Meta Mind
Explore Knowledge - Open Future

BUỔI 1 : Tổng quan về Khoa học Dữ liệu

AGENDA

Meta Mind
Explore Knowledge - Open Future

- Giới thiệu về môn học
- Tổng quan về Khoa học Dữ liệu
- Cài đặt Anaconda / Jupyter Notebook
- Markdown trên Jupyter
- Phương pháp luận Khoa học dữ liệu
- Git/Github
- Tổng kết

GIỚI THIỆU VỀ MÔN HỌC

Meta Mind
Explore Knowledge - Open Future

Mục tiêu môn học:

- Nắm vững các khái niệm về Khoa học dữ liệu
- Nắm vững các công cụ dành cho Khoa học dữ liệu
- Nắm vững về phương pháp luận trong Khoa học dữ liệu và có thể vận dụng vào quy trình giải quyết dự án về Khoa học dữ liệu
- Vận dụng được Python cơ bản, cấu trúc dữ liệu bằng Python
- Vận dụng được làm việc với dữ liệu văn bản thông qua Numpy, Pandas
- Hiểu rõ cách làm việc với dữ liệu hình ảnh, âm thanh trong Python
- Vận dụng được làm việc với dữ liệu Web thông qua Regex và Json

Thời lượng học ước tính của môn: 6 buổi

Checklist chuẩn đầu ra của môn học: Checklist Module 1

TỔNG QUAN VỀ KHOA HỌC DỮ LIỆU

Meta Mind
Explore Knowledge - Open Future

Định nghĩa về Khoa học Dữ liệu:

- Là quá trình sử dụng dữ liệu để hiểu về sự vật, hiện tượng, hiểu thế giới.
- Tồn tại một mô hình hoặc giả thiết cho một vấn đề, bạn cố gắng xác nhận giả thiết hoặc mô hình đó với dữ liệu của bạn.
- Là nghệ thuật khám phá hiểu biết và xu hướng ẩn sau của dữ liệu.
- Diễn giải dữ liệu thành một câu chuyện và sử dụng khả năng kể chuyện để kiến tạo hiểu biết về sự vật, hiện tượng; cung cấp các lựa chọn chiến lược cho công ty hoặc tổ chức.
- Là lĩnh vực về các quy trình và hệ thống trích xuất dữ liệu từ các dạng dữ liệu khác nhau.
- Là nghiên cứu về dữ liệu, nỗ lực làm việc với dữ liệu, trả lời câu hỏi mà chúng ta muốn khám phá.

KHDL Là sự kết hợp của

Dữ liệu

Sự tò mò

Thao tác DL

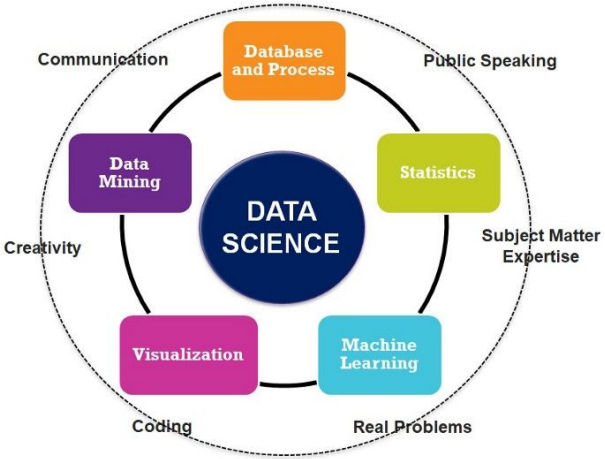
Khám phá DL

Phân tích DL

TỔNG QUAN VỀ KHOA HỌC DỮ LIỆU

Phẩm chất của nhà khoa học dữ liệu:

- ***Tính tò mò***: Giúp biết cách sử dụng dữ liệu và đưa ra các hành động phù hợp.
- ***Thích tranh luận và đưa ra ý kiến riêng***: Không bảo thủ, giúp củng cố giả thuyết và nâng cao chất lượng của quá trình phân tích.
- ***Khả năng linh hoạt với các nền tảng phân tích***: Sự quen thuộc và linh hoạt với công cụ, phần mềm, và nền tảng là yếu tố quan trọng.
- ***Khả năng kể truyện***: Truyền đạt hiểu biết thông qua câu chuyện, giúp người tiếp nhận dễ hiểu và áp dụng thông tin.
- ***Xác định lợi thế cạnh tranh***: Đòi hỏi sự hiểu biết sâu rộng về lĩnh vực và khả năng nhận diện những điểm mạnh riêng để tạo lợi thế cạnh tranh.

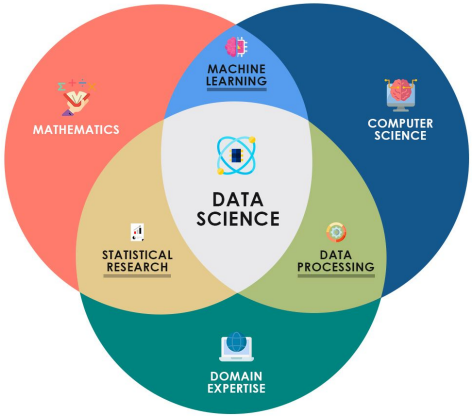


TỔNG QUAN VỀ KHOA HỌC DỮ LIỆU

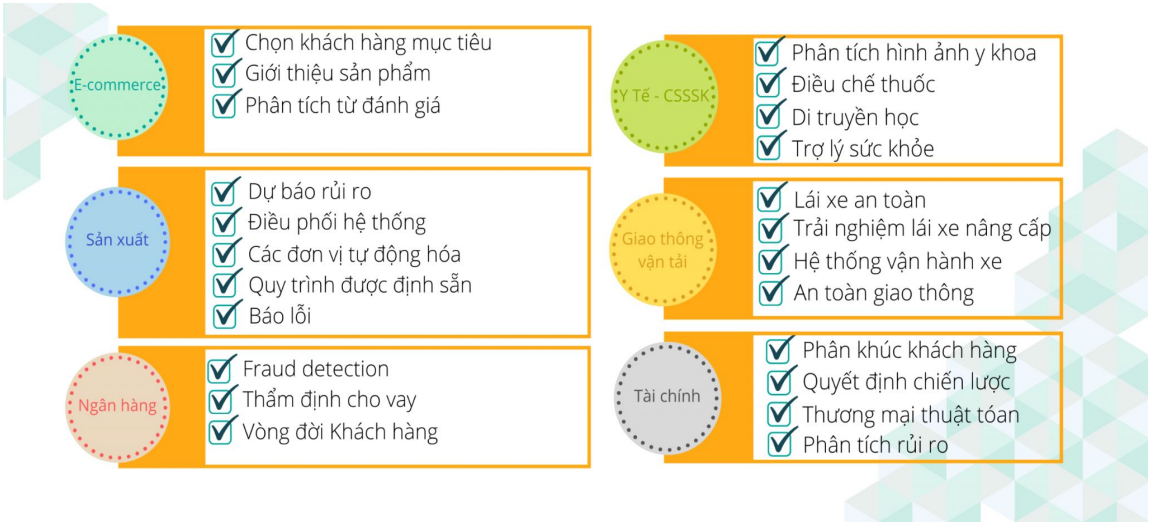
Meta Mind
Explore Knowledge - Open Future

Kỹ năng cần có của nhà KHDL:

- Lập trình / tư duy máy tính
- Đại số & Hình học
- Giải tích cơ bản
- Xác suất và thống kê cơ bản
- Cơ sở dữ liệu



6 Ứng dụng của Data Science:



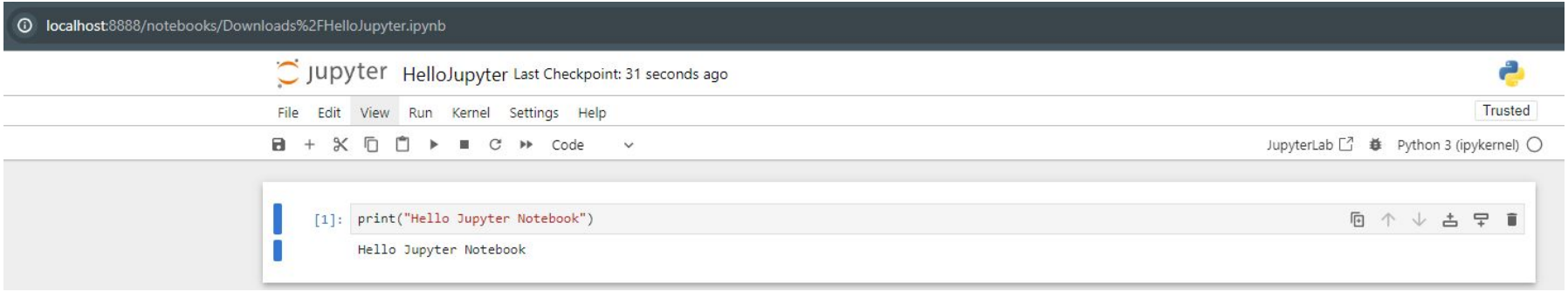
CÀI ĐẶT ANACONDA / JUPYTER NOTEBOOK

Meta Mind
Explore Knowledge - Open Future

Hướng dẫn cài đặt Anaconda (khuyến khích): Hướng dẫn cài đặt Anaconda

Hướng dẫn chỉ cài đặt Jupyter Notebook: Hướng dẫn cài đặt Jupyter Notebook

Chạy thử Jupyter Notebook



Thực hành thao tác Anaconda cơ bản: Thực hành thao tác Anaconda

MARKDOWN TRÊN JUPYTER NOTEBOOK

Meta Mind
Explore Knowledge - Open Future

Markdown trên Jupyter là một công cụ định dạng văn bản đơn giản và trực quan được tích hợp trong Jupyter Notebook. Nó cho phép bạn tạo các tài liệu đẹp mắt và dễ đọc với nhiều kiểu văn bản, hình ảnh, bảng biểu, và hơn thế nữa.

Ưu điểm:

- **Dễ sử dụng:** Cú pháp Markdown đơn giản và dễ học, giúp bạn tạo tài liệu nhanh chóng và hiệu quả.
- **Đa dạng:** Markdown hỗ trợ nhiều kiểu văn bản, hình ảnh, bảng biểu, và các yếu tố khác để tạo tài liệu phong phú.
- **Tích hợp:** Markdown được tích hợp sẵn trong Jupyter Notebook, giúp bạn tạo tài liệu liền mạch với mã code.
- **Dễ chia sẻ:** Tài liệu Markdown có thể được chia sẻ và xem trên nhiều nền tảng khác nhau.

Hướng dẫn sử dụng Markdown: Hướng dẫn sử dụng Markdown

Thực hành viết Markdown: Bài tập thực hành viết Markdown

PHƯƠNG PHÁP LUẬN KHOA HỌC DỮ LIỆU

Meta Mind
Explore Knowledge - Open Future

Phương pháp này giúp nhà khoa học dữ liệu tự tin và hướng dẫn họ về bước tiếp theo mà không gặp khó khăn. Có 10 câu hỏi cơ bản được chia thành 3 phần chính, đảm bảo sự chính xác và đúng đắn trong công việc của nhà khoa học dữ liệu.

Từ bài toán đến hướng tiếp cận giải quyết (xác định vấn đề và hướng tiếp cận) :

1. Vấn đề mà bạn đang cố gắng giải quyết là gì?
2. Làm thế nào bạn có thể sử dụng dữ liệu để trả lời câu hỏi?

Làm việc với dữ liệu (tổ chức dữ liệu):

3. Bạn cần dữ liệu nào để trả lời câu hỏi?
4. Dữ liệu bạn cần đến từ đâu (xác định tất cả các nguồn) và bạn sẽ lấy nó như thế nào?
5. Dữ liệu bạn thu thập có giải quyết cho vấn đề cần giải quyết không?
6. Cần thêm việc gì để thao tác và làm việc với dữ liệu?

Tìm ra giải pháp (xác nhận dữ liệu và phương pháp luận được thiết kế):

7. Dữ liệu được trực quan hóa theo cách nào giải quyết câu trả lời được yêu cầu?
8. Mô hình có thực sự trả lời được câu hỏi ban đầu hay không? Cần điều chỉnh gì hay không?
9. Bạn có thể đẩy mô hình của bạn vào thực tiễn hay không?
10. Bạn có thể nhận được những phản hồi mang tính xây dựng để trả lời câu hỏi không?

PHƯƠNG PHÁP LUẬN KHOA HỌC DỮ LIỆU

Meta Mind
Explore Knowledge - Open Future

Từ bài toán đến hướng tiếp cận giải quyết

- Business Understanding (Hiểu biết về doanh nghiệp): Xác định mục tiêu, yêu cầu của doanh nghiệp.
- Analytical Approach (Phương pháp phân tích): Xác định và áp dụng các phương pháp phân tích dữ liệu.

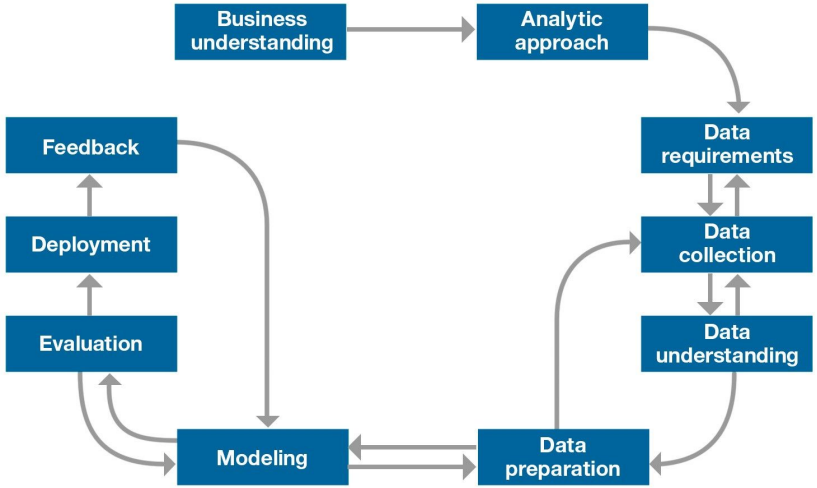
Làm việc với dữ liệu:

- Data Requirements (Yêu cầu về dữ liệu): Xác định loại dữ liệu cần thiết để phân tích và đáp ứng yêu cầu bài toán.
- Data Collection (Thu thập dữ liệu): Thu thập dữ liệu từ nhiều nguồn khác nhau.
- Data Understanding (Hiểu dữ liệu): Hiểu cấu trúc, tính chất và đặc điểm của dữ liệu.
- Data Preparation (Chuẩn bị dữ liệu): Tiền xử lý dữ liệu, bao gồm xử lý dữ liệu thiếu, loại bỏ nhiễu, và chuyển đổi dữ liệu

Tìm ra giải pháp:

- Modeling (Mô hình): Xây dựng mô hình dự đoán hoặc mô tả dữ liệu.
- Evaluation (Đánh giá): Đánh giá hiệu suất của mô hình.
- Development (Phát triển): Phát triển các giải pháp hoặc sản phẩm dựa trên kết quả đánh giá mô hình.
- Feedback (Phản hồi): Cung cấp phản hồi và điều chỉnh quy trình theo kết quả.

Vòng đời của dự án Khoa học dữ liệu



Git & Github

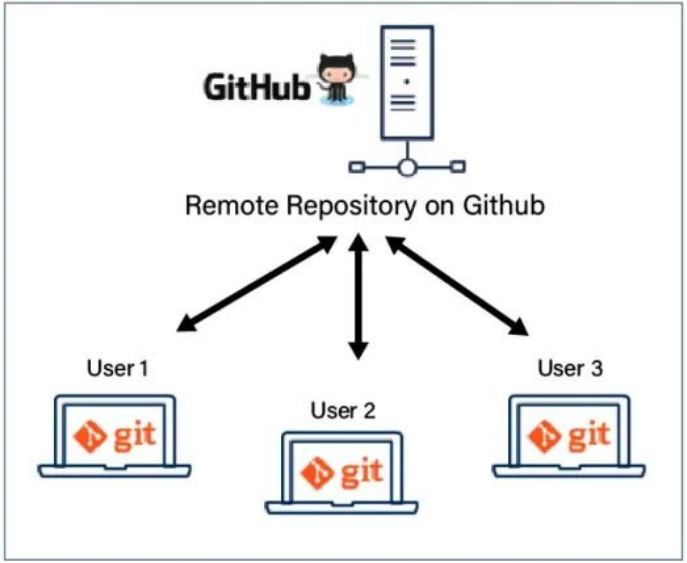
Git là phần mềm mã nguồn mở miễn phí, là một hệ thống kiểm soát phiên bản phân tán, có nghĩa là người dùng ở mọi nơi trên thế giới đều có thể có một bản sao của dự án của bạn trên máy tính của riêng họ. Khi họ thực hiện các thay đổi, họ có thể đồng bộ hóa phiên bản của họ với một máy chủ từ xa để chia sẻ nó với bạn.

Ưu điểm của Git:

- Theo dõi thay đổi của mã nguồn dễ dàng.
- Hỗ trợ cộng tác hiệu quả, dễ dàng quay lại các phiên bản trước đó.
- Phân nhánh và hợp nhất code linh hoạt.
- Mã nguồn mở và miễn phí.

Lưu trữ dữ liệu:

- Git lưu trữ dữ liệu dưới dạng các snapshot (ảnh chụp), còn gọi là các trạng thái của mã nguồn tại một điểm thời gian cụ thể.
- Mỗi khi có sự thay đổi trong mã nguồn, Git chỉ lưu trữ những phần thay đổi (diff) thay vì lưu trữ toàn bộ file mới.
- Nhờ vậy, Git giúp tiết kiệm dung lượng lưu trữ và cho phép truy cập nhanh chóng đến các phiên bản trước đó.

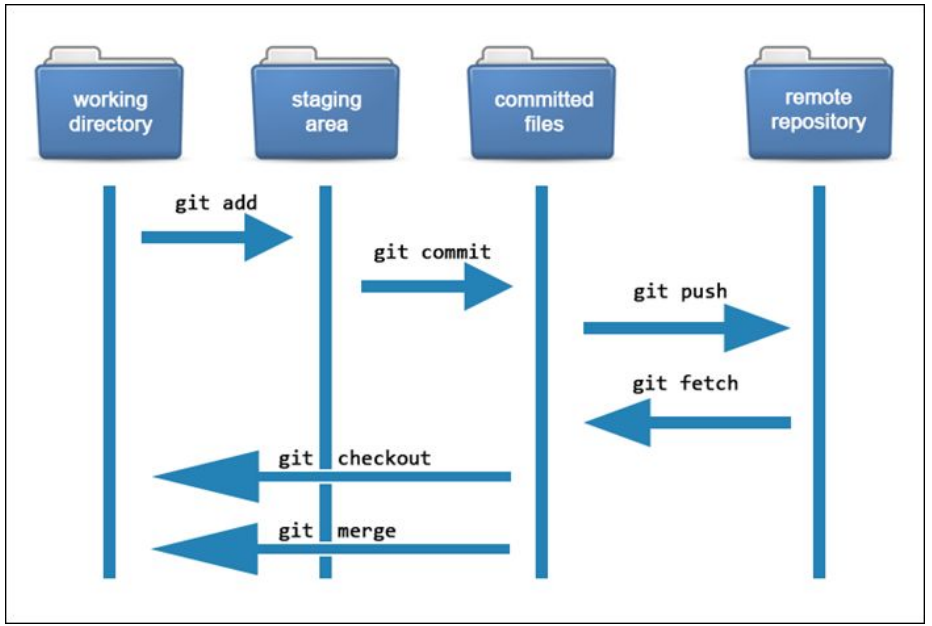


Khu vực làm việc:

- **Working Tree:** Nơi chứa các tệp mã nguồn đang được chỉnh sửa.
- **Staging Area (Index):** Nơi lưu trữ các tệp đã được chọn để commit.
- **Local Repository:** Nơi lưu trữ tất cả các snapshot của dự án.

Quá trình làm việc:

- **Sửa đổi:** Thay đổi mã nguồn trong Working Tree.
- **Thêm vào Staging Area:** Chọn các tệp đã thay đổi và thêm vào Staging Area bằng lệnh git add.
- **Commit:** Lưu trữ các thay đổi trong Staging Area vào Local Repository bằng lệnh git commit.
- **Push:** Gửi các thay đổi đã commit lên kho lưu trữ remote (trên Github) bằng lệnh git push.



Xem thêm: Git/Github

Thuật ngữ	Giải thích
Branch	Nhánh, phiên bản cụ thể trong kho lưu trữ tách ra từ project
Commit	Thời điểm cụ thể trong lịch sử thực hiện code
Check out	Chuyển đổi giữa các branch
Master/main	Nhánh chính trong kho lưu trữ
Merge	Bổ sung các thay đổi từ nhánh này sang nhánh khác
Pull	Yêu cầu thay đổi cho nhánh chính
Push	Cập nhật các branch từ xa
Repository	Kho lưu trữ Git

Lệnh	Tác dụng
git clone	Sao chép kho lưu trữ Git từ xa
git status	Xem trạng thái các thay đổi trong khu vực làm việc
git add	Thêm tệp tin vào khu vực Staging Area
git commit	Lưu lại các thay đổi đã được thêm vào Staging Area
git push	Gửi các thay đổi đã commit lên kho lưu trữ từ xa
git branch	Liệt kê các nhánh hiện có
git checkout	Chuyển đổi sang nhánh khác
git merge	Hợp nhất các thay đổi từ một nhánh khác vào nhánh hiện tại

Thực hành với Github:

- Làm quen với Github
- Tạo nhánh và gộp nhánh Github
- Hướng dẫn nộp bài tập lên Git

TỔNG KẾT

Meta Mind
Explore Knowledge - Open Future

Hoàn thành Form điểm danh sau (10'): Form điểm danh

Bài tập sau buổi học:

- Lab 2
- Lab 3
- Lab 4
- Lab 5

Yêu cầu:

- Đặt tên folder theo mẫu DSMMC1_<Buổi học>_<Tên> (Ví dụ: *DSMMC1_Buoi1_DangNH*)
- Hoàn thành các bài lab sau đó nộp lên Git chung của lớp: Link Git , sau đó submit link folder git trên Form bài tập
- Hạn nộp: Trước 12 giờ của buổi học tiếp theo

THANK YOU !