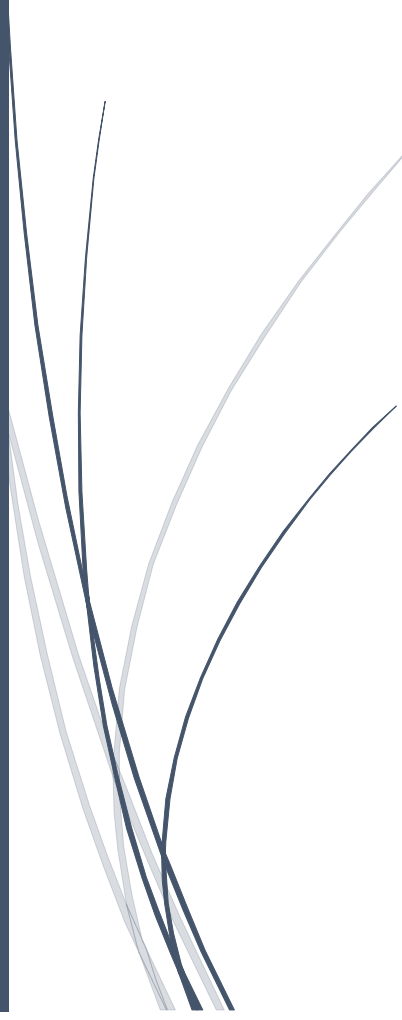
Học viện công nghệ bưu chính viễn thông
Khoa công nghệ thông tin

11/2/2024

Báo cáo bài tập lớn môn Python

Họ tên:	Ngô Thảo Nguyên
Mã sinh viên:	B22DCCN590
Lớp:	D22CQCN02-B
Nhóm:	Python Nhóm 11
Giảng viên:	Kim Ngọc Bách



Bài 1:

Bài code hoàn chỉnh gồm 3 phần :

a. Tiêu đề :
   Mục đích: Để đặt tên và cấu trúc các cột dữ liệu, giúp xác định các chỉ số cần lấy từ trang web và đảm bảo dữ liệu được sắp xếp chính xác khi ghi vào file CSV, dễ dàng điều chỉnh các chỉ số khi cần :
b. Player:
   Mục đích : Thu thập thông tin chi tiết của từng cầu thủ.
   Các bước làm :
   Bước 1: Phân tích yêu cầu và cấu trúc dữ liệu

1. Xác định các chỉ số cần thu thập dựa trên yêu cầu.
2. Xác định cấu trúc file results.csv, trong đó mỗi cột tương ứng với một chỉ số, và các cầu thủ sẽ được sắp xếp theo tên và độ tuổi.

   Bước 2: Lập trình thu thập dữ liệu từ website fbref.com

1. **Thư viện cần thiết**: Sử dụng thư viện requests để gửi yêu cầu HTTP và BeautifulSoup để phân tích HTML,pandas để xử lý dữ liệu và ghi file CSV.
2. **Gửi yêu cầu đến trang web**: Truy cập trang thống kê các cầu thủ Ngoại hạng Anh mùa giải 2023-2024 trên trang fbref.com.
3. **Phân tích HTML để tìm các bảng dữ liệu**: Xác định các thẻ HTML chứa bảng dữ liệu thống kê của cầu thủ.

   Bước 3: Thu thập dữ liệu và xử lý các chỉ số

1. **Lọc cầu thủ**: Chỉ thu thập dữ liệu của các cầu thủ có số phút thi đấu lớn hơn 90 phút.
2. **Thu thập các chỉ số yêu cầu**:
   - **Thông tin cơ bản**: Nation, Team, Position, Age,…
   - **Chỉ số chơi bóng và hiệu suất**: Như non-Penalty Goals, Penalty Goals, Assists, Yellow Cards, Red Cards,…
   - **Chỉ số mong đợi và tiến trình**: xG, npxG, PrgC, PrgP, PrgR,…
   - **Sắp xếp cầu thủ**: Theo tên và độ tuổi từ lớn đến nhỏ nếu trùng tên.
3. **Điền giá trị thiếu**: Nếu một chỉ số không có dữ liệu, điền là "N/a".

   Bước 4: Ghi dữ liệu vào file results.csv

1. Sử dụng pandas.DataFrame để tạo bảng từ dữ liệu đã thu thập.

c. Sắp xếp
   Mục đích: Sắp xếp cầu thủ theo tên và tuổi

Bài 2:

Bước 1: Đọc dữ liệu và chuẩn bị xử lý

Đọc dữ liệu từ file results.csv.

Bước 2: Tính toán trung vị, trung bình, độ lệch chuẩn

Dùng các hàm thống kê của pandas và numpy để tính toán trung vị, trung bình và độ lệch chuẩn cho từng chỉ số.

a) Tính trung vị, trung bình và độ lệch chuẩn cho toàn giải

b) Tính trung bình và độ lệch chuẩn cho từng đội

Nhóm dữ liệu theo đội và tính toán từng chỉ số theo từng đội.

Bước 3: Tìm top 3 cầu thủ cao nhất và thấp nhất cho mỗi chỉ số

Duyệt qua từng cột chỉ số và tìm ra top 3 cầu thủ có giá trị cao nhất và thấp nhất.

Bước 4: Ghi kết quả vào file results2.csv

Ghi kết quả thống kê và top 3 vào file CSV.

Bước 5: Vẽ histogram phân bố cho mỗi chỉ số

Dùng matplotlib để vẽ biểu đồ histogram cho từng chỉ số. Điều này sẽ giúp bạn hình dung phân bố điểm số.

Bước 6: Xác định đội bóng có phong độ tốt nhất

So sánh các chỉ số trung bình của từng đội để tìm đội có phong độ tốt nhất dựa trên các chỉ số.

Bước 7: Ghi nhận và phân tích phong độ tốt nhất

Dựa trên các chỉ số quan trọng như bàn thắng và chỉ số xG, xác định đội có phong độ tốt nhất.

Bài 3.

Phần 1: Phân loại cầu thủ bằng K-Means

1. Xác định số lượng cụm: Tự chọn n=3
2. Áp dụng K-means

Phần 2: Nhận xét nhóm

Phần 3: Giảm số chiều bằng PCA và vẽ phân cụm trên mặt phẳng 2D

1. Giảm chiều dữ liệu bằng PCA:
    - Áp dụng PCA để giảm chiều dữ liệu từ nhiều chỉ số xuống 2 chiều.
2. Vẽ phân cụm trong không gian 2D:

- Dùng Matplotlib để vẽ các điểm dữ liệu (các cầu thủ) trong không gian 2D, với mỗi cụm được đánh dấu màu khác nhau.

Phần 4: Viết chương trình vẽ biểu đồ radar so sánh cầu thủ

Sử dụng Matplotlib để tạo biểu đồ radar.