

# BÁO CÁO ASSIGNMENT 1

Bài 1:

Để thực hiện yêu cầu của bài, em sử dụng các thư viện:

```
from bs4 import BeautifulSoup
from bs4 import Comment
import pandas as pd
import requests
import os
```

Trong đó:

- +BeautifulSoup: được sử dụng để phân tích cú pháp HTML và trích xuất dữ liệu từ các thẻ HTML
- +pandas: dùng để xử lý dữ liệu, cho phép lưu trữ và phân tích dữ liệu
- +requests: dùng để thực hiện các yêu cầu HTTP và nhận nội dung từ trang web
- +os: dùng để tương tác với hệ điều hành.

Các chương trình con:

```
def crawler(url, stat_id, table_id):
```

-Chương trình dùng để nhận url của trang web và id để thu thập dữ liệu.

-Các bước chương trình thực thi:

- +Đầu tiên, chương trình nhận URL của trang web, ID thống kê và số hiệu bảng để thu thập dữ liệu
- +Sử dụng BeautifulSoup để phân tích HTML và tìm kiếm bảng thống kê với id theo stat\_id được nhận
- +Trích xuất dữ liệu từ các hàng tr(table row) của bảng và lưu vào 1 dictionary(res)
- +Chuyển đổi dictionary thành DataFrame của pandas và lưu thành tệp csv.

```
def cleaner():
```

-Chương trình dùng để hợp nhất các bảng dữ liệu thu được thành 1 bảng result.csv

-Các bước chương trình thực thi:

- +Đầu tiên, đọc bảng table1.csv làm bảng cơ sở
- +Duyệt các bảng từ table3.csv đến table 10.csv, loại bỏ các cột trùng lặp và hợp nhất với bảng cơ sở
- +Kết hợp dữ liệu với table2.csv dựa trên các cột chung
- +Lọc ra các cầu thủ có số phút chơi từ 90 phút trở lên
- +Sắp xếp kết quả theo tên cầu thủ và tuổi, lưu lại vào tệp result.csv.

```
def delete_tables():
```

-Xóa các tệp csv từ table1 đến table 10 để tránh gây trùng lặp cho lần chạy sau và để thư mục gọn gàng.

```
def delete_result():
```

-Xóa tệp result.csv để tránh gây trùng lặp cho lần chạy sau.

```
if __name__ == '__main__':
```

-Trong hàm main:

- +Đầu tiên, gọi delete\_result để xóa tệp result.csv nếu có tồn tại.
  - +Tiếp theo, khai báo mảng các URL và ID cần thu thập dữ liệu
  - +Thực hiện request đến trang web để lấy thông tin cầu thủ và trích xuất dữ liệu vào table1.csv
  - +Gọi hàm crawler() cho từng URL trong mảng đã khai báo bên trên và lưu dữ liệu lần lượt vào các file từ table2.csv cho đến table10.csv
  - +Cuối cùng, gọi hàm cleaner() để xử lý và tổng hợp dữ liệu, và delete\_tables() để xóa các bảng table.csv đã tạo.
- Ta nhận được được file result.csv khi chạy chương trình như trong hình:

