2. Giới thiệu Pandas

Pandas mà chúng ta đang viết trong chương này không liên quan gì đến những chú gấu trúc dễ thương, và chúng không phải là những gì đọc giả của chúng ta mong muốn trong một hướng dẫn Python. Pandas là một mô-đun Python, được làm mạnh mẽ lên các khả năng của Numpy, Scipy và Matplotlab. Từ pandas là một từ viết tắt có nguồn gốc từ "Python and data analysis" and "panel data".

Có một số nhầm lẫn về việc Pandas có thay thế cho Numpy, SciPy và Matplotlib hay không. Sự thật là nó được xây dựng trên nền Numpy. Điều này có nghĩa là Numpy là lõi của pandas. Scipy và Matplotlib mặt khác không đòi hỏi cần trong pandas nhưng chúng rất hữu ích. Đó là lý do tại sao dự án Pandas liệt kê chúng là "tùy chọn phụ thuộc".

Pandas là một thư viện phần mềm viết cho ngôn ngữ lập trình Python. Nó được sử dụng cho thao tác và phân tích dữ liệu. Nó cung cấp cấu trúc dữ liệu đặc biệt và hoạt động cho các thao tác của các bảng số liệu và chuỗi thời gian. Pandas là phần mềm miễn phí được phát hành theo giấy phép BSD.

Python với Pandas được sử dụng trong nhiều lĩnh vực bao gồm lĩnh vực học thuật và thương mại bao gồm tài chính, kinh tế, thống kê, phân tích, v.v.

Các đặc điểm nổi bật của pandas:

Các công cụ load dữ liệu vào bộ nhớ từ nhiều định dạng khác nhau.

Liên kết dữ liệu và tích hợp xử lý dữ liệu bị thiếu.

Xoay và chuyển đổi chiều của dữ liệu dễ dàng.

Tách, đánh chỉ mục và chia nhỏ các tập dữ liệu lớn dựa trên nhãn.

Cột dữ liệu có thể bị xóa hoặc thêm mới.

Có thể nhóm dữ liệu cho các mục đích aggregation và transformations.

Hợp nhất và nhập dữ liệu hiệu quả.

Là công cụ làm việc với series data hiệu quả.

Cài đặt:

Gói pandas không đi kèm với bộ cài đặt python gốc. Nếu chúng ta sử dụng bộ cài Anaconda2 thì pandas đã tích hợp sẵn. Nếu các bạn đang dùng bộ cài Python nguyên gốc thì có thể cài đặt gói pandas qua công cụ pip như sau: **pip install pandas**

Cấu trúc dữ liệu:

Pandas có ba cấu trúc dữ liệu và nó được xây dựng dựa trên thư viện Numpy vậy nên chúng hoạt động rất nhanh và hiệu quả:

Series

DataFrame

Panel

Trong ba kiểu dữ liệu, DataFrame là kiểu dữ liệu được sử dụng rộng rãi nhất, trong khi Panel lại ít được sử dụng nhất.

Cách tốt nhất để có một hình dung về ba cấu trúc dữ liệu trong pandas này là cấu trúc dữ liệu có chiều cao hơn chứa cấu trúc dữ liệu có chiều thấp hơn. Ví dụ: dataframe chứa tập các Series, Panel chứa tập các dataframe.

Cấu trúc dữ liệu	Chiều	Miêu tả
Series	1	Cấu trúc dạng mảng 1D đồng nhất, có kích thước cố định
Dataframe	2	Cấu trúc dạng bảng 2D, kích thước có thể thay đổi được với các cột đã được tạo không đồng nhất. Dữ liệu một cột là đồng nhất.
Panel	3	Cấu trúc mảng 3D, kích thước có thể thay đổi được.

Ví dụ về ba loại cấu trúc dữ liệu phía trên:

Series: Ví dụ chuỗi các số nguyên

12	34	21	34	45	67	76	34	26	90

DataFrame: Các bạn nhận thấy dữ liệu là đồng nhất trong một cột nhưng có thể không đồng nhất giữa các cột, ví dụ “Countryname” kiểu string trong khi ‘TotalMedals’ và ‘Population’ là kiểu integer.

Bảng “Total Medals vs Population“.

CountryName	**TotalMedals**	**Population**
Great Britain	802	63
France	765	66
Norway	451	5
Japan	435	127
Canada	423	35

South Korea	288	50
Austria	287	9
Brazil	108	220
New Zealand	100	5

Panel: Panel là một cấu trúc dữ liệu ba chiều với dữ liệu không đồng nhất. Rất khó để đại diện cho bảng để biểu diễn đồ họa. Nhưng bằng dataframe có thể được minh họa dưới dạng tập của các dataframe.

		Open	Close	High	Low
major	minor				
3/31/2015	IBM	0.023602	0.132903	0.180478	0.90085
	APPL	0.421412	0.212665	0.434623	0.808586
	CVX	0.568055	0.409201	0.575276	0.398886
	BHP	0.487414	0.515413	0.919127	0.175972
4/30/2015	IBM	0.150868	0.457895	0.337286	0.643918
	APPL	0.204729	0.957179	0.637748	0.839661
	CVX	0.090679	0.888687	0.440256	0.776386
	BHP	0.831527	0.714202	0.931162	0.896876
5/31/2015	IBM	0.788582	0.922422	0.263036	0.344814
	APPL	0.329716	0.304964	0.792955	0.158162
	CVX	0.36578	0.981508	0.61547	0.182513

	BHP	0.313848	0.882293	0.540126	0.004455

Kết luận

Pandas là một bộ thư viện hỗ trợ trong Python, với chức năng chính là phân tích và thao tác trên dữ liệu. Pandas được xây dựng dựa trên nền thư viện Numpy nên nó kế thừa khả năng tính toán nhanh và hiệu quả trên các mảng và ma trận lớn, thích hợp để sử dụng cho lĩnh vực data analysis. Trong bài này chúng tôi có giới thiệu ba cấu trúc dữ liệu trong Pandas là Serial, DataFrame và Panel. Mà hai trong số chúng được sử dụng khá rộng rãi và sẽ được chúng tôi đề cập chi tiết hơn trong những bài học tiếp theo.