Thời gian

Đồ án sẽ kéo dài trong vòng 1 tuần. Chi tiết về thời gian cụ thể có thể được tìm thấy trên Moodle.

Yêu cầu nộp đồ án

- Sinh viên hoặc nhóm sinh viên nếu có cần nén thư mục bài làm thành định dạng zip và nộp qua Moodle.
- Tên thư mục bài làm:
    1. Nếu chỉ có một sinh viên: <MSSV>
    2. Nếu có 2-3 sinh viên: <MSSV1>_<MSSV2>_<MSSV3>
- Thư mục bài làm bao gồm các phần sau:
    1. Báo cáo trả lời các câu hỏi tự luận: Định dạng PDF, trang đầu tiên ghi thông tin nhóm, tỉ lệ thực hiện của mỗi thành viên và các câu hỏi chưa làm được.
    2. Mã nguồn chương trình cài đặt: Đặt trong thư mục "Source", bao gồm các file mã nguồn liên quan trong bài tập lập trình. Ngôn ngữ sử dụng là Python 3.

Tổng quan về Dự án

Lưu ý: Hiện tại, khóa học này chỉ có sẵn miễn phí, vì vậy bạn sẽ không thể gửi công việc của mình để được xem xét. Chúng tôi khuyến khích bạn sử dụng các thông số kỹ thuật và công cụ đánh giá để hoàn thành nó, sau đó tự đánh giá và tìm kiếm phản hồi từ gia đình, bạn bè và mạng xã hội của bạn. Sử dụng phản hồi từ họ để cải thiện và bạn sẽ có một ví dụ tuyệt vời về công việc của mình để thể hiện bất cứ lúc nào!

Trong dự án này, bạn sẽ phân tích một tập dữ liệu và sau đó truyền đạt những phát hiện của mình về nó. Bạn sẽ sử dụng các thư viện Python như NumPy, Pandas và Matplotlib để giúp việc phân tích của bạn dễ dàng hơn.

Tôi cần cài đặt những gì?

Bạn sẽ cần cài đặt Python, cùng với các thư viện sau:

pandas
numpy
matplotlib
csv hoặc unicodecsv

Chúng tôi khuyên bạn nên cài đặt Anaconda, đi kèm với tất cả các gói cần thiết cũng như sổ ghi chép IPython. Bạn có thể tìm thấy hướng dẫn cài đặt tại đây.

Tại sao Dự án này quan trọng?

Dự án này sẽ giới thiệu cho bạn quy trình phân tích dữ liệu. Trong dự án này, bạn sẽ trải qua toàn bộ quy trình để biết tất cả các phần khớp với nhau như thế nào. Các khóa học khác trong Nhà phân tích dữ liệu tập trung vào các phần riêng lẻ của quy trình phân tích dữ liệu. Trong dự án này, bạn cũng sẽ có cơ hội thực hành sử dụng các thư viện Python như NumPy, Pandas và Matplotlib, giúp việc viết mã phân tích dữ liệu bằng Python trở nên dễ dàng hơn rất nhiều!

Bạn sẽ học được gì?

Sau khi hoàn thành dự án này, bạn sẽ:Hiểu rõ tất cả các bước trong quy trình phân tích dữ liệu điển hình. Thoải mái đặt ra những câu hỏi có thể được trả lời bằng một tập dữ liệu nhất định và sau đó trả lời những câu hỏi đó. Biết cách điều tra các vấn đề trong tập dữ liệu và sắp xếp dữ liệu thành định dạng bạn có thể sử dụng. Có kinh nghiệm truyền đạt kết quả của phân tích của bạn. Có khả năng sử dụng các thao tác vector hóa trong NumPy và Pandas để tăng tốc mã phân tích dữ liệu của bạn. Quen thuộc với các đối tượng Series và DataFrame của Pandas, giúp bạn truy cập dữ liệu của mình một cách thuận tiện hơn

Biết cách sử dụng Matplotlib để tạo ra các biểu đồ thể hiện những phát hiện của bạn

Tại sao điều này quan trọng đối với sự nghiệp của bạn?

Dự án này sẽ thể hiện nhiều kỹ năng phân tích dữ liệu khác nhau, giúp bạn chứng minh cho các nhà tuyển dụng tiềm năng thấy rằng bạn biết cách thực hiện toàn bộ quy trình phân tích dữ liệu.

Giới thiệu

Cho dự án cuối cùng, bạn sẽ tiến hành phân tích dữ liệu của riêng mình và tạo một tập tin để chia sẻ những phát hiện của bạn. Bạn nên bắt đầu bằng cách xem xét tập dữ liệu của mình và ý thức ra những câu hỏi mà bạn có thể trả lời bằng nó. Sau đó, bạn nên sử dụng Pandas và NumPy để trả lời những câu hỏi bạn quan tâm nhất và tạo một báo cáo chia sẻ những câu trả lời đó. Bạn sẽ không cần phải sử dụng thống kê hoặc học máy để hoàn thành dự án này, nhưng bạn nên làm rõ trong giao tiếp của mình rằng những phát hiện của bạn chỉ là tạm thời. Dự án này là mở cửa trong việc không tìm kiếm một câu trả lời đúng duy nhất.

Bước một - Chọn Bộ dữ liệu của Bạn

Chọn một trong các bộ dữ liệu sau để phân tích cho dự án của bạn:

- Dữ liệu Titanic - Chứa thông tin về dân số và hành khách từ 891 trong số 2224 hành khách và phi hành đoàn trên tàu Titanic. Bạn có thể xem mô tả của bộ dữ liệu này trên trang web Kaggle, nơi dữ liệu được lấy.

Chọn phiên bản phân cách bằng dấu phẩy, chứa các tập tin CSV.

Bước hai - Tổ chức

Cuối cùng, bạn sẽ muốn chia sẻ dự án của mình với bạn bè, gia đình và nhà tuyển dụng. Hãy tổ chức trước khi bắt đầu. Chúng tôi khuyến nghị tạo một thư mục duy nhất mà sau này sẽ chứa:

- Báo cáo chia sẻ những phát hiện của bạn
- Mọi mã Python bạn viết như một phần của phân tích của bạn
- Tập dữ liệu bạn đã sử dụng (mà bạn sẽ không cần phải gửi)

Bạn có thể muốn sử dụng sổ ghi chép IPython, trong trường hợp đó, bạn có thể chia sẻ cả mã bạn đã viết và báo cáo về những phát hiện của bạn trong cùng một tài liệu. Nếu không, bạn sẽ cần lưu trữ báo cáo và mã riêng biệt.

Bước ba - Phân tích Dữ liệu của Bạn

Xem xét một số câu hỏi mà bạn có thể trả lời bằng tập dữ liệu bạn đã chọn, sau đó bắt đầu trả lời những câu hỏi đó. Dưới đây là một số ý tưởng để bạn bắt đầu:

- Dữ liệu Titanic
    - Những yếu tố nào làm cho người ta có khả năng sống sót cao hơn?

Hãy chắc chắn bạn sử dụng NumPy và Pandas khi cần thiết!

Bước bốn - Chia sẻ Những Phát Hiện của Bạn

Khi bạn đã hoàn thành phân tích dữ liệu, tạo một báo cáo chia sẻ những phát hiện mà bạn thấy thú vị nhất. Bạn có thể muốn sử dụng sổ ghi chép IPython để chia sẻ những phát hiện của mình cùng với mã bạn đã sử dụng để thực hiện phân tích, nhưng bạn cũng có thể sử dụng một công cụ khác nếu bạn muốn.