R VÀ PYTHON TRONG PHÂN TÍCH DỮ LIỆU
R AND PYTHON FOR DATA ANALYTICS

Th.S Nguyễn Thị Hoài Linh
Trường Đại học Kinh tế - Tài chính thành phố Hồ Chí Minh, linhnth@uef.edu.vn

Tóm tắt: *Trong bối cảnh công nghệ thông tin và truyền thông xã hội ngày càng phát triển và ứng dụng rộng rãi trong sản xuất và kinh doanh, việc phân tích dữ liệu trở thành một yếu tố then chốt cho các doanh nghiệp và tổ chức. Bài nghiên cứu này nhằm giới thiệu và đánh giá hai ngôn ngữ được sử dụng nhiều trong phân tích dữ liệu là R và Python. Bài viết bao gồm 3 phần chính: giới thiệu về bài toán phân tích dữ liệu, giới thiệu về hai ngôn ngữ R và Python; đồng thời so sánh một số đặc điểm của hai ngôn ngữ.*

Từ khóa*: Dữ liệu, Khoa học dữ liệu, Phân tích dữ liệu, Python, R,*

Abstract*: In the context of information technology and social communication increasingly developing and widely applied in production and business, data analytics becomes a key factor for enterprises and organizations. This study aims to introduce and evaluate two languages that are widely used in data analysis, namely R and Python. The article consists of three main parts: introduction to the problem of data analysis, introduction to two languages R and Python; at the same time compare the popularity of the two languages.*

Keywords*: Data, Data Science, Data Analytics, Python, R*

1. Giới thiệu

Trong thời đại công nghệ thông tin và cách mạng công nghệ 4.0, công nghệ thông tin được áp dụng rộng rãi trong các lĩnh vực kinh tế. Đặc biệt, sự phát triển của dữ liệu lớn, khoa học dữ liệu và các ngôn ngữ phân tích dữ liệu đã góp phần thúc đẩy sự phát triển nhanh chóng của phân tích dữ liệu trong kinh tế. Các ngôn ngữ lập trình phân tích dữ liệu cùng các gói mô đun đơn giản, linh hoạt, tương thích với nhiều loại dữ liệu và ngữ cảnh khác nhau, không chỉ cho ra các kết quả phân tích dữ liệu chính xác mà còn hỗ trợ trong các bài toán dự báo và xử lý các dữ liệu thiếu sót.

2. Cơ sở lý thuyết

2.1 Phân tích dữ liệu

Phân tích dữ liệu (data analytics) là quá trình tìm ra, giải thích và trình bày các mô hình có ý nghĩa trong dữ liệu để tạo ra các báo cáo, các bản thống kê tổng hợp theo yêu cầu của người dùng. Các tổ chức doanh nghiệp có thể sử dụng phân tích dữ liệu kinh doanh để biểu diễn, dự báo và nâng cao hiệu quả kinh doanh.

Bài báo gồm 5 phần. Phần 1 giới thiệu tổng quan. Phần 2 trình bày cơ sở lý thuyết. Phần 3 giới thiệu R và Python đồng thời phân tích ưu nhược điểm của hai ngôn ngữ trên. Phân tích mức độ phổ biến của hai ngôn ngữ được trình bày ở phần 4. Phần 5 kết luận.

2.2 Bài toán phân tích dữ liệu

Phân tích dữ liệu trong các tổ chức, doanh nghiệp là bài toán đã được chú trọng trong những năm gần đây. Bài toán phân tích dữ liệu không phải là bài toán mới nhưng các ngôn ngữ hỗ trợ phân tích dữ liệu ngày càng được phát triển và đa dạng nhằm phù hợp với nhiều ngữ cảnh và nhiều kiểu dữ liệu khác nhau từ dữ liệu văn bản, dữ liệu số đến các dữ liệu đa phương tiện, các kiểu dữ liệu có yếu tố thời gian, hoặc phân tích dữ liệu theo thời gian thực. Một quy trình phân tích dữ liệu thông thường có 3 giai đoạn gồm: Tìm hiểu thu thập dữ liệu mà tổ chức, doanh nghiệp cần; Phân tích dữ liệu và phân loại dữ liệu; Cuối cùng là tạo các báo cáo và xác định các chiến lược kinh doanh dựa trên dữ liệu phân tích.

3. R và Python

Python và R đều là những ngôn ngữ lập trình tuyệt vời cho khoa học dữ liệu và phân tích dữ liệu.

Các đặc điểm chung của R và Python:

- Đều là ngôn ngữ lập trình đa mô hình, hỗ trợ lập trình hướng đối tượng, lập trình mệnh lệnh, lập trình thủ tục,...
- Đều là ngôn ngữ thông dịch và ngôn ngữ lập trình cấp cao.
- Được sử dụng để phát triển các thuật toán.
- Là mã nguồn mở và miễn phí.
- Được tích hợp với các cơ sở dữ liệu như MySQL, Oracle, v.v.
- Hỗ trợ các tệp khác nhau như tệp CSV, tệp Excel, tệp XML và tệp JSON.
- Dễ sử dụng và dễ học.

Mặc dù R vs Python phổ biến trong phân tích dữ liệu và học máy. Tuy nhiên cả hai ngôn ngữ đều có các tính năng khác nhau. Mỗi ngôn ngữ cung cấp những ưu và nhược điểm khác nhau.

3.1 Ngôn ngữ Python

Từ khi phát hành vào năm 1991, Python đã trở nên phổ biến và được sử dụng rộng rãi trong xử lý dữ liệu [1].

Python là một ngôn ngữ đa mục đích, giống như C ++ và Java, với cú pháp dễ đọc, dễ học. Các lập trình viên sử dụng Python để đi sâu vào phân tích dữ liệu hoặc sử dụng học máy trong các môi trường sản xuất mở rộng.

Python thường được sử dụng khi các tác vụ phân tích dữ liệu cần được tích hợp với các ứng dụng web hoặc nếu mã thống kê cần được tích hợp vào cơ sở dữ liệu sản xuất. Vì nó là một ngôn ngữ lập trình đầy đủ, Python là một công cụ tốt để triển khai các thuật toán áp dụng trong sản xuất.

Ưu điểm của Python

- Dễ học: Python có cấu trúc code dễ đọc và dễ hiểu, dù là người mới bắt đầu.
- Tính linh hoạt: Python là một trong những ngôn ngữ linh hoạt nhất. Code Python gọn gàng, không phức tạp để sử dụng và có cấu trúc tốt. Tính linh hoạt của Python giúp việc phân tích dữ liệu trở nên dễ dàng.
- Mã nguồn mở: Python có thể được tải xuống dễ dàng. Đồng thời Python có những diễn đàn hỗ trợ tích cực như pydev.vn, python.org, … , và bất kỳ ai cũng có thể đóng góp vào việc cải thiện các thư viện và chức năng của Python.
- Thư viện: Python có nhiều thư viện cần thiết để thực hiện các chức năng chính liên quan đến phân tích dữ liệu như: thư viện *Matplotlib* giúp tạo ra các đồ thị và biểu đồ cơ bản, thư viện *Seaborn* cho phép vẽ đồ họa thống kê, thư viện *Pandas* hỗ trợ lọc, sắp xếp và phân tích dữ liệu chỉ trong vài giây…
- Năng suất: Khả năng tích hợp và kiểm soát của Python được tăng cường và tiết kiệm rất nhiều thời gian.
- Có thể nhúng: Mã Python có thể được tích hợp với các ngôn ngữ lập trình khác như C ++.

Nhược điểm của Python

- Tốc độ: Python là một ngôn ngữ thông dịch và do đó tương đối chậm hơn các ngôn ngữ lập trình khác.
- Môi trường di động: Python không phù hợp với môi trường Android và iOS.
- Tiêu thụ bộ nhớ: Python tiêu thụ một lượng RAM đáng kể. Quá trình này trở nên chậm hơn khi cần truy cập nhiều đối tượng hơn.
- Lớp truy cập cơ sở dữ liệu: Các lớp truy cập cơ sở dữ liệu của Python kém phát triển so với Kết nối cơ sở dữ liệu Java (JDBC) và Kết nối cơ sở dữ liệu mở (ODBC).
- Threading: Threading hoặc flow của nhiều hàm cùng một lúc là một nhược điểm trong Python do Global Interpreter Lock (GIL) của nó.

3.2 Ngôn ngữ R

Phát hành lần đầu năm 1995, R là công cụ rất mạnh cho máy học, thống kê và phân tích dữ liệu [2].

R được xây dựng bởi các nhà thống kê và tập trung nhiều vào các mô hình thống kê và phân tích chuyên biệt. Các nhà khoa học dữ liệu sử dụng R để phân tích thống kê chuyên sâu, R chỉ chỉ sử dụng một vài dòng mã tuy nhiên có thể cho kết quả với hình ảnh hóa dữ liệu đẹp mắt. R đi kèm với một loạt các kỹ thuật thống kê như mô hình tuyến tính, mô

hình phi tuyến tính, kiểm tra thống kê, phân cụm, v.v.

Ưu điểm của ngôn ngữ R

- Mã nguồn mở: R là một ngôn ngữ mã nguồn mở và được tải xuống và sử dụng miễn phí. Người dùng cũng có thể đóng góp bằng cách tối ưu hóa mã nguồn của nó.
- Nền tảng độc lập: R độc lập với nền tảng và có thể hoạt động trên tất cả các hệ điều hành như UNIX, Windows và Mac.
- Data Wrangling: Thông qua các gói của R như *readr* và *dplyr*, R có khả năng chuyển đổi một mã lộn xộn thành một mã có cấu trúc.
- Biểu đồ và Đồ thị: R được chuẩn bị tốt để trực quan hóa dữ - một bước cần thiết trong báo cáo phân tích dữ liệu. R có hàng nghìn thư viện để trực quan hóa dữ liệu ví dụ như *ggplot2, lattice, plotly*...
- Tính khả dụng của gói: R có nhiều gói dành riêng cho việc phát triển học máy, phân tích dữ liệu và các dự án thống kê.

Nhược điểm của ngôn ngữ R

- Bộ nhớ: R tiêu tốn nhiều bộ nhớ hơn khi tất cả các đối tượng được lưu trữ trong bộ nhớ vật lý. Theo thời gian, khi chương trình có dữ liệu lớn hơn, quá trình này sẽ chậm lại.
- Bảo mật: R thiếu bảo mật cơ bản khiến việc nhúng vào các ứng dụng web thực tế trở nên khó khăn.
- Khó học: Không giống như Python, R là một ngôn ngữ phức tạp và khó học cho người mới bắt đầu.
- Thời gian chạy chậm: R là ngôn ngữ xử lý chậm. So với các ngôn ngữ khác như MATLAB và Python, phải mất nhiều thời gian hơn để đưa ra kết quả.
- Xử lý dữ liệu: Xử lý dữ liệu trong R rất tẻ nhạt vì nR yêu cầu tất cả dữ liệu phải ở một nơi. Điều này làm cho R không lý tưởng cho Big Data.

4. Mức độ phổ biến

4.1 Ứng dụng của ngôn ngữ Python

Python là một ngôn ngữ đa mục đích và linh hoạt, nên nó có thể được sử dụng trong nhiều lĩnh vực chính vì vậy Python có rất nhiều ứng dụng khác nhau như:

- Phát triển phần mềm: Python là một ngôn ngữ lập trình mạnh mẽ và đa năng, thích hợp cho việc phát triển phần mềm từ ứng dụng desktop đơn giản đến hệ thống phức tạp. Nó cung cấp các framework như Django và Flask để phát triển ứng dụng web.
- Khoa học dữ liệu và máy học: Python là một trong những ngôn ngữ phổ biến nhất trong lĩnh vực khoa học dữ liệu và máy học. Các thư viện như NumPy, Pandas và Matplotlib hỗ trợ trong việc xử lý dữ liệu, phân tích và hiển thị kết quả. Thư viện machine learning như Scikit-learn và TensorFlow cũng được sử dụng rộng rãi trong việc xây dựng mô hình và huấn luyện.
- Web development: Python có thể được sử dụng để phát triển các ứng dụng web động và tĩnh. Django và Flask là hai framework phổ biến giúp xây dựng các ứng dụng web mạnh mẽ và linh hoạt.
- Điều khiển và tự động hóa: Python được sử dụng rộng rãi trong việc viết các script để điều khiển và tự động hóa các tác vụ. Với các thư viện như Selenium, bạn có thể tự động hoá việc điều khiển trình duyệt và thao tác trên các trang web.
- Game development: Python cung cấp các thư viện như Pygame để phát triển các trò chơi đơn giản. Nó cũng được sử dụng trong phát triển trò chơi trên nền tảng web.
- Ứng dụng di động: Python cung cấp các framework như Kivy và BeeWare để phát triển ứng dụng di động đa nền tảng.
- Hacking và bảo mật: Python cũng được sử dụng trong lĩnh vực hacking và bảo mật, đặc biệt trong việc phân tích và khai thác dữ liệu, viết các công cụ tấn công và bảo vệ hệ thống.

4.2 Ứng dụng của ngôn ngữ R

R là một ngôn ngữ phổ biến trong lĩnh vực thống kê và phân tích dữ liệu, một số ứng dụng phổ biến của R:

- Phân tích dữ liệu và thống kê: R được sử dụng rộng rãi trong lĩnh vực phân tích dữ liệu và thống kê. Các gói phần mềm

như ggplot2, dplyr và tidyr cung cấp các công cụ mạnh mẽ để khám phá và xử lý dữ liệu, tạo biểu đồ và thực hiện các phân tích thống kê phức tạp.

- Khoa học dữ liệu: R là công cụ phổ biến trong nghiên cứu khoa học và phân tích dữ liệu. Nó được sử dụng để xử lý dữ liệu, tạo biểu đồ, thực hiện kiểm định thống kê, xây dựng mô hình và hiển thị kết quả.
- Machine learning: R cung cấp nhiều gói phần mềm phổ biến để triển khai và huấn luyện các mô hình machine learning. Các gói phổ biến bao gồm caret, randomForest, xgboost và keras. R cũng hỗ trợ việc thực hiện cross-validation và tinh chỉnh mô hình.
- Bioinformatics: R là công cụ quan trọng trong lĩnh vực sinh học thông tin, nơi nó được sử dụng để xử lý và phân tích dữ liệu sinh học, bao gồm các loại dữ liệu như chuỗi DNA, dữ liệu gen và dữ liệu về cấu trúc protein.
- Visualizations: R cung cấp các gói phần mềm như ggplot2 và plotly để tạo biểu đồ và trực quan hóa dữ liệu. Điều này giúp người dùng hiểu và trình bày dữ liệu một cách rõ ràng và hấp dẫn.
- Data mining: R cung cấp các gói phần mềm như arules và rpart để thực hiện khai phá dữ liệu và phát hiện mẫu trong dữ liệu lớn.
- Điều khiển và tự động hóa: R có thể được sử dụng để viết các script và chương trình để tự động hóa các tác vụ phân tích dữ liệu và thống kê, giúp tiết kiệm thời gian và nâng cao hiệu suất công việc.

4.3 So sánh độ phổ biến giữa Python và R

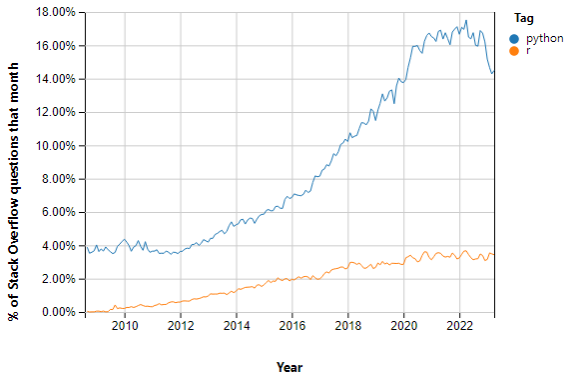
Theo báo cáo của The Importance Of Being Earnest (TOIBE) [3]. Python dẫn đầu bảng xếp hạng độ phổ biến của các ngôn ngữ lập trình với xếp hạng 13.45% trong báo cáo tháng 5 năm 2023, thể hiện mức tăng 0.74% so với tháng 5 năm 2022. Các ưu điểm của Python như tính linh hoạt, dễ sử dụng và có cộng đồng sử dụng rộng lớn đã góp phần vào sự phát triển to lớn này. Một lí do khác làm cho Python trở nên phổ biến hơn là cơ sở người dùng phong phú, đa dạng của Python; bao gồm các nhà phát triển và lập trình viên. Tuy nhiên Python chủ yếu phát triển trong lĩnh vực sản xuất.

R ít phổ biến hơn Python. Theo báo cáo tháng 5 năm 2023 của TIOBE, nó được xếp hạng là ngôn ngữ phổ biến thứ 16 với xếp hạng 0.82%, tương ứng với mức thay đổi -0,39% trong vòng 1 năm, giảm 3 hạng so với tháng 5 năm 2022. Lý giải cho điều này là vì cơ sở người dùng của R chủ yếu thuộc lĩnh vực học thuật, bao gồm các nhà khoa học dữ liệu và nghiên cứu và phát triển (R&D), những người thực hiện phân tích dữ liệu.

May 2023	May 2022	Change	Programming Language	Ratings	Change
1	1		Python	13.45%	+0.71%
2	2		C	13.35%	+1.76%
3	3		Java	12.22%	+1.22%
4	4		C++	11.96%	+3.13%
5	5		C#	7.43%	+1.04%
6	6		Visual Basic	3.84%	-2.02%
7	7		JavaScript	2.44%	+0.32%
8	10	↑	PHP	1.59%	+0.07%
9	9		SQL	1.48%	-0.39%
10	8	↓	Assembly language	1.20%	-0.72%
11	11		Delphi/Object Pascal	1.01%	-0.41%
12	14	↑	Go	0.99%	-0.12%
13	24	⇈	Scratch	0.95%	+0.29%
14	12	↓	Swift	0.91%	-0.31%
15	20	⇈	MATLAB	0.88%	+0.06%
16	13	↓	R	0.82%	-0.39%

Hình 4.1: Chỉ số TIOBE xếp hạng mức độ phổ biến của các ngôn ngữ lập trình

Số liệu từ Stack Overflow cũng cho thấy Python phổ biến hơn R về số lượng câu hỏi được hỏi trên tháng từ năm 2010 tới năm 2022 [4].



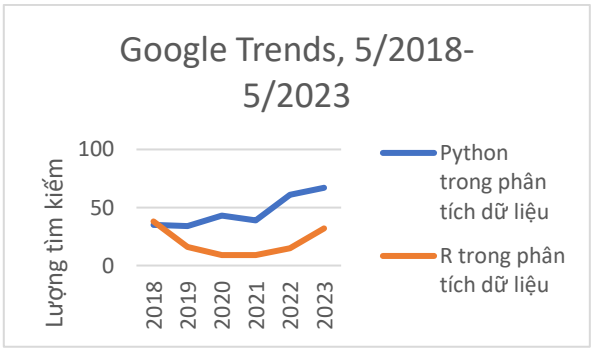
Hình 4.2: Stack Overflow tỷ lệ tìm kiếm từ khóa R và Python

Biểu đồ ở hình 4.2 cho thấy, trong toàn bộ cộng đồng nhà phát triển, Python dường như được quan tâm hơn R; tuy nhiên, điều quan trọng cần lưu ý là Python là ngôn ngữ lập trình có mục đích chung trong khi R chuyên về tính toán thống kê, điều đó có nghĩa là sự so sánh này không đánh giá mức độ phổ biến của Python hơn R trong lĩnh vực phân tích dữ liệu [6].

Theo số liệu từ Google Trends, trong 5 năm gần đây, tỷ lệ tìm kiếm Python cho phân tích dữ liệu cao hơn so với tìm kiếm R cho phân tích dữ liệu [5].

Python data analytics: 45

R data analytics: 23



Hình 4.3: Google Trends tỷ lệ tìm kiếm R và Python trong phân tích dữ liệu

5. Kết luận

Cả hai ngôn ngữ R và Python đều có ưu và nhược điểm riêng. Ngoài lĩnh vực phân tích dữ liệu, Python hầu như phổ biến trong lĩnh vực phát triển ứng dụng vì nó là ngôn ngữ đa mục đích, trong khi đó R được phát triển để phân tích thống kê. Cả hai ngôn ngữ đều cung cấp nhiều loại thư viện ứng dụng trong các trường hợp khác nhau. Do đó, việc lựa chọn sử dụng ngôn ngữ nào hoặc chọn kết hợp cả hai hoàn toàn phụ thuộc vào người dùng như kinh nghiệm lập trình, môi trường, vấn đề cần giải quyết, mức độ quan trọng của biểu đồ và đồ thị….

Tài liệu tham khảo

[1] Advantages of python programming language in hydrological model development, Milan T, Dragoljub B, Vesna RV, Dusan P, 6/2022.

[2] Advantages of R as a tool for data Analysis and Visualization in Social Sciences, MI Fernasndez Lizana, 8/2020.

[3] https://www.tiobe.com/tiobe-index/

https://online.codegym.vn/2022/06/19/phan-tich-du-lieu-bang-python-nen-hay-khong/

https://www.kdnuggets.com/2017/08/python-overtakes-r-leader-analytics-data-science.html

Ngày truy cập: 16/05/2023

[4]https://insights.stackoverflow.com/trends?tags=python%2Cr

Ngày truy cập: 16/05/2023

[5]https://trends.google.com/trends/explore?date=2018-04-16%202023-05-16&q=Python%20data%20analytics,R%20data%20analytics

Ngày truy cập: 16/05/2023

[6]https://www.linkedin.com/pulse/python-vs-r-who-really-ahead-data-science-machine-piatetsky-shapiro

Ngày truy cập: 16/05/2023