**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: **ĐÀO XUÂN HOÀNG TUẤN- 21114651**

**LÊ THANH PHÚC- 21025861**

Lớp **: 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: **ĐÀO XUÂN HOÀNG TUẤN- 21114651**

**LÊ THANH PHÚC- 21025861**

Lớp **: 420301410401**

Khoá  **: 17**

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

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023**

LỜI CẢM ƠN

Bên cạnh việc trình bày nội dung báo cáo final project ở bên trên, thì nhóm chúng em xin gửi lời cảm ơn sâu sắc trước hết là đến Mr.Hung vì đã tạo điều kiện cho thuận lợi cho sự phát triển của chúng em trong ngành Khoa học dữ liệu và hướng dẫn tận tình chúng em trong khoảng thời gian học tập trên lớp cũng như những bài tập về nhà chất lượng bên cạnh đó còn cho chúng em được thử sức với việc làm đồ án từ sớm, trong việc bắt nhóm giữa hai thành viên, chúng em đã có những sự tương tác, thấu hiểu, đặc biệt là lắng nghe và tôn trọng ý kiến của teammates để có một đề tài cuối kì thật chỉnh chu. Xin cảm ơn chân thành đến những bạn sinh viên cả trong và ngoài lớp Khoa học dữ liệu 17A vì đã hỗ trợ việc tham gia điền form thu thập dữ liệu của nhóm. Lời cuối cùng nhóm em xin chúc Thầy thật nhiều sức khỏe để có thể dẫn dắt thêm nhiều thế hệ sinh viên trong thời gian sắp tới và chúc cho đề tài cuối kì của các nhóm thành công rực rỡ!.

**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 tôi / 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)*

*Đào Xuân Hoàng Tuấn*

*Lê Thanh Phúc*

PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

TÓM TẮT

Vấn đề nghiên cứu của nhóm bao gồm 3 bài toán lớn, nhóm em sẽ trình bày ngay dưới đây:

Vấn đề 1: Phân tích ý kiến người dùng là cho biết ý kiến của người dùng về một chủ đề nào đó. Chúng ta có thể phân tích ý kiến của người dùng theo nhiều thang điểm khác nhau. Thường chúng ta hay chia thành 3 lớp chính: 1 (Positive -Tích cực), 0 (Neutral- Trung tính) và -1 (Negative -Tiêu cực). Hãy thu thập 60 dòng dữ liệu về 5 chủ đề (như bảng dưới), mỗi chủ đề 12 ý kiến, 4 ý kiến mỗi lớp liên quan đến đánh giá phản hồi của sinh viên ở IUH và lưu thành file csv

Đối với vấn đề này nhóm sẽ làm form khảo sát online vào khảo sát các bạn sinh viên sau đó từ form khảo sát đó trích xuất ra file csv. Kết quả là nhóm đã thu thập được hơn 60 ý kiến của các bạn sinh viên và đáp ứng đầy đủ các yêu cầu trên

Vấn đề 2: Từ các chủ đề ở vấn đề 1, đặt ra 2 câu hỏi lớn và từ 2 câu hỏi lớn đó rút ra ít nhất 10 câu hỏi nhỏ và làm phiếu khảo sát các bạn sinh viên, sau đó thì phân tích dữ liệu đó để giải quyết các câu hỏi.

Đối với vấn đề này nhóm em đã làm phiếu khảo sát để khảo sát các bạn sinh viên trong đó có 13 câu hỏi nhỏ bao quanh 2 câu hỏi lớn, và sau đó nhóm em sẽ số hóa dữ liệu và bắt đầu phân tích, kết quả là thu thập được 73 phiếu khảo sát của các bạn sinh viên thuộc nhiều chuyên ngành và khóa khác nhau. Dựa vào đó nhóm em đã phân tích dữ liệu để giải quyết 2 câu hỏi lớn.

Vấn đề 3: Từ website sưu tầm những câu nói của người nổi tiếng, viết code thể thu thập dữ liệu và làm giàu dữ liệu đó, sau đó khai phá, xử lý, khám phá dữ liệu, trích xuất đặc trưng và cuối cùng là suy luận để giải quyết 2 yêu cầu: dự đoán tên của người nổi tiếng theo câu nói và 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 vấn đề này nhóm em đã thực hiện cào dữ liệu từ website đó về sau đó khai phá dữ, xử lý, khám phá những dữ liệu đó từ những kiến thức đã có và áp dụng kiến thức về học máy đề giải quyết 2 yêu cầu về suy luận.

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc387692905)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc387692906)

[TÓM TẮT iv](#_Toc387692907)

[MỤC LỤC 1](#_Toc387692908)

[DANH MỤC CHỮ VIẾT TẮT 3](#_Toc387692909)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 3](#_Toc387692909)

[CHƯƠNG 1 – Thu thập dữ liệu 4](#_Toc387692910)

[CHƯƠNG 2 – Khảo sát và phân tích ý kiến của sinh viên khoa CNTT](#_Toc387692917)

2[.1 Câu hỏi đặt ra](#_Toc387692918)

2[.2 Xây dựng câu hỏi khảo sát](#_Toc387692919)

2.3 Phân tích

2.3.1 Kết quả khảo sát

2.3.2 Phân tích dữ liệu

2.4 Kết luận

[CHƯƠNG 3 – BÀI 3 **Error! Bookmark not defined.**](#_Toc387692917)

3.1 Thu thập dữ liệu

3.2 Khai phá dữ liệu

3.2.1. Xử lý dữ liệu- Data Imputation

3.2.2. Khám phá dữ liệu- Data Exploration

3.2.3. Trích xuất đặc trưng- Feature Extraction

3.2.4. Suy luận

LÀM VIỆC NHÓM 2

TỰ ĐÁNH GIÁ 6

**DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT**

**CÁC KÝ HIỆU**

**CÁC CHỮ VIẾT TẮT**

CSVC Cơ sở vật chất

CTĐT Chương tình đào tạo

DANH MỤC CÁC HÌNH VẼ

Hình 2.1 Biểu đồ phân tích số lượng sinh viên tham gia khảo sát theo chuyên ngành

Hình 2.2 Biểu đồ phân tích số lượng sinh viên tham gia khảo sát theo niên khóa

Hình 2.3 Biểu đồ phân tích đối tượng tham gia khảo sát

Hình 2.4 Mô tả dữ liệu

Hình 2.5 Hình ảnh các thống kê căn bản dữ liệu từng cột (từng câu hỏi)

Hình 2.6 Biểu đồ trực quan hóa kết quả của câu hỏi 2

Hình 2.7 Biểu đồ đánh giá sự chưa tốt của trang thiết bị IUH

Hình 2.8 Biểu đồ phân tích kết quả đánh giá về CTĐT và ĐNGV của sinh viên IUH

Hình 3.1 Cấu trúc trang web quotes.toscrape.com

Hình 3.2 Hiển thị giá trị biến 'result’ ra màn hình

Hình 3.3 Hiển thị kết quả các dữ liệu chứa nhãn “small” với class “author”

Hình 3.4 Lấy đường link tác giả và kết quả

Hình 3.5 Hàm tacgiaLink()

Hình 3.6 Kết quả khi in ra màn hình của tacgiaLink()

Hình 3.7 File Quote.csv

Hình 3.8 Lưu file và kiểm tra dữ liệu

Hình 3.9 Thêm trường STT

Hình 3.10 Nơi chứa dữ liệu ngày mất của tác giả

Hình 3.11 Phương pháp và kết quả cào ngày mất tác giả

Hình 3.12 Phương pháp và kết quả tính tuổi tác giả

Hình 3.13 Biểu đồ phân tích độ tuổi của các tác giả

Hình 3.14 Biểu đồ phân tích từ được sử dụng nhiều nhất trong các câu quotes

Hình 3.15 Biểu đồ top 10 tác giả các nhiều quotes nhất

Hình 3.16 Biểu đồ phân tích về đồ dài câu quote của tác giả

Hình 3.17 Biểu đồ phân tích về số từ trong một quotes của tác giả

Hình 3.18 Trực quan mối quan hệ giữa giữa tác giả và câu nói nổi tiếng

Hình 3.19 Kết quả huấn luyện mô hình Machine learning

Hình 3.20 Độ tương đồng phong cách của các tác giả

DANH MỤC CÁC BẢNG

CHƯƠNG 1

**THU THẬP DỮ LIỆU**

CHƯƠNG 2

**PHÂN TÍCH CÁC Ý KIẾN PHẢN HỒI CỦA SINH VIÊN KHOA CÔNG NGHỆ THÔNG TIN VỀ CƠ SỞ VẬT CHẤT, ĐỘI NGŨ GIẢNG VIÊN VÀ CHƯƠNG TRÌNH ĐÀO TẠO CỦA TRƯỜNG**

2.1 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 |

Nhóm em đã đặt ra 2 câu hỏi lớn liên quan đến 5 chủ đề trên đó là:

1. *Trải qua thời gian làm quen và học tập trong các chuyên ngành của ngành Công nghệ thông tin ở IUH thì bạn có cảm thấy hài lòng về cơ sở vật chất của trường không?*
2. *Trong thời gian học tập trong ngành CNTT ở IUH thì bạn có cảm nhận như thế nào về chương trình đào cùng sự dẫn dắt của đội ngũ giảng viên trong ngành?*

Lý do nhóm em đặt ra 2 câu hỏi này đó là trong khoảng thời gian học tập ở trường Đại học Công nghiệp Thành phố Hồ Chí Minh em đã nhận được rất nhiều ý kiến, phản ánh và quan tâm của các bạn sinh viên (bạn bè, trên các group facebook, bạn cùng lớp) và ngay cả bản thân em cũng muốn làm rõ nên đã đặt 2 câu hỏi về cơ sở vật chất, đội ngũ giảng viên và chương trình đào tạo của nhà trường.

Từ câu hỏi mà nhóm chúng em đặt ra, và kết quả của phần thu thập ý kiến em mong là nhà trường, ban lãnh đạo, những người có khả năng sẽ có thêm thông tin để có thể giải quyết những vấn đề mà sinh viên đang gặp phải, cũng như để phát triển IUH tới gần với sự yêu mến của sinh viên, tạo môi trường học tập tốt, sinh viên đạt thành tích cao, góp phần phát triển đất nước.

2.2 Xây dựng câu hỏi khảo sát

Từ 2 câu hỏi lớn nhóm em đã xây dựng hơn 10 câu hỏi nhỏ để khảo sát và sau đó lên website canva.com để thiết kế mẫu khảo sát sao cho đẹp mắt và dễ dàng nhất dành cho người được khảo sát

Đối với câu hỏi trong khảo sát, nhóm em đã thống nhất với nhau sẽ trình bày theo 2 dạng:

Dạng 1: Một câu hỏi và câu trả lời sẽ là một trong 4 đáp án (Tốt, tạm ổn, chưa tốt, tệ)

Dạng 2: Một câu hỏi và câu trả lời sẽ là nhiều đáp án trong đó mỗi đáp án là những đối tượng khác nhau để sinh viên lựa chọn, và nhờ đó thu thập được nhiều thông tin hơn

Dưới đây nhóm sẽ liệt kê các câu hỏi khảo sát và giải thích chi tiết cho từng câu hỏi khảo sát:

Câu hỏi 1: Theo bạn, cơ sở (các khu vực học tập, không gian) của trường (ngành) hiện đang trong tình trạng như thế nào?

* Câu này nhóm em thu thập ý kiến đánh giá của các bạn về cơ sở của nhà trường các khu vực học tập hoặc không gian học

Câu hỏi 2: Gạch chân những nơi bạn nghĩ cơ sở vật chất kém, cần cải thiện, và giữ nguyên những nơi có cơ sở vật chất tốt, tạm ổn?

* Câu này nhóm em sẽ đưa ra nhiều đối tượng như nhà xe, thư viện, căn tin,… để các bạn lựa chọn, mục đính là để thu được ý kiến đánh giá của sinh viên về những nơi cơ sở vật chất kém

Câu hỏi 3: Theo bạn, các trang thiết bị của trường (ngành) hiện đang trong tình trạng như thế nào?

* Câu này nhóm em thu thập ý kiến đánh giá của các bạn về các trang thiết bị của nhà trường (các công cụ để phục vụ việc dạy học)

Câu hỏi 4: Gạch chân những trang thiết bị bạn nghĩ cần cải thiện, và giữ nguyên nếu trang thiết bị đó đã đủ tốt?

* Câu này nhóm em sẽ đưa ra nhiều đối tượng như wifi, đèn, quạt, máy tính,… để các bạn lựa chọn, mục đính là để thu được ý kiến đánh giá của sinh viên về những trang thiết bị mà nhà trường cần cải thiện

Câu hỏi 5: Theo bạn, các website online của trường (ngành) hiện đang

trong tình trạng như thế nào, có hoạt động tốt không (trang sv, lms, ...) ?

* Câu này nhóm em thu thập ý kiến đánh giá của các bạn về các website online của trường.

Câu hỏi 6: Bạn cảm thấy chương trình đào tạo được thiết kế phù hợp với công việc sau này như thế nào?

* Câu này nhóm em thu thập ý kiến đánh giá của các bạn về chương trình đào tạo của trường.

Câu hỏi 7: Bạm cảm thấy các giảng viên dạy áp dụng kiến thức để làm các dự án thực tiễn trong cuộc sống như thế nào?

* Câu này nhóm em thu thập ý kiến đánh giá của các bạn về các kiến thức thực tiễn của giảng viên dạy cho sinh viên.

Câu hỏi 8: Thiết kế chương trình đào tạo, khối lượng kiến thức đã hợp lý và phù hợp với năng lực của bạn?

* Câu này nhóm em thu thập ý kiến đánh giá của các bạn về khối lượng kiến thức trong chương trình đào tạo của trường.

Câu hỏi 9: Bạn cảm thấy thế nào về trình độ chuyên môn của đội ngũ giảng viên (kiến thức sâu rộng trong ngành)?

* Câu này nhóm em thu thập ý kiến đánh giá của các bạn về trình độ chuyên môn của đội ngũ giảng viên.

Câu hỏi 10: Bạm cảm thấy thế nào về phương pháp giảng dạy của giảng viên (sự nhiệt tình giúp đỡ sinh viên, cách truyền đạt, giao bài tập)?

* Câu này nhóm em thu thập ý kiến đánh giá của các bạn về phương pháp giảng dạy của giảng viên.

Câu hỏi 11: Với cách truyền đạt của giảng viên ở mỗi giờ lên lớp, thì mức

hiểu bài của bạn như nào?

* Câu này nhóm em thu thập ý kiến đánh giá của các bạn về độ hiểu bài của các bạn sinh viên.

Câu hỏi 12: Bạn cảm thấy thế nào về các hoạt động nghiên cứu khoa học

các cuộc thi và các buổi training trong chuyên ngành của mình?

* Câu này nhóm em thu thập ý kiến đánh giá của các bạn về các hoạt động như nghiên cứu khoa học các cuộc thi và các sự kiện ngoài việc học của sinh viên.

Câu hỏi 13: Tới thời điểm hiện tại, bạn cảm thấy như thế nào về chuyên

ngành của bạn?

* Câu này nhóm em thu thập ý kiến đánh giá của các bạn về ngành học của mình.

2.3 Phân tích

2.3.1 Kết quả khảo sát

Về các thức tiến hành khảo sát, nhóm em đã tranh thủ những giờ đầu tiết học hoặc những lúc giải lao của các bạn sinh viên để khảo sát.

Về thời gian thì nhóm em đã khảo sát vào 18/04 nhóm em đã gấp rút triển khai khảo sát sớm để có dữ liệu phân tích sau này.

Về số lượng, đối tượng nhóm em đã tạo ra 75 phiếu khảo sát và mục tiêu là khảo sát được nhiều sinh viên nhất có thể và đa dạng các đối tượng trong khoa công nghệ thông tin.

Kết quả:

Nhóm em đã thu được 73 phiếu khảo sát của nhiều sinh viên thuộc đủ các chuyên ngành khác nhau trong khoa công nghệ thông tin và bao phủ đủ các khóa.

Ngoài ra còn thu được khảo sát của một số sinh viên khoa khác như: Kỹ thuật điện, Kỹ thuật điện- điện tử, thương mại điện tử,…

A picture containing text, screenshot, diagram, font

Description automatically generated

Hình 2.1 Biểu đồ phân tích số lượng sinh viên tham gia khảo sát theo chuyên ngành

Khoa học dữ liệu: 36 sinh viên

Kỹ thuật phần mền: 13 sinh viên

Công nghệ thông tin: 6 sinh viên

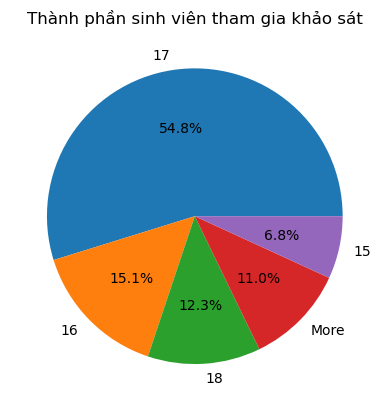
Thương mại điện tử: 4 sinh viên

Khoa học máy tính: 3 sinh viên

Kỹ thuật điện: 3 sinh viên

Kỹ thuật điện – điện tử: 1 sinh viên

Các chuyên ngành khác: 7 sinh viên



Hình 2.2 Biểu đồ phân tích số lượng sinh viên tham gia khảo sát theo niên khóa

K17: 40 sinh viên

K16: 11 sinh viên

K18: 9 sinh viên

K15: 5 sinh viên

Khác: 8 sinh viên

A picture containing text, screenshot, diagram, plot

Description automatically generated

Hình 2.3 Biểu đồ phân tích đối tượng tham gia khảo sát

2.3.2 Phân tích dữ liệu

Trình bày cách tiền xử lý dữ lý dữ liệu và

Kết quả phân tích 10 câu hỏi khảo sát ở trên

Chú ý tập trung vào các vấn đề đề bài gợi ý:

Mỗi phân tích đều bao gồm các nội dung theo yêu cầu sau:

* 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

**Tiền xử lý dữ liệu**

Ban đầu đối với dữ liệu đã được số hóa từ các phiếu khảo sát, nhóm em cần tiền xử lý dữ liệu để làm sạch dữ liệu trước khi phân tích, và nhóm em cần giải quyết những vấn đề sau:

Vấn đề 1: Hiện tượng thiếu dữ liệu ở một số cột thông tin sinh viên, do sinh viên ngại điền hoặc các vấn đề cá nhân

Vấn đề 2: Ngoài các trường dữ liệu khác là số thì có 2 trường dữ liệu sau khi nhóm em số hóa là ở dạng text, và ở dạng text thì khi số hóa dữ liệu có thể nhập dữ liệu thừa kí tự dấu cách làm cho phân loại đối tượng sinh viên sai.

Để giải quyết vấn đề 1 thì nhóm em sẽ thay thế những dữ liệu thông tin trống bằng các thông tin khác chứ không muốn xóa những dữ liệu quý giá đó nên ví dụ như dòng nào thiếu cột chuyên ngành (major) thì nhóm em sẽ thay thế bằng “More” nghĩa là chuyên ngành khác, còn thiếu cột tên (name) thì thay thế bằng “Unknown” và thiếu cột niên khóa (School year) thì sẽ thay thế bằng “More”.

Đối với vấn đề 2 thì nhóm em sẽ sử dụng các hàm python và xóa dấu cách ở đầu và cuối mỗi ô dữ liệu.

**Mô tả dữ liệu**

File dữ liệu của nhóm em gồm:

Cột Name: Tên của sinh viên làm khảo sát (Giá trị string)

Cột major: Chuyên ngành của sinh viên làm khảo sát (Các chuyên ngành của khoa công nghệ thông tin) (Giá trị string)

Cột School year: Khóa của sinh viên làm khảo sát (K15 → K18) (Giá trị integer từ 15 → 18 đại diện cho mỗi khóa)

Những cột còn lại mang quy tắt như sau:

Nếu cột có dạng *PxCy*: có nghĩa đây là câu hỏi thuộc phần *x* câu hỏi thứ *y* (Giá trị *x, y* là integer từ 1 → 4: trong đó 1 là *Tệ*, 2 là *Chưa tốt*, 3 là *Tạm ổn*, 4 là *Tốt*)

Nếu cột có dạng *PxCy(Subject name)* : có nghĩa đây là câu hỏi thuộc phần *x* câu hỏi thứ *y* và là đối tượng (ví dụ: thang máy, máy tính,...) (Giá trị là 0 hoặc 1: 0 là *Không đồng ý,* 1 là *Đồng ý*) dành cho câu hỏi có nhiều đáp án.

Và mỗi hàng là một phiếu đánh giá của sinh viên.

A screenshot of a computer

Description automatically generated with medium confidence

Hình 2.4 Mô tả dữ liệu

**Kết quả phân tích các câu hỏi**

A screenshot of a computer

Description automatically generated with low confidence

Hình 2.5 Hình ảnh các thống kê căn bản dữ liệu từng cột (từng câu hỏi)

Chi tiết phân tích cho từng câu hỏi như sau:

Câu hỏi 1: Theo bạn, cơ sở (các khu vực học tập, không gian) của trường (ngành) hiện đang trong tình trạng như thế nào?

Trung bình: 3.22 (Trên mức tạm ổn)

Trung vị: 3.0 (Tạm ổn)

Độ lệch chuẩn: 0.58

Từ thống kê trên ta thấy rằng các khu vực học tập, không gian của trường đang ở mức tạm ổn có thể chấp nhận được.

Câu hỏi 2: Gạch chân những nơi bạn nghĩ cơ sở vật chất kém, cần cải thiện, và giữ nguyên những nơi có cơ sở vật chất tốt, tạm ổn?

A picture containing text, screenshot, diagram, plot

Description automatically generated

Hình 2.6 Biểu đồ trực quan hóa kết quả của câu hỏi 2

Nhà xe:

Trung bình: 0.81

Trung vị: 1.0

Độ lệch chuẩn: 0.4

Thư viện

Trung bình: 0.08

Trung vị: 0.0

Độ lệch chuẩn: 0.28

Phòng lý thuyết

Trung bình: 0.3

Trung vị: 0.0

Độ lệch chuẩn: 0.46

Phòng thực hành

Trung bình: 0.43

Trung vị: 0.0

Độ lệch chuẩn: 0.5

Thang máy

Trung bình: 0.6

Trung vị: 1.0

Độ lệch chuẩn: 0.49

Căn tin

Trung bình: 0.31

Trung vị: 0.0

Độ lệch chuẩn: 0.46

Hội trường

Trung bình: 0.07

Trung vị: 0.0

Độ lệch chuẩn: 0.25

Khuôn viên

Trung bình: 0.27

Trung vị: 0.0

Độ lệch chuẩn: 0.45

Nhà vệ sinh

Trung bình: 0.3

Trung vị: 0.0

Độ lệch chuẩn: 0.46

Văn phòng

Trung bình: 0.04

Trung vị: 0.0

Độ lệch chuẩn: 0.2

Từ những thống kê và biểu đồ trên ta có thể thấy được những yếu tố quan trọng ảnh hưởng đến CSVC của IUH đó là nhà xe, thang máy, phòng thực hành là những CSVC quan trọng mà nhà trường cần cải thiện.

Câu hỏi 3: Theo bạn, các trang thiết bị của trường (ngành) hiện đang trong tình trạng như thế nào?

Trung bình: 2.86 (Dưới mức tạm ổn)

Trung vị: 3.0 (Tạm ổn)

Độ lệch chuẩn: 0.63

Từ thống kê trên của câu hỏi 3, ta có thể thấy rằng các trang thiết bị của trường hiện đang trong tình trạng chưa tốt, dưới mức tạm ổn và cần cải thiện thêm, và để biết nên cải thiện những gì thì nhóm em sẽ trả lời ở câu hỏi 4.

Câu hỏi 4: Gạch chân những trang thiết bị bạn nghĩ cần cải thiện, và giữ nguyên nếu trang thiết bị đó đã đủ tốt?

A picture containing text, screenshot, diagram, plot

Description automatically generated

Hình 2.7 Biểu đồ đánh giá sự chưa tốt của trang thiết bị IUH

Điều hòa, quạt

Trung bình: 0.58

Trung vị: 1.0

Độ lệch chuẩn: 0.5

Đèn

Trung bình: 0.15

Trung vị: 0.0

Độ lệch chuẩn: 0.36

Wifi

Trung bình: 0.84

Trung vị: 1.0

Độ lệch chuẩn: 0.37

Bàn, ghế

Trung bình: 0.22

Trung vị: 0.0

Độ lệch chuẩn: 0.42

Thang máy

Trung bình: 0.56

Trung vị: 1.0

Độ lệch chuẩn: 0.5

Máy tính

Trung bình: 0.55

Trung vị: 1.0

Độ lệch chuẩn: 0.5

Giáo trình

Trung bình: 0.32

Trung vị: 0.0

Độ lệch chuẩn: 0.47

Máy chiếu, bảng

Trung bình: 0.29

Trung vị: 0.0

Độ lệch chuẩn: 0.46

Từ biểu đồ và các thống kê trên ta có thể thấy rằng các yếu tố chính ảnh hưởng đến trang thiết bị học tập của sinh viên IUH đó là Wifi, điều hòa, quạt, máy tính, giáo trình và nhà trường nên đầu tư cải thiện những thứ này.

Câu hỏi 5: Theo bạn, các website online của trường (ngành) hiện đang trong tình trạng như thế nào, có hoạt động tốt không (trang sv, lms, ...) ?

Trung bình: 2.73 (Dưới mức tạm ổn)

Trung vị: 3.0 (Tạm ổn)

Độ lệch chuẩn: 0.82

Từ thống kê trên ta có thể thấy rằng các website online của trường hiện đang ở dưới mức tạp ổn, cần được cải thiện, nhất là những lúc sinh viên đăng kí học phần.

*Từ các thống kê và biểu đồ trên ta thấy các yếu tố chính ảnh hưởng xấu đến sinh viên IUH đó chính là nhà xe, thang máy, wifi, hệ thống làm mát, giáo trình học, máy tính, phòng học thực hành. Và để cải thiện chất lượng giáo dục nhà trường nên giải quyết những vấn đề này.*

(Những câu hỏi tiếp theo là về Chương trình đào tạo và đội ngũ giảng viên)

Câu hỏi 6: Bạn cảm thấy chương trình đào tạo được thiết kế phù hợp với công việc sau này như thế nào?

Trung bình: 3.18 (Trên mức tạm ổn)

Trung vị: 3.0 (Tạm ổn)

Độ lệch chuẩn: 0.63

Đối với câu hỏi này, sinh viên đánh giá khá tốt đó là trên mức tạm ổn, nên có thể thấy rằng chương trình đào tạo vẫn đang tốt đối với sinh viên.

Câu hỏi 7: Bạm cảm thấy các giảng viên dạy áp dụng kiến thức để làm các dự án thực tiễn trong cuộc sống như thế nào?

Trung bình: 3.19 (Trên mức tạm ổn)

Trung vị: 3.0 (Tạm ổn)

Độ lệch chuẩn: 0.59

Đối với câu hỏi này cũng được sinh viên đánh giá khá tốt, trên mức tạm ổn nên có thể thấy rằng các kiến thức giảng viên dạy đang đáp ứng được các dự án thực tiễn.

Câu hỏi 8: Thiết kế chương trình đào tạo, khối lượng kiến thức đã hợp lý và phù hợp với năng lực của bạn?

Trung bình: 3.14 (Trên mức tạm ổn)

Trung vị: 3.0 (Tạm ổn)

Độ lệch chuẩn: 0.63

Câu hỏi 9: Bạn cảm thấy thế nào về trình độ chuyên môn của đội ngũ giảng viên (kiến thức sâu rộng trong ngành)?

Trung bình: 3.42 (Trên mức tạm ổn)

Trung vị: 3.5 (Trên mức tạm ổn)

Độ lệch chuẩn: 0.67 (Tạm ổn)

Từ thống kê trên ta có thể thấy rằng sinh viên đáng giá rất tốt về trình độ chuyên môn của các giảng viên với trung bình cao nhất trong số các câu hỏi, nên có thể yên tâm về trình độ chuyên môn của đội ngủ giảng viên

Câu hỏi 10: Bạm cảm thấy thế nào về phương pháp giảng dạy của giảng viên (sự nhiệt tình giúp đỡ sinh viên, cách truyền đạt, giao bài tập)?

Trung bình: 3.29 (Trên mức tạm ổn)

Trung vị: 3.0 (Tạm ổn)

Độ lệch chuẩn: 0.61

Đối với câu hỏi này cũng được sinh viên đáng giá rất tốt, từ đó có thể kết luận phương pháp giảng dạy của đội ngũ giảng viên đang rất là ổn.

Câu hỏi 11: Với cách truyền đạt của giảng viên ở mỗi giờ lên lớp, thì mức

hiểu bài của bạn như nào?

Trung bình: 3.08 (Trên mức tạm ổn)

Trung vị: 3.0 (Tạm ổn)

Độ lệch chuẩn: 0.62

Với câu hỏi này thì có thể thấy sự đánh giá của sinh viên về cách truyền đạt của trang viên đang làm tạm ổn, có thể chấp nhận được

Câu hỏi 12: Bạn cảm thấy thế nào về các hoạt động nghiên cứu khoa học

các cuộc thi và các buổi training trong chuyên ngành của mình?

Trung bình: 3.41 (Trên mức tạm ổn)

Trung vị: 3.0 (Tạm ổn)

Độ lệch chuẩn: 0.62

Từ thống kê của câu hỏi 12 thì sinh viên có thể yên tâm về các hoạt động ngoại khóa của khoa công nghệ thông tin khi được đánh giá rất tốt của sinh viên

Câu hỏi 13: Tới thời điểm hiện tại, bạn cảm thấy như thế nào về chuyên

ngành của bạn?

Trung bình: 3.18 (Trên mức tạm ổn)

Trung vị: 3.0 (Tạm ổn)

Độ lệch chuẩn: 0.69

Từ khảo sát và thống kê số liệu câu hỏi này ta thấy phần đông các sinh viên vẫn đang đánh giá tốt về chuyên ngành mà mình đang học.

Dưới đây là biểu đồ chi tiết hơn về kết quả của các câu hỏi thuộc CTĐT và ĐNGV

A picture containing text, screenshot, diagram, parallel

Description automatically generated

Hình 2.8 Biểu đồ phân tích kết quả đánh giá về CTĐT và ĐNGV của sinh viên IUH

*Từ biểu đồ trên ta có thể kết luận rằng ở IUH thì đội ngũ giảng viên và chương trình đạo tạo rất phù hợp đối với sinh viên, mọi khảo sát điều trên mức tạm ổn.*

2.4 Kết luận

Sau khi thực hiện bài tập này thì nhóm em có thể kết luận được những nguyên nhân ảnh hưởng đến chất lượng giáo dục của nhà trường (Đại học Công nghiệp Thành phố Hồ Chí Minh) đa số thuộc về vấn đề CSVC cụ thể đó là nhà xe, thang máy, wifi, hệ thống làm mát, giáo trình học, máy tính, phòng học thực hành.

Những hạn chế của bài nhóm em đó là số lượng khảo sát vẫn đang còn ít, và còn đa số là của khoa Công nghệ thông tin chưa thể đánh giá tổng thể được mọi phương diện của nhà trường.

Nên trong tương lai định hướng phát triển của nhóm em là thu thập thêm nhiều thông tin hơn, càng nhiều càng tốt về mọi lĩnh vực của nhà trường và đối tượng khảo sát là toàn bộ sinh viên của trường và khảo sát một cách trung thực để đưa ra được những kết luận đúng đắn nhất.

CHƯƠNG 3

**KHAI PHÁ DỮ LIỆU TỪ CÂU NÓI CỦA NGƯỜI NỔI TIẾNG**

3.1 Thu thập dữ liệu

Đầu tiên để thu thập dữ liệu nhóm em đã sử dụng các thư viện như urllib, beautifulSoup, requests, selenium, … Từ các công cụ trên đầu tiên nhóm em sẽ dùng urllib, requests, beautifulSoup để crawl html của trang web quotes.toscrape.com về và kết quả được lưu vào file kq.txt và từ đó em mô tả cấu trúc trang web như hình bên dưới.

Diagram

Description automatically generated

Hình 3.1 Cấu trúc trang web quotes.toscrape.com

Tiếp theo với dữ liệu vừa cào về, nhóm em thực hiện bốn công việc:

Đầu tiên là đọc tất cả các thẻ html (div) với lớp là "quote" bằng thuộc tính select và lưu nó trong biến 'result’, hiển thị giá trị biến 'result’ ra màn hình.

Text

Description automatically generated

Hình 3.2 Hiển thị giá trị biến 'result’ ra màn hình

Thứ hai là tìm trong biến 'result’ vừa rồi các dữ liệu có chứa nhãn "small" với class là "author" bằng thuộc tính select và in kết quả ra màn hình.

Letter

Description automatically generated with medium confidence

Hình 3.3 Hiển thị kết quả các dữ liệu chứa nhãn “small” với class “author”

Thứ ba là viết hàm tacgiaLink() để lấy nội dung của mỗi tác giả. Với mỗi tác giả in ra màn hình các nội dung:

* Tên tác giả
* Đường link của tác giả
* Ngày tháng năm sinh
* Và câu nói nổi tiếng của tác giả

Với tên nhiệm vụ này đầu tiên em sẽ trích ra đường link của tác giả trước bằng cách trích đuôi link sử dụng select và gắn với link web từ đó tạo ra được link tác giả

Graphical user interface, text, application

Description automatically generated

Hình 3.4 Lấy đường link tác giả và kết quả

Sau khi có link tác giả em sẽ vào link đó và lấy tên từng tác giả bằng cách trích chuỗi từ thẻ h3 có class là “author-title”, và lấy luôn cả ngày sinh tác giả bằng cách select thẻ span có class là “author-born-date” và sử dụng kiểu dữ liệu datetime trong pandas để có thể phân loại được ngày tháng năm sinh của tác giả

Từ các dữ liệu đã có ở trên nhóm em viết thêm hàm get\_quotes\_author() để lấy được các câu nói nổi tiếng của tác giả bằng cách search tên tác giả trong biến result (trong đó chứa tất cả các câu nói nổi tiếng) đã xử lý ở trên, và dưới đây là hình ảnh chi tiết của cách thực hiện.

Text

Description automatically generated

Hình 3.5 Hàm tacgiaLink()

Và kết quả khi in ra màn hình là đây:

Graphical user interface, text

Description automatically generated

Hình 3.6 Kết quả khi in ra màn hình của tacgiaLink()

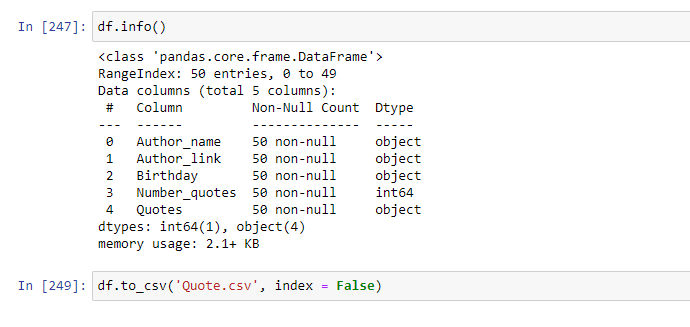
Thứ tư là lưu kết quả vào file Quote.csv tương ứng, với mỗi tác giả là 1 dòng dữ liệu và dưới đây là hình ảnh file Quote.csv của nhóm em.

Graphical user interface, text, application, email

Description automatically generated

Hình 3.7 File Quote.csv

Kết quả mà nhóm em đạt được bằng phương pháp trên là rất tốt, không có dữ liệu bị thiếu, cào về cũng rất đầy đủ (đủ 100 quotes và 50 tác giả, tất cả các trường dữ liệu đều được lấp đầy).



Hình 3.8 Lưu file và kiểm tra dữ liệu

3.2 Khai phá dữ liệu

**3.2.1. Xử lý dữ liệu- Data Imputation**

Với yêu cầu đầu tiên là Trường STT và điền tự động dữ liệu của trường thì tương đối đơn giản, nhóm em đã giải quyết nhiệm vụ này bằng cách gán cột STT bằng index của hàm reset\_index() và tăng nó lên một.

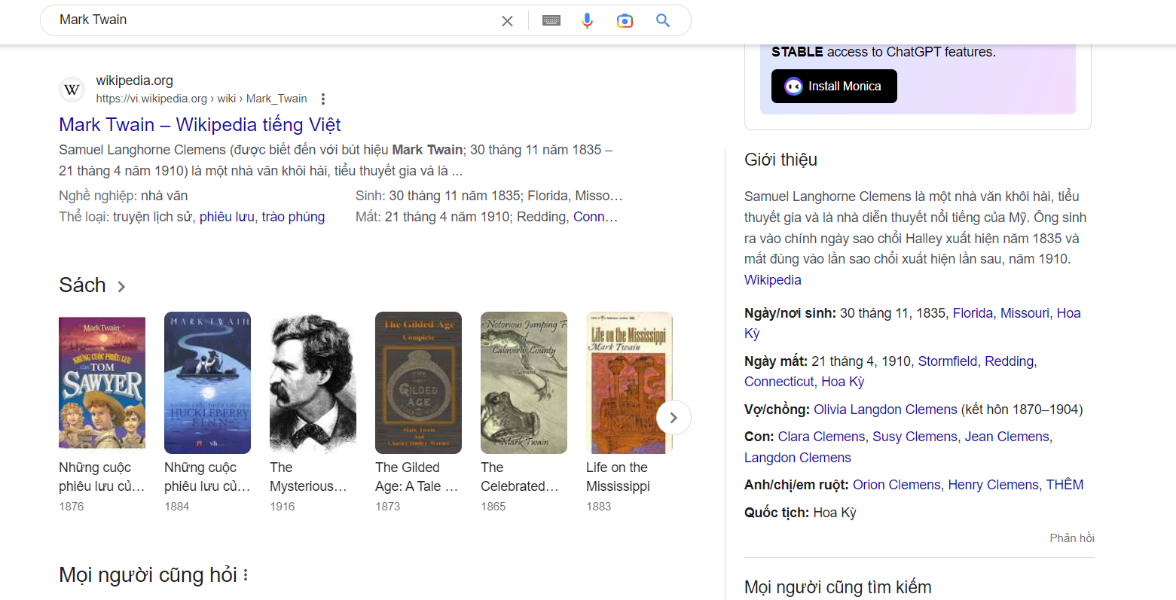
Graphical user interface, text, application

Description automatically generated with medium confidence

Hình 3.9 Thêm trường STT

Tiếp theo là giải quyết vấn đề trường ngày sinh chưa có thì trong quá trình cào dữ liệu với phương pháp của nhóm em thì không bị thiếu dữ liệu, nhưng nếu mà có bị thiếu dữ liệu em sẽ đề xuất 2 phương pháp sau đây:

- Phương pháp 1: Search google mà lấy ngày tháng năm sinh từ wikipedia (vì tất cả các tác giả trong dữ liệu đề là tác giả nổi tiếng, em có kiểm tra thì tất cả tác giả điều có tên trên wikipedia và điều có ngày sinh, ngày mất, nên từ đó em có thể crawl được ngày sinh của tất cả các tất giả, đây là phương pháp em cũng dụng để crawl ngày mất tác giả và từ đó tính tuổi tác giả



Hình 3.10 Nơi chứa dữ liệu ngày mất của tác giả

- Phương pháp 2: Thay vì sử dụng google search thì em sẽ dùng API của ChatGPT bởi vì là tác giả nổi tiếng nên mọi thông tin của tác giả chúng ta đều có thể trích xuất từ ChatGPT

Cuối cùng là phải thêm được trường “Tuổi” của tác giả, như đã nói ở trên em sẽ dùng phương pháp 1 trong 2 phương pháp em đã nêu ra đó là sử dụng google search để crawl thêm được ngày mất của tác giả và từ đó tính được tuổi tác giả

Graphical user interface, text, application

Description automatically generated

Hình 3.11 Phương pháp và kết quả cào ngày mất tác giả

Đối với phương pháp này những trường hợp wikipedia không có ngày mất của tác giả thì em kiểm tra tất cả trường hợp là tác giả còn sống, nên phương pháp tính tuổi của tác giả sẽ là lấy năm mất trừ ngày sinh (nếu có ngày mất) còn không thì sẽ lấy năm hiện tại trừ cho năm sinh.

Graphical user interface, text, application

Description automatically generated

Hình 3.12 Phương pháp và kết quả tính tuổi tác giả

Kết quả của những phương pháp này rất tốt, gần như đáp ứng tuyệt đối yêu cầu đề bài

**3.2.2. Khám phá dữ liệu- Data Exploration**

Đầu tiên là nhóm em thống kê, phân tích và trực quan hóa về các tác giả:

Năm sinh:

Trung bình năm sinh: 1903.54

Trung vị năm sinh: 1914.0

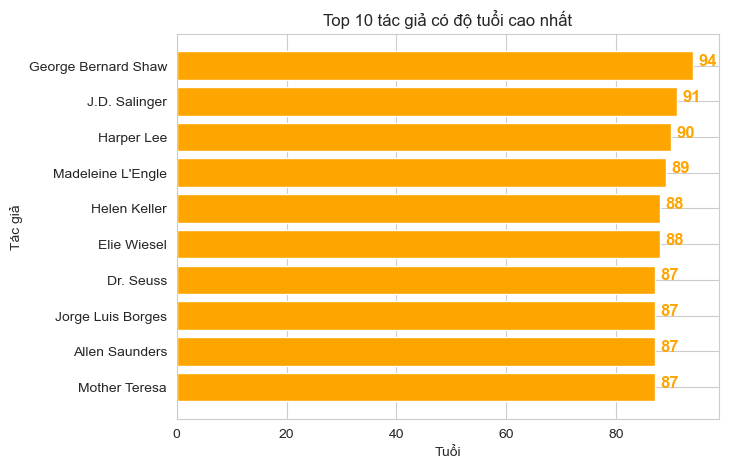
Độ lệch chuẩn năm sinh: 46.96

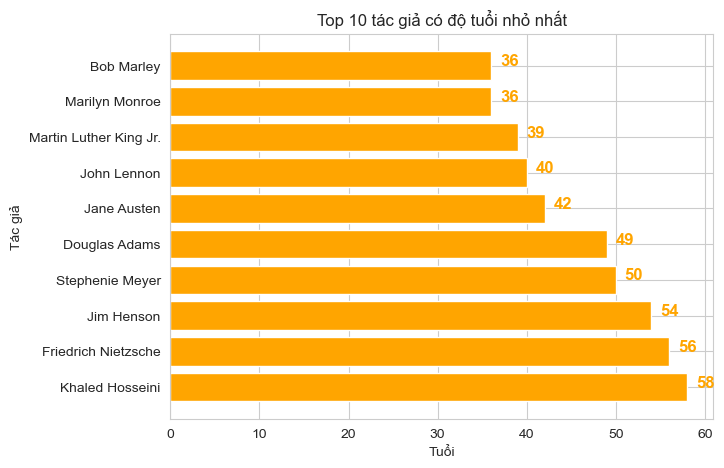
Tuổi:

Trung bình tuổi: 70.76

Trung vị tuổi: 75.0

Độ lệch chuẩn tuổi: 15.52

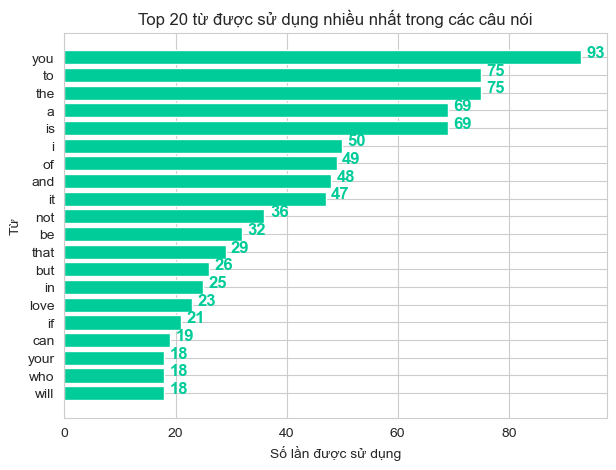




Hình 3.13 Biểu đồ phân tích độ tuổi của các tác giả

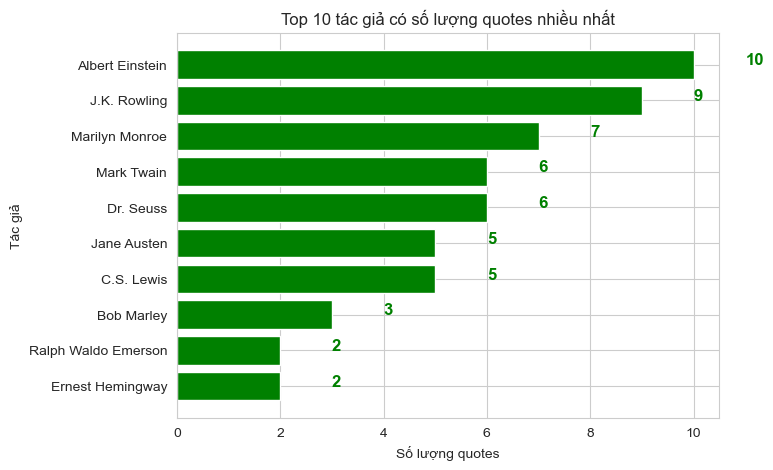
Nhìn vào biểu đồ trên ta có thể thấy rằng độ tuổi của các tác giả giao động từ 36 đến 94 tuổi trẻ nhất là Bob Marley và lớn tuổi nhất là George Bernard Shaw

Tiếp theo là nhóm em sẽ phân tích về sự liên hệ giữa các tác giả và quotes:



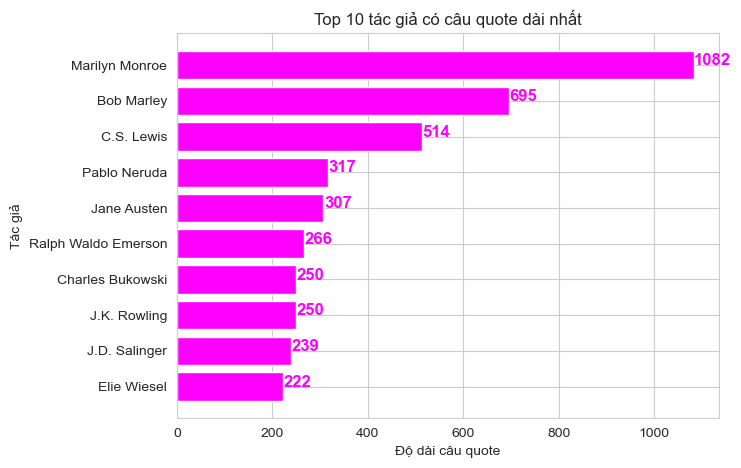
Hình 3.14 Biểu đồ phân tích từ được sử dụng nhiều nhất trong các câu quotes

Nhìn vào biểu đồ trên thì có thể thấy rằng từ sử dụng trong các câu quotes nhiều nhất là “you” sau đó đến “to”, “the”, “a”, …



Hình 3.15 Biểu đồ top 10 tác giả các nhiều quotes nhất

Nhìn vào biểu đồ phân tích này thì chúng ta có thể thấy tác giả có số lượng quotes nhiều nhất là Albert Einstein sau đó đến J.K. Rowlling.



A picture containing text, screenshot, font, number

Description automatically generated

Hình 3.16 Biểu đồ phân tích về đồ dài câu quote của tác giả

Từ biểu đồ phân tích trên ta có thể thấy câu quote dài nhất là 1082 kí tự và câu quote có ngắn nhất là 32 kí tự, lần lượt thuộc về Marilyn Monroe và William Nicholson.

A picture containing text, screenshot, number, line

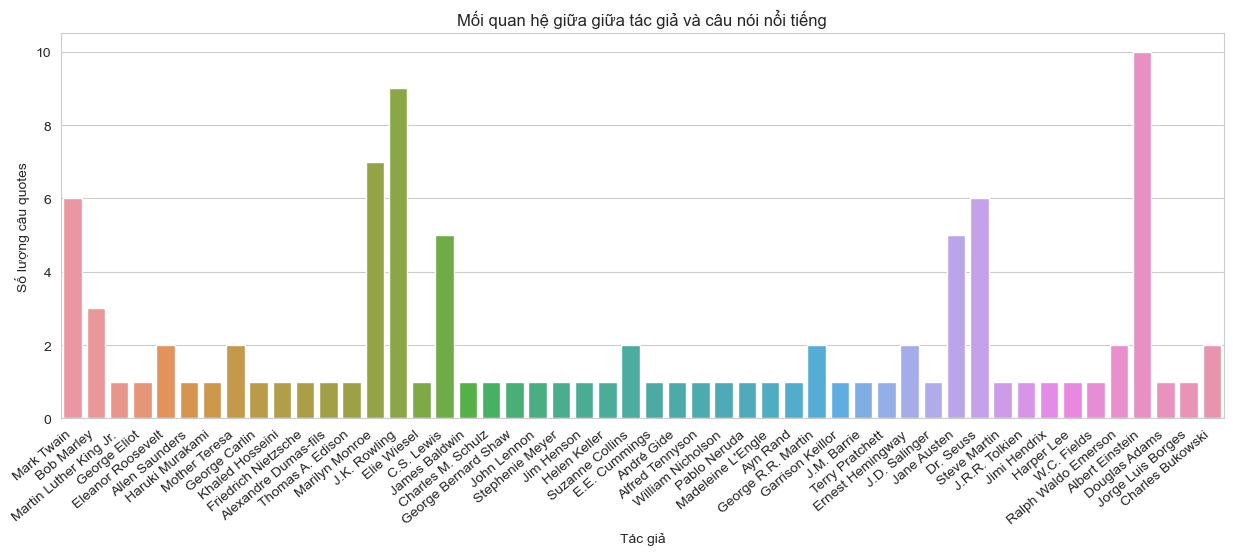
Description automatically generated

A picture containing text, screenshot, number, font

Description automatically generated

Hình 3.17 Biểu đồ phân tích về số từ trong một quotes của tác giả

Đối với biểu đồ trên ta có thể nhìn ra được là câu quote nhiều từ nhất của Marilyn Monroe và ít từ nhất là của J.R.R. Toikien.



Hình 3.18 Trực quan mối quan hệ giữa giữa tác giả và câu nói nổi tiếng

Từ biểu đồ trên ta có thể thấy được mối quan hệ giữa tác giả và câu nói nổi tiếng, số lượng câu quote của mỗi tác giả.

**3.2.3. Trích xuất đặc trưng- Feature Extraction**

Có nhiều phương pháp trích xuất đặc trưng từ văn bản được sử dụng trong xử lý ngôn ngữ tự nhiên (NLP) như: Bag-of-Words, TF-IDF, Word Embeddings, CountVectorizer. Và cách nhóm em đề xuất sử dụng đó là CountVectorizer và TF-IDF, lý do là vì:

CountVectorizer là một công cụ tuyệt vời được cung cấp bởi thư viện scikit-learn trong Python. Nó được sử dụng để chuyển đổi một văn bản nhất định thành một vectơ trên cơ sở tần suất (số lượng) của mỗi từ xuất hiện trong toàn bộ văn bản. Điều này hữu ích khi chúng tôi có nhiều văn bản như vậy và chúng tôi muốn chuyển đổi từng từ trong mỗi văn bản thành vectơ (để sử dụng trong phân tích văn bản tiếp theo)[1].

TF-IDF là viết tắt của thuật ngữ term frequency-inverse document frequency và nó là thước đo, được sử dụng trong các lĩnh vực truy xuất thông tin (IR) và học máy, có thể định lượng tầm quan trọng hoặc mức độ liên quan của các biểu diễn chuỗi (từ, cụm từ, v.v.) trong một tài liệu giữa một bộ sưu tập các tài liệu (còn được gọi là kho văn bản). Ưu điểm lớn nhất của TF-IDF đến từ mức độ đơn giản và dễ sử dụng của nó. Nó đơn giản để tính toán, nó rẻ về mặt tính toán và nó là điểm khởi đầu đơn giản để tính toán độ tương tự (thông qua vector hóa TF-IDF + độ tương tự cosine)[2].

Đầu tiên trước khi trích xuất đặc trưng thì nhóm em sẽ tiền xử lý dữ liệu trước. quá trình như sau:

* Xóa các kí tự đặc biệt chỉ dữ lại chữ cái, chữ số
* Xóa các khoảng trắng thừa
* Chuyển toàn bộ chữ về viết thường
* Loại bỏ các từ không có ý nghĩa (stop word)

Sau đó nhóm em sẽ trích xuất đặc trưng của trường dữ liệu quote đã được làm sạch từ bước trên, và sử dụng CountVectorizer hoặc TF-IDF để chuyển văn bản thành vector đặc trưng trong machine learning.

**3.2.4. Suy luận**

Đối với yêu cầu 1 là 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 thì nhóm em giải quyết như sau:

Dựa vào các đặc trưng đã trích xuất ở trên thì nhóm em chia bộ dữ liệu và random ra thành 2 tập

* Tập train: tỉ lệ 80%
* Tập test: tỉ lệ 20%

Sau đó từ 2 bộ dữ liệu trên nhóm em tiến hành huấn luyện trên 4 mô hình Machine learning đó là: Multinomial Naive Bayes, Linear Support Vector Machine, Random Forest Classifier và Logistic Regression để dự đoán tên người nổi tiếng theo câu nói. Nhóm em huấn luyện 4 mô hình này bằng tập train sau đó sử dụng tập test để đánh giá bằng các độ đo như Accuracy, F1-score.

Kết quả thu được như sau:

Multinomial Naive Bayes có Accuracy Score: 0.05

Multinomial Naive Bayes có F1 score: 0.025

Linear Support Vector Machine có Accuracy Score: 0.05

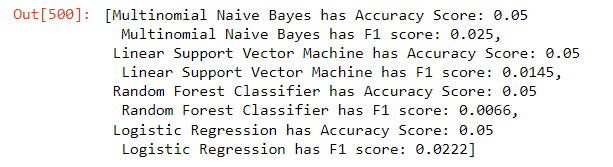
Linear Support Vector Machine có F1 score: 0.0145

Random Forest Classifier có Accuracy Score: 0.05

Random Forest Classifier có F1 score: 0.0066

Logistic Regression có Accuracy Score: 0.05

Logistic Regression có F1 score: 0.0222



Hình 3.19 Kết quả huấn luyện mô hình Machine learning

Sau khi đã huấn luyện xong mô hình nhóm em bắt đầu dự đoán tác giả dựa trên câu nói nổi tiếng, ví dụ như sau:

A screenshot of a computer screen

Description automatically generated with low confidence

Hình 3. Dự đoán tên tác giả từ 4 mô hình machine learning

Từ hình trên ta có thể thấy tác giả mà 4 mô hình Multinomial Naive Bayes, Linear Support Vector Machine, Random Forest Classifier và Logistic Regression lần lượt là Marilyn Monroe, Albert Einstein, Albert Einstein, Albert Einstein cho câu quote “nothing is impossible as long as you believe you can do it you can do it”

Bởi vì dữ liệu quá ít nên ở đây nhóm em chỉ trình bày về cách giải quyết bài toán chứ không đi sâu vào kết quả.

Đối với yêu cầu 2 là 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 thì nhóm em đã giải quyết như sau:

Các bước làm sạch dữ liệu hay trích xuất đặc trưng thì nhóm em cũng làm giống yêu cầu 1, lần này nhóm em sử dụng phương pháp TF-IDF để chuyển văn bản thành vector đặc trưng sau đó dùng phương pháp Cosine similarity để đo độ tương tự giữa hai vectơ của một không gian tích bên trong. Nó được đo bằng cosin của góc giữa hai vectơ và xác định xem hai vectơ có chỉ theo cùng một hướng hay không. Nó thường được sử dụng để đo độ tương tự của tài liệu trong phân tích văn bản.[3]

Và kết quả đạt được của nhóm như sau:

A screenshot of a computer

Description automatically generated

Hình 3.20 Độ tương đồng phong cách của các tác giả

Và đáp án về các tác giả có độ tương đồng lớn nhất của nhóm em là: Haruki Murakami và E.E. Cummings với độ tương đồng: 0.3358.

**LÀM VIỆC NHÓM**

Nhóm làm việc theo 2 hình thức online và offline

Phân chia công việc: Bạn Phúc sẽ tập trung thu thập dữ liệu lên ý tưởng, hỗ trợ, bạn Tuấn sẽ code chính, xử lý dữ liệu, cào dữ liệu, cả hai cùng viết paper

Tổng số lần gặp nhau: 5

Tổng thời gian gặp nhau: 12 giờ

**TÀI LIỆU THAM KHẢO**

1. Khushali verma (2022) Using CountVectorizer to Extracting Features from Text, link: <https://www.geeksforgeeks.org/using-countvectorizer-to-extracting-features-from-text/>
2. Anirudha Simha (2021) Understanding TF-IDF for Machine Learning, link: <https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/>
3. Jiawei Han, ... Jian Pei, Cosine Similarity (2012), link: <https://www.sciencedirect.com/topics/computer-science/cosine-similarity>

**PHỤ LỤC**

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 | 5 |  |
| 2.2 Xây dựng câu hỏi khảo sát | 5 | 5 |  |
| **2.3 Phân tích** |  |  |  |
| **2.3.1 Kết quả khảo sát** | 10 | 10 |  |
| **2.3.2 Phân tích dữ liệu** | 15 | 15 |  |
| 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 | 10 |  |
| **3.2.3. Trích xuất đặc trưng** | 5 | 4 |  |
| 3.2.4. Suy luận | 15 | 10 |  |
| 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 | 10 |  |
| Tổng điểm (120) | | | 112 |  |
| Đôi qua thang điểm 10 (120=10 điểm) | | | 9 |  |