BÁO CÁO BÀI TẬP LỚN MÔN LẬP TRÌNH PYTHON

GIẢNG VIÊN: KIM NGỌC BÁCH

HỌ VÀ TÊN SV: NGUYỄN HỒNG QUANG

MSV: B23DCVT361

LỚP: D23CQCE04 - B

**🧠 1. Mục tiêu bài tập lớn**

Mục tiêu của bài là xây dựng một quy trình phân tích dữ liệu cầu thủ Ngoại hạng Anh từ trang FBref và FootballTransfers:

* **Thu thập dữ liệu chi tiết về các cầu thủ** qua web scraping.
* **Xử lý và phân tích thống kê** để rút ra các đặc trưng nổi bật.
* **Áp dụng học máy (Machine Learning)** để phân cụm (clustering) cầu thủ.
* **Ghép dữ liệu định giá cầu thủ** từ nguồn khác và xuất ra file tổng hợp.

**📦 2. Tóm tắt từng phần code**

**🔹 B1.py – Thu thập & xử lý dữ liệu từ FBref**

**Chức năng:**

* Sử dụng Selenium và BeautifulSoup để scrape dữ liệu từ trang FBref của 20 đội Ngoại hạng Anh.
* Lấy 10 bảng thống kê (standard, shooting, passing, defense, etc.) và hợp nhất theo các cột chung như "Player", "Nation", "Squad", "Pos".
* Làm sạch dữ liệu: xóa cột thừa, chuẩn hóa tên quốc gia, chỉ giữ cầu thủ đá trên 90 phút.
* Kết quả được lưu thành file results.csv.

**Thư viện sử dụng:**

* Selenium, BeautifulSoup (web scraping)
* Pandas, StringIO, time

**🔹 B2.py – Phân tích thống kê và trực quan hóa dữ liệu**

**Chức năng:**

* Đọc results.csv, xác định các cột số để phân tích.
* Tìm **Top 3 cao nhất / thấp nhất** cho từng chỉ số → ghi vào top\_3.txt.
* Tính **trung vị, trung bình, độ lệch chuẩn** cho từng đội và toàn bộ giải → ghi vào results2.csv.
* Vẽ **biểu đồ histogram** cho từng chỉ số, cả toàn giải và từng đội → lưu vào thư mục histograms/.
* Xác định **cầu thủ có chỉ số cao nhất** mỗi cột → ghi vào best\_players.txt.

**Thư viện sử dụng:**

* Pandas, NumPy, Matplotlib, re, os

**🔹 B3.py – Phân cụm cầu thủ với KMeans & PCA**

**Chức năng:**

* Đọc lại results.csv, chuẩn hóa dữ liệu số.
* Dùng **Elbow method** để xác định số lượng cụm tối ưu → lưu hình elbow\_method.png.
* Áp dụng **KMeans** để chia cụm (ví dụ: 3 cụm).
* Dùng **PCA** để giảm chiều dữ liệu và trực quan hóa kết quả phân cụm trong không gian 2D → lưu hình player\_clusters.png.

**Thư viện sử dụng:**

* Pandas, NumPy, Matplotlib, Seaborn
* Sklearn (KMeans, StandardScaler, PCA)

**🔹 B4.py – Thu thập giá trị cầu thủ và ghép dữ liệu**

**Chức năng:**

* Dùng Selenium để thu thập danh sách cầu thủ có giá trị cao từ trang FootballTransfers (22 trang).
* Làm sạch tên cầu thủ (hàm normalize\_name), dùng **fuzzy matching** để ghép tên giữa hai bảng dữ liệu.
* Lọc ra những cầu thủ đã đá > 900 phút, ghép với giá trị thị trường → lưu file transfer\_player.csv.

**Thư viện sử dụng:**

* Pandas, Selenium, BeautifulSoup, RapidFuzz, re, time, Pathlib

**🛠️ 3. Các công nghệ & thư viện sử dụng chính**

| **Thư viện/Công nghệ** | **Mục đích sử dụng** |
| --- | --- |
| **Selenium** | Truy cập và tương tác với website để scrape dữ liệu |
| **BeautifulSoup** | Phân tích HTML từ trang web |
| **Pandas / NumPy** | Xử lý, phân tích và thao tác dữ liệu bảng |
| **Matplotlib / Seaborn** | Vẽ biểu đồ phân tích và phân cụm |
| **Scikit-learn** | Chuẩn hóa, phân cụm (KMeans), giảm chiều (PCA) |
| **RapidFuzz** | So khớp tên cầu thủ bằng fuzzy string matching |
| **Regex, OS, Pathlib** | Hỗ trợ xử lý chuỗi, quản lý file hệ thống |

**📌 4. Tổng kết**

Dự án của bạn là một ví dụ thực tiễn và chuyên nghiệp về cách kết hợp kỹ năng lập trình, xử lý dữ liệu và machine learning để phân tích dữ liệu thể thao. Quy trình được chia thành 4 bước rõ ràng:

1. **Thu thập dữ liệu** (Web scraping từ nhiều nguồn)
2. **Làm sạch & xử lý** (filter, chuẩn hóa, loại bỏ thiếu)
3. **Phân tích thống kê & trực quan hóa**
4. **Kết hợp học máy & nâng cao dữ liệu**