LAB 5

I. Thông tin chung

- Thực hành chương 3: Crawling website, hoàn thiện search engine với dữ liệu bài báo đã crawl được từ các webstie.
- Sinh viên lưu file với tên theo định dạng sau: Lab5_MSSV.ipynb,
- Sinh viên nộp lên Elearning của lớp học tại buổi thực hành 5.
- Deadline: 17h00, ngày 14/03/2023

II. Yêu cầu

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

III. Nội dung

1. Chuẩn bị

1.1. Thực hiện

B1. Tạo new directory trong project lab 4, đặt tên directory là json_file.

B2. Module utils.py, tạo các function như sau:

- **add_news(): để thêm dữ liệu bài báo thu thập được vào database**

```
#define functin Add bài báo được crawling từ web vào server sử dụng newspaper3k
#tạo query để insert bài báo craw được vào table news, gọi conn.commit() để add vào database
def add_news(conn,url,category_id):
    query = """
    INSERT INTO news(subject,description,image,original_url,category_id)
    VALUES(?,?,?,?,?)
    """
    article = Article(url)
    article.download()
    article.parse()

    conn.execute(query, (article.title,article.text,article.top_image,article.url,category_id))
    conn.commit()
```

- **get_news_url(): để build link chi tiết các bài báo từ trang chủ các website được lưu trong category**

```
# với mỗi category được lấy, sẽ tiến hành lấy id, url, sau đó gọi method build của newspaper\
# để build link.
# với mỗi link sẽ gọi function add_news để add nội dung bài báo craw được vào database,
# sử dụng try - except để bỏ qua một số trường hợp không thể parse.
def get_news_url():
    cats = get_all("SELECT * FROM category") # lấy danh mục web từ DB
    conn = sqlite3.connect("DATA/data.db") # tạo connect tới DB
    for cat in cats: # với mỗi trang trong DB
        cat_id = cat[0] # lấy id
        url = cat[2] # lấy link gốc
        cat_paper = newspaper.build(url) # dùng phương thức build của newspaper tạo ra các link cần craw
        for article in cat_paper.articles:
            try:
                print("===",article.url)
                add_news(conn,article.url,cat_id)# gọi add_news để add link dl vào DB
            except Exception as ex:
                print("ERROR:" + str(ex))
                pass

    conn.close() # đóng kết nối
```

B3. Module render_templates.py, tạo các function như sau:

- **get_news(): load các news từ database ghi vào file json: news.json trong directory json_file.**

```
# read news
def get_news():
    rows = utils.get_all("SELECT * FROM news")
    data = []
    for r in rows:
        data.append(
        {
            "id": r[0],
            "subject": r[1],
            "description": r[2],
            "image": r[3],
            "original_url": r[4],
            "category_id": r[5]
        })
    with open("json_file/news.json", "w", encoding="utf8") as f:
        json.dump(data, f)
```

- **read_news(keyword = none): đọc các news từ file json news.json để tạo ra dữ liệu cho news.html**

```
def read_news(keywords = None):
    with open("json_file/news.json", encoding="utf8") as f:
        news = json.load(f)
    if keywords:
        news = [n for n in news if n["subject"].lower().find(keywords.lower()) >= 0]
    return news
```

B4. Module api.py: define endpoint để render dữ liệu từ news.html thành dữ liệu website, nếu không có keyword sẽ load tất cả news, nếu có keyword sẽ load các news chứa keyword tương ứng

```
# get tất cả news from database
@app.route("/news", methods = ["GET"]) # define endpoint để thực thi API
def get_news():
    kw = request.args.get("keywords", None)
    return render_template("news.html", data=render_templates.read_news(kw))
```

B5: Thiết lập file news.html

```
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <title> DANH MỤC BÀI BÁO </title>
</head>
<body>
<h1> DANH MỤC CÁC BÀI BÁO </h1>
<p>
    <form>
        <div class = "form-group">
            <label> Tìm kiếm theo tiêu đề </label>
            <input type = "text" name = "keywords"> <br>
            <input type = "submit" value = "Tìm">
        </div>
    </form>
</p>
```

```
<table class ="table">
        <tr>
            <th> ID </th>
            <th> Tiêu đề </th>
            <th> Link gốc bài </th>
            <th> ID danh mục </th>
        </tr>
        {% for new in data %}
        <tr>
            <td>
                {{new.id}}
            </td>
            <td>
                {{new.subject}}
            </td>
            <td>
                {{new.original_url}}
            </td>
            <td>
                {{new.category_id}}
            </td>
        </tr>
        {% endfor %}
    </table>
</body>
</html>
```