07/2023

# Yêu cầu

* Sử dụng các IDE (anaconda, pycharm, google colab,…) để tạo file thực hành.

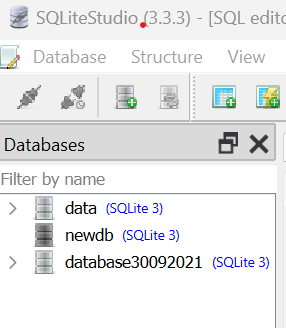
# Nội dung

### Chuẩn bị

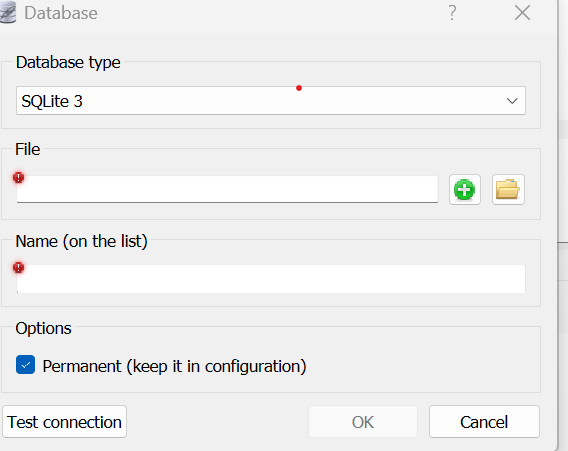
* 1. **Cài đặt thư viện**
* Cài đặt sqlite3 studio
* Tạo project mới
* Install thư viện: thực hiện install hai thư viện sau
  + pip install flask
  + pip install newspaper3k #dùng để crawl data từ danh mục web
  1. **Thực hiện**

B1. Trong project, Tạo thư mục DATA => tạo file database có tên crawlingDB.db

B2. Trong SQLiteStudio, tạo new DB

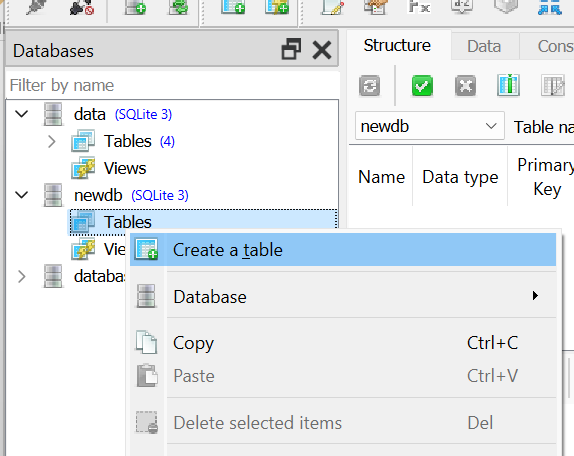


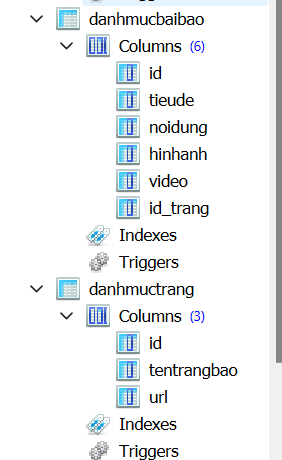
Click vào đây để tạo new DB



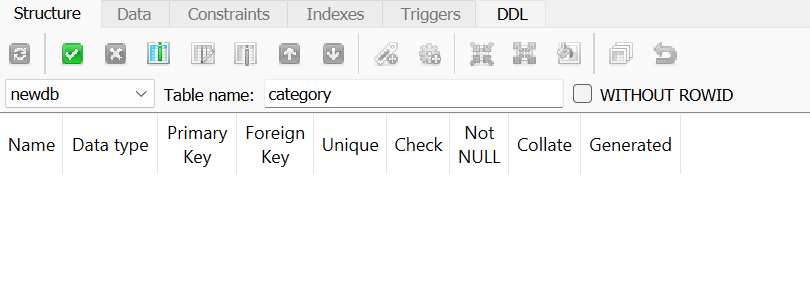
Click vào đây để link tới file crawlingDB.db trong project

Trong new DB, click chuột phải vào tables => create table để tạo ra hai bảng category (danh mục trang) và news (danh mục bài báo) có cấu trúc như sau:





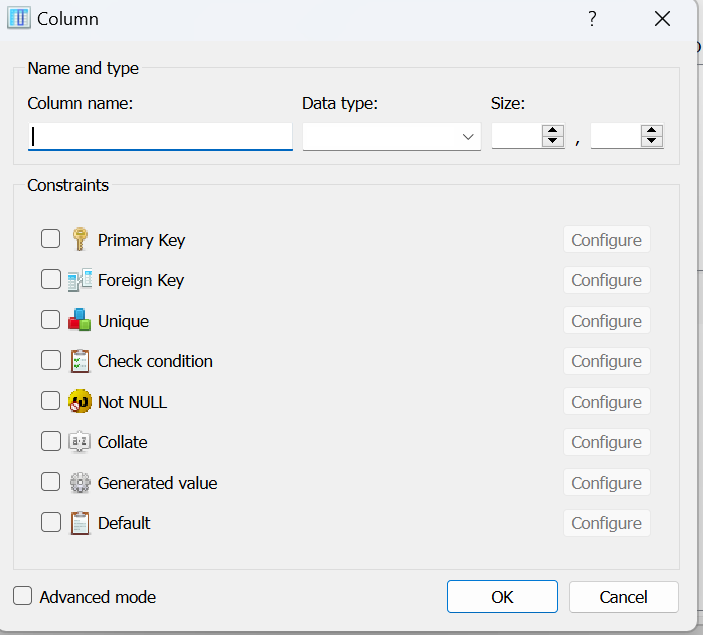
Để tạo các column cho table, thực hiện:



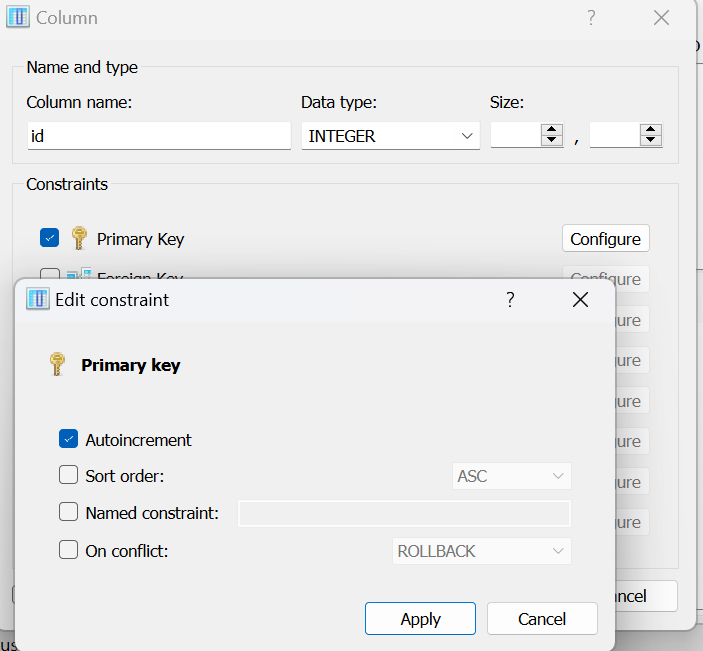
Đặt tên cho cột, chọn kiểu dữ liệu cho cột,

Click vào đây để đặt tên cho các table.

Double click vào khoảng trắng để tạo ra các column cho các table.



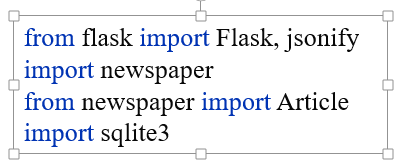
Check vào primary key để đặt khóa chính cho cột. Click configure để tùy chỉnh cho khóa



mở bảng category =>thêm danh mục các trang báo sẽ tiến hành crawl dl.

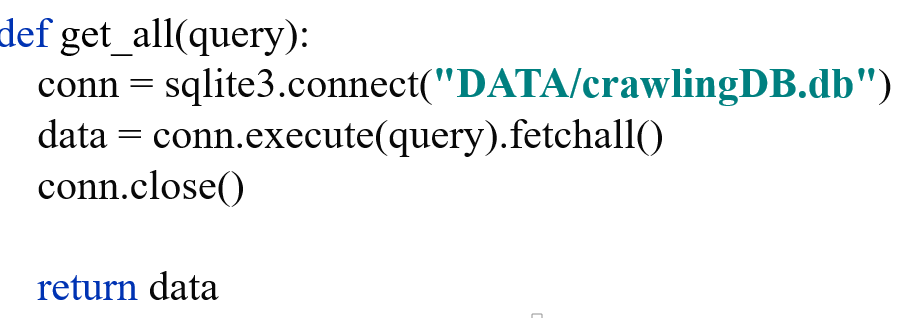
B3.Tạo hai module:

* utils.py: import các thư viện



* Define các hàm trong utils

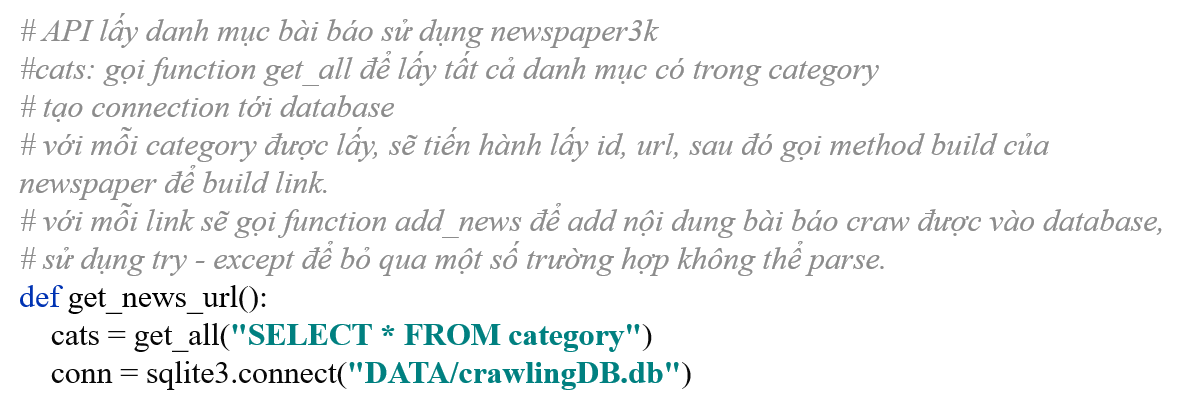
**get\_all():**

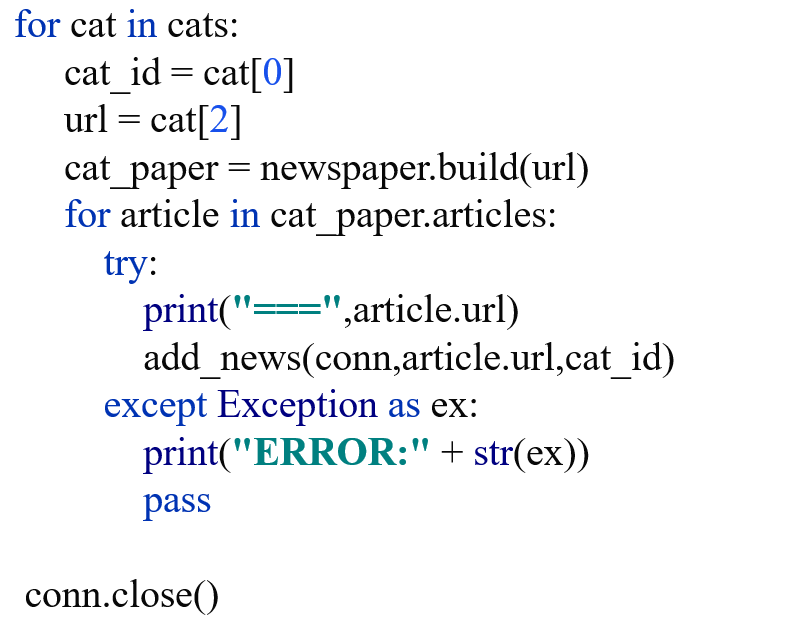


**add\_news()**



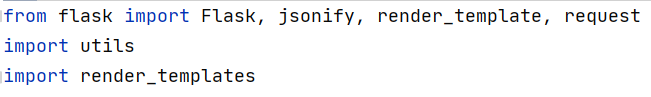
**get\_news\_url()**

****

****

**Module api:**

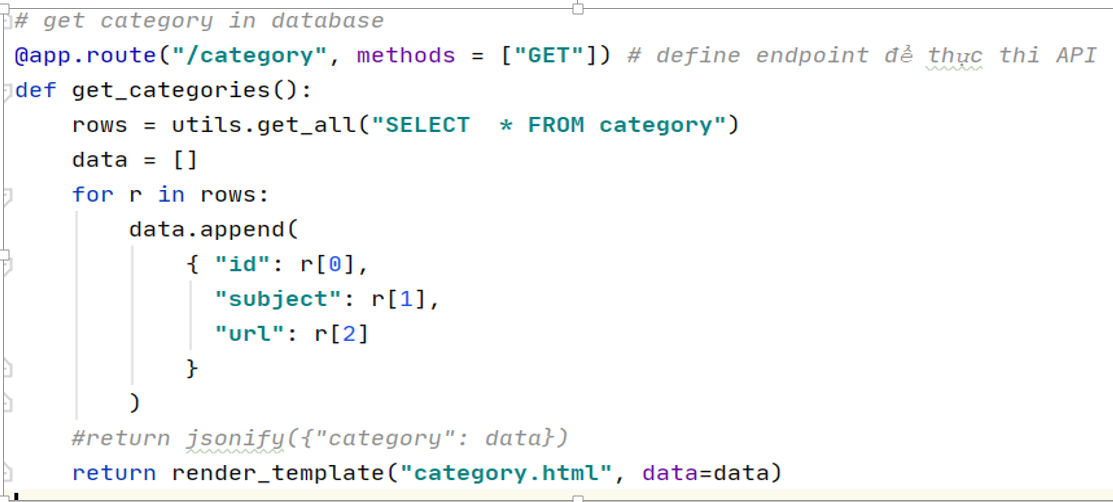
* api.py: để define các API, import các thư viện sau



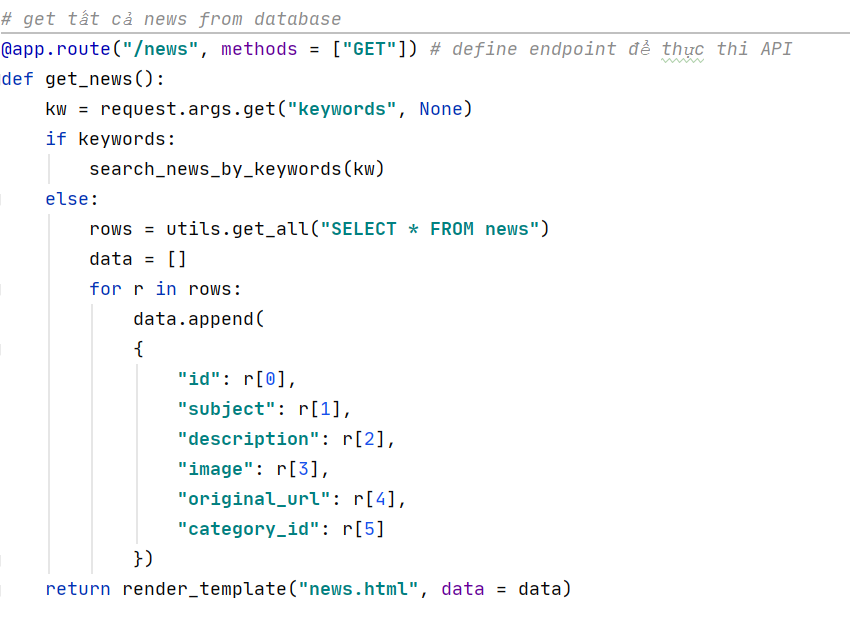
* Define các hàm trong api
  + Route cho trang chủ của báo cáo



Route cho trang danh mục các website dùng để thu thập dữ liệu



**Route hiển thị các tin tức thu thập được**

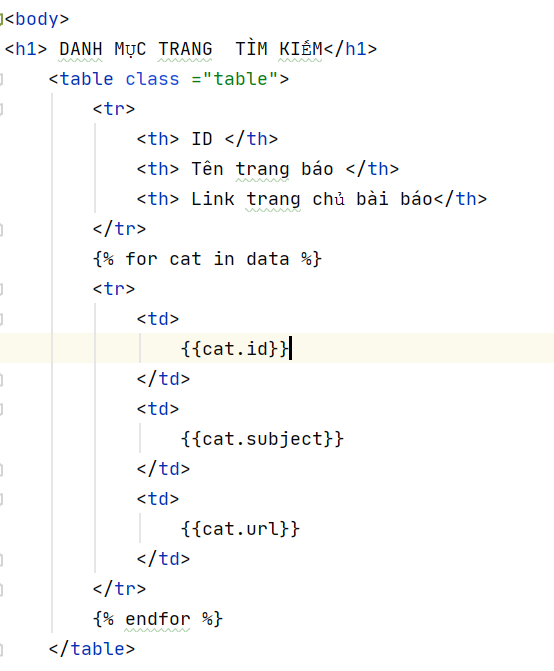
****

**Thư mục templates**

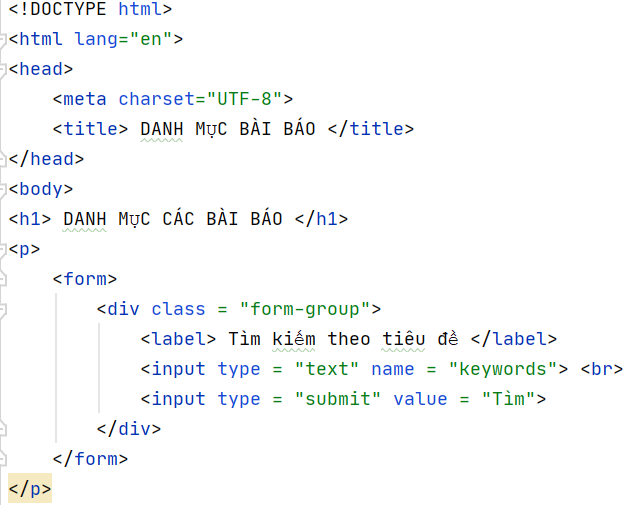
Tạo thư mục templates, sau đó tạo các file **category.html** (để hiển thị danh mục các trang chứa dữ liệu cần thu thập) , **news.html** (để hiển thị dữ liệu các bài báo thu thập được) và index.html (trang chủ của báo cáo).

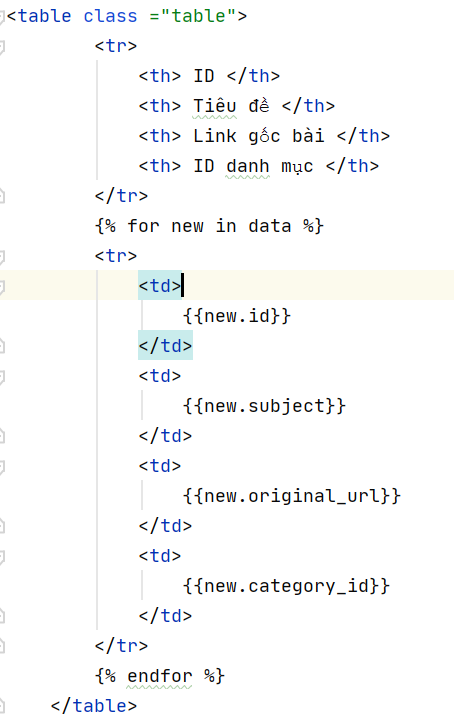
Cấu trúc các file html như sau:

**Category.html:**

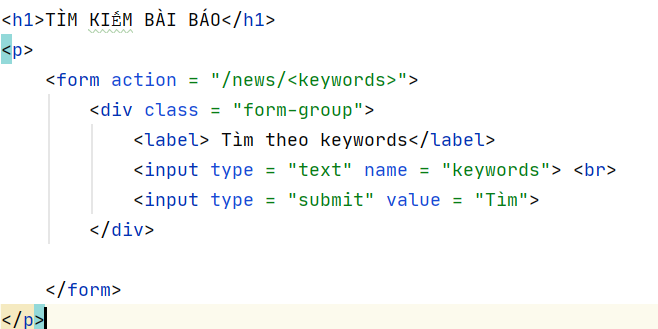


**News.html**

****

****

**Index.html**

****

Sau khi hoàn thành việc lấy dữ liệu và render lên website, thực hiện gắn CSS cho website.