Equipment with new general education program, Volume 1, Issue 324 (November 2024)
ISSN 1859 - 0810

Nghiên cứu và ứng dụng Google Colab trong trực quan hóa dữ liệu với Python

Vũ Văn Huân*, Trương Mạnh Đạt*

**ThS. Khoa CNTT,T rường ĐH Tài nguyên và Môi trường Hà Nội*

Received: 18/10/2024; Accepted: 28/10/2024; Published:4/11/2024

Abstract: *Data visualization is the representation of data in a visual way through graphics. One of those visualization forms is visualization through images, charts or graphs. Complex data will become more understandable when represented visually. Currently, with the development of science and technology, the era of the 4th industrial revolution (CMCN 4.0), the era of big data, in addition to processing large, complex data, data visualization is an essential factor in the data science process. Data visualization helps highlight important data, distilling it into an easy-to-understand format. In this article, we will focus on researching and visualizing data with the Python programming language on the Google Colab environment.*

Keywords: *Google Colab, Data Visualization, Data Science, Python Programming.*

1. Đặt vấn đề

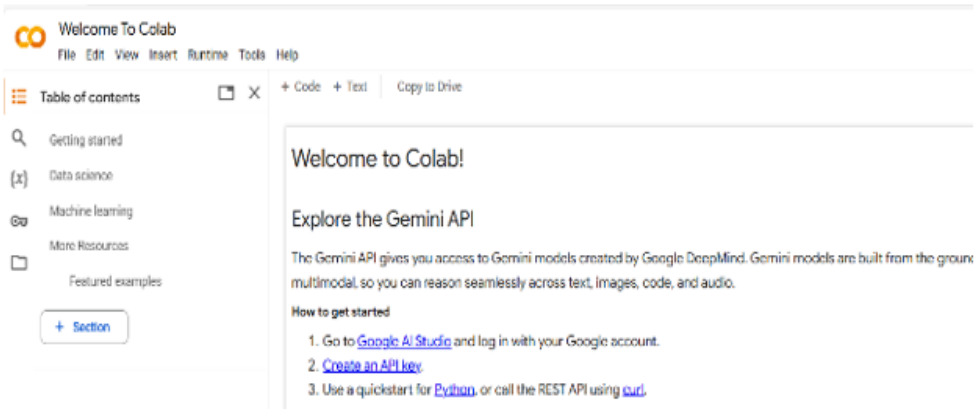
Trực quan hóa dữ liệu là biểu diễn dữ liệu theo cách trực quan thông qua đồ họa. Một trong những hình thức trực quan đó là trực quan thông qua các hình ảnh, biểu đồ hoặc đồ thị. Dữ liệu phức tạp sẽ trở nên dễ hiểu hơn khi được biểu diễn một cách trực quan hóa[1][2]. Hiện nay, với sự phát triển của khoa học công nghệ, thời đại cuộc cách mạng công nghiệp lần thứ 4 (CMCN 4.0), thời đại của dữ liệu lớn, ngoài việc xử lý dữ liệu lớn, phức tạp thì việc trực quan hóa dữ liệu là một yếu tố thiết yếu quan trọng trong quy trình khoa học dữ liệu. Trực quan hóa dữ liệu giúp làm nổi bật những dữ liệu quan trọng, chắt lọc thành định dạng dễ hiểu. Trong bài báo này sẽ tập trung nghiên cứu, trực quan hóa dữ liệu với ngôn ngữ lập trình Python trên môi trường Google Colab.

2. Nội dung nghiên cứu

2.1. Công cụ Google Colab

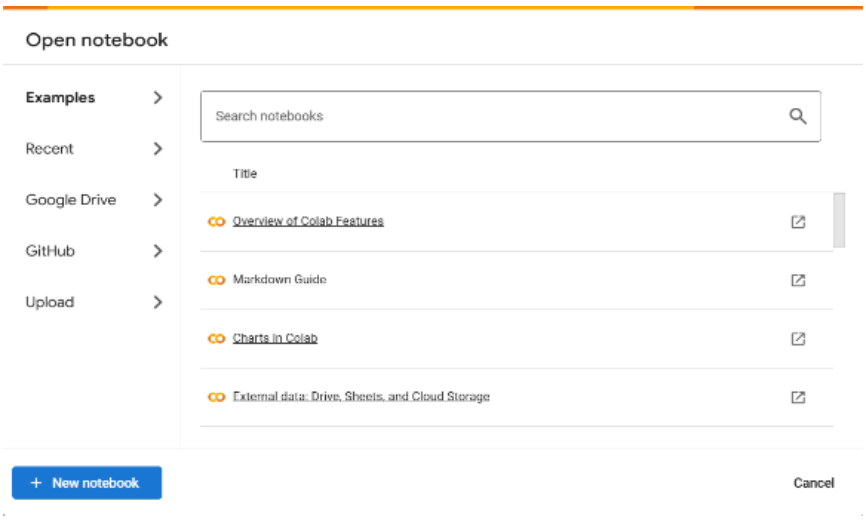
Google Colab được cung cấp bởi Goole. Google Colab là một dịch vụ cung cấp môi trường cho phép thực thi các mã lệnh Python trên nền tảng đám mây. Google Colab không yêu cầu thiết lập, cài đặt hay cần cấu hình máy tính để sử dụng mà cung cấp quyền truy cập miễn phí vào các tài nguyên điện toán, bao gồm GPUs và TPUs. Google Colab phù hợp trong nghiên cứu học máy và khoa học dữ liệu. Đồng thời, cũng hỗ trợ tốt cho các dự án về AI [3].

- Giao diện của Google Colab (hình 2.1).



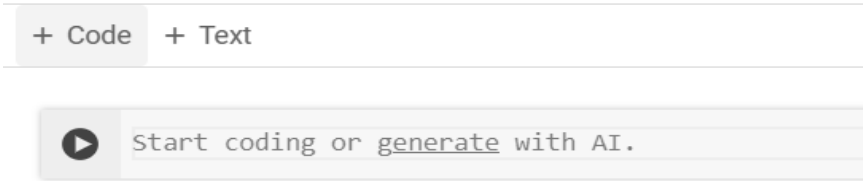
Hình 2.1. Giao diện của Google Colab

- Google Colab cho phép thao tác với tệp tin và thư mục như tạo mới một notebook, mở hoặc upload một file từ máy tính cá nhân và tải lên file notebook. Đặc biệt, Google Colab cho phép mở một file từ Google Drive hoặc cho phép kết nối với GitHub (hình 2.2).



Hình 2.2. Thao tác với File và Folder

- Google Colab cho phép thêm hoặc copy/paste Code cell (thực thi lệnh) và Text cell (văn bản) rất thuận lợi (hình 2.3).



Hình 2.3. Thêm, chỉnh sửa mã lệnh

Đồng thời, Google Colab cho phép tùy biến hiển thị như xem lại lịch sử các dòng lệnh đã thực thi (Executed Code History), thu gọn các nội dung (Collapse sections) thuận lợi.

2.2. Thực thi Python trong môi trường Google Colab [2][4]

- Python là một trong những ngôn ngữ lập trình được sử dụng rộng rãi trong lĩnh vực khoa học dữ liệu. Python hỗ trợ nhiều thư viện cho phép tạo trực quan hóa dữ liệu như: Matplotlib, Seaborn, Keras, Pandas, SciPy,…

- Trên Google Colab cho phép cài đặt các thư viện mới không có sẵn trong Google Colab để thực thi mã lệnh Python. Sử dụng câu *lệnh pip install package_name* để cài đặt một thư viện không có sẵn.

Ví dụ: Để cài đặt thư viện seaborn, sử dụng câu lệnh như sau: pip install seaborn

- Hoặc sử dụng câu lệnh *import package_name as something* để import một thư viện bất kỳ vào Google Colab.

Ví dụ: Để nhập (import) thư viện và tải (load) dữ liệu "tips", sử dụng câu lệnh sau:

```
import seaborn as sns
tips = sns.load_dataset("tips")
```

- Các mã lệnh trên Google Colab có thể được thực thi toàn bộ (Run all) hoặc thực thi các dòng lệnh xuất hiện trước hoặc sau code cell đang chọn hoặc thực thi code cell đang chọn. Bên cạnh đó, trong quá trình thực thi mã lệnh Google Colab cho phép dừng thực thi mã lệnh hoặc khởi động lại môi trường.

- Để tận dụng sức mạnh GPUs và TPUs trên môi trường đám mây (Cloud) google Colab cho phép lựa chọn Cloud Runtime hoặc sử dụng Local Runtime giúp không phải cài đặt hoặc thiết lập lại thư viện cần thiết sau mỗi lần sử dụng Google Colab.

2.3. Trực quan hóa dữ liệu với Python trong Google Colab

2.3.1. Bộ dữ liệu và các bước thực hiện

Bộ dữ liệu mẫu Iris Dataset chứa thông tin kích thước của các phần khác nhau trên bông hoa[5]. Bộ dữ liệu Iris gồm có 5 biến: Sepal.Length: chiều dài đài hoa, Sepal.Width: chiều rộng đài hoa, Petal.Length: chiều dài cánh hoa, Petal.Width : chiều rộng cánh hoa, Species: Loài hoa. Trực quan hóa dữ liệu với tập dữ liệu Iris trong môi trường Google Colab sử dụng Python được thực hiện như sau:

Bước 1: Tạo một Colab Notebooks

Bước 2: Upload bộ dữ liệu Iris.csv

Bước 3: Import các thư viện seaborn, matplotlib, pandas

Bước 4: Đọc dữ liệu Iris.csv

Bước 5: Trực quan hóa tập dữ liệu Iris.csv thông qua một số loại biểu đồ như biểu đồ phân tán, biểu đồ hộp,biểu đồ tương quan phức hợp,…

2.3.2. Một số biểu đồ trực quan hóa dữ liệu

a. Hiển thị dữ liệu

```
iris = pd.read_csv("Iris.csv")
iris.head()
iris["Species"].value_counts()
```

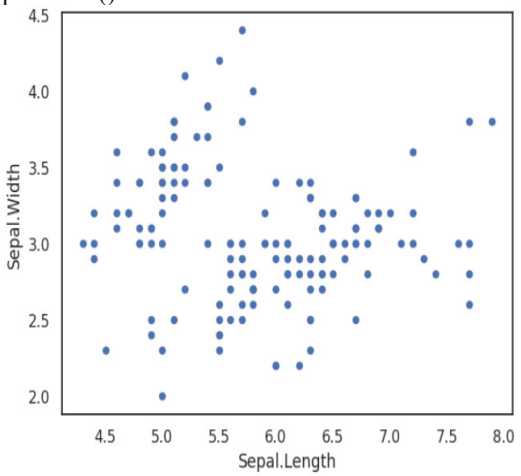
rownames	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa

Species	count
setosa	50
versicolor	50
virginica	50

Hình 2.4. Thông tin tập dữ liệu Iris.csv và số lượng mỗi loài hoa

b. Biểu đồ phân tán (Scatter): Biểu đồ dạng phân tán thể hiện mối tương quan giữa chiều dài, chiều rộng của cánh hoa

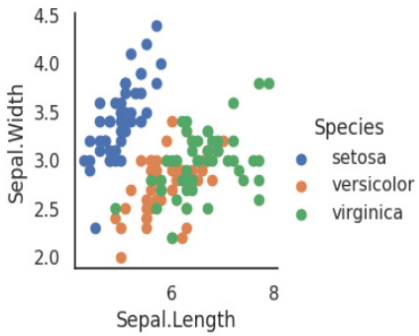
```
iris.plot(kind="scatter", x="Sepal.Length", y="Sepal.Width")
plt.show()
```



Hình 2.5. Biểu đồ phân tán thể hiện mối tương quan giữa chiều dài, chiều rộng đài

c. Biểu đồ tương quan phức hợp (FacetGrid): Biểu đồ phức hợp thể hiện mối tương quan giữa chiều dài, chiều rộng của cánh hoa

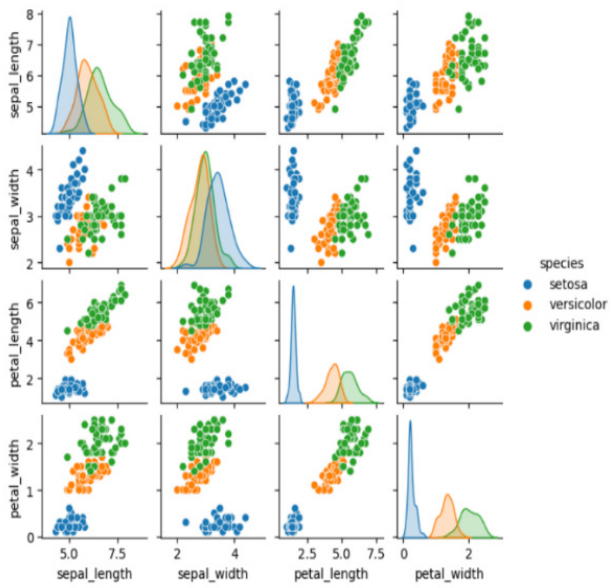
```
sns.FacetGrid(iris, hue="Species")\
        .map(plt.scatter, «Sepal.Length», «Sepal.Width»)\
    .add_legend()
plt.show()
```



Hình 2.6. Biểu đồ phức hợp thể hiện mối tương quan giữa chiều dài, chiều rộng đài

c. Biểu đồ cặp (pairplot): Biểu đồ dạng cặp thể hiện mối tương quan giữa các chiều dữ liệu theo từng cặp.

```
sns.pairplot(iris.drop("rownames", axis=1), hue="Species", size=4)
plt.show()
```



Hình 2.7. Biểu đồ thể hiện mối tương quan giữa các chiều dữ liệu theo từng cặp

3. Kết luận

Trực quan hóa dữ liệu là một bước thiết yếu trong khoa học dữ liệu. Trực quan hóa dữ liệu là cách biểu diễn hay thể hiện dữ liệu thông qua đồ họa dạng các đồ thị hoặc biểu đồ. Trực quan hóa làm nổi bật, chắt lọc thông tin hữu ích, giúp người đọc hiểu rõ hơn về dữ liệu. Công cụ Google Colab là một trong những môi trường được ứng dụng phổ biến trong lĩnh vực học máy và khoa học dữ liệu, hỗ trợ tốt ngôn ngữ lập trình Python và các thư viện hỗ trợ trong xử lý dữ liệu lớn, trực quan hóa dữ liệu. Google Colab có thể coi là một trong những công cụ hỗ trợ đắc lực cho các nhà nghiên cứu về khoa học dữ liệu nói chung và trong trực quan hóa dữ liệu nói riêng.

Tài liệu tham khảo

[1]. Chun-houh Chen, Wolfgang Härdle (2008), Antony Unwin, *Handbook of Data Visualization,* 2008 Springer-Verlag Berlin Heidelberg.

[2]. Wes McKinney (2013), *Python for Data Analysis,* Published by O'Reilly Media, Inc.

[3]. https://colab.research.google.com.

[4]. https://www.datacamp.com.

[5].https://archive.ics.uci.edu/dataset/53/iris