TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN CUỐI KÌ

NHẬP MÔN KHOA HỌC DỮ LIỆU

Người thực hiện: **Nguyễn Văn Nam – 21017711**

Nguyễn Tấn Đạt – 21022831

Lớp **: DHKHDL17A - 420301410401**

Khoá **: 17**

Người hướng dẫn: **TS BÙI THANH HÙNG**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023

TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN CUỐI KÌ

NHẬP MÔN KHOA HỌC DỮ LIỆU

Người thực hiện: **Nguyễn Văn Nam - 21017711**
Nguyễn Tấn Đạt - 21022831
Lớp **:DHKH DL17A - 420301410401**
Khoá **: 17**
Người hướng dẫn: **TS. BÙI THANH HÙNG**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023

LỜI CẢM ƠN

Chúng tôi xin gửi lời cảm ơn đến TS.Bùi Thanh Hùng thầy là người hướng dẫn định hướng, và chỉ điểm chúng tôi để chúng tôi có thể hoàn thành đồ án này, chúng tôi chân thành cảm ơn thầy đã giúp đỡ và chia sẻ những kiến thức bổ ích để chúng tôi có thể thuận lợi hoàn thành đồ án kết thúc môn học này.

Một lần nữa chúng tôi chân thành cảm ơn thầy TS.Bùi Thanh Hùng rất nhiều.

Trân trọng,

ĐỒ ÁN ĐƯỢC HOÀN THÀNH
TẠI TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH

Tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được sự hướng dẫn của TS. Bùi Thanh Hùng. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình. Trường đại học Công nghiệp TP Hồ Chí Minh không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày tháng năm

Tác giả

(ký tên và ghi rõ họ tên)

Nguyễn Văn Nam

Nguyễn Tấn Đạt

PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN

_____
_____
_____
_____
_____
_____
_____

Tp. Hồ Chí Minh, ngày tháng năm
(kí và ghi họ tên)

TÓM TẮT

Trình bày tóm tắt vấn đề nghiên cứu, các hướng tiếp cận, cách giải quyết vấn đề và một số kết quả đạt được, những phát hiện cơ bản trong vòng 1 -2 trang của từng Bài toán, mỗi bài toán là 1 đoạn văn bản khác nhau.

Câu 1 :

Để thu thập 60 dòng dữ liệu về 5 chủ đề (mỗi chủ đề 12 ý kiến, 4 ý kiến mỗi lớp), ta có thể thực hiện như sau:

Tạo một form google trình duyệt web và tìm kiếm các trang web liên quan đến đánh giá của sinh viên về các chủ đề trong trường đại học.

Đọc và phân tích các bài đánh giá trên các trang web đó để tìm ra các ý kiến về cơ sở vật chất, đội ngũ giảng viên, chương trình đào tạo, công tác sinh viên và tư vấn tuyển sinh.

Lựa chọn 12 ý kiến cho mỗi chủ đề, bao gồm 4 ý kiến tích cực, 4 ý kiến trung tính và 4 ý kiến tiêu cực.

Lưu ý rằng, các ý kiến này phải đảm bảo tính khách quan và không mang tính chất cá nhân. Tạo một bảng dữ liệu với cấu trúc như sau:

Mã chủ đề	**Đánh giá**	**Ý kiến**
CSVC	1	...
CSVC	1	...
CSVC	1	...
CSVC	0	...
CSVC	0	...
CSVC	0	...

CSVC	-1	...
CSVC	-1	...
CSVC	-1	...
...	...	...

Lưu bảng dữ liệu vào một tập tin CSV để có thể sử dụng lại sau này.

Câu 2 :
Đầu tiên tìm kiếm câu hỏi liên quan đến quá trình khảo sát , tiếp theo in phiếu khảo sát và tiến hành khảo sát, sau khi khảo sát xong chúng tôi tiến hành mã hoá dữ liệu về dạng (-1,0,1) đại diện cho 3 mức độ tích cực, trung tính, tiêu cực sau khi mã hoá chúng tôi có file dữ liêu 50 dòng được lấy ý kiến từ các sinh viên IUH sau đó tiến hành phân tích dữ liệu bằng các công cụ điển hình được sử dụng trong bài này là ngôn ngữ lập trình python với các thư viện numpy, pandas ,…

Câu 3:
Đầu tiên sử dụng thư viện có sẵn trong python để thu thập dữ liệu từ trang web , sau đó lưu thành file csv để có dữ liệu phân tích, tiến hành làm sạch dữ liệu (các dữ liệu có giá Nan sẽ được loại bỏ) sau đó chúng tôi tiến hành phân tích dữ liệu bằng python với các thư viện có sẵn như pandas, numpy,…. kèm với một số thư viện vẽ đồ thị như matplotlib,seaborn,…. cùng với đó để phân tích bằng học máy chúng tôi sử dụng thư viện sklearn,.. để giải quyết vấn đề mà bài toán đưa rá

MỤC LỤC

CHƯƠNG 1.....3
THU THẬP DỮ LIỆU.....3
1.1 Yêu cầu thu thập dữ liệu.....3
1.2 Cách thức tiến hành thu thập dữ liệu.....3
1.3 Kết quả.....3
CHƯƠNG 2.....5
ĐẶT TÊN NGẮN GỌN LIÊN QUAN ĐẾN CÂU HỎI ĐẶT RA.....5
(Ví dụ Phân tích các ý kiến phản hồi về Chương trình Đào tạo ngành Khoa học dữ liệu ở Đại học Công nghiệp TP Hồ Chí Minh).....5
2.1 Câu hỏi đặt ra.....5
2.2 Xây dựng câu hỏi khảo sát.....7
2.3 Phân tích.....10
2.3.1 Kết quả khảo sát.....10
2.3.2 Phân tích dữ liệu.....13
2.3.2.1 Mô tả về dữ liệu.....14
2.3.2.2 Thực hiện các thống kê căn bản.....15
2.3.2.3 Tìm mối tương quan giữa các câu hỏi khảo sát và kết quả.....15
2.3.2.4 Trực quan hóa dữ liệu và kết quả.....18
1. Phân tích cơ sở vật chất.....18
2. Phân tích đội ngũ giảng viên.....22
2.4 Kết luận.....22
CHƯƠNG 3.....23
KHAI PHÁ DỮ LIỆU TỪ CÂU NÓI CỦA NGƯỜI NỔI TIẾNG.....23
3.1 Thu thập dữ liệu.....23
3.2 Khai phá dữ liệu.....24
3.2.1. Xử lý dữ liệu- Data Imputation.....24
3.2.1.1 Bạn hãy thêm vào Trường STT và điền tự động dữ liệu của trường này?.....24

3.2.1.2 Một số giá trị của dữ liệu Trường ngày sinh chưa có, bạn hãy đề xuất cách điền?.....24
3.2.1.3 Bạn hãy thêm vào Trường Tuoi (Tuổi) và đề xuất cách điền tuổi của các tác giả?..... 25
3.2.2. Khám phá dữ liệu- Data Exploration..... 26
1. Top 10 người nổi tiếng có số lượng câu nói nhiều nhất và ít nhất....26
2. Đặt ra câu hỏi vậy người có nhiều câu nói nhất có phải là người có từ trong câu nhiều nhất và ít nhất.....27
3. Thống kê về các từ được sử dụng trong các câu nói?.....29
4. Vậy những người nổi tiếng có tuổi tập trung ở những khoảng nào? 29
5. Mối quan hệ giữa số lượng câu nói với tuổi của tác giả ?..... 30
3.2.3. Trích xuất đặc trưng- Feature Extraction..... 30
3.2.4. Suy luận..... 31

CHƯƠNG 1

THU THẬP DỮ LIỆU

1.1 Yêu cầu thu thập dữ liệu

Yêu cầu thu thập dữ liệu là thu thập ý kiến của sinh viên đại học IUH về 5 chủ đề khác nhau. Mỗi chủ đề gồm 12 ý kiến, bao gồm 4 ý kiến tích cực (positive), 4 ý kiến tiêu cực (negative) và 4 ý kiến trung tính (neutral).

1.2 Cách thức tiến hành thu thập dữ liệu

Để thu thập dữ liệu, chúng ta có thể sử dụng các phương pháp như đăng ký trực tuyến và thu thập dữ liệu thông qua mẫu câu hỏi hoặc phỏng vấn trực tiếp các sinh viên. Tuy nhiên, trong trường hợp này, chúng ta sẽ tạo ra một bản khảo sát online bằng google form và gửi nó đến những sinh viên IUH để khảo sát.

1.3 Kết quả

Kết quả đạt được:

Sau khi thu thập xong chúng ta sẽ có bộ dữ liệu có 60 dòng dữ liệu gồm 5 chủ đề mỗi chủ đề CSVC, GV, CTDT, CTSV, TVTS, mỗi chủ đề có 12 dòng dữ liệu bao gồm 4 dòng dữ liệu đánh giá tích cực (được biểu diễn là 1), 4 dòng dữ liệu đánh giá trung tính (được biểu diễn là 0) và 4 dòng dữ liệu đánh giá tiêu cực (được biểu diễn là -1)

Cách đánh giá tính chính xác của bộ dữ liệu:

Để đánh giá bằng cách đọc từng dòng dữ liệu, trước hết ta cần xác định các tiêu chí đánh giá để có thể đánh giá độ chính xác và tính đúng đắn của dữ liệu. Các tiêu chí đánh giá có thể bao gồm:

1. Sự phù hợp của đánh giá với nội dung ý kiến: điểm đánh giá phải phản ánh đúng ý nghĩa và tình huống được mô tả trong ý kiến.
2. Sự thống nhất của đánh giá trong các ý kiến liên quan: các ý kiến liên quan đến cùng một chủ đề và có nội dung tương tự nên có điểm đánh giá tương đồng với nhau.
3. Độ chính xác của đánh giá: điểm đánh giá phải phù hợp với thực tế và đánh giá phản ánh đúng quan điểm của người dùng.

Sau khi xác định các tiêu chí đánh giá, ta có thể thực hiện đánh giá từng dòng dữ liệu bằng cách đọc và hiểu nội dung của từng ý kiến, sau đó đưa ra một điểm đánh giá theo ba mức độ -1,0,1. Với mỗi ý kiến, ta cần đánh giá x chúng tôi điểm đánh giá có phù hợp với nội dung ý kiến hay không, và có thể so sánh với các ý kiến khác để đánh giá tính thống nhất của dữ liệu.

Tuy nhiên, phương pháp này có thể gặp khó khăn khi đánh giá các ý kiến phức tạp hoặc khó hiểu, đặc biệt là trong trường hợp có sự không đồng nhất giữa các ý kiến liên quan đến cùng một chủ đề.

CHƯƠNG 2

ĐẶT TÊN NGẮN GỌN LIÊN QUAN ĐẾN CÂU HỎI ĐẶT RA

(Ví dụ Phân tích các ý kiến phản hồi về Chương trình Đào tạo ngành Khoa học dữ liệu ở Đại học Công nghiệp TP Hồ Chí Minh)

2.1 Câu hỏi đặt ra

Trình bày ít nhất 2 câu hỏi có liên quan đến 5 chủ đề dưới đây, lý giải tại sao lại đặt ra câu hỏi này, ý nghĩa của câu hỏi đặt ra.

STT	Mã chủ đề	Diễn giải
1	CSVC	Cơ sở vật chất
2	GV	Đội ngũ giảng viên
3	CTDT	Chương trình đào tạo
4	CTSV	Công tác sinh viên
5	TVTS	Tư vấn tuyển sinh

- Hai câu hỏi:

+ Cơ sở vật chất của trường đáp ứng các tiêu chuẩn để đảm bảo chất lượng giảng dạy và học tập?

+ Đội ngũ giảng viên của trường có được đào tạo chuyên môn và nghiệp vụ đầy đủ để giảng dạy cho sinh viên?

⇒ Lý do để đặt 2 câu hỏi này là: Để đánh giá một cách khách quan và nhanh nhất về chất lượng đào tạo của trường có tốt hay không thì các câu hỏi đó liên

quan trực tiếp đến chất lượng đào tạo của trường, bao gồm các yếu tố quan trọng như cơ sở vật chất và đội ngũ giảng viên. Cơ sở vật chất và đội ngũ giảng viên đều là những yếu tố quan trọng đóng vai trò quyết định đến chất lượng đào tạo của trường. Vì vậy, việc đặt ra các câu hỏi liên quan đến cơ sở vật chất và đội ngũ giảng viên sẽ giúp đánh giá được mức độ đáp ứng của trường đối với những yêu cầu về chất lượng đào tạo.

- Ý nghĩa

+ Cơ sở vật chất của trường đáp ứng các tiêu chuẩn để đảm bảo chất lượng giảng dạy và học tập?

Ý nghĩa của câu hỏi này là để xác định x chúng tôi cơ sở vật chất của trường có đáp ứng được các tiêu chuẩn đảm bảo chất lượng giảng dạy và học tập hay không. Nếu không đạt được tiêu chuẩn, việc cải thiện cơ sở vật chất sẽ giúp nâng cao chất lượng giảng dạy và học tập của sinh viên.

+ Đội ngũ giảng viên của trường có được đào tạo chuyên môn và nghiệp vụ đầy đủ để giảng dạy cho sinh viên?

Ý nghĩa của câu hỏi này là để đánh giá x chúng tôi đội ngũ giảng viên của trường có đủ năng lực và kiến thức chuyên môn để giảng dạy cho sinh viên hay không. Nếu không đạt được yêu cầu, việc đào tạo và cập nhật kiến thức cho giảng viên sẽ giúp nâng cao chất lượng đào tạo của trường. Từ đó, ta có thể suy ra được chất lượng đào tạo của trường đang ở mức độ nào và những điểm cần cải thiện để nâng cao chất lượng đào tạo của trường.

--> Từ đó suy ra chất lượng đào tạo của trường đang như thế nào ?

2.2 Xây dựng câu hỏi khảo sát

Trình bày cách thức xây dựng ít nhất 10 câu hỏi câu hỏi khảo sát để trả lời cho 2 câu hỏi đặt ra ở 2.1

- Bước 1: Xác định mục tiêu của câu hỏi khảo sát. Bạn muốn tìm hiểu về những gì trong chủ đề đó? Muốn trả lời những câu hỏi gì?
    + Ở đây chúng tôi khảo sát về "chất lượng đào tạo của trường Đại Học Công Nghiệp Thành Phố Hồ Chí Minh", vì vậy chúng tôi đã chọn những câu hỏi khảo sát về cơ sở vật và giảng viên là 2 yếu tố then chốt để có thể đánh giá một cách khách quan nhất về vấn đề đã đặt ra
- Bước 2: Liệt kê ra các câu hỏi cơ bản về chủ đề đó, từ đó phân tích để xác định các câu hỏi chi tiết hơn.
    + Để có thể lập ra bảng câu hỏi chi tiết chúng tôi tiến hành xác định 2 câu hỏi lớn, để có thể từ đó đưa ra những câu hỏi khảo sát nhỏ, để trả lời 2 câu hỏi lớn và đưa ra kết quả một cách chính xác nhất.
- Bước 3: Lựa chọn các dạng câu hỏi phù hợp để trả lời cho các mục tiêu của bạn. Có thể sử dụng câu hỏi đa lựa chọn, câu hỏi mở, câu hỏi đánh giá, câu hỏi dành cho nhóm đối tượng cụ thể, vv.
    + Để có thể đánh giá về cơ sở vật chất chúng tôi tiến hành đặt những câu hỏi khảo sát liên quan đến phòng thực hành, phòng học, các dịch vụ có sẵn trong trường, ...
    + Để đánh giá về giảng viên chúng tôi tiến hành đặt những câu hỏi khảo sát liên quan đến đánh giá của sinh viên về thái độ giảng dạy của giảng viên, sự công bằng của giảng viên,...
- Bước 4: Chuẩn bị các câu trả lời khả thi cho các câu hỏi.
    + Ở bước này chúng tôi đánh giá câu trả lời theo 3 mức độ là : tích cực, trung tính và tiêu cực.

- Liệt kê ít nhất 10 câu hỏi câu hỏi khảo sát và giải thích chi tiết các câu hỏi khảo sát này:
- Ở đây khi khảo sát chúng tôi chỉ cần tên sinh viên không yêu cầu điền mã sinh viên để không can thiệp sâu quá vào tính riêng tư của sinh viên.

Câu hỏi khảo sát 1: Cơ sở vật chất của trường có đáp ứng các tiêu chuẩn để đảm bảo chất lượng giảng dạy và học tập?

1. Các phòng học, phòng thực hành và các thiết bị trong trường có đáp ứng đầy đủ nhu cầu học tập của bạn không?
   ⇒ Đây là câu hỏi đánh giá x chúng tôi cơ sở vật chất của trường có đủ và đáp ứng được nhu cầu học tập của sinh viên hay không.
2. Bạn có được hỗ trợ tốt từ trường trong việc sử dụng các thiết bị, phòng học hay không?
   ⇒ Câu hỏi này nhằm đánh giá mức độ hỗ trợ từ trường đối với sinh viên trong việc sử dụng các thiết bị, phòng học, giúp cho sinh viên có thể sử dụng đầy đủ tiện nghi mà trường đang cung cấp.
3. Bạn có hài lòng với các dịch vụ hỗ trợ của trường như thư viện, phòng học, khu vực sinh hoạt chung, khu vực ăn uống, vv. không?
   ⇒ Câu hỏi này đánh giá mức độ hài lòng của sinh viên với các dịch vụ hỗ trợ mà trường cung cấp như thư viện, khu vực sinh hoạt, khu vực ăn uống, vv.
4. Bạn có cảm thấy an toàn, thoải mái khi sử dụng cơ sở vật chất của trường không?
   ⇒ Câu hỏi này nhằm đánh giá mức độ an toàn, thoải mái của sinh viên khi sử dụng các tiện nghi của trường.
5. Bạn cảm thấy trường đáp ứng đủ về cơ sở vật chất hay không?
   ⇒ Câu hỏi này tổng hợp lại những đánh giá của sinh viên về cơ sở vật chất của trường để xác định x chúng tôi trường có đáp ứng đủ về cơ sở vật chất hay không.

Câu hỏi khảo sát 2: Đội ngũ giảng viên của trường có được đào tạo chuyên môn và nghiệp vụ đầy đủ để giảng dạy cho sinh viên?

6. Bạn có cảm thấy giảng viên có đủ kiến thức chuyên môn để truyền đạt cho bạn không?
   ⇒ Câu hỏi này giúp đánh giá năng lực chuyên môn của giảng viên, x chúng tôi liệu họ có đủ kiến thức và kinh nghiệm để giảng dạy một cách hiệu quả hay không.
7. Bạn có được hỗ trợ đầy đủ từ giảng viên khi gặp khó khăn trong quá trình học tập không?
   ⇒ Câu hỏi này giúp đánh giá phương pháp giảng dạy của giảng viên, x chúng tôi liệu họ có cân nhắc đến nhu cầu và khả năng của sinh viên hay không, giúp sinh viên học tập hiệu quả hơn.
8. Giảng viên có sử dụng các phương pháp giảng dạy phù hợp với nhu cầu và khả năng của sinh viên không?
   ⇒ Câu hỏi này đánh giá tác phong của giảng viên, x chúng tôi liệu họ có tôn trọng và đối xử công bằng với sinh viên hay không, giúp sinh viên cảm thấy tự tin và thoải mái trong quá trình học tập.
9. Bạn có cảm thấy giảng viên tôn trọng và đối xử công bằng với sinh viên không?
   ⇒ Câu hỏi này đánh giá tác phong của giảng viên, x chúng tôi liệu họ có tôn trọng và đối xử công bằng với sinh viên hay không, giúp sinh viên cảm thấy tự tin và thoải mái trong quá trình học tập.
10. Bạn có cảm thấy hài lòng với đội ngũ giảng viên của trường không?
    ⇒ Câu hỏi này giúp đánh giá tổng thể về chất lượng giảng viên của trường, x chúng tôi liệu sinh viên có hài lòng với đội ngũ giảng viên của trường hay không. Điều này giúp quản lý trường cải thiện chất lượng giảng dạy và đảm bảo sự hài lòng của sinh viên.

2.3 Phân tích

2.3.1 Kết quả khảo sát

Trình bày kết quả khi tiến hành khảo sát ít nhất 10 câu hỏi khảo sát trên với ít nhất 50 bạn trong trường về các nội dung như: Cách thức tiến hành khảo sát, thời gian, số lượng, đối tượng, kết quả thu được, …

Thời gian	Công việc
05/04/2023	- Tiến hành làm câu hỏi xác - Câu hỏi được xây dựng theo dạng: câu hỏi khảo sát với các mức độ khác nhau

06/04/2023 - 07/04/2023	- Tiến hành khảo xác
	- Đối tượng:
	  - Sinh viên K15, K16 và có cả K18
	- Địa điểm khảo sát:
	  Lab h2.2
	  Lab h3.2
	  Tiết học Đại số tuyến tính của sinh viên DS K17
	- Số lượng sinh viên: 50 người

Kết quả đạt:

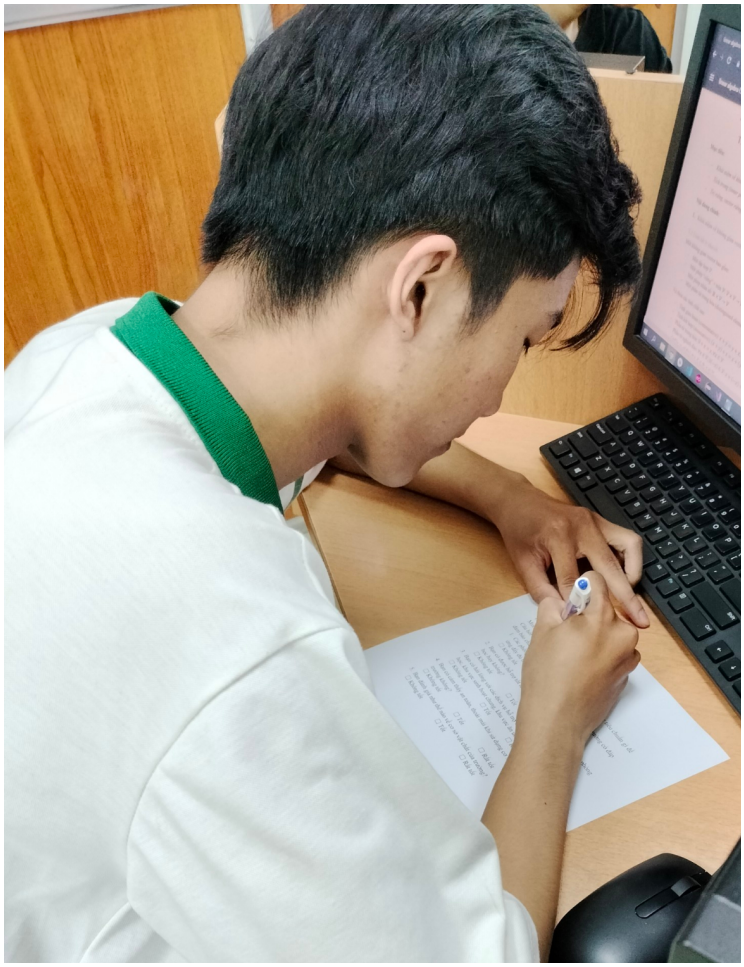
- Thu thập được 50 phiếu khảo sát trong đó:
  - Lab H2.2 chiếm khoảng 30 % phiếu khảo sát.

Hình 2.1 Hình ảnh minh chứng ở lab H2.2

- Lab h3.2 chiếm khoảng 20 % phiếu khảo sát.

Hình 2.2 Hình ảnh minh chứng ở lab H3.2

- Lớp DS K17 chiếm khoảng 50 % phiếu khảo sát.



Hình 2.3 Hình ảnh minh chứng ở lab H3.2

2.3.2 Phân tích dữ liệu

Trước khi phân tích dữ liệu chúng tôi đã tiến hành mã hóa các câu trả lời thành các số:

Mã hóa chuỗi	
Rất đồng tình	1
Đồng tình	0
Không đồng tình	-1
Có	1
Không	0

Trước khi tiến hành phân tích chúng tôi có đặt ra cái vấn đề mình phần phân tích là gì ?

Vấn đề chính chúng ta cần phải trả lời đó là:

1. Cơ sở vật chất của trường đáp ứng các tiêu chuẩn để đảm bảo chất lượng giảng dạy và học tập?
2. Đội ngũ giảng viên của trường có được đào tạo chuyên môn và nghiệp vụ đầy đủ để giảng dạy cho sinh viên?

Từ đó suy ra chất lượng đào tạo của trường đang như thế nào?

2.3.2.1 Mô tả về dữ liệu

<table>
<tr><th colspan="2">Bảng mô tả về dữ liệu</th></tr>
<tr><td>Dữ liệu gồm các cột nào?</td><td>Gồm các cột: 'Tên', '1', '2', '3', '4', '5', '6', '7', '8', '9', '10' các số tương ứng với vị trí câu hỏi trên phiếu khảo sát</td></tr>
<tr><td>Dữ liệu có kích thước như nào?</td><td>50 dòng và 11 cột</td></tr>
<tr><td>Có kiểu dữ liệu như nào?</td><td>int64 (do đã được mã hóa thành số)</td></tr>
<tr><td>Dữ liệu có trùng lặp không?</td><td>Không có dòng nào trùng nhau</td></tr>
<tr><td>Vậy dữ liệu ở các trường có giá null không?</td><td>Không có cột nào có giá trị null</td></tr>
<tr><td>Các trường có các loại dữ liệu nào và số lượng?</td><td>
<table>
<tr><th>Type</th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th><th>8</th><th>9</th><th>10</th></tr>
<tr><td>1</td><td>2</td><td>8</td><td>17</td><td>9</td><td>9</td><td>17</td><td>14</td><td>15</td><td>15</td><td>15</td></tr>
<tr><td>0</td><td>39</td><td>31</td><td>26</td><td>24</td><td>24</td><td>32</td><td>27</td><td>28</td><td>24</td><td>24</td></tr>
<tr><td>-1</td><td>9</td><td>11</td><td>7</td><td>17</td><td>0</td><td>1</td><td>9</td><td>7</td><td>11</td><td>0</td></tr>
</table>

Hình 2.4 Bảng mô tả các thể loại dữ liệu của các trường
Nhìn vào bảng này chúng ta có thể thấy:
+ Các dữ liệu trừ 5, và 10 thì tất cả đều có 3 loại là -1, 0, 1 tương ứng với “Không đồng tình”, “Đồng tình”, “Rất đồng tình”. Còn cột 5 và 10 có 2 loại đó là 0, 1 tương ứng với là No và Yes
</td></tr>
</table>

2.3.2.2 Thực hiện các thống kê căn bản

Tiến hành hiển thị mô tả tổng quan về dữ liệu như count, standard deviation, variance, ..

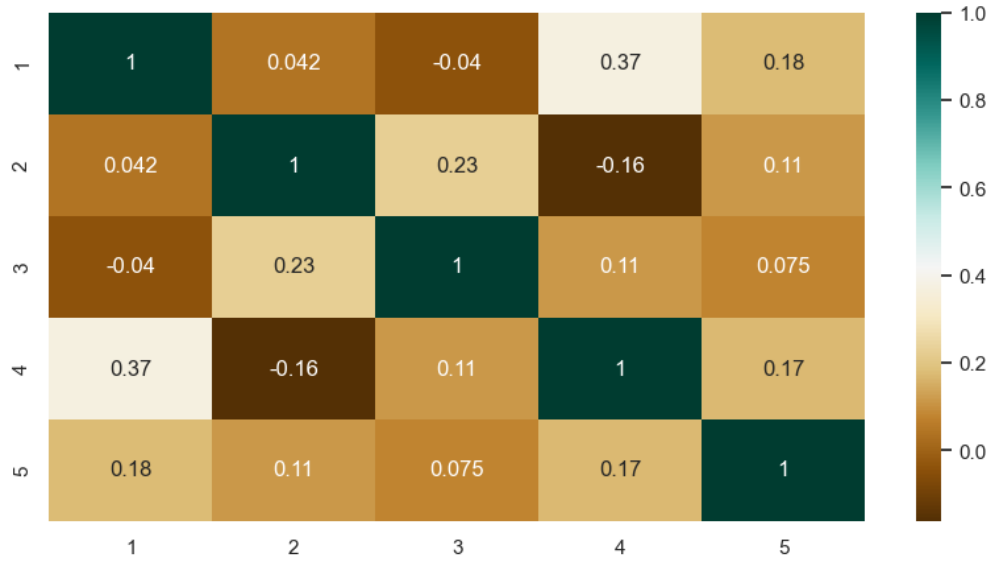
	1	2	3	4	5	6	7	8	9	10
count	50.000000	50.000000	50.000000	50.000000	50.000000	50.000000	50.000000	50.000000	50.000000	50.000000
mean	-0.140000	-0.060000	0.200000	-0.160000	0.800000	0.320000	0.100000	0.160000	0.080000	0.920000
std	0.452205	0.619743	0.670059	0.710274	0.404061	0.512696	0.677631	0.650275	0.723935	0.274048
min	-1.000000	-1.000000	-1.000000	-1.000000	0.000000	-1.000000	-1.000000	-1.000000	-1.000000	0.000000
25%	0.000000	0.000000	0.000000	-1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	1.000000
50%	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	1.000000
75%	0.000000	0.000000	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Hình 2.5 Thực hiện các thống kê căn bản

2.3.2.3 Tìm mối tương quan giữa các câu hỏi khảo sát và kết quả

Ở bước này chúng tôi tiến hành mô tả mức độ tương quan theo 3 nhóm chính

1. Phân tích mức độ tương quan của các câu trả lời cho câu hỏi "Cơ sở vật chất của trường đáp ứng các tiêu chuẩn để đảm bảo chất lượng giảng dạy và học tập? "



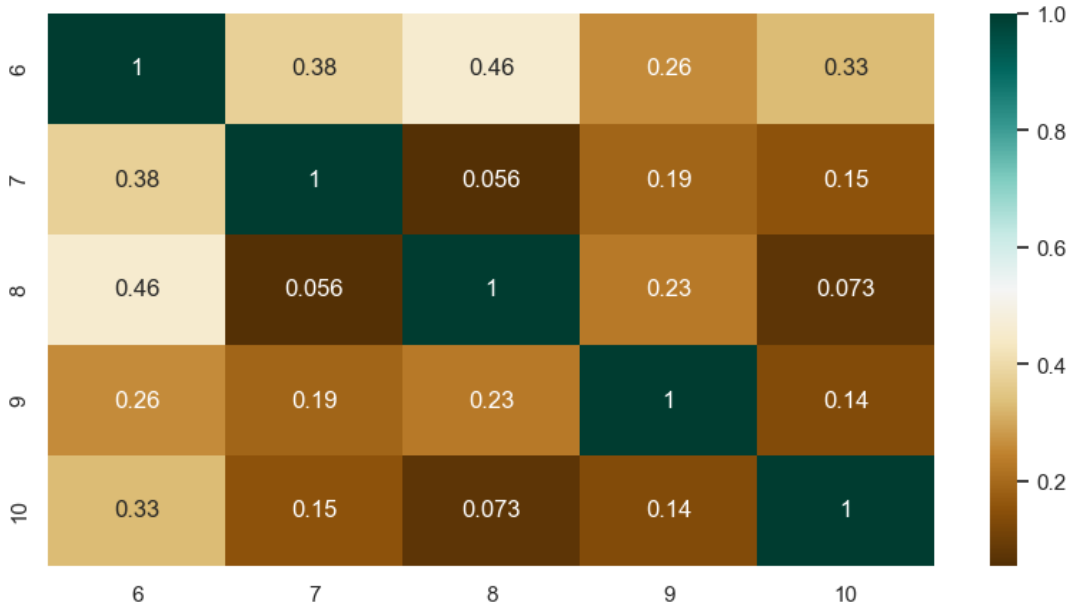
Hình 2.5 Mô tả tương quan của 5 câu hỏi đầu

Nhìn vào bảng này chúng tôi xác định lại một điều trong câu 1 có 5 câu và 4 câu đầu là ý để trả lời cho câu 5 chính vì vậy câu 5 chính là biến mục tiêu.

Nhận xét:

+ Mức độ tương quan của ở các trường khá thấp thấp nhất là 0,075 của cột (5, 3) và cao nhất là (5, 4) và (5, 1).
+ Mức độ tương quan khả quan để phân tích là (4, 5), (2, 5), (1, 5).

2. Phân tích mức độ tương quan của các câu trả lời cho câu hỏi "Đội ngũ giảng viên của trường có được đào tạo chuyên môn và nghiệp vụ đầy đủ để giảng dạy cho sinh viên? "

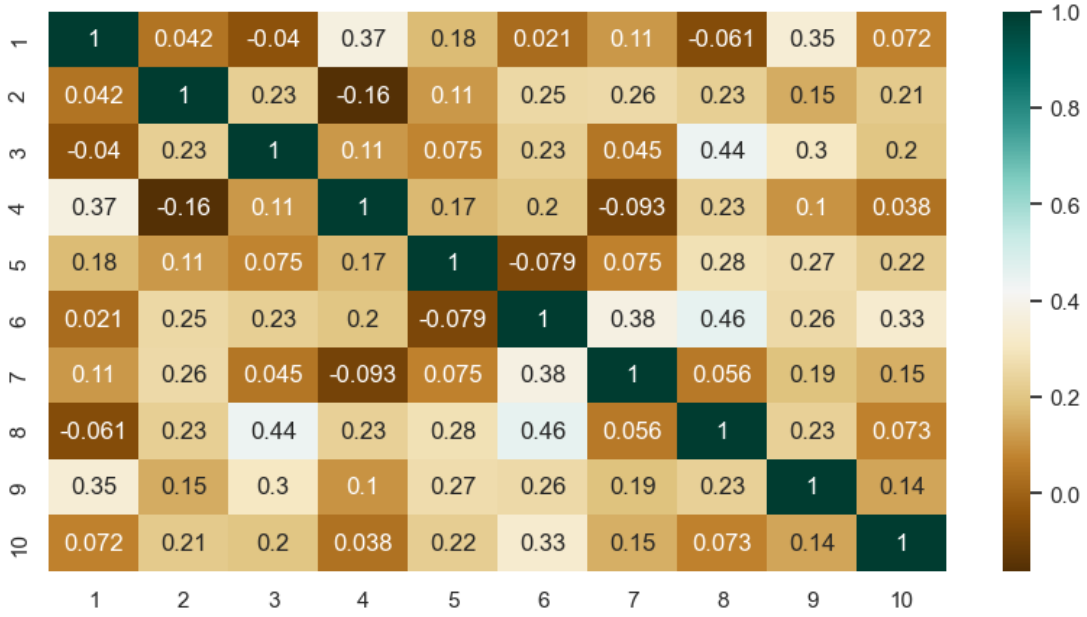


Hình 2.6 Mô tả tương quan của 5 câu hỏi cuối

Nhận xét cau 2:

1. Mức độ tương quan của ở các trường khá cao, cao nhất là (6, 10)
2. Mức độ tương quan khả quan để phân tích là (6, 10), (7, 10), (8, 10), (9, 10)

3. Phân tích mức độ tương quan một cách tổng thể của 10 câu hỏi



Hình 2.7 Mô tả tương quan

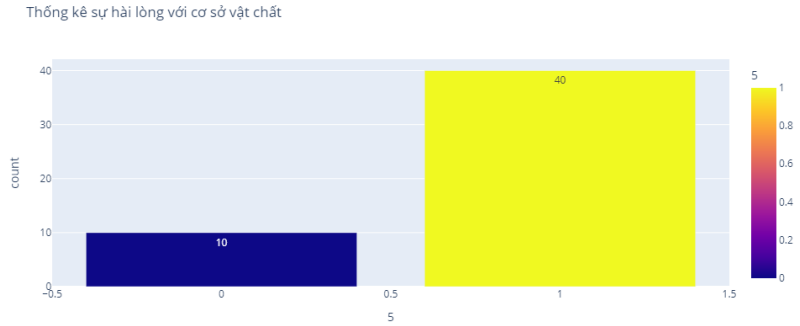
Khi nhìn vào tổng thể chúng ta thấy được rằng mối tương quan giữa câu 10 và câu 5 khá cao khoảng 22% điều này thể hiện khi câu 10 được đánh giá cao thì câu 5 cũng sẽ có xu hướng được đánh giá cao.

- Xác định các yếu tố quan trọng ảnh hưởng đến kết quả

  Sau khi x chúng tôi qua mức độ tương quan thì chúng tôi sẽ chọn các câu sau là yếu tố quan trọng ảnh hưởng đến kết quả tức là câu 5, và câu 10 đó là các cột: '1', '2', '4', '5', '6','7', '8', '9', '10'

2.3.2.4 Trực quan hóa dữ liệu và kết quả

1. Phân tích cơ sở vật chất



Hình 2.8 Thống kê sự hài lòng với cơ sở vật chất

Nhìn vào biểu đồ này chúng ta cũng đang thấy được rằng là cơ sở vậy chất đa số vẫn đang đáp ứng đủ nhu cầu của sinh viên?

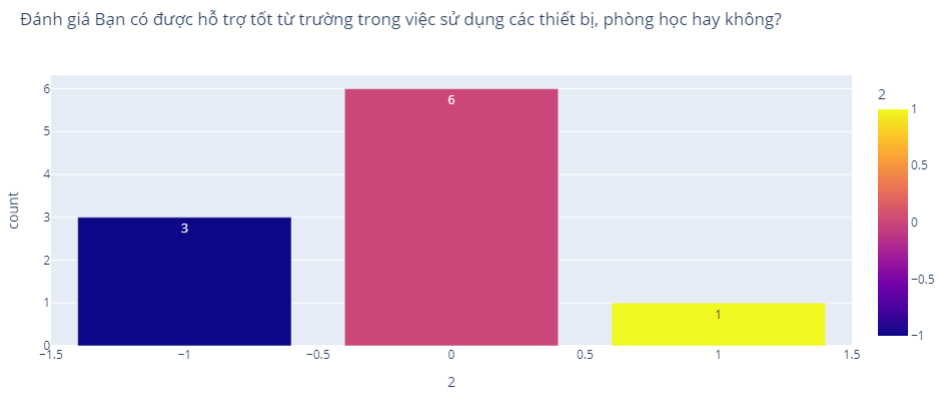
Nhưng vẫn có 20% sinh viên cảm thấy cơ sở vật chất vẫn chưa đáp ứng đủ?
Vậy tại sao lại sinh viên lại không đồng tình?

A. Về các phòng học, phòng thực hành và các thiết bị trong trường?

Hình 2.9 Đánh giá phòng học, phòng thực hành và các thiết bị trong trường

Ở đây chúng ta thấy được gần một nữa số sinh viên không đồng ý cảm thấy các phòng học, phòng thực hành và các thiết bị trong trường chưa đáp ứng đủ. Từ đó chúng ta thấy được các phòng học, phòng thực hành và các thiết bị chưa đáp ứng đủ nhu cầu học tập.

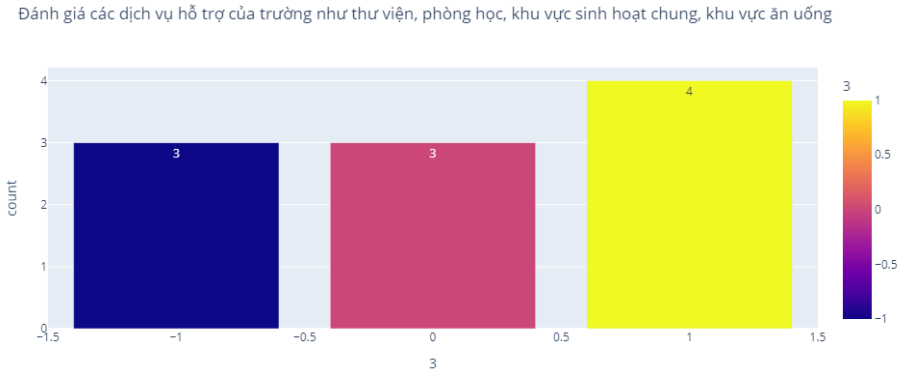
B. Bạn có được hỗ trợ tốt từ trường trong việc sử dụng các thiết bị, phòng học hay không?



Hình 2.10 Đánh giá sự hỗ trợ tốt từ trường trong việc sử dụng các thiết bị, phòng học hay không?

Ở đây chúng ta thấy được 70% số sinh viên không đồng ý cảm thấy hỗ trợ tốt từ trường trong việc sử dụng các thiết bị, phòng học khá là tốt nhưng vẫn chỉ có 1 sinh viên đánh giá rất tốt điều đó thể hiện là sự hỗ trợ vẫn chưa đáp ứng được nhu cầu của sinh viên với 30% sinh viên cảm thấy sự hỗ trợ tệ. Từ đó chúng ta thấy được hỗ trợ tốt từ trường trong việc sử dụng các thiết bị, phòng học cần được cải thiện.

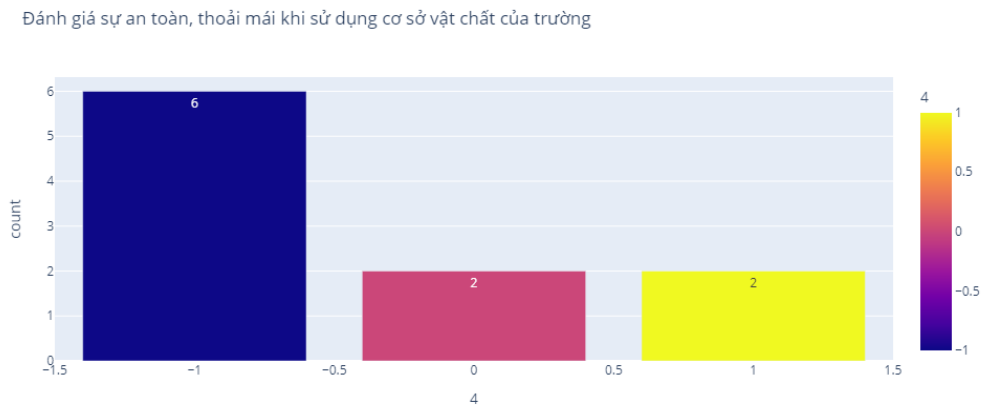
C. Bạn có hài lòng với các dịch vụ hỗ trợ của trường như thư viện, phòng học, khu vực sinh hoạt chung, khu vực ăn uống, vv. không?



Hình 2.11 Đánh giá các dịch vụ hỗ trợ của trường

Với 40% sinh viên đánh giá rất tốt và 30% sinh viên đánh giá tốt thì dịch vụ hỗ trợ của trường như thư viện, phòng học, khu vực sinh hoạt chung, khu vực ăn uống đang hỗ trợ sinh viên rất tốt.

D. Bạn có cảm thấy an toàn, thoải mái khi sử dụng cơ sở vật chất của trường không?

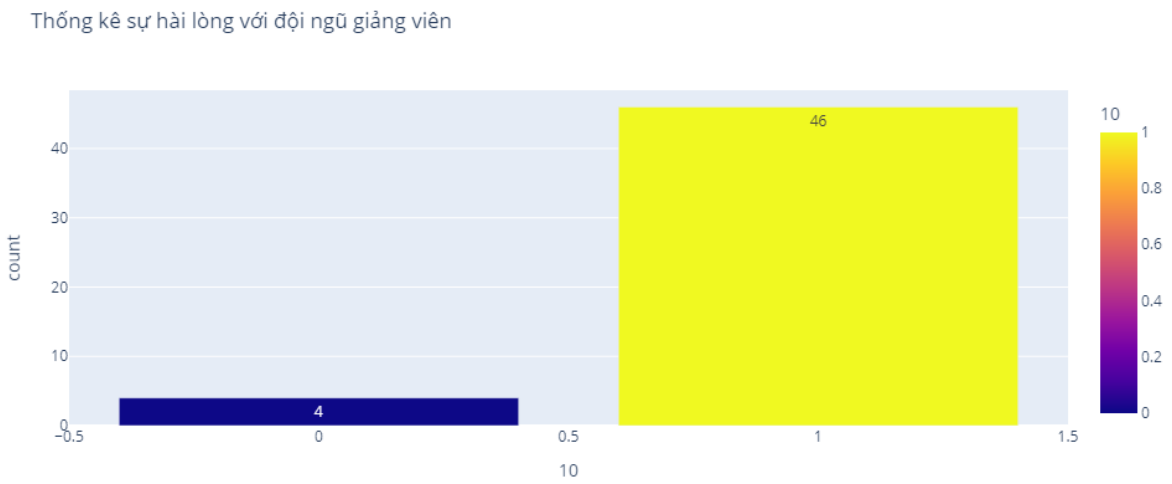


Hình 2.12 Đánh giá sự an toàn, thoải mái khi sử dụng cơ sở vật chất của trường

Chiếm 60% sinh viên đánh giá tệ, điều đó chứng tỏ sự an toàn khi sử dụng các cơ sở vật chất nhà trường đang thực sự tệ về độ an toàn và thoải mái

Tổng quan, sinh viên nhà trường cần phải cải thiện chất lượng các thiết bị hơn nữa và độ an toàn của các thiết bị đang là vấn đề chính cho sinh viên cảm thấy cơ sở vật chất không ổn.

2. Phân tích đội ngũ giảng viên



Hình 2.13 Thống kê sự hài lòng với đội ngủ giảng viên

Về đội ngũ giảng viên của trường chiếm 92% là tốt điều đó thể hiện đội ngũ giảng viên của trường đang rất tốt

2.4 Kết luận

Trình bày kết luận sau khi thực hiện bài tập này:

a) Qua phân tích cho thấy cơ sở vật chất của trường đang có một số vấn đề cần được cải thiện và mặt giảng viên thì đang rất tốt.

b) Sau khi phân tích đánh giá thí chúng tôi thấy rằng để trả lời cho câu hỏi chất lượng đào tạo của trường đang như thế nào? thì bắt buộc 2 câu hỏi

+ Cơ sở vật chất của trường đáp ứng các tiêu chuẩn để đảm bảo chất lượng giảng dạy và học tập? phải là Yes

+ Đội ngũ giảng viên của trường có được đào tạo chuyên môn và nghiệp vụ đầy đủ để giảng dạy cho sinh viên? phải là Yes

Từ đó ta với suy ra được chất lượng đào tạo của trường đang tốt còn ngược lại thì không tốt

chúng tôi có sử dụng model Decision tree để dự đoán chất lượng đào tạo của trường ở trong code.

Những điểm còn hạn chế và hướng phát triển trong tương lai?

+ Vấn đề chính của trường đó chính là cơ sở vật chất, cần phải được cải thiện trong tương lai.

CHƯƠNG 3

KHAI PHÁ DỮ LIỆU TỪ CÂU NÓI CỦA NGƯỜI NỔI TIẾNG

3.1 Thu thập dữ liệu

Trình bày cách thu thập dữ liệu và kết quả dữ liệu thu thập được

Đầu tiên để thu thập dữ liệu từ trang web **http://quotes.toscrape.com/** ta sử dụng thư viện requests để lấy dữ liệu HTML và BeautifulSoup (bs4) để phân tích cú pháp HTML và trích xuất thông tin.

Mô tả cấu trúc của trang web:

- Đầu tiên là header của trang, bao gồm tên trang và thanh tìm kiếm.
- Tiếp theo là phần chính của trang, bao gồm danh sách các câu nói của các tác giả nổi tiếng, được định dạng theo thẻ div với class là "quote".

- Mỗi câu nói bao gồm thông tin về tác giả (thẻ small với class là "author"), ngày tháng năm sinh của tác giả (thẻ span với class là "author-born-date") và nội dung câu nói (thẻ span với class là "text").
- Dưới phần danh sách câu nói là các nút phân trang (thẻ ul với class là "pager").

Sử dụng requests để lấy html của trang web, sau đó dùng BeautifulSoup để parse html thành đối tượng soup. Từ đây ta có thể sử dụng các class như find, find_all để tìm kiếm các thẻ có chứa thông tin cần trích xuất

3.2 Khai phá dữ liệu

3.2.1. Xử lý dữ liệu- Data Imputation

3.2.1.1 Bạn hãy thêm vào Trường STT và điền tự động dữ liệu của trường này?

Thêm trường STT chúng tôi sử dụng số range() để tạo ra các dãy số liên tiếp tự động

```
meta['STT'] = list(range(len(meta)))
```

3.2.1.2 Một số giá trị của dữ liệu Trường ngày sinh chưa có, bạn hãy đề xuất cách điền?

Khi nhìn vào vấn đề này chúng tôi có một số đánh giá các ngành sinh chắc chắn nhỏ hơn năm 2021. Sau đó kiểm tra lại bộ dữ liệu của mình có thực sự thiếu hay không, sau khi kiểm tra thì không có thiếu.

Nếu trường hợp thiếu trường dữ liệu Namsinh thì chúng tôi sẽ có một số để xuất sau:

+ Cách 1 chúng tôi thấy dữ liệu này là tên của những người nổi tiếng vì thế thông tin của nó là đa số sẽ được xuất bản trên mạng. Trường hợp nếu trên mạng không có thì chúng tôi sẽ lấy dải phân cách của ngày sinh đã có

+ Cách 2 sử dụng chat GPT api hiện tại để lấy thông tin (10.000 token/tháng)

-> Nếu gặp sự cố này, nhóm chúng tôi sẽ chọn cách thứ 1

3.2.1.3 Bạn hãy thêm vào Trường Tuoi (Tuổi) và đề xuất cách điền tuổi của các tác giả?

Khi nhìn vào vấn đề tuổi, thì chúng tôi có tra trên mạng thử các cái tên thì chợt nhận ra họ rất nổi tiếng vì thế thông tin tuổi của họ có sẵn trên mạng:

Và làm sao để bổ sung trường dữ liệu Age thì chúng tôi có 3 giải pháp sau:

+ Cách thứ 1 Đối với việc thay thế các dữ liệu thiếu thì có rất nhiều cách để điền điển hình như là mean, median nhưng với 2 tham số này thì rất dễ bị bias vì thế chúng ta có thêm các là lấy trimmed mean, trimmed median thì sẽ giảm bị bias.

+ Cách thứ 2 chúng tôi thấy dữ liệu này là tên các người nổi tiếng vì thế thông tin của nó đa số sẽ được publish ở trên mạng. Trường hợp nếu trên mạng không có thì chúng tôi sẽ lấy trimmed median của các độ tuổi đã có

+ Cách thứ 3 sử dụng selenium để lấy tuổi

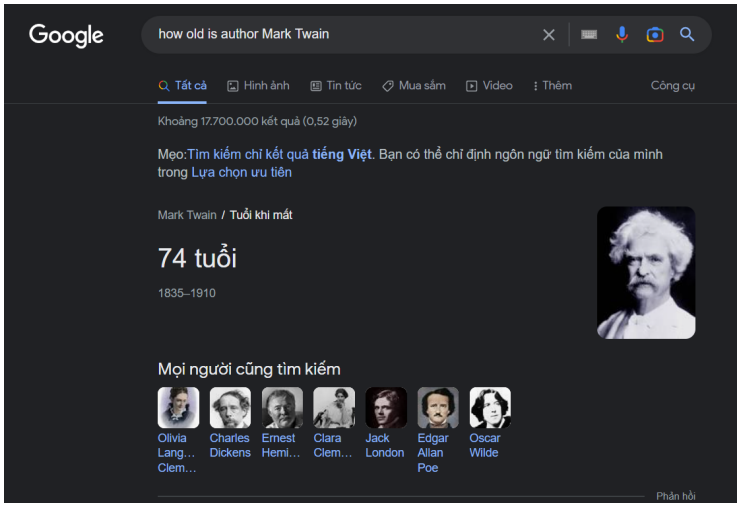
--> Nếu gặp tình huống này nhóm chúng tôi sẽ chọn cách thứ 3

Sử dụng selenium để lấy tuổi:

Cú pháp tìm kiếm: how old is author {Ten tác giả}

Phương pháp tìm kiếm: chúng tôi sử dụng selenium để lấy tuổi của tác giả

Với cú pháp như trên thì kết quả tìm kiếm sẽ là:



Hình 3.1 Kết quả khi sử dụng cú pháp tìm kiếm

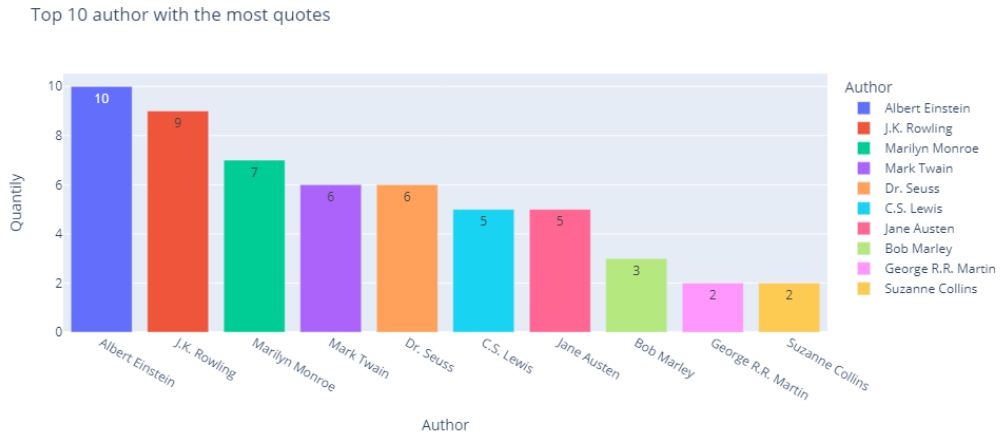
Từ đây chúng tôi nhận thấy, chúng ta có thể sử dụng selenium để trỏ đến tuổi của tác giả.Nhưng một vấn đề đã xảy ra là làm sao để biết ông đó chắc chắn là tác giải chúng ta cần tìm ? Thì giải pháp tương đối của chúng tôi đó là sẽ trỏ vào năm sinh và so sánh với năm sinh trên file csv, điều này làm tăng tính chính xác:

Kết quả:

chúng tôi đã thu thập đủ tuổi các tác giả và không ai lệch ngày sinh.

3.2.2. Khám phá dữ liệu- Data Exploration

1. ***Top 10 người nổi tiếng có số lượng câu nói nhiều nhất và ít nhất***



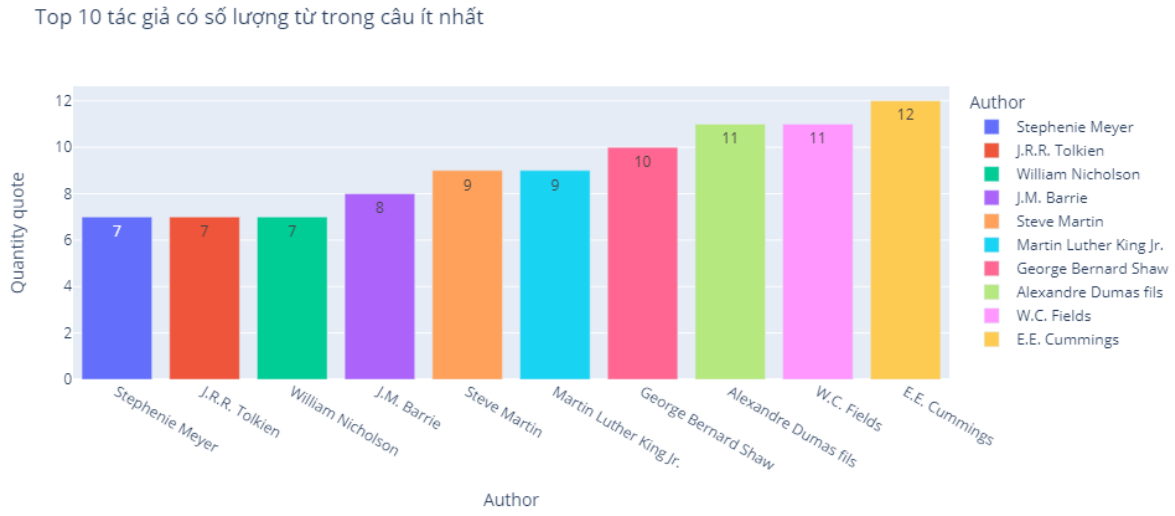
Hình 3.2 Top 10 tác giả có nhiều câu nói nhất

Ở biểu đồ cho chúng ta thấy số câu nói nhiều nhất là 10 và thuộc về Albert Einstein.

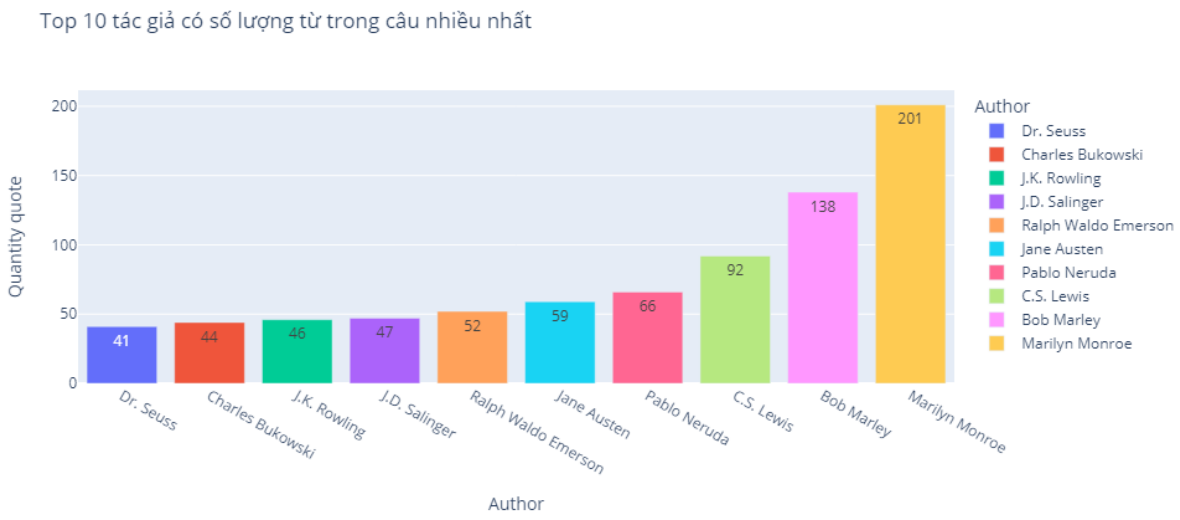
Hình 3.3 Top 10 tác giả có ít câu nói nhất

Số lượng câu nói ít nhất đa số là 1 câu.

2. ***Đặt ra câu hỏi vậy người có nhiều câu nói nhất có phải là người có từ trong câu nhiều nhất và ít nhất***



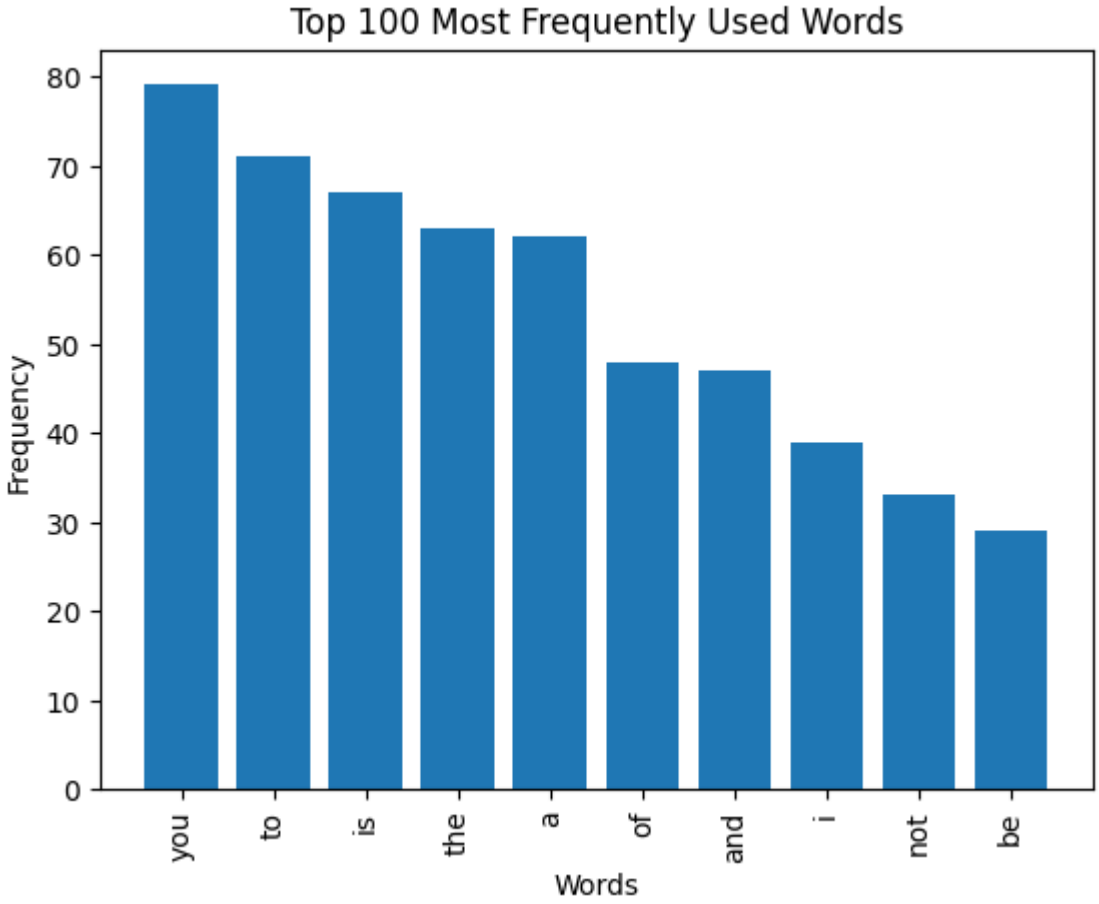
Hình 3.4 Top 10 tác giả có số lượng từ trong câu ít nhất



Hình 3.5 Top 10 tác giả có số lượng từ trong câu nhiều nhất

Từ đó ta suy ra, tác giải có số lượng câu nói nhiều nhất không phải là người có số từ nhiều nhất.

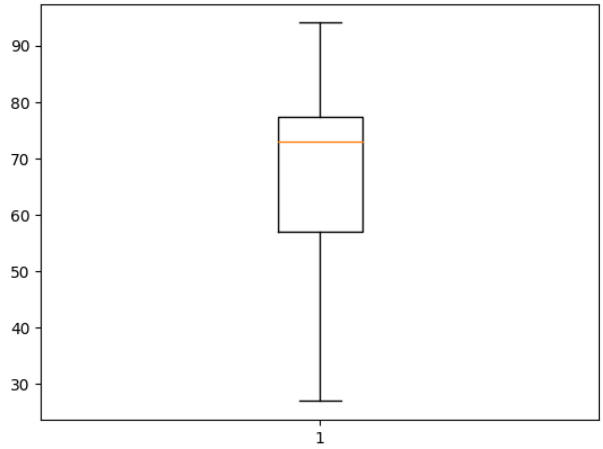
3. *Thống kê về các từ được sử dụng trong các câu nói?*



Hình 3.6 Top 10 từ được sử dụng nhiều nhất

Có thể thấy các từ You được rất nhiều tác giả sử dụng rất nhiều

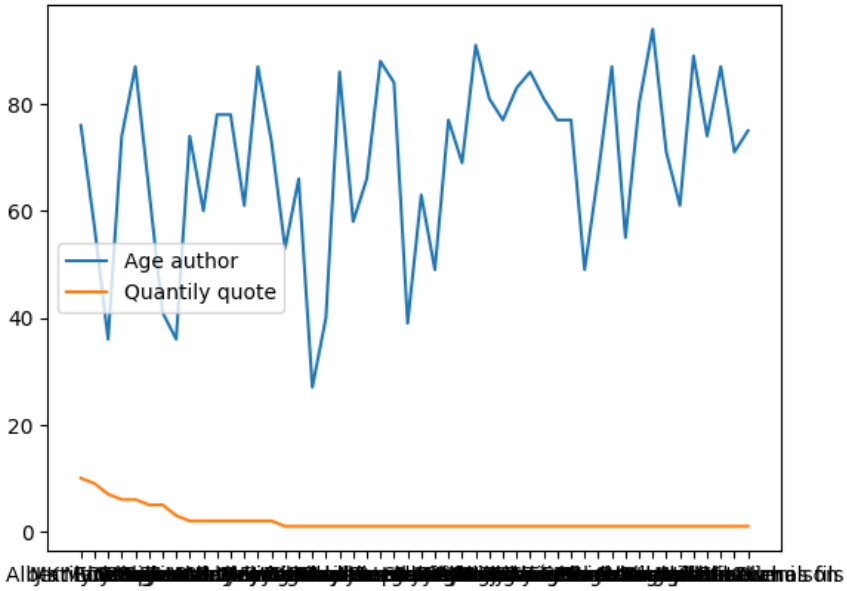
4. ***Vậy những người nổi tiếng có tuổi tập trung ở những khoảng nào?***



Hình 3.7 Thống kê tuổi

Vậy độ tuổi của người nổi tiếng thường tập trung vào khoảng 55 đến 80 tuổi là chủ yếu

5. ***Mối quan hệ giữa số lượng câu nói với tuổi của tác giả ?***



Hình 3.8 Mô tả mối quan hệ giữa số lượng câu nói với tuổi

Việc có nhiều câu nói không phụ thuộc nhiều vào độ tuổi khoảng từ 25 trở lên.

3.2.3. Trích xuất đặc trưng- Feature Extraction

Suy khi phân tích thì chúng tôi nhận thấy rằng mỗi tác giả đều có một đặc điểm riêng và mỗi đặc điểm riêng được được thể hiện ở trong các câu nói chứ không phải ở tuổi hay ngày sinh Vì tuổi hay ngày sinh đều có thể giống nhau nhưng đặc điểm câu nói câu từ thì không ai trùng nhau. Chính vì vậy chúng tôi sẽ lựa chọn Tacgia và Quote làm đặc trưng.

Suy khi lựa chọn feature thì để phụ trợ cho train model đòi hỏi dữ liệu phải ở dạng số thì chúng tôi tiến hành mã hóa chuỗi ở 2 trường:

Tacgia:

+ Sử dụng LabelEncoder là một class trong scikit-learn để mã hóa tên tác giải.

Quote:

Sử dụng công cụ xử lý văn bản CountVectorizer để chuyển đổi các câu nói thành các ma trận đếm các từ. Qua đó chúng tôi có một ma trận đếm.

Sau đó, chúng tôi tiến hành concat ma trận đã được mã hóa vào dữ liệu chính

quantity_quotes	0	1	2	...	653	654	655	656	657	658	659	6
21	0	0	0	...	0	0	0	0	0	0	0	
16	0	0	1	...	0	0	0	0	0	0	0	

Hình 3.9 Dữ liệu sau khi được mã hóa

Sau khi thực hiện mã hóa, thì chúng tôi tiến hàng xóa các cột không liên quan và dữ lại các cột chính.

	Tacgia	0	1	2	3	4	5	6	7	8	...	653	654	655	656	657	658	659	660	661	662
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	26	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1	0	0
3	30	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

Hình 3.10 Kết quả Feature Extraction cuối cùng

3.2.4. Suy luận

- Hãy dự đoán tên của người nổi tiếng theo câu nói dựa trên các đặc trưng bạn trích xuất ở trên và đánh giá trên bộ dữ liệu đã cho với tỉ lệ Train/Test và các độ đo phù hợp?

Nhận diện bài toán: Bài toán phân loại

Chính vì vậy nên chúng tôi sẽ sử dụng 2 mô hình là: DecisionTreeClassifier và RandomForestClassifier. Với bộ dữ liệu chỉ có 100 mẫu thì chúng tôi sử dụng Accuracy để đánh giá mô hình.

Sau đó tiến hành chia dữ liệu theo tỉ lệ 80 train và 20 tests. Sau khi chia tập dữ liệu, chúng tôi sử dụng 2 model trên kết quả cho thấy:

Model Decision-Tree accuracy score = 0.15

Model Random Forest accuracy score = 0.2

Khi thấy kết quả chưa được tốt chúng tôi tiến hành chia tập dữ liệu theo mới. Đầu tiên với 50 tác giả chúng tôi sẽ lấy 50 câu nói của từng người bỏ vào bộ train và lấy thêm ngẫu nhiên 30 mẫu còn lại vào bộ train, khi đó bộ train sẽ có 80 mẫu và phần còn lại là của bộ test sau khi train thì kết quả:

Model Decision-Tree accuracy score = 0.2

Model Random Forest accuracy score = 0.2

Chỉ số của Decision-Tree đã được cải thiện nhưng chưa phải vượt qua ngưỡng 0,2 và cuối cùng chúng tôi sử dụng Random Forest là model chính.

- Hãy đề xuất cách tính độ tương đồng phong cách nói giữa các tác giả và tìm ra các tác giả có phong cách nói tương đồng nhau nhất?

Đối với câu hỏi này thì để đánh giá được phong cách thì chính là đánh giá về mức độ trùng nhau giữa các câu chữ của 2 tác giả

Để tìm ra được tác giả có phong cách nói tương đồng nhau nhất chúng tôi cosine similarity để tính độ tương đồng:

Bước 1: Tiền xử lý dữ liệu (đã làm ở phần Trích xuất đặc trưng- Feature Extraction)

Loại bỏ các dấu câu và các ký tự đặc biệt không cần thiết khác trong các câu nói.

Chuyển đổi các từ sang dạng viết thường để tránh sự khác biệt trong viết hoa/thường.

Loại bỏ các từ dừng (stopwords) như "a", "an", "the",... vì chúng không mang nhiều ý nghĩa trong việc phân tích phong cách nói.

Bước 2: Tạo ma trận từ (đã làm ở phần Trích xuất đặc trưng- Feature Extraction)

Sử dụng CountVectorizer để tạo ma trận từ vựng và tính toán số lần xuất hiện của các từ trong câu nói.

Bước 3: Tính toán độ tương đồng

Sử dụng cosine similarity để tính toán độ tương đồng giữa các tác giả.

Sau khi tính toán độ tương đồng, ta có thể tìm ra các tác giả có phong cách nói tương đồng nhau nhất bằng cách sắp xếp các độ tương đồng theo thứ tự tăng dần hoặc giảm dần.

Kết quả:

Mức độ tương đồng cao nhất là: 0.518
('Albert Einstein', 'J.K. Rowling')
('Jane Austen', 'C.S. Lewis')
('Eleanor Roosevelt', 'Helen Keller')
('Eleanor Roosevelt', 'J.D. Salinger')

LÀM VIỆC NHÓM

-Trình bày tóm tắt cách thức làm việc nhóm
Làm việc online
Phân chia công việc của các thành viên trong nhóm

Câu 1:
Làm form ,tìm câu hỏi Đạt

Hoàn thành khảo sát (Nam , Đạt)

Lọc ra đúng đk của ông thầy (Nam , Đạt)

Câu 2:

Thực hiện các yêu cầu sau:

2.1 Hãy đặt ít nhất 2 câu hỏi có liên quan đến 5 chủ đề ở trên (Đạt)

2.2 Xây dựng ít nhất 10 câu hỏi khảo sát cho việc trả lời 2 câu hỏi ở câu hỏi 1(Đạt)

2.3 Thực hiện khảo sát ít nhất 50 bạn trong trường bằng phiếu khảo sát đã thiết kế ở 2.2, nộp lại các phiếu này) (Nam , Đạt)

2.4 Số hóa dữ liệu thu được (Nam, Đạt)

2.4 Phân tích dữ liệu trên bằng Python. Mỗi phân tích đều bao gồm các nội dung theo yêu cầu sau: (Nam)

- Mô tả về dữ liệu
- Thực hiện các thống kê căn bản
- Tìm mối tương quan giữa các câu hỏi khảo sát và kết quả
- Xác định các yếu tố quan trọng ảnh hưởng đến kết quả
- Trực quan hóa dữ liệu và kết quả

Câu 3:

Đạt

- Làm 3.1

Nam

- Làm 3.2

Viết báo cáo : Nam & Đạt (ai làm phần nào thì biết báo cáo phần đó)

Tổng số lần gặp nhau (tính theo buổi)
- 4 buổi

Tổng thời gian gặp nhau (tính theo giờ)
- 4 tiếng

TỰ ĐÁNH GIÁ

Câu	Nội dung	Điểm chuẩn	Tự chấm	Ghi chú
Chương 1 (15 điểm)	Thu thập dữ liệu	15	15	
Chương 2 (35 điểm)	**2.1 Câu hỏi đặt ra**	5	4	
	2.2 **Xây dựng câu hỏi khảo sát**	5	4	
	2.3 Phân tích			
	2.3.1 Kết quả khảo sát	10	10	
	2.3.2 Phân tích dữ liệu	15	14	
Chương 3 (50 điểm)	**3.1 Thu thập dữ liệu**	15	15	
	3.2 Khai phá dữ liệu			
	3.2.1. Xử lý dữ liệu	3	3	
	3.2.2. Khám phá dữ liệu	12	11	
	3.2.3. Trích xuất đặc trưng	5	4	
	3.2.4. Suy luận	15	14	
Báo cáo	**(**chú ý các chú ý 2,3,4,6 ở trang trước, nếu sai sẽ bị trừ điểm nặng**)**	10đ	10	
Làm việc nhóm	Chú ý trả lời đúng 4 yêu cầu trong phần làm việc nhóm	10đ	9	
Tổng điểm (120)			114	
Đôi qua thang điểm 10 (120=10 điểm)				