**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 KHẮC LUẬT – 21099741**

**HOÀNG THANH TÚ – 21105251**

Lớp **: DHKHDL17A**

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 KHẮC LUẬT- 21099741**
**HOÀNG THANH TÚ- 21105251**
Lớp **: DHKHDL17A**
Khoá **: 17**
Người hướng dẫn: **TS. BÙI THANH HÙNG**


**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023**

# LỜI CẢM ƠN

Đồ án được sự thực hiện bởi nhóm 15 gồm 2 thành viên (Hoàng Thanh Tú và Nguyễn Khắc Luật) lớp DHKHDL17A dưới sự hướng dẫn của Thầy Bùi Thanh Hùng giảng viên môn Nhập môn Khoa học dữ liệu bộ môn Khoa học dữ liệu. Thầy đã tạo điều kiện và giúp đỡ về kiến thức và hỗ trợ các thắc mắc về các vấn đề trong quá trình thực hiện để chúng em hoàn thành đồ án; sự hỗ trợ và tham gia nhiệt tình của các bạn sinh viên trường Đại học Công nghiệp thành phố Hồ Chí Mình để chúng em thực hiện và hoàn thành tốt đồ án cuối kì này. Chúng em xin chân thành cảm ơn các Thầy và các bạn.

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

Chúng em xin cam đoan đây là sản phẩm đồ án của riêng chúng em 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 chúng em 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 Khắc Luật*

*Hoàng Thanh Tú*

# PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN

_____

_____

_____

_____

_____

_____

_____

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

# TÓM TẮT

Bài 1 chúng em xây dựng các câu hỏi để khảo sát đối với trải nghiệm và nhìn nhận trong quá trình học tập tại IUH của sinh viên về các vấn đề *Cơ sở vật chất (CSVC), Đội ngũ giảng viên (GV), Chương trình đào tạo (GV), Công tác sinh viên (CTDT), Tư vấn tuyển sinh (TVTS)* của trường [1]. Sau khi thu thập dư lượng dữ liệu cần cho phần của đồ án, chúng em thực hiện sàng lọc các ý kiến trả lời của các bạn sinh viên, chúng em chọn ra 60 ý kiến phù hợp với câu hỏi trong đồ án để thực hiện các bước tiếp theo. Tiếp đó, chúng em phân tích các ý kiến được các bạn sinh viên trả lời rồi đánh giá các mức độ theo thang điểm: *Tích cực (1), Trung tính (0) và Tiêu cực (-1)*. Sau cùng chúng em thực hiện theo yêu cầu của đồ án, chúng em xử lý dữ liệu trên Excel một cách hoàn chỉnh và hoàn tất việc chuyển định dạng file sang .csv theo yêu cầu của đồ án.

Bài 2 chúng em nghiên cứu tập trung vào việc khảo sát ý kiến của 60 bạn trong trường Đại học liên quan về 5 chủ đề cụ thể của bài 1. Các hướng tiếp cận nghiên cứu bao gồm xây dựng câu hỏi khảo sát, thu thập dữ liệu từ ít nhất 60 bạn sinh viên bằng phiếu khảo sát đã thiết kế [1], số hóa dữ liệu và phân tích dữ liệu bằng ngôn ngữ lập trình Python [2]. Để giải quyết vấn đề của đề bài nhóm em đã đặt ít nhất 2 câu hỏi liên quan đến 5 chủ đề nghiên cứu của bài 1: “Đánh giá của bạn về chất lượng dịch vụ nhà trường như thế nào?” và “Đánh giá của bạn về các hoạt động công tác hỗ trợ sinh viên như thế nào?”. Từ đó xây dựng 10 câu hỏi khảo sát để trả lời cho 2 câu hỏi đã đặt ở trên, các câu hỏi bao gồm:

Nhóm câu hỏi 1:

- Bạn có đồng ý rằng dịch vụ thọ viện, tra cứu tài liệu đáp ứng đủ nhu cầu tìm kiếm tài liệu của sinh viên không?
- Dịch vụ căn tin an toàn, vệ sinh sạch sẽ đáp ứng đủ nhu cầu của sinh viên, bạn có đồng ý điều đó không?
- Bạn có đồng ý rằng dịch vụ bãi đổ xe, di chuyển học tập dành cho sinh viên luôn đáp ứng đủ nhu cầu sinh viên không?

- Công tác đảm bảo an ninh, đảm bảo trật tự trong trọờng luôn an toàn, bạn có đồng ý điều đó không?

- Bạn có đồng ý rằng những đãi ngộ, chính sách dành cho sinh viên (chương trình học bổng, chính sách học phí cho sinh viên thuộc diện chính sách của nhà nước, …) luôn được nhà trường quan tâm không?

Nhóm câu hỏi 2:

- Cán bộ nhân viên nhiệt tình, vui vẻ, thân thiện với sinh viên, bạn có đồng ý điều đó không?

- Bạn có đồng ý rằng các hoạt động tư vấn học tập đáp ứng nhu cầu học tập và nghiên cứu của sinh viên không?

- Các hoạt động tư vấn hướng nghiệp, định hướng việc đều làm đáp ứng nhu cầu sau khi trường của sinh viên, bạn có đồng ý điều đó không?

- Bạn có đồng ý rằng các khiếu nại (đăng kí học phần, phúc khảo, …) của sinh viên đọợc giải quyết nhanh chóng và thỏa đáng không?

- Thủ tục hành chính (học phí, học bổng, …) liên quan đến sinh viên đọợc giải quyết kịp thời, bạn có đồng ý điều đó không?

Chúng em thực hiện khảo sát trên 60 bạn sinh viên trong trường bằng phiếu khảo sát đã thiết kế ở trên. Số liệu từ phiếu khảo sát này sẽ được nộp lại và số hóa để phục vụ cho việc phân tích dữ liệu. Sau đó phân tích dữ liệu thu được bằng ngôn ngữ lập trình Python, thống kê mô tả cơ bản tần số, tỷ lệ phần trăm, áp dụng xác suất để tính trung bình, phương sai, độ lệch chuẩn, trung vị, 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 [3]. Trực quan hóa dữ liệu và kết quả bằng các biểu đồ,biểu đồ cột, biểu đồ tròn, các bảng để hỗ trợ việc trình bày và phân tích kết quả của nghiên cứu [4]. Một số kết quả đạt được tù khảo sát là tần số và tỉ lệ đánh giá của các sinh viên đều ở mức độ đồng ý rất cao. trung bình thang điểm ở mức tốt, đa số trung vị ở mức 4. Tất cả các câu đều có tương quan dương cho thấy một mối quan hệ đồng biến giữa chúng. Đưa ra các yếu tố quan trọng

ảnh hưởng đến kết quả như: thiết kế khảo sát, chất lượng mẫu khảo sát, quá trình xử lý dữ liệu, các yếu tố nhiễu và sai số, độ tin cậy và độ chính xác của công cụ đo lường, ngữ cảnh và thời gian thu thập dữ liệu.

Bài 3, đầu tiên chúng em thực hiện việc thu thập dữ liệu về những câu nói của Những người nổi tiếng trên thế giới có ở đường link: http://quotes.toscrape.com/:. Chúng em thống nhất sử dụng thư viện *Beautifulsoup* và *requests* để viết code python tiến hành cào toàn bộ dữ liệu có trong trang web như: kiểm tra website có bao nhiêu trang tất cả, tên tác giả, đường link của tác giả, ngày tháng năm sinh, câu quote của tác giả,… sau đó lưu vào 1 file *"kq.txt"* [5]. Đông thời chúng em mô tả ngắn gọn về câu trúc của trang web bằng cách in ra cấu trục code html trang đầu tiên của trang web đã cào. Dựa vào dữ liệu vừa cào về ở file *"kq.txt"*. Chúng em tiếp tục thực hiện tiếp tục đọc tất cả các thẻ html (div) với lớp là "quote" và lưu nó trong biến 'result', sau đó hiển thị giá trị biến 'result' ra màn hình. Đông thời. tìm trong biến 'result' vừa rồi các dữ liệu có chứa nhãn "small" với class là "author" và in kết quả ra màn hình. Tiếp theo chúng em viết hàm tacgiaLink() để lấy nội dung của mỗi tác giả bao gồm: Tên tác giả; Đường link của tác giả; Ngày tháng năm sinh; Và câu nói nổi tiếng của tác giả. Cuối cùng lưu kết quả của hàm tacgiaLink vào file Quote.csv tương ứng, với mỗi tác giả là 1 dòng dữ liệu. Chúng em đã thu thập được 100 câu nói nổi tiếng từ trang web trên một cách tự động theo code của các ý trên.

3.2. Khai phá dữ liệu:

Phần này chúng em thực hiện tất cả dựa trên file "Quote.csv" đã lưu ở phần 3.1.2d. Phần này chúng em sử dụng nhiều về thư viện pandas và matplotlib.pyplot để thực hiện tạo bảng và vẽ biểu đồ. [6]

3.2.1. Xử lý dữ liệu – Data Imputation:

Ở phần này chúng em chỉ sử dụng thư viện pandas

Chúng em sử dụng thư viện pandas để tạo bảng rồi chúng em thêm trường "STT" và điền tự động số tăng dần vào trước vị trí tên tác giả

Chúng em đã kiểm tra và không phát hiện giá trị nào của trường ngày sinh bị thiếu trong bảng dữ liệu

Tiếp đó, dựa vào trường trường ngày sinh đã kiểm tra không bị thiếu, chúng tối tiếp tục với tìm tuổi của các tác giả. Chúng em xem lại trang web thấy rằng trang web không cung cấp thông tin năm mất của các tác giả, nên chúng em dùng phương pháp thủ công là tra cứu Internet về tuổi của tác giả. Chúng tối phát hiện thấy có những tác giả đã mất và có tuổi là con số nhất định và có những tác giả còn sống, chưa xác định được tuổi. Vì vậy, chúng em thống nhất tạo 1 list tuổi các tác giả, điền tuổi đã xác đinh của các tác giả vào list còn những tác giả chưa xác định được tuổi, chúng em lưu dưới chuỗi 'age' (với 'age' là số tuổi cần tìm = năm hiện tại – năm sinh của tác giả đó). Chúng em sử dụng vòng for trong list để xác định vị trí của các 'age' trong list, chúng em tạo tiếp 1 list rỗng để chứa các năm sinh của các tác giả cần tìm chúng em tách năm sinh và lưu vào list rỗng đó. Sau đó chúng em tạo thêm 1 list nữa, chúng em thực hiện tìm tuổi của tác giả bằng công thức "số tuổi = cần tìm = năm hiện tại – năm sinh của tác giả đó". Tiếp theo, chúng em gộp list tuổi có chứa chuỗi 'age' và list tuổi của tác giả còn sống đã tìm được. Cuối cùng chúng em thêm trường "Tuổi" vào bảng dữ liệu bằng thư viện pandas.

3.2.2. Khám phá dữ liệu- Data Exploration:

Ở phần này, chúng em sử dụng kết hợp cả 2 thư viện pandas và matplotlib.pyplot để thực hiện

Trước tiên chúng em hiển thị một số thông tin thống của tập dữ liệu đã cho: lấy số dòng với số cột có trong bộ dữ liệu, xem thông tin về Index, kiểu dữ liệu và dung lượng của dữ liệu, tổng kết thông tin thống kê cho các cột có kiểu dữ liệu là số.

Chúng em thực hiện thống kê số lượng tác giả có trong web và in ra; các câu nói nổi tiếng, tác giả và câu quote tương ứng của họ trong trang web, thống kê số lượng câu nói nổi tiếng của mỗi tác giả.

Tiếp đó chúng em thống kê về năm sinh và độ tuổi của tác giả

Thống kê về câu nổi tiếng dài nhất, ngắn nhất, số từ,chiều dài của các câu quote… Và vẽ các biểu đồ về top 10 câu dài nhất, top 10 câu ngắn nhất [7]

# MỤC LỤC

# DANH MỤC CÁC HÌNH VẼ

# DANH MỤC CÁC BẢNG

# CHƯƠNG 1
# THU THẬP DỮ LIỆU

## 1.1. Yêu cầu thu thập dữ liệu

Thu thập 60 dòng dữ liệu về 5 chủ đề (CSVC, GV, CTDT, CTSV, TVTS), 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.

## 1.2. Cách thức tiến hành thu thập dữ liệu

Chúng em tạo một mẫu Google Form với các câu hỏi về các chủ đề trên. Sau đó, chúng em tạo 1 QR từ link Google Form đã tạo. Chúng em lên thư viện trường IUH – nơi có lượng sinh viên tập trung đông đảo học tập, làm việc nhóm… Nhờ mọi người quét QR đã tạo và tiến hành trả lời các câu hỏi khảo sát được tạo trong Google Form của chúng em. [1]

## 1.3. Kết quả

Khi các câu trả lời của các bạn được khảo sát, chúng sẽ được ghi nhận và lưu trữ vào 1 file Excel bằng Google Sheet mà chúng em liên kết.

Đánh giá tính chính xác của bộ dữ liệu: Chúng em xem xét các câu trả lời nghiêm túc, cảm nhận của các bạn sinh viên, sàng lọc và xóa bỏ các câu trả lời đùa cợt xúc phạm trường, làm ảnh hưởng xấu đến hình ảnh trường của những bạn thực hiện không nghiêm túc.

# CHƯƠNG 2
# PHÂN TÍCH KHẢO SÁT VỀ CHẤT LƯỢNG DỊCH VỤ VÀ CÁC HOẠT ĐỘNG HỖ TRỢ SINH VIÊN Ở ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH

## 2.1 Câu hỏi đặt ra

Hai câu hỏi mà nhóm đặt ra là "Đánh giá của bạn về chất lượng dịch vụ nhà trường như thế nào?" thuộc chủ đề về cơ sở vật chất và "Đánh giá của bạn về các hoạt động công tác hỗ trợ sinh viên như thế nào?" thuộc chủ đề Công tác sinh viên [1]. Việc đặt ra hai câu hỏi trên nhằm mục đích đánh giá chất lượng dịch vụ và hoạt động hỗ trợ sinh viên của nhà trường. Đối với câu hỏi đầu tiên về các dịch vụ, việc đánh giá này có ý nghĩa rất lớn đối với học sinh và sinh viên khi muốn lựa chọn trường học phù hợp. Chất lượng dịch vụ là yếu tố quan trọng ảnh hưởng đến quá trình học tập và nghiên cứu của sinh viên, cũng như là yếu tố thu hút được đông đảo học sinh đăng ký học tại trường. Đối với câu hỏi thứ hai về hoạt động hỗ trợ sinh viên, đây là một yếu tố quan trọng đối với sinh viên, giúp họ có môi trường học tập và phát triển tốt nhất. Việc đánh giá này có thể giúp nhà trường cải thiện và nâng cao chất lượng các hoạt động hỗ trợ sinh viên, giúp cho sinh viên cảm thấy thoải mái, hài lòng và có được sự phát triển tốt nhất trong quá trình học tập tại trường. Vì vậy, việc đặt ra hai câu hỏi này giúp cho nhà trường có cái nhìn tổng quan hơn về chất lượng dịch vụ và hoạt động hỗ trợ sinh viên của mình. Đồng thời, việc đánh giá này cũng giúp cho sinh viên có thể có sự lựa chọn phù hợp hơn khi đăng ký học tại trường.

## 2.2 Xây dựng câu hỏi khảo sát

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: [1]

- Xác định mục đích của khảo sát: thu thập ý kiến đánh giá của sinh viên về chất lượng dịch vụ và các hoạt động hỗ trợ sinh viên ở Đại học Công nghiệp TP Hồ Chí Minh.

- Hình thức khảo sát: khảo sát trực tiếp bằng bảng hỏi để thu thập thông tin.
- Đối tượng khảo sát: Sinh viên ở Đại học Công nghiệp TP Hồ Chí Minh
- Loại câu hỏi: câu hỏi khảo sát đơn lựa chọn.
- Sử dụng các câu hỏi cụ thể và rõ ràng nêu ra quan các nhận định để người khảo sát lựa chọn các đánh giá.
- Sử dụng thang đo quảng 5 để xây dựng các lựa chọn.

Dưới đây là 10 câu hỏi khảo sát:

- "Bạn có đồng ý rằng dịch vụ thư viện, tra cứu tài liệu đáp ứng đủ nhu cầu tìm kiếm tài liệu của sinh viên không?". Câu hỏi này đánh giá mức độ hài lòng của sinh viên đối với dịch vụ thư viện và tra cứu tài liệu trên trường học. Dịch vụ này có đáp ứng được nhu cầu tìm kiếm tài liệu của sinh viên hay không và sinh viên có gặp khó khăn gì trong quá trình sử dụng dịch vụ này hay không.
- "Dịch vụ căn tin an toàn, vệ sinh sạch sẽ đáp ứng đủ nhu cầu của sinh viên, bạn có đồng ý điều đó không?". Câu hỏi này đánh giá mức độ hài lòng của sinh viên đối với dịch vụ căn tin trên trường học. Dịch vụ này có đảm bảo an toàn vệ sinh sạch sẽ hay không, đáp ứng được nhu cầu ăn uống của sinh viên hay không và sinh viên có gặp khó khăn gì trong quá trình sử dụng dịch vụ này hay không.
- "Bạn có đồng ý rằng dịch vụ bãi đổ xe, di chuyển học tập dành cho sinh viên luôn đáp ứng đủ nhu cầu sinh viên không?". Câu hỏi này đánh giá mức độ hài lòng của sinh viên đối với dịch vụ bãi đỗ xe và di chuyển trên trường học. Dịch vụ này có đáp ứng được nhu cầu di chuyển của sinh viên hay không và sinh viên có gặp khó khăn gì trong quá trình sử dụng dịch vụ này hay không.
- "Công tác đảm bảo an ninh, đảm bảo trật tự trong trọờng luôn an toàn, bạn có đồng ý điều đó không?". Câu hỏi này đánh giá mức độ hài lòng của sinh viên đối với công tác đảm bảo an ninh và trật tự trên trường học. Công tác này có đảm bảo an ninh, trật tự trên trường học hay không và sinh viên có gặp khó khăn gì trong quá trình học tập do không đảm bảo an ninh, trật tự hay không.

- “Bạn có đồng ý rằng những đãi ngộ, chính sách dành cho sinh viên (chọơng trình học bổng, chính sách học phí cho sinh viên thuộc diện chính sách của nhà nước, …) luôn được nhà trường quan tâm không?”. Câu hỏi này liên quan đến các chính sách và đãi ngộ mà trường đang cung cấp cho sinh viên nhằm giúp đỡ họ trong quá trình học tập. Điều này bao gồm các chương trình học bổng, chính sách học phí ưu đãi cho sinh viên thuộc diện chính sách của nhà nước như miễn giảm học phí, hỗ trợ chi phí sinh hoạt, vật chất, thực phẩm, chỗ ở... Ngoài ra, câu hỏi này cũng đánh giá sự hiệu quả và tính hợp lý của các chính sách và đãi ngộ này, xem chúng có đáp ứng được nhu cầu của sinh viên hay không.
- “Cán bộ nhân viên nhiệt tình, vui vẻ, thân thiện với sinh viên, bạn có đồng ý điều đó không?”. Câu hỏi này đánh giá mức độ hài lòng của sinh viên về cán bộ nhân viên của trường, bao gồm các nhân viên trong phòng hành chính, cán bộ thư viện, giáo viên, nhân viên bảo vệ, đội ngũ y tế... Mục đích của câu hỏi là để đánh giá mức độ chuyên nghiệp, tận tâm và thân thiện của các cán bộ nhân viên này đối với sinh viên, điều này cũng ảnh hưởng đến chất lượng dịch vụ mà trường cung cấp.
- “Bạn có đồng ý rằng các hoạt động tư vấn học tập đáp ứng nhu cầu học tập và nghiên cứu của sinh viên không?”. Câu hỏi này đánh giá mức độ hài lòng của sinh viên về hoạt động tư vấn học tập tại trường, bao gồm các hoạt động như tư vấn chọn ngành, lựa chọn học phần, tư vấn nghiên cứu khoa học, hướng dẫn tìm kiếm tài liệu, thư viện điện tử... Mục đích của câu hỏi này là để đánh giá tính hữu ích và chất lượng của các hoạt động này, xem chúng có đáp ứng được nhu cầu của sinh viên và giúp ích cho việc học tập của họ hay không.
- “Các hoạt động tới vấn hướng nghiệp, định hướng việc đều làm đáp ứng nhu cầu sau khi trường của sinh viên, bạn có đồng ý điều đó không?”. Câu hỏi này đề cập đến việc đánh giá hoạt động tư vấn hướng nghiệp và định hướng việc làm của trường đại học. Nhu cầu của sinh viên sau khi ra trường rất đa dạng và phức tạp, bao gồm việc tìm

kiếm việc làm, xây dựng kế hoạch nghề nghiệp và phát triển kỹ năng mềm để trở thành một ứng viên tốt.

- "Bạn có đồng ý rằng các khiếu nại (đăng kí học phần, phúc khảo, …) của sinh viên được giải quyết nhanh chóng và thỏa đáng không?". Câu hỏi này đánh giá mức độ hài lòng của sinh viên đối với quy trình giải quyết khiếu nại của trường đại học. Khiếu nại có thể bao gồm các vấn đề liên quan đến đăng ký học phần, điểm số, phúc khảo điểm, hay các vấn đề khác liên quan đến chất lượng dịch vụ giáo dục của trường.
- "Thủ tục hành chính (học phí, học bổng, …) liên quan đến sinh viên đợợc giải quyết kịp thời, bạn có đồng ý điều đó không?". Câu hỏi này đánh giá mức độ hài lòng của sinh viên đối với quy trình giải quyết các thủ tục hành chính của trường đại học, bao gồm việc đóng học phí, nộp đơn xin học bổng hay các vấn đề khác liên quan đến chính sách và thủ tục hành chính của trường.

## 2.3 Phân tích

### 2.3.1 Kết quả khảo sát

Dữ liệu được thu thập trực tiếp bằng bảng hỏi từ 60 sinh viên thuộc năm nhất và năm hai của nhiều chuyên ngành khác nhau tại trường Đại học Công nghiệp ở Thành phố Hồ Chí Minh vào tháng 4 năm 2023 bằng cách sử dụng phiếu khảo sát về chất lượng dịch vụ và các hoạt động hỗ trợ sinh viên của nhà trường. Kết quả thu được gồm 60 bảng khảo sát dữ liệu đầy đủ của 10 trường tương ứng với 10 câu hỏi khảo sát, kiểu số nguyên trong phạm vi từ thang đo 1 đến 5.

### 2.3.2 Phân tích dữ liệu

Tiền xử lý dữ lý dữ liệu

Từ 60 bảng khảo sát trước khi phân tích dữ liệu ta tiến hành tiền xử lí dữ liệu. Đầu tiên sắp xếp lại các bảng khảo sát, loại bỏ dữ liệu không hợp lệ, kiểm tra tính đầy đủ của dữ liệu, Chuẩn hóa dữ liệu về thang đo quãng 5. Kiểm tra và loại bỏ dữ liệu nhiễu và các giá trị bị khuyết [2].

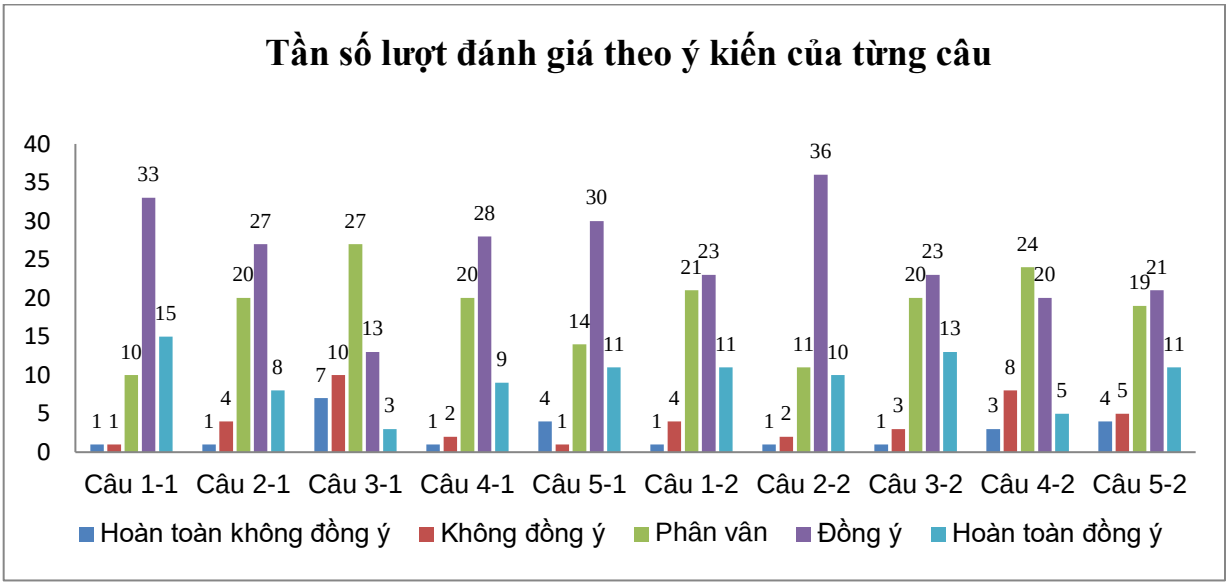
**Phân tích 10 câu hỏi khảo sát trên**

**- Mô tả dữ liêu:** dữ liệu được thu thập từ 60 sinh viên thuộc năm nhất và năm hai của nhiều chuyên ngành khác nhau tại trường Đại học Công nghiệp ở Thành phố Hồ Chí Minh vào tháng 4 năm 2023 bằng cách sử dụng phiếu khảo sát về chất lượng dịch vụ và các hoạt động hỗ trợ sinh viên của nhà trường. Nội dung gồm hai phần chính: (1) đánh giá chất lượng dịch vụ nhà trường và (2) đánh giá các hoạt động hỗ trợ sinh viên của nhà trường. Mỗi phần có năm câu hỏi, mỗi câu hỏi được đánh giá trên một đơn vị đo lường là thang điểm từ 1 đến 5 (thang đo Likert quãng 5) với các mức độ đánh giá: hoàn toàn không đồng ý (1), không đồng ý (2), trung lập (3), đồng ý (4) và hoàn toàn đồng ý (5).Dữ liệu thu được gồm 60 dòng dữ liệu đầy đủ của 10 trường, kiểu số trong phạm vi từ 1 đến 5.

**- Thực hiện các thông kê căn bản:**

*** Thống kê tần số lượt đánh giá theo ý kiến của từng câu** [2] [4]

| **Dữ liệu thống kê** | **1** | **2** | **3** | **4** | **5** |
|---|---|---|---|---|---|
| **Câu 1-1** | 1 | 1 | 10 | 33 | 15 |
| **Câu 2-1** | 1 | 4 | 20 | 27 | 8 |
| **Câu 3-1** | 7 | 10 | 27 | 13 | 3 |
| **Câu 4-1** | 1 | 2 | 20 | 28 | 9 |
| **Câu 5-1** | 4 | 1 | 14 | 30 | 11 |
| **Câu 1-2** | 1 | 4 | 21 | 23 | 11 |
| **Câu 2-2** | 1 | 2 | 11 | 36 | 10 |
| **Câu 3-2** | 1 | 3 | 20 | 23 | 13 |
| **Câu 4-2** | 3 | 8 | 24 | 20 | 5 |
| **Câu 5-2** | 4 | 5 | 19 | 21 | 11 |

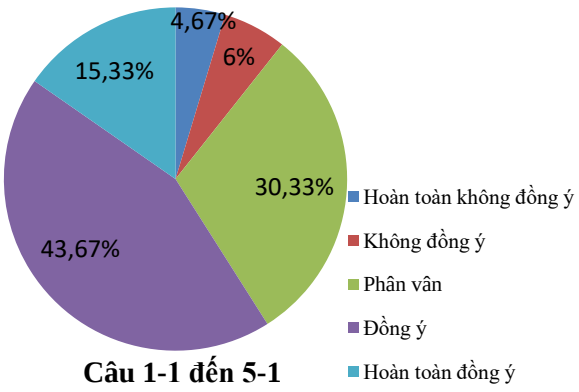



Biểu đồ 2.1 Tần số lượt đánh giá theo ý kiến của từng câu

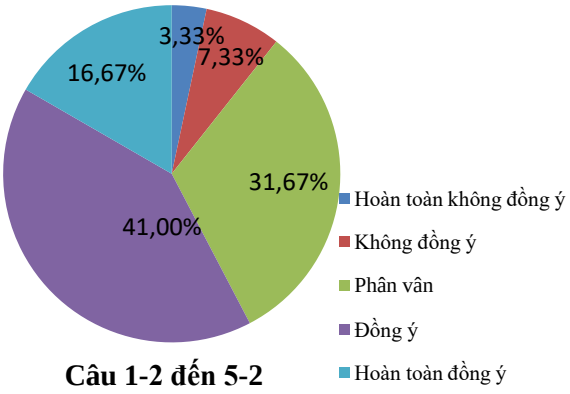
Nhận xét: Thang điểm 4 có tần số đánh giá nhiều nhất và thang điểm 1 có tần số đánh giá ít nhất trong các đánh giá của các câu. Dữ liệu cho thấy sự khác biệt giữa các câu, với một số câu có tần số đánh giá thang điểm cao hơn hoặc thấp hơn so với các câu khác.

*** Tỉ lệ phần trăm từng đánh giá của hai nhóm câu hỏi** [2] [4]

| **Đánh giá** | **Câu 1-1 đến 5-1** | **Câu 1-2 đến 5-2** |
|---|---|---|
| **Hoàn toàn không đồng ý** | 4.67% | 3.33% |
| **Không đồng ý** | 6% | 7.33% |
| **Phân vân** | 30.33% | 31.67% |
| **Đồng ý** | 43.67% | 41.00% |
| **Hoàn toàn đồng ý** | 15.33% | 16.67% |




**Câu 1-1 đến 5-1**




**Câu 1-2 đến 5-2**

0

Nhận xét: Tỷ lệ phần trăm đồng ý của nhóm câu 1 và nhóm của 2 là cao nhất lần lượt là 43.67% và 41.00%, tỉ lệ phần trăm hoàn toàn không đồng ý của nhóm câu 1 và nhóm câu 2 lâ thấp nhất lần lượt là 4,67% và 3.33%, trong các tỷ lệ phần trăm của tổng 2 nhóm câu cho thấy chất lượng dịch vụ và các hoạt động hỗ trợ sinh viên của nhà trường được đánh giá khá cao. Dữ liệu cho thấy sự khác biệt giữa số lượng các mức đánh giá, với một số mức có tỉ lệ phần trăm đánh giá cao hơn hoặc thấp hơn so với các mức khác.

*** Thống kê giá trị trung bình từng ý kiến của từng câu hỏi** [3] [4]

- Công thức:

$$\acute{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \quad (2.1)$$

Trong đó:

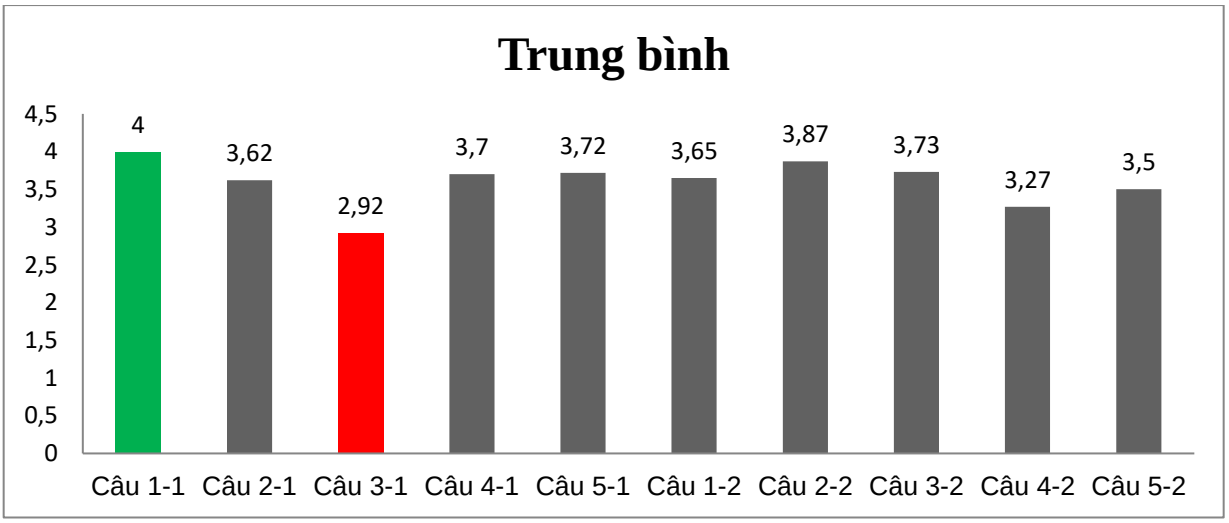
$\acute{x}$ là giá trị trung bình

$x_i$ là giá trị của cột tương ứng trong bảng dữ liệu

$w_i$ là trọng số tương ứng với giá trị $x_i$. Trong trường hợp này, trọng số được tính bằng cách sử dụng các số nguyên từ 1 đến số lượng phần tử của cột.

| | Câu 1-1 | Câu 2-1 | Câu 3-1 | Câu 4-1 | Câu 5-1 | Câu 1-2 | Câu 2-2 | Câu 3-2 | Câu 4-2 | Câu 5-2 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Trung bình** | 4 | 3.62 | 2.92 | 3.7 | 3.72 | 3.65 | 3.87 | 3.73 | 3.27 | 3.5 |

0




0

Nhận xét: Câu 1-1 có giá trị trung bình là 4.0, cao nhất trong các giá trị trung bình của các câu cho thấy về dịch vụ thư viện, tra cứu tài liệu đáp ứng đủ nhu cầu tìm kiếm tài liệu của sinh viên. Câu 3-2 có giả trị trung bình là 2.92, thấp nhất trong các giá trị trung bình của các câu cho thấy các hoạt động tư vấn hướng nghiệp, định hướng việc đều làm đáp ứng nhu cầu sau khi trường của sinh viên được sự đồng thuận chưa cao. Các giá trị trung bình của các câu nằm trong khoảng từ 2.92 đến 4.00, cho thấy sự biến

động về độ chênh lệch giữa các câu. Dữ liệu cho thấy sự khác biệt giữa các câu, với một số câu có giá trị trung bình cao hơn hoặc thấp hơn so với các câu khác.

*** Thống kể giá trị phương sai từng ý kiến của từng câu hỏi** [3] [4]

- Công thức:

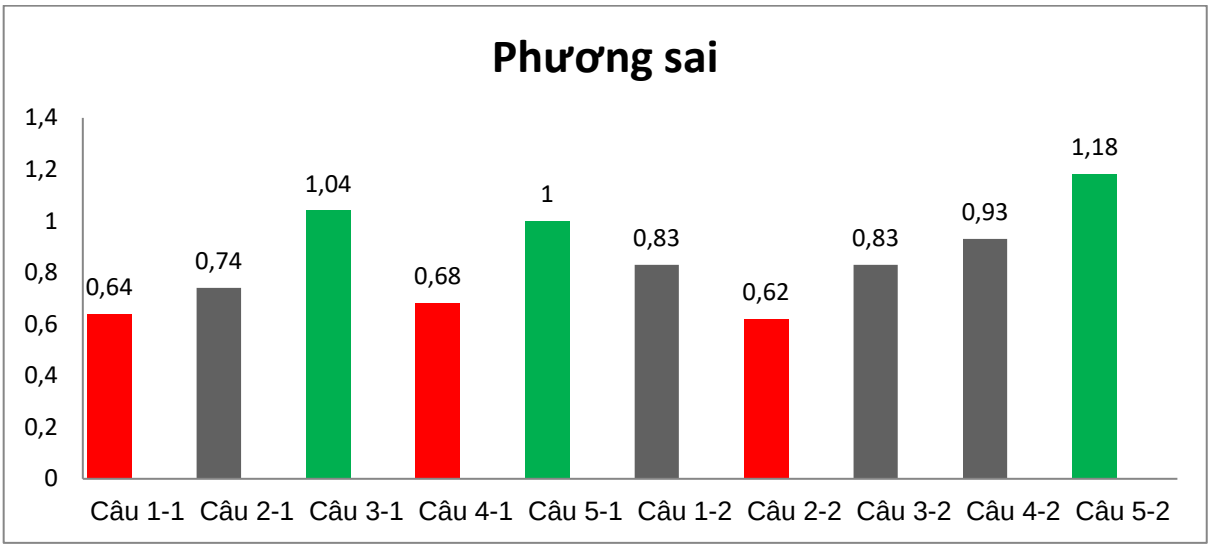
$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n} (2.2)$$

Trong đó:

$\sigma^2$ là phương sai $n$ là số lượng quan sát

$x_i$ là giá trị của quan sát thứ $i$ $\mu$ là giá trị trung bình của tập dữ liệu.

| | **Câu 1-1** | **Câu 2-1** | **Câu 3-1** | **Câu 4-1** | **Câu 5-1** | **Câu 1-2** | **Câu 2-2** | **Câu 3-2** | **Câu 4-2** | **Câu 5-2** |
|---|---|---|---|---|---|---|---|---|---|---|
| **Phương sai** | 0.64 | 0.74 | 1.04 | 0.68 | 1 | 0.83 | 0.62 | 0.83 | 0.93 | 1.18 |




02 Phương sai đánh giá của từng câu hỏi

Nhận xét: Câu 3-1, câu 5-1 và câu 5-2 có phương sai cao, cho thấy mức độ biến động của đánh giá trong các câu này là lớn so với các câu khác. Điều này có thể cho thấy sự đa dạng trong các đánh giá của các câu này. Câu 1-1, câu 4-1 và câu 2-2 có phương sai thấp, cho thấy mức độ biến động của đánh giá trong các câu này là thấp so với các câu khác. Điều này có thể cho thấy sự đồng đều trong các đánh giá của các câu này.

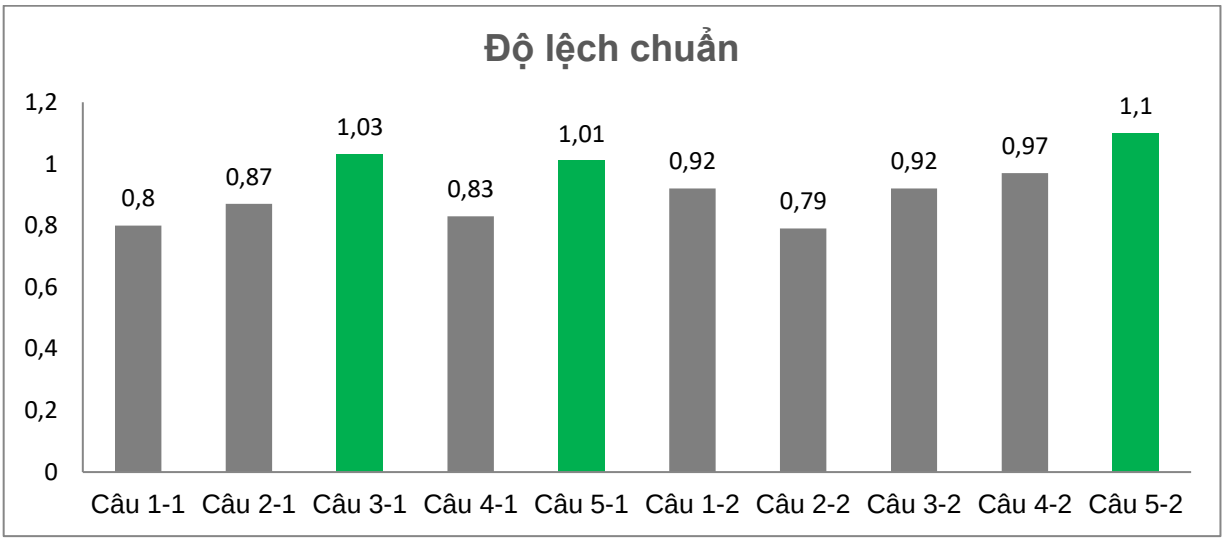
*** Thống kê độ lệch chuẩn** [3] [4]

- Công thức:

$$Std = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \acute{x})^2}{n-1}} \quad (2.3)$$

Trong đó, $x_i$ là giá trị của mẫu thứ $i$, $x$ là giá trị trung bình của mẫu $n$ là số lượng mẫu.

| | **Câu 1-1** | **Câu 2-1** | **Câu 3-1** | **Câu 4-1** | **Câu 5-1** | **Câu 1-2** | **Câu 2-2** | **Câu 3-2** | **Câu 4-2** | **Câu 5-2** |
|---|---|---|---|---|---|---|---|---|---|---|
| **Độ lệch chuẩn** | 0.8 | 0.87 | 1.03 | 0.83 | 1.01 | 0.92 | 0.79 | 0.92 | 0.97 | 1.1 |




03 Đồ lệch chuẩn theo từng câu hỏi

Nhận xét: Dữ liệu này cho thấy độ biến động, độ phân tán của các câu đánh giá trong các câu hỏi khác nhau, với câu 3-1,câu 5-1 và câu 5.2 có độ lệch chuẩn cao và các câu khác có độ lệch chuẩn dao động từ thấp đến trung bình.

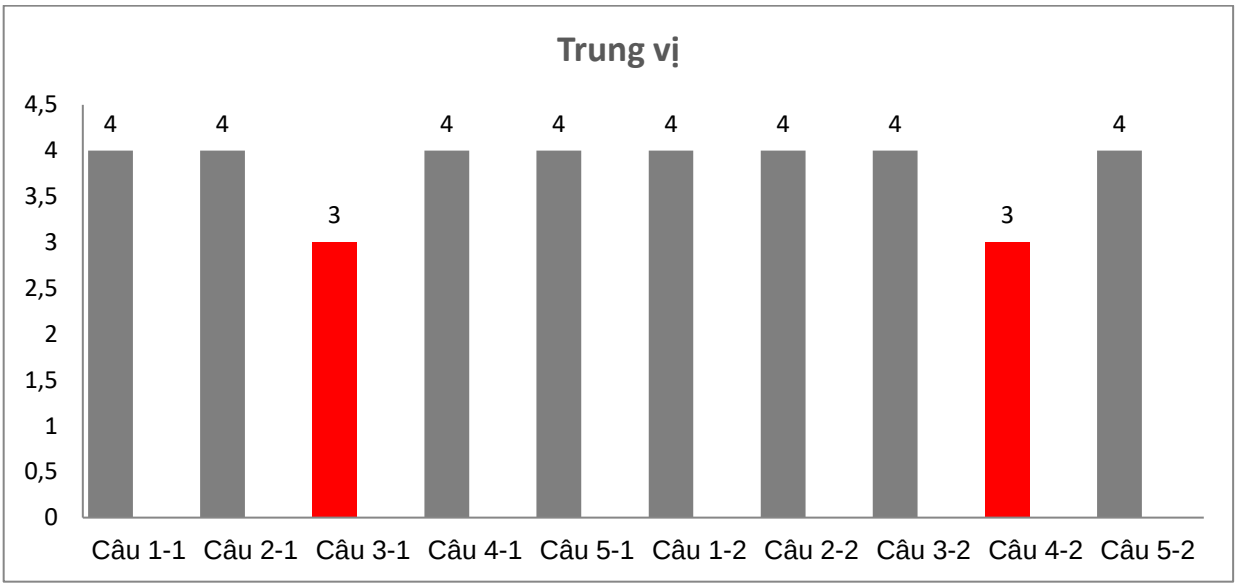
*** Thống kê trung vị theo từng ý kiến và từng câu** [3] [4]

- Công thức:

$$\text{Trung vị} = \text{median}(\text{data}) \quad (2.4)$$

| | **Câu 1-1** | **Câu 2-1** | **Câu 3-1** | **Câu 4-1** | **Câu 5-1** | **Câu 1-2** | **Câu 2-2** | **Câu 3-2** | **Câu 4-2** | **Câu 5-2** |
|---|---|---|---|---|---|---|---|---|---|---|
| **Trung vị** | 4 | 4 | 3 | 4 | 4 | 4 | 4 | 4 | 3 | 4 |

01 Trung vị của từng câu hỏi




004 Trung vị từng câu hỏi0

Nhận xét**:** Đa số các câu có giá trị trung vị là 4.0 , chỉ có câu 3-1 và câu 4-2 có giá trị là 3.0 thấp hơn các câu khác.

- **Mối tương quan giữa các câu hỏi khảo sát và kết quả** [3] [4]

+ Công thức:

$$r(X,Y) = \frac{cov(X,Y)}{std(X)\cdot std(Y)} \quad (2.5)$$

Trong đó:

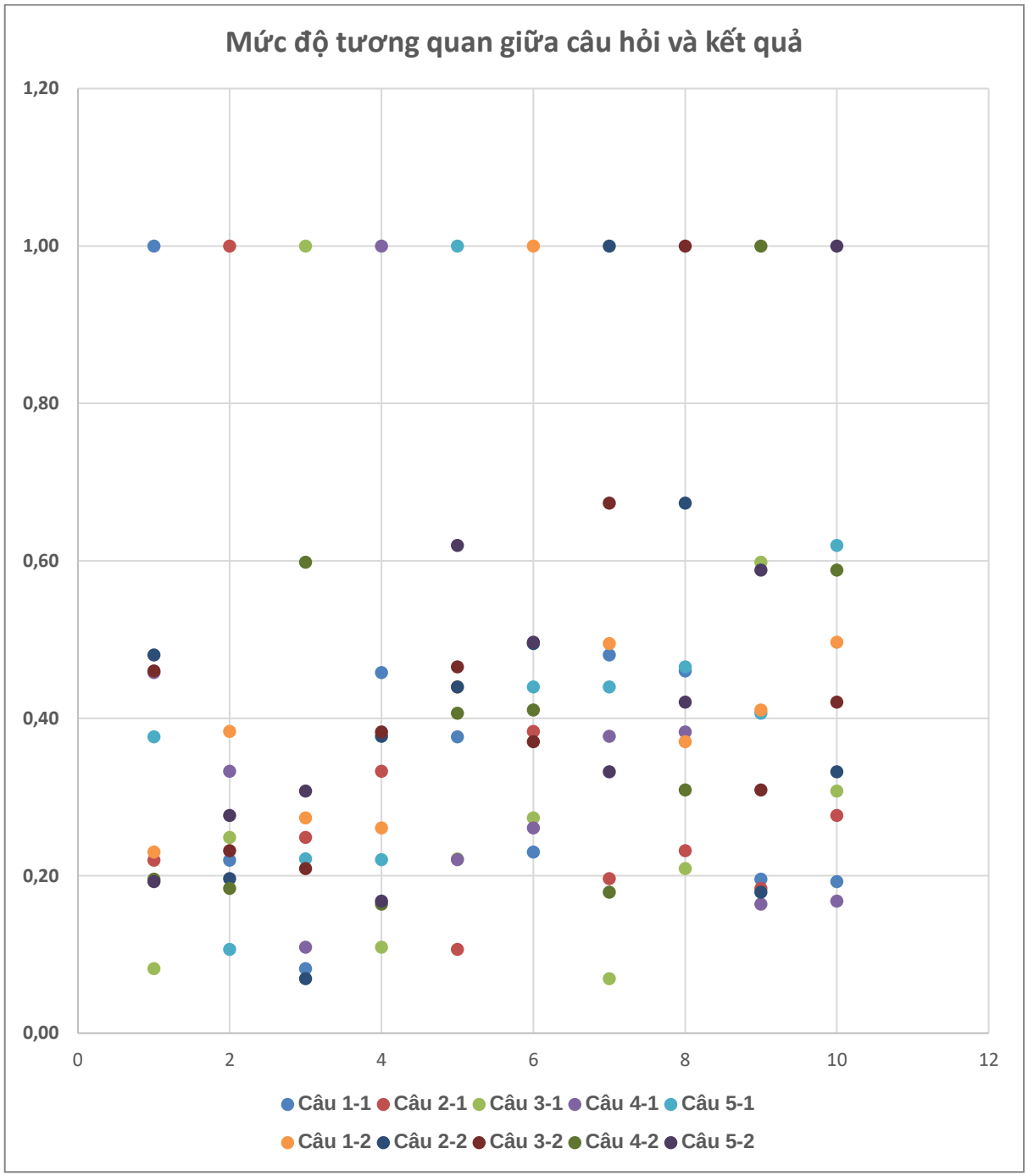
r(X,Y) là độ tương quan Pearson giữa hai biến ngẫu nhiên X và Y

cov(X,Y) là hiệp phương sai giữa X và Y

std(X) và std(Y) lần lượt là độ lệch chuẩn của X và Y

| | **Câu 1-1** | **Câu 2-1** | **Câu 3-1** | **Câu 4-1** | **Câu 5-1** | **Câu 1-2** | **Câu 2-2** | **Câu 3-2** | **Câu 4-2** | **Câu 5-2** |
|---|---|---|---|---|---|---|---|---|---|---|
| **Câu 1-1** | 1.00 | 0.22 | 0.08 | 0.46 | 0.38 | 0.23 | 0.48 | 0.46 | 0.20 | 0.19 |
| **Câu 2-1** | 0.22 | 1.00 | 0.25 | 0.33 | 0.11 | 0.38 | 0.20 | 0.23 | 0.18 | 0.28 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Câu 3-1** | 0.08 | 0.25 | 1.00 | 0.11 | 0.22 | 0.27 | 0.07 | 0.21 | 0.60 | 0.31 |
| **Câu 4-1** | 0.46 | 0.33 | 0.11 | 1.00 | 0.22 | 0.26 | 0.38 | 0.38 | 0.16 | 0.17 |
| **Câu 5-1** | 0.38 | 0.11 | 0.22 | 0.22 | 1.00 | 0.44 | 0.44 | 0.47 | 0.41 | 0.62 |
| **Câu 1-2** | 0.23 | 0.38 | 0.27 | 0.26 | 0.44 | 1.00 | 0.50 | 0.37 | 0.41 | 0.50 |
| **Câu 2-2** | 0.48 | 0.20 | 0.07 | 0.38 | 0.44 | 0.50 | 1.00 | 0.67 | 0.18 | 0.33 |
| **Câu 3-2** | 0.46 | 0.23 | 0.21 | 0.38 | 0.47 | 0.37 | 0.67 | 1.00 | 0.31 | 0.42 |
| **Câu 4-2** | 0.20 | 0.18 | 0.60 | 0.16 | 0.41 | 0.41 | 0.18 | 0.31 | 1.00 | 0.59 |
| **Câu 5-2** | 0.19 | 0.28 | 0.31 | 0.17 | 0.62 | 0.50 | 0.33 | 0.42 | 0.59 | 1.00 |




05 Độ tương quan giữa các câu hỏi khảo sát và kết quả

Nhận xét:

- Tất cả đều có tương quan dương cho thấy một mối quan hệ đồng biến giữa chúng, thể hiện mối quan hệ tuyến tính dương, điều đó có thể cho thấy sự tương quan tích cực giữa các câu hỏi.

- Có một số cặp câu có hệ số tương quan cao (gần 1), thể hiện mối quan hệ giữa các câu hỏi là rất mạnh và có thể được sử dụng để dự đoán giá trị của một câu hỏi dựa trên giá trị của câu hỏi khác. Như câu 5-1 và Câu 5-2 có hệ số tương quan là 0.62, Câu 2-2 và Câu 3-2 có hệ số tương quan là 0.67.

- Có một số cặp câu có hệ số tương quan thấp (gần 0), cho thấy mối quan hệ giữa các câu hỏi không mạnh và có thể bị ảnh hưởng bởi những yếu tố khác. Như câu 1-1 và Câu 3-1 có hệ số tương quan là 0.08, Câu 3-1 và Câu 2-2 có hệ số tương quan là 0.07.

**- Các yếu tố quan trọng ảnh hưởng đến kết quả trong các phân tích trên:**

| | | | |
|---|---|---|---|
| ========== | | | |
| Dep. Variable: | y | R-squared: | 1.000 |
| Model: | OLS | Adj. R-squared: | 1.000 |
| Method: | Least Squares | F-statistic: | 2.231e+29 |
| Date: | Fri, 19 May 2023 | Prob (F-statistic): | 0.00 |
| Time: | 07:41:11 | Log-Likelihood: | 1927.5 |
| No. Observations: | 60 | AIC: | -3833. |
| Df Residuals: | 49 | BIC: | -3810. |
| Df Model: | 10 | | |
| Covariance Type: | nonrobust | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| ========== | | | | | | |
| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
| ---------- | | | | | | |
| const | -8.882e-16 | 2.7e-15 | -0.329 | 0.744 | -6.31e-15 | 4.54e-15 |
| Câu 1-1 | 0.1000 | 6.19e-16 | 1.62e+14 | 0.000 | 0.100 | 0.100 |
| Câu 2-1 | 0.1000 | 5.32e-16 | 1.88e+14 | 0.000 | 0.100 | 0.100 |
| Câu 3-1 | 0.1000 | 4.87e-16 | 2.05e+14 | 0.000 | 0.100 | 0.100 |
| Câu 4-1 | 0.1000 | 5.62e-16 | 1.78e+14 | 0.000 | 0.100 | 0.100 |
| Câu 5-1 | 0.1000 | 5.48e-16 | 1.82e+14 | 0.000 | 0.100 | 0.100 |
| Câu 1-2 | 0.1000 | 5.78e-16 | 1.73e+14 | 0.000 | 0.100 | 0.100 |
| Câu 2-2 | 0.1000 | 7.63e-16 | 1.31e+14 | 0.000 | 0.100 | 0.100 |
| Câu 3-2 | 0.1000 | 6.28e-16 | 1.59e+14 | 0.000 | 0.100 | 0.100 |
| Câu 4-2 | 0.1000 | 6.06e-16 | 1.65e+14 | 0.000 | 0.100 | 0.100 |
| Câu 5-2 | 0.1000 | 5.48e-16 | 1.83e+14 | 0.000 | 0.100 | 0.100 |
| ========== | | | | | | |
| Omnibus: | | 4.683 | Durbin-Watson: | | 0.224 | |
| Prob(Omnibus): | | 0.096 | Jarque-Bera (JB): | | 2.551 | |
| Skew: | | 0.261 | Prob(JB): | | 0.279 | |
| Kurtosis: | | 2.136 | Cond. No. | | 81.7 | |
| ========== | | | | | | |

03 Kết quả phân tích hồi quy tuyến tính

Nhận xét:

- Mô hình đã giải thích được 100% phương sai của biến phụ thuộc y, cho thấy rằng các biến độc lập đã giải thích được tất cả sự biến động của biến phụ thuộc.
- Các hệ số hồi quy (coef) đều bằng 0.1, cho thấy rằng tất cả các biến độc lập đều có ảnh hưởng tương đương đến biến phụ thuộc y.
- Giá trị p (P>|t|) của tất cả các hệ số đều bằng 0, cho thấy rằng các hệ số hồi quy là có ý nghĩa thống kê.
- Giá trị F-statistic rất lớn, kèm theo giá trị P của F-statistic bằng 0, cho thấy rằng mô hình tuyến tính đã giải thích được sự biến động của biến phụ thuộc với mức ý nghĩa thống kê cao.
- Trong kết quả của mô hình hồi quy tuyến tính, các hệ số (coef) ứng với các biến độc lập (Câu 1-1 đến Câu 5-2) đều bằng 0.1, tức là khi các biến độc lập tăng lên 1 đơn vị, biến phụ thuộc (y) cũng tăng lên 0.1 đơn vị. Điều này cho thấy tất cả các yếu tố được nghiên cứu đều ảnh hưởng tích cực đến kết quả.
- Tuy nhiên, hệ số cho biến hằng (const) là rất nhỏ, không đáng kể (gần bằng 0). Điều này cho thấy biến độc lập không có ảnh hưởng đáng kể đến biến phụ thuộc khi các biến độc lập đều bằng 0.
- Với R-squared và Adj. R-squared đều bằng 1.0, mô hình hoàn toàn giải thích được sự biến động của biến phụ thuộc bằng các biến độc lập. Tuy nhiên, với giá trị F-statistic rất lớn (2.231e+29) và giá trị p-value bằng 0.00, cho thấy mô hình hoàn toàn không ngẫu nhiên, và giả thuyết rằng tất cả các hệ số bằng 0 được bác bỏ.

**Một số yếu tố khác bên ngoài**

+ Thiết kế khảo sát: bao gồm việc chọn đúng đối tượng nghiên cứu, định nghĩa rõ ràng các biến đo lường, lựa chọn phương pháp thu thập dữ liệu phù hợp, thiết kế câu hỏi, độ

dài khảo sát, thứ tự các câu hỏi, và việc đảm bảo tính nhất quán và độ tin cậy của câu hỏi.

+ Mẫu khảo sát: cần được lựa chọn một cách ngẫu nhiên hoặc ngẫu nhiên hợp lý, đại diện cho đúng đối tượng nghiên cứu và đảm bảo tính khả thi trong việc thu thập dữ liệu. Kích thước mẫu cũng cần đủ lớn để đạt được độ chính xác và độ tin cậy.

+ Phương pháp khảo sát: Phương pháp khảo sát phải được lựa chọn sao cho phù hợp với mục đích của nghiên cứu và đảm bảo tính đáng tin cậy của kết quả khảo sát.

+ Câu trả lời của các bạn sinh viên: bao gồm mức độ chính xác, độ trung thực, khả năng hiểu và đọc hiểu câu hỏi, cũng như sự đồng nhất trong cách đánh giá và trả lời.

+ Thời gian khảo sát: Thời gian khảo sát phải đảm bảo đủ thời gian để thu thập đầy đủ thông tin cần thiết và đáp ứng mục tiêu nghiên cứu.

+ Xử lý dữ liệu: là bước quan trọng để đảm bảo tính chính xác của kết quả khảo sát. + Việc kiểm tra, lọc và chuyển đổi dữ liệu, tính toán các chỉ số, đánh giá tính đúng đắn và độ tin cậy của kết quả là yếu tố quan trọng ảnh hưởng đến kết quả của khảo sát.

+ Nhiễu và sai số: bao gồm sai số trong quá trình thu thập dữ liệu, sai số có thể làm giảm tính chính xác và độ tin cậy của kết quả khảo sát.

+ Độ tin cậy và độ chính xác của công cụ đo lường.

**2.4 Kết luận**

**- Khi hoàn thành xong bài 2 qua sô hóa dữ liệu và các thống kê căn bản em có một số kết luận như sau:**

+ Dữ liệu thu đa phần được có sự đồng nhất về kết quả đánh giá ở đây đa số sinh viên đều đồng ý(thang điểm 4) cho thấy các dự liệu có liên quan mật thiết.

+ Một một kết quả có sự lệch so với các kết quả khác cho thấy mức độ biến động trong câu hỏi và cách trả lời, cần chú trọng vào xử lí các câu này.

+ Kết quả thu nhận cuối cùng cho thấy rằng về chất lượng dịch vụ và các hoạt động hỗ sinh viên ở Đại học Công nghiệp TP Hồ Chí Minh khá tốt.

**- Những điểm còn hạn chế:**

**+** Thời gian thu thập và xử lí dữ liệu mất nhiều thời gian.

+ Các phân tích chỉ có thể phân tích được các dũ liệu cục bộ không gian khảo sát chưa toàn diện.

+ Các phương thức chỉ có thể đưa ra các kết luận vào một thời điểm và chưa dự đoán được từ các dũ liệu đã có.

+ Chưa có nhiều các nơi để lưu trữ các dữ liệu như vậy để có thể thuận tiện trích xuất và phân tích.

**- Hướng phát triển trong tương lai:**

+ Tham khảo thêm một số mô hình phương pháp mới có thể thực hiện việc phân tích dữ liệu một cách nhanh chóng và chính xác.

+ Xây dựng một cơ sở dữ liệu để tiện cho việc trích xuất.

+ Mở rộng đối tượng khảo sát để đưa ra các kết quả tốt hơn.

# CHƯƠNG 3
# KHAI PHÁ DỮ LIỆU TỪ CÂU NÓI CỦA NGƯỜI NỔI TIẾNG

## 3.1 Thu thập dữ liệu

Dữ liệu về những câu nói của Những người nổi tiếng trên thế giới có ở đường link: http://quotes.toscrape.com/ [5].

Chúng em sử dụng thư viện BeautifulSoup và request để thu thập dữ liệu của trang web theo yêu cầu của đồ án

## 3.2 Khai phá dữ liệu

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

Trước tiên, chúng em đọc file “Quote.csv” đã  lưu từ ý 3.1.2d bằng lệnh [5]

```
import pandas as pd
df_data = pd.read_csv('Quote.csv')
```

**- Thêm vào Trường STT và điền tự động dữ liệu của trường này**

Chúng em thực hiện thêm Trường STT và điền tự động dữ liệu của trường này bằng lệnh [6].

```
df_data['STT'] = list(range(1, 101))
df_data = df_data[['STT', 'Tên tác giả', 'Đường link của tác giả', 'Ngày tháng năm sinh', 'Câu nói nổi tiếng của tác giả']]
```

**- Đề xuất cách điền một số giá trị của dữ liệu Trường ngày sinh chưa có**

Chúng em kiểm tra trường ngày sinh của các tác giả bằng đoạn code.

```
df_data['Ngày tháng năm sinh'].isnull().values.any()
```

→ Sau khi chạy code này, kết quả trả lại được là 'False'. Điều này cho thấy trường ngày sinh không bị thiếu dòng nào nên chúng em thực hiện các yêu cầu tiếp theo của đồ án.

**- Thêm vào Trường Tuoi (Tuổi) và đề xuất cách điền tuổi của các tác giả**

Chúng em nhận thấy trong trang web cào không có dữ liệu về ngày mất của các tác giả nên chúng em sử dụng phương pháp tra cứu trên Internet về tuổi của người đã mất và tính tuổi của người còn sống theo năm hiện tại bằng cách "lấy năm hiện tại – năm sinh của tác giả".

Chúng em lấy dữ liệu ngày tháng năm sinh có trong bảng đọc từ file 'Quote.csv' và chuyển nó thành 1 list để xử lí [6].

```
author_born = df_data['Ngày tháng năm sinh'] # lấy dữ liệu cột ngày tháng năm sinh
list_born = author_born.tolist() # chuyển dữ liệu vừa lấy thành 1 list để dễ xử lí
```

Sau đây là bảng tuổi của các tác giả được thu thập từ Internet, các tác giả còn sống cần tính tuổi theo công thức "lấy năm hiện tại – năm sinh của tác giả" thì chúng em lưu dưới chuỗi 'age'.

```
list_age = [76, 'age', 76, 41, 36, 76, 82, 84, 78, 'age',
```

```
        36, 'age', 76, 36, 87, 49, 87, 56, 74, 86,
        69, 78, 87, 'age', 54, 87, 76, 'age', 76, 36,
        87, 'age', 36, 87, 'age', 78, 'age', 76, 87, 61,
        'age', 64, 36, 36, 76, 36, 36, 39, 'age', 63,
        41, 78, 36, 76, 'age', 71, 'age', 62, 88, 54,
        74, 'age', 'age', 64, 81, 'age', 62, 78, 75, 87,
        83, 74, 67, 87, 91, 71, 40, 66, 77, 75,
        76, 41, 'age', 41, 41, 64, 64, 75, 75, 64,
        'age', 28, 77, 68, 'age', 90, 89, 75, 87, 'age']
list_find_age = []
# tìm vị trí tuổi của tác giả còn sống được lưu dưới biến 'age'
for i in range(len(list_age)):
    if list_age[i] == 'age':
        list_find_age.append(i)
```

Sau đó tách năm sinh của tác giả để tìm tuổi của tác giả còn sống, chúng em sử dụng thư viện datetime.

```
from datetime import datetime
# tạo 1 list chứa các năm sinh của các tác giả cần tìm
borns_format_year = []
for i in range(len(list_born)):
    born_format = datetime.strptime(list_born[i], '%B %d, %Y')
    year = born_format.year
    borns_format_year.append(year)
```

Lấy năm của thời điểm hiện tại để tìm tuổi tác giả còn sống.

```
import time
x = time.localtime()
```

Đoạn code tìm tuổi của tất cả tác giả còn sống.

```
age_list_author = []
```

```
for i in range(len(borns_format_year)):
    age = x[0] - borns_format_year[i]
    age_list_author.append(age)
age_list_author
# lấy tuổi của các tác giả còn sống
age_found_all = []
for i in list_find_age:
    age_found = age_list_author[i]
    age_found_all.append(age_found)
```

Thống nhất, gộp lại tuổi đã tìm được của tất cả các tác giả.

```
list_find_age
age_found_all
k = 0
for i in list_find_age:
    if list_age[i]:
        list_age[i] = age_found_all[k]
        k+=1
        if k >99:
            break
```

Sau khi tìm được tuổi của tất cả các tác giả, chúng em thực hiện thêm trường Tuổi và điền dữ liệu vào trường này bằng code sau

```
df_data['Tuổi'] = list_age

df_data = df_data[['STT', 'Tên tác giả', 'Đường link của tác giả',

                   'Ngày tháng năm sinh', 'Câu nói nổi tiếng của tác giả', 'Tuổi']]
```

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

Trước tiên, chúng em thực hiện các bước khám phá cơ bản về bảng dữ liệu [6]

```
# Lấy số dòng với số cột có trong bộ dữ liệu
a = df_data.shape
# Xem thông tin về Index, kiểu dữ liệu và dung lượng của dữ liệu
df_data.info()
# Tổng kết thông tin thống kê cho các cột có kiểu dữ liệu là số
df_data.describe()
```

**- Thống kê về tác giả và câu nói nổi tiếng có trong bộ dữ liệu**

+ Thống kê số lượng tác giả và in ra:

```
# lấy dữ liệu cột tên tác giả
author_name = df_data['Tên tác giả']
# chuyển dữ liệu vừa lấy thành 1 list để xử lí
list_name = author_name.tolist()
from collections import Counter
author_counts = Counter(list_name)
# author_counts
print(f"Có {len(author_counts)} tác giả trong bộ dữ liệu:")
new_author = dict(author_counts)
# liệt kê tác giả
# count = 0
for x in new_author:
    print(f"\t{x}")
```

- Kết quả: nhận được có 50 tác giả

+ Các tác giả và câu quote tương ứng của họ trong bộ dữ liệu [6]

```
# các tác giả và câu quote tương ứng của họ
k = 0
for i in range(len(list_name)):
    print(f"{i+1}. {list_name[i]}:  {list_quote[k]} ")
    k+=1
```

+ Thống kê số lượng câu nói nổi tiếng của mỗi tác giả trong bộ dữ liệu: [3]

```
# chuyển Counter thành 1 dict có tên là new_dict
new_dict = dict(author_counts)
# liệt kê tác giả có bao nhiêu câu nói nổi tiếng trong bộ dữ liệu
for x in new_dict:
   print(f"{x} có {new_dict.get(x)} câu nói nổi tiếng trong bộ dữ liệu")
```

**- Thống kê về năm sinh và độ tuổi của các tác giả**

+ Thống kê về năm sinh theo code: [3]

```
list_name
# chuyển thành dict
age_level = dict(zip(list_name, borns_format_year))
age_level
```

→ code này sẽ hiển thị 1 dict gồm tên và tuổi của các người nổi tiếng dưới dạng:

{“Ten_tac_gia_1”: Tuổi,

“Ten_tac_gia_2”: Tuổi, …} với Ten_tac_gia_1, Ten_tac_gia_2 … là dưới kiểu dữ liệu string, Tuổi dưới dạng kiểu dữ liệu int

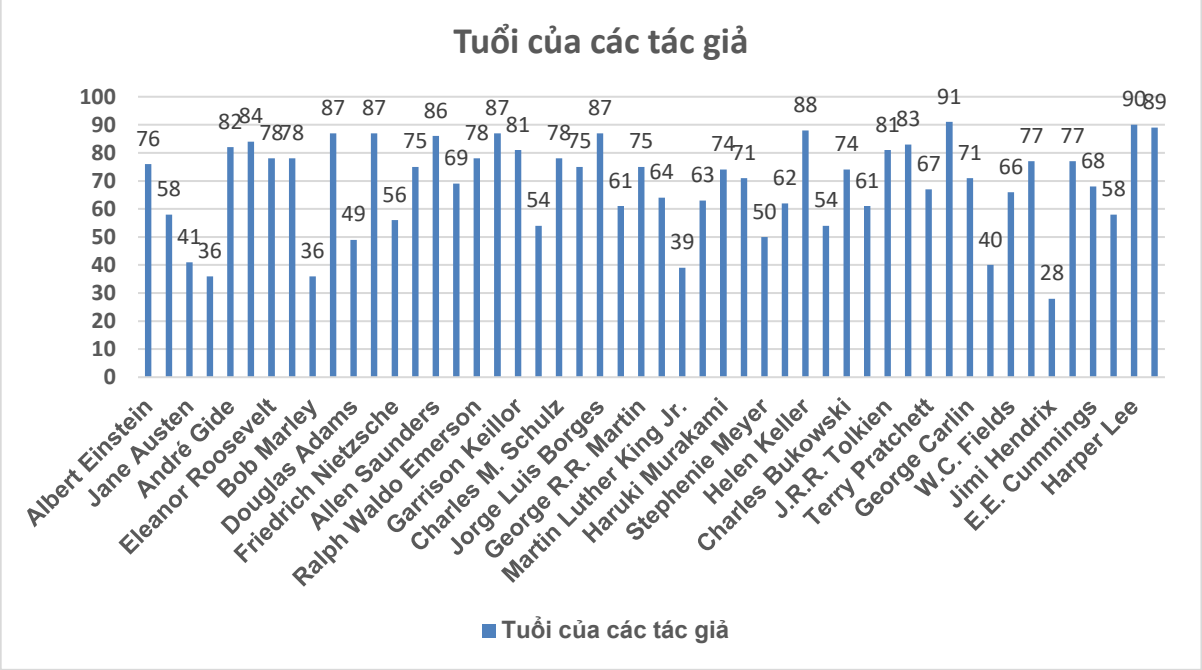
+ Biểu đồ về tuổi của các tác giả, chúng em thực hiện vẽ bằng thư viện ‘matplotlib.pyplot’ bằng code sau: [4]

```
import matplotlib.pyplot as plt


# Tạo list các tác giả và tuổi tương ứng
authors = list(age_level.keys())
ages = list(age_level.values())


# Vẽ biểu đồ
plt.figure(figsize=(20,6))
plt.bar(authors, ages, color='green')
plt.title('Biểu đồ thể hiện tuổi của các tác giả')
plt.xlabel('Tác giả')
```

```
plt.ylabel('Độ tuổi')
# Thêm chỉ số tuổi trên đầu thanh bar
for i, age in enumerate(ages):
    plt.text(i, age+1, str(age), ha='center', va='bottom')
plt.xticks(rotation=90)
plt.show()
```




Biểu đồ 301 Tuổi của các tác giả

+ Tìm tác giả có tuổi lớn nhất bằng code: [7]

```
max_age = max(ages)
def loc(ages):
    for i in range(len(ages)):
        if ages[i] == max_age:
            return i
a= loc(ages)
print(f'Tác giả lớn tuổi nhất là {authors[a]}: {max_age}')
```

+ Tìm tác giả có tuổi nhỏ nhất bằng code:

```
min_age = min(ages)
def loc(ages):
    for i in range(len(ages)):
        if ages[i] == min_age:
            return i
b= loc(ages)
print(f'Tác giả nhỏ tuổi nhất là {authors[b]}: {min_age}')
```

**- Thống kê về các câu nói nổi tiếng như: câu dài nhất, ngắn nhất, số từ, …** [7]

+ Câu dài nhất

```
quote_length = []
for i in range(len(list_quote)):
    quote_length.append(len(list_quote[i]))
a = max(quote_length)
for i in range(len(list_quote)):
    if len(list_quote[i]) == max(quote_length):
        # print(list_quote_new[i])
        print(list_quote[i])
        break
```

+ Câu ngắn nhất

```
quote_length
min_sen = min(quote_length)
for i in range(len(list_quote)):
    if len(list_quote[i]) == min_sen:
        print(list_quote[i])
        break
```

+ Chiều dài của các câu quote:

*Độ dài của các câu quote:

```
list_quote
lengh_quote = []
for quote in list_quote:
    lengh_quote.append(len(quote))
# so_tu_quote
dict_lengh_quote = dict(zip(list_quote, lengh_quote))
dict_lengh_quote
```

*Số từ trong các câu quote:

Top 10 câu quote dài nhất, chúng em sử dụng biểu đồ cột để thể hiện được 10 câu quote dài nhất: [4] [8]

```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv('Quote.csv')
# Tính độ dài của mỗi câu quote
df['QuoteLen'] = df['Câu nói nổi tiếng của tác giả'].apply(lambda x: len(x))
# Sắp xếp các câu quote theo thứ tự giảm dần độ dài
sorted_quotes = df.sort_values(by='QuoteLen', ascending=False)
# Lấy 10 câu quote đầu tiên
top_10 = sorted_quotes.head(10)
# Tạo figure object với figsize
fig, ax = plt.subplots(figsize=(20, 8))
# Vẽ biểu đồ Top 10 câu quote dài nhất
plt.bar(top_10['Tên tác giả'], top_10['QuoteLen'])
# hiển thị chỉ số độ dài của câu quote
for i, v in enumerate(top_10['QuoteLen']):
    plt.text(i, v, str(v), ha='center', va='bottom')
# Thêm tiêu đề cho biểu đồ
```

```
plt.title('Top 10 câu quote dài nhất của tác giả')
# Thêm tên cho trục x và trục y
plt.xlabel('Tên tác giả')
plt.ylabel('Độ dài của câu quote')
# Hiển thị biểu đồ
plt.show()
```


Biểu đồ 3.2 Top 10 câu quote dài nhất của tác giả

Top 10 câu quote ngắn nhất, chúng em sử dụng biểu đồ cột để thể hiện được 10 câu quote ngắn nhất:

```
import pandas as pd
import matplotlib.pyplot as plt
# Đọc dữ liệu từ file csv
df = pd.read_csv('Quote.csv')
# Tính độ dài của mỗi câu quote
df['QuoteLen'] = df['Câu nói nổi tiếng của tác giả'].apply(lambda x: len(x))
```

```
# Sắp xếp các câu quote theo thứ tự giảm dần độ dài
sorted_quotes = df.sort_values(by='QuoteLen', ascending=True)
# Lấy 10 câu quote đầu tiên
top_10 = sorted_quotes.head(10)
# Tạo figure object với figsize
fig, ax = plt.subplots(figsize=(20, 8))
# Vẽ biểu đồ Top 10 câu quote dài nhất
plt.bar(top_10['Tên tác giả'], top_10['QuoteLen'])
# hiển thị chỉ số độ dài của câu quote
for i, v in enumerate(top_10['QuoteLen']):
    plt.text(i, v, str(v), ha='center', va='bottom')
# Thêm tiêu đề cho biểu đồ
plt.title('Top 10 câu quote ngắn nhất thuộc về các tác giả')
# Thêm tên cho trục x và trục y
plt.xlabel('Tên tác giả')
plt.ylabel('Độ dài của câu quote')
# Hiển thị biểu đồ
plt.show()
```


Biểu đồ 3.3 Top 10 câu quote ngắn nhất thuộc về các tác giả

**- Thống kê về các từ được sử dụng trong các câu nói [3] [4]**

+ Đếm số lượng từ xuất hiện trong các câu nói:

```
import collections
list_quote
# Tạo danh sách rỗng để lưu các từ
word_list = []
# Duyệt qua các câu nói, tách các câu nói thành các từ và thêm các từ đó vào danh sách
for quote in list_quote:
    words = quote.split()
    word_list.extend(words)
# Sử dụng collections.Counter() để đếm số lượng các từ trong danh sách
word_counts = collections.Counter(word_list)
print(word_counts)
```

+ Từ được xuất hiện nhiều nhất được sử dụng trong các câu quote của các tác giả:

```
word_counts.most_common(1)
```

Kết quả: “you” với 76 lần xuất hiện trong tất cả các câu quote.

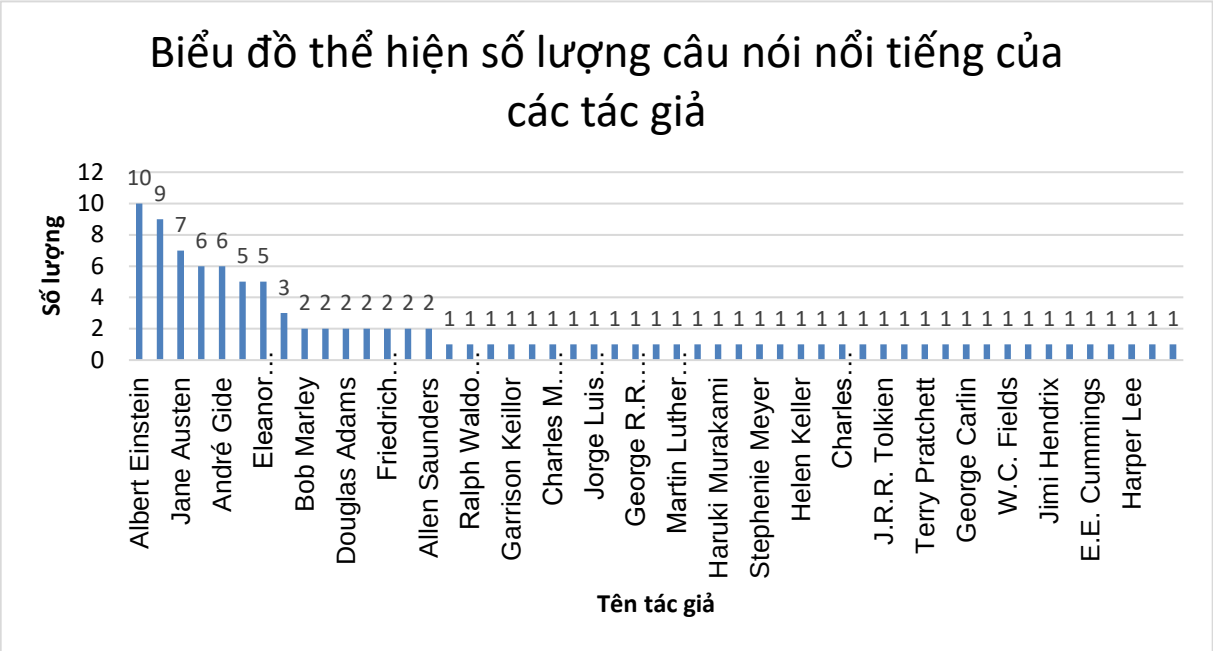
+ Top 10 các từ được sử dụng nhiều nhất:

```
top_10_word = []
top_10_amount = []
top_words = word_counts.most_common(10)
# In ra 10 các từ được sử dụng nhiều nhất
print("Các từ được sử dụng nhiều nhất:")
for word, count in top_words:
    print(f'\t{word}: {count}')
    top_10_word.append(word)
    top_10_amount.append(count)
```

**- Phân tích, trực quan mối quan hệ giữa các tác giả và câu nói nổi tiếng**

+ Số lượng câu nói nổi tiếng của các tác giả:

```
import pandas as pd
df = pd.read_csv('Quote.csv')
amount_quotes = df['Tên tác giả'].value_counts()
# plt.bar(freq_table.index, freq_table.values)
plt.figure(figsize=(20, 8))
plt.bar(amount_quotes.index, amount_quotes.values)
plt.xticks(rotation=90)
for i, amount_quote in enumerate(amount_quotes):
    plt.text(i, amount_quote, str(amount_quote), ha='center', va='bottom')
plt.title('Số lượng câu nói nổi tiếng của các tác giả', fontweight = 'bold')
plt.xlabel('Tên tác giả', fontweight = 'bold', color = 'red')
plt.ylabel('Số lượng')
plt.show()
amount_quote_author = amount_quotes.sort_values(ascending=True)
```




Biểu đồ 304 Số lượng câu nói nổi tiếng của các tác giả

+ Đồ thị biểu diễn 10 từ được sử dụng nhiều nhất trong các câu quote của các tác giả (biểu thị sự ưa thích sử dụng từ trong các câu quote của các tác giả)

```
plt.bar(x = top_10_word, height = top_10_amount)
# Hiển thị chỉ số lượng các từ
for i, v in enumerate(top_10_amount):
    plt.text(i, v, str(v), ha='center', va='bottom')
plt.title('Top 10 các từ được sử dụng nhiều nhất trong các câu quote của các tác giả')
plt.xlabel('Các từ được sử dụng nhiều nhất')
plt.ylabel('Số lượng')
```


Biểu đồ 305 Top 10 từ được sử dụng nhiều nhất trong các câu quote của các tác giả

**- Phân tích, trực quan mối quan hệ giữa các tác giả với nhau: [6] [8]**

```
# Tạo một list copy lưu trữ độ tuổi của các tác giả
authors_age = age_level.copy()
# Tạo các biến đếm cho các khoảng độ tuổi
age_20s = []
age_30s = []
age_40s = []
age_50s = []
age_60s = []
age_70s = []
age_80s = []
age_90s = []
# Duyệt qua từng tác giả và kiểm tra độ tuổi của họ nằm trong khoảng nào
for author, age in authors_age.items():
    if age >= 20 and age < 30:
```

```
        age_20s.append(author)
    elif age >= 30 and age < 40:
        age_30s.append(author)
    elif age >= 40 and age < 50:
        age_40s.append(author)
    elif age >= 50 and age < 60:
        age_50s.append(author)
    elif age >= 60 and age < 70:
        age_60s.append(author)
    elif age >= 70 and age < 80:
        age_70s.append(author)
    elif age >= 80 and age < 90:
        age_80s.append(author)
    elif age >= 90 and age < 100:
        age_90s.append(author)
# In ra các tác giả theo khoảng độ tuổi
# chúng em sử dụng dòng lệnh print(f'\t\033[94m{author}\033[0m') để hiện thị màu xanh cho tên các tác giả, tạo sự khác biệt về màu để có thể nhìn một cách trực quan và dễ đọc hơn
print('Các tác giả trong khoảng độ tuổi 20-29:')
for author in age_20s:
    print(f'\t\033[94m{author}\033[0m')
print('Các tác giả trong khoảng độ tuổi 30-39:')
for author in age_30s:
    print(f'\t\033[94m{author}\033[0m')
print('Các tác giả trong khoảng độ tuổi 40-49:')
for author in age_40s:
    print(f'\t\033[94m{author}\033[0m')
print('Các tác giả trong khoảng độ tuổi 50-59:')
```

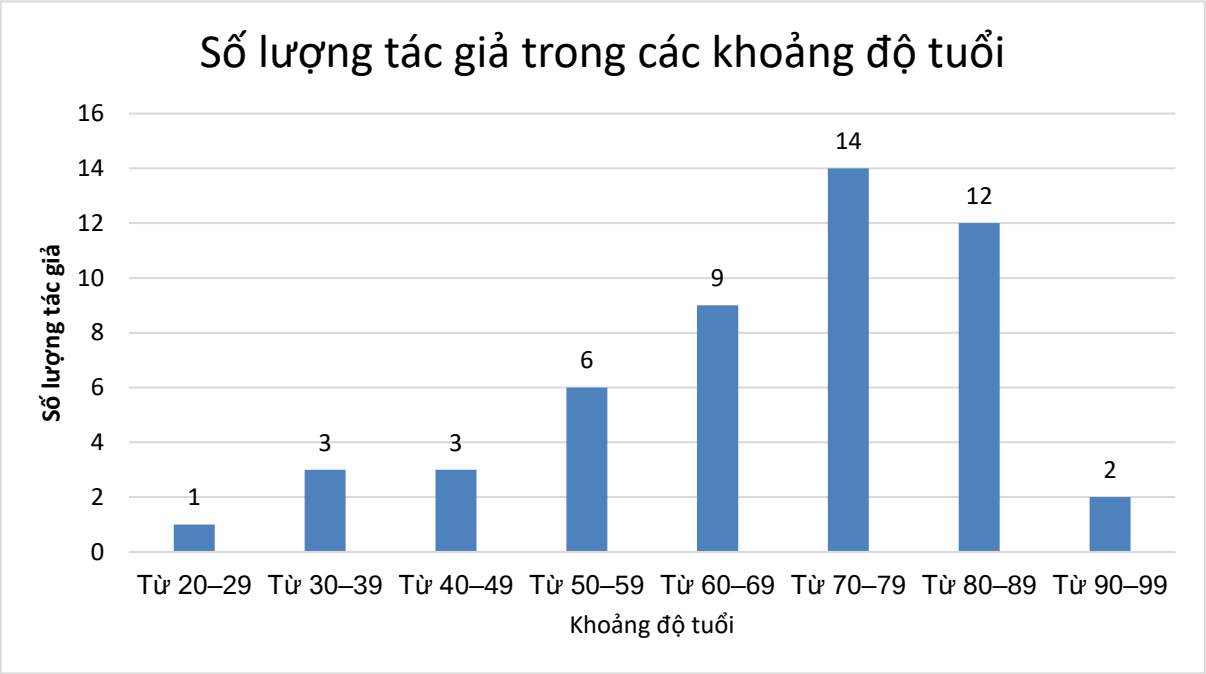
```
for author in age_50s:
    print(f'\t\033[94m{author}\033[0m')
print('Các tác giả trong khoảng độ tuổi 60-69:')
for author in age_60s:
    print(f'\t\033[94m{author}\033[0m')
print('Các tác giả trong khoảng độ tuổi 70-79:')
for author in age_70s:
    print(f'\t\033[94m{author}\033[0m')
print('Các tác giả trong khoảng độ tuổi 80-89:')
for author in age_80s:
    print(f'\t\033[94m{author}\033[0m')
print('Các tác giả trong khoảng độ tuổi 90-99:')
for author in age_90s:
    print(f'\t\033[94m{author}\033[0m')
```

+ Vẽ biểu đồ thể hiện số lượng tác giả theo độ tuổi:

```
import matplotlib.pyplot as plt
age_20s = 0
age_30s = 0
age_40s = 0
age_50s = 0
age_60s = 0
age_70s = 0
age_80s = 0
age_90s = 0
# Duyệt qua từng tác giả và kiểm tra độ tuổi của họ nằm trong khoảng nào
for age in authors_age.values():
    if age >= 20 and age < 30:
        age_20s += 1
```

```
        elif age >= 30 and age < 40:
            age_30s += 1
        elif age >= 40 and age < 50:
            age_40s += 1
        elif age >= 50 and age < 60:
            age_50s += 1
        elif age >= 60 and age < 70:
            age_60s += 1
        elif age >= 70 and age < 80:
            age_70s += 1
        elif age >= 80 and age < 90:
            age_80s += 1
        elif age >= 90 and age < 100:
            age_90s += 1
# Tạo list các khoảng độ tuổi và list số lượng tác giả tương ứng
age_ranges = ['Từ 20-29', 'Từ 30-39', 'Từ 40-49', 'Từ 50-59',
              'Từ 60-69', 'Từ 70-79', 'Từ 80-89', 'Từ 90-99']
author_counts = [age_20s, age_30s, age_40s,
                 age_50s, age_60s, age_70s, age_80s, age_90s]
fig, ax = plt.subplots(figsize=(10, 6))
# Vẽ đồ thị số lượng tác giả trong các khoảng độ tuổi
ax.bar(age_ranges, author_counts)
ax.set_xlabel('Khoảng độ tuổi')
ax.set_ylabel('Số lượng tác giả')
ax.set_title('Số lượng tác giả trong các khoảng độ tuổi')
# Hiển thị số lượng tác giả của mỗi khoảng độ tuổi trên thanh bar chart
for i, v in enumerate(author_counts):
    plt.text(i, v, str(v), ha='center', va='bottom')
```

```
# Hiển thị đồ thị
plt.show()
```




Biểu đồ 3.6 Số lượng tác giả trong các khoảng độ tuổi

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

Cách tiếp cận: trước tiên chúng em lưu 1 file quote mới "Quote_1.csv" có đầy đủ trường Tuổi của các các giả để thực hiện tiếp theo của đề bài. [9]

```
df_data.to_csv('Quote_1.csv')
```

Trong file quote mới "Quote_1.csv" em nhận thấy trong file Quote_1.csv của tập dữ liệu thu thập được có những trường "Tên tác giả", "Đường link của tác giả", "Ngày tháng năm sinh" và "Tuổi" có các giá trị trùng lặp với nhau của các tác giả. Riêng trường "Câu nói nổi tiếng của tác giả" thì không trùng lặp.

Vì vậy nên chúng em viết code Python thực hiện gộp các trường trùng lặp lại với nhau thành 1, trường không trùng lặp thì em gộp chung vào 1 ô dữ liệu. Làm vậy để trích lọc được những dữ liệu cần thiết và bảng dữ liệu trở nên trực quan hơn. [9]

```
import pandas as pd
df = pd.read_csv('Quote_11.csv')
# Nhóm các dòng theo tên tác giả và sử dụng hàm agg để lấy giá trị đầu tiên của các cột trùng lặp
df_merged = df.groupby('Tên tác giả', as_index=False).agg({'Đường link của tác giả': 'first',
                                                           'Ngày tháng năm sinh': 'first',
                                                           'Câu nói nổi tiếng của tác giả': '\n'.join,
                                                           'Tuổi': 'first'})
# code này để hiển thị bảng đã thực hiện trích xuất đặc trưng
df_merged
```

Kết quả chúng em nhận được 1 bảng đã được trích xuất sau đó thực hiện lưu nó bằng 1 file mới "new.csv"

```
df_merged.to_csv('new.csv')
```

### 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?**

Để có thể dự đoán tên của người nổi tiếng theo câu nói nổi tiếng, chúng ta cần phân tích câu nói nổi tiếng để hiểu các vấn đề và chủ đề mà nó đề cập. Xem xét các từ khóa, ý nghĩa và thông điệp chính trong câu nói được sử dụng trong các câu nói của mỗi người. Sau đó, dựa trên thông tin và phân tích, suy luận và đưa ra dự đoán về tác giả có thể đã nói câu nói nổi tiếng đó. Xem xét các tác giả có những quan điểm, ý tưởng hay tư tưởng tương tự.

Ví dụ: Albert Einstein – ngài có phong cách nói chủ yếu hướng đến về một cuộc sống tốt hơn, sự tạo hình thế giới, quan niệm sống, giá trị cá nhân, sự hiểu biết, sự sáng tạo, tưởng tượng, sự nhận thức, sự cân bằng, sự học hỏi, thử thách, cái đẹp của vật lí trong cuộc sống, khuyên nhủ mọi người hãy không người cố gắng… Từ đó ta có thể dự đoán được tên của tác giả từ câu nói nổi tiếng của họ.

**- 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.**

Để tính độ tương đồng phong cách nói giữa các tác giả, ta có thể sử dụng phương pháp Vector Space Model (mô hình không gian vector) và độ đo Cosine Similarity. Dưới đây là một phân tích cụ thể về cách áp dụng phương pháp này [10]:

- *Tiền xử lý dữ liệu:*

+ Thu thập và tạo bộ dữ liệu chứa các câu nói của các tác giả

+ Loại bỏ các ký tự đặc biệt, số và dấu câu không cần thiết từ các câu nói.

+ Chuyển đổi các câu nói về dạng viết thường (lowercase) để đảm bảo tính nhất quán trong việc so sánh.

- *Trích xuất đặc trưng:*

+ Tạo một từ điển (vocabulary) từ tất cả các từ xuất hiện trong các câu nói của tác giả.

+ Đếm tần số xuất hiện của từng từ trong từ điển trong mỗi câu nói của từng tác giả.

+ Tạo vector đặc trưng cho mỗi tác giả bằng cách sắp xếp các tần số từng từ vào một vector.

- *Tính toán độ tương đồng:*

+ Sử dụng độ đo Cosine Similarity để tính toán độ tương đồng giữa các vector đặc trưng của tác giả.

+ Độ đo Cosine Similarity tính toán cosin của góc giữa hai vector trong không gian vector.

+ Kết quả nằm trong khoảng từ -1 đến 1, với 1 là độ tương đồng tuyệt đối và -1 là độ tương đồng hoàn toàn đối ngược.

- *Tìm các tác giả có phong cách nói tương đồng nhau nhất:*

+ Dựa vào ma trận độ tương đồng, xác định ngưỡng (threshold) để quyết định tác giả có phong cách nói tương đồng nhau.

+ Có thể sử dụng ngưỡng dựa trên một giá trị cố định hoặc dựa trên phân phối tự nhiên của các giá trị độ tương đồng.

+ Tìm các cặp tác giả có độ tương đồng vượt qua ngưỡng và hiển thị danh sách các tác giả có phong cách nói tương đồng nhau nhất.

- *Tìm các tác giả có phong cách nói tương đồng nhau nhất:*

+ Dựa vào ma trận độ tương đồng (similarity matrix), xác định ngưỡng (threshold) để quyết định tác giả có phong cách nói tương đồng nhau.

+ Có thể sử dụng ngưỡng dựa trên một giá trị cố định hoặc dựa trên phân phối tự nhiên của các giá trị độ tương đồng. Ví dụ, bạn có thể đặt ngưỡng là 0.8, nghĩa là nếu độ tương đồng giữa hai tác giả vượt qua 0.8, thì được coi là phong cách nói tương đồng.

+ Duyệt qua ma trận độ tương đồng và tìm các cặp tác giả có độ tương đồng vượt qua ngưỡng đã định.

+ Hiển thị danh sách các tác giả có phong cách nói tương đồng nhau nhất.

# LÀM VIỆC NHÓM

- Nhóm chúng em được hình thành với mục tiêu chung là cùng nhau trao đổi, chia sẻ, thực hiện và hoàn thành bài đồ án cuối kì môn Nhập môn khoa học dữ liệu một cách chỉnh chu và tốt nhất.

- Công việc được phân chia bình đẳng, cụ thể và phù hợp với năng lực mỗi người cụ thể như sau:

+ Bài 1, bài 2 chúng em cùng tham gia nghiên cứu, thiết kế câu hỏi, khảo sát và tiến hành thu thập dữ liệu sau đó bài 1 dữ liệu được giao cho bạn Nguyễn Khắc Luật xử lí và hoàn thành, bài 2 dữ được giao cho bạn Hoàng Thanh Tú nhập, xử lí và đưa ra các mô tả thống kê để hoàn thành. Bạn Luật tham gia đánh giá và đóng góp ý kiến.

+ Bài 3, từ đầu đến hết câu 3.1, ý 2 câu 3.2.4 được giao cho bạn Hoàng Thanh Tú cào, xử lí và suy luận dữ liệu, phần còn lại được giao cho bạn Nguyễn Khắc Luật khám phá, xử lí và suy luận dữ liệu.

+ Bài báo cáo nhóm cùng làm, cụ thể lời cảm ơn, làm việc nhóm và nội dung, trực quan về nhiệm vụ đã được phân công ở 3 bài trên được bạn Hoàng Thanh Tú hoàn thiện, các mục lục được và nội dung, trực quan về nhiệm vụ đã được phân công ở 3 bài trên được giao cho bạn Nguyễn Khắc Luật hoàn thành.

- Nhóm có 9 buổi họp trực tiếp và 2 buổi họp online trong khoảng 8 tuần, tổng số giờ họp là 52 giờ.

+ Buổi 1 họp để thống nhất nguyên tắc và nghiên cứu, bàn bạc, lên kế hoạch, phân công nhiệm vụ và thiết kế và thực hiện bài 1 và bài 2, tổng thời lượng trong 4 giờ.

+ Buổi 2 họp để thực hiện thu thập dữ liệu bài 1 và bài 2, sau đó giao dữ liệu theo phân công, tổng thời lượng trong 5 giờ.

+ Buổi 3 họp online thực hiện việc phân tích tích toán và thống nhất phương pháp xử lí dữ liệu và tiến hành xử lí dữ liệu của bài 1 và bài 2 trong 4 giờ.

+ Buổi 4 họp để kiểm tra lại tiến độ thực hiện bài 1, bài 2 và kết quả nhiệm vụ để có sự trao đổi, hỗ trợ, điều chỉnh phù hợp, tổng thời lượng là 5 giờ.

+ Buổi 5 họp online kiểm tra lại kết quả của bài 1, bài 2 và thống nhất kết quả đó. Tiến hành trao đổi và thực hiện Bài 3 câu 3.1 và câu 3.2.1, tổng thời lượng là 5 giờ.

+ Buổi 6 và 7 họp để hoàn thành phần còn lại của câu 3 và thực hiện điều chỉnh phù hợp trong 5 giờ/buổi.

+ Buổi 8 cùng nhau trao đổi và viết báo cáo các phần như lời mở đầu, tóm tắc, chương 1, chương 2 trong 5 tiếng.

+ Buổi 9 và 10 cùng nhau hoàn thành chương 3 và điều chỉnh bổ sung các phần còn lại khoàng 4 giờ/buổi.

+ Buổi 11 tổng hợp lại kết quả làm được kiểm tra và tiến hành trình bày với nhau những gì mình đã làm được, tổng thời lượng là 6 giờ.

# TÀI LIỆU THAM KHẢO


[1] Nguyễn Thị Kim Ngọc, Trần Minh Hiển, Kỹ năng đặt câu hỏi, Phòng Kết nối Khoa học với Công chúng Đơn vị Nghiên cứu Lâm sàng Đại học Oxford (OUCRU), 2020.

[2] W. McKinney, Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, O'Reilly Media, 2017.

[3] N. Quyết, Giáo Trình Xác Suất Và Thống Kê Cơ Bản, NXB Kinh Tế TP.HCM, 2015.

[4] C. N. Knaflic, Storytelling With Data – Kể chuyện thông qua dữ liệu, NXB Thế Giới, 2015.

[5] R. Mitchell, Web Scraping with Python: Collecting More Data from the Modern Web, O'Reilly Media, 2018.

[6] Trần Hùng Cường, Trần Thanh Hùng, Giáo trình khai phá dữ liệu, NXB Thống kê , 2017.

[7] F. Nelli, Python Data Analytics: Data Analysis and Science Using Pandas, matplotlib, and the Python Programming Language, Apress, 2015.

[8] D. Y. Chen, Pandas for Everyone: Python Data Analysis (Addison-Wesley Data & Analytics Series), Addison-Wesley Professional, Addison-Wesley Professional, 2022.

[9] Phan Thanh Sơn, Dương Tử Cường, Trích chọn các tham số đặc trưng tiếng nói cho hệ thống tổng hợp tiếng Việt dựa vào mô hình Markov ẩn, Tạp chí Tin học và Điều khiển học, T.29, S.1, 55-65, 2013.

[10] K. Erk, Vector Space Models of Word Meaning and Phrase Meaning: A Survey, O'Reilly Media, 2012.

# PHỤ LỤC

Bảng khảo sát về chất lượng dịch vụ nhà trường và các hoạt động hỗ trợ sinh viên của nhà trường phục vụ cho bài tập 2.

| | |
|---|---|
| INDUSTRIAL UNIVERSITY OF HOCHIMINH CITY | **TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP. HỒ CHÍ MINH** |

# KHẢO SÁT VỀ CHẤT LƯỢNG DỊCH VỤ NHÀ TRƯỜNG VÀ CÁC HOẠT ĐỘNG HỖ TRỢ SINH VIÊN CỦA NHÀ TRƯỜNG

Chúng em là: Hoàng Thanh Tú và Nguyễn Khắc Luật, đến từ DHKHDL17A chuyên ngành Khoa học dữ liệu khoa Công nghệ thông tin, chúng em thu thập thông tin để hoàn thành đề tài cuối kì, các thông tin thu thập từ các bạn được sử dụng để làm đề tài và không có mục đích sử dụng nào khác.

Rất mong nhận được câu trả lời một cách thẳng thắn và khách quan của bạn cho các câu hỏi đặt ra trong phiếu.

**Họ vtên:**.....

**Mã số sinh viên:** .....

Bạn vui lòng đánh dấu (X) vào ô thể hiện đúng quan điểm của bạn nhất vào các ô theo các mức độ đánh giá như sau:

*(1). Hoàn toàn không đồng ý.* *(2). Không đồng ý* *(3). Trung lập*
*(4). Đồng ý.* *(5). Hoàn toàn đồng ý*

| STT | Các tiêu chí đánh giá | Mức độ đánh giá | | | | |
|---|---|---|---|---|---|---|
| | | (1) | (2) | (3) | (4) | (5) |
| **Đánh giá của bạn về chất lượng dịch vụ nhà trường như thế nào?** | | | | | | |
| 1 | Bạn có đồng ý rằng dịch vụ thư viện, tra cứu tài liệu đáp ứng đủ nhu cầu tìm kiếm tài liệu của sinh viên không? | | | | | |
| 2 | Dịch vụ căn tin an toàn, vệ sinh sạch sẽ đáp ứng đủ nhu cầu của sinh viên, bạn có đồng ý điều đó không? | | | | | |
| 3 | Bạn có đồng ý rằng dịch vụ bãi đỗ xe, di chuyển học tập dành cho sinh viên luôn đáp ứng đủ nhu cầu sinh viên không? | | | | | |
| 4 | Công tác đảm bảo an ninh, đảm bảo trật tự trong trường luôn an toàn, bạn có đồng ý điều đó không? | | | | | |
| 5 | Bạn có đồng ý rằng những đãi ngộ, chính sách dành cho sinh viên (chương trình học bổng, chính sách học phí cho sinh viên thuộc diện chính sách của nhà nước, …) luôn được nhà trường quan tâm không? | | | | | |
| **Đánh giá của bạn về các hoạt động công tác hỗ trợ sinh viên như thế nào?** | | | | | | |
| 1 | Cán bộ nhân viên nhiệt tình, vui vẻ, thân thiện với sinh viên, bạn có đồng ý điều đó không? | | | | | |
| 2 | Bạn có đồng ý rằng các hoạt động tư vấn học tập đáp ứng nhu cầu học tập và nghiên cứu của sinh viên không? | | | | | |
| 3 | Các hoạt động tư vấn hướng nghiệp, định hướng việc đều làm đáp ứng nhu cầu sau khi trường của sinh viên, bạn có đồng ý điều đó không? | | | | | |
| 4 | Bạn có đồng ý rằng các khiếu nại (đăng kí học phần, phúc khảo, …) của sinh viên được giải quyết nhanh chóng và thỏa đáng không? | | | | | |
| 5 | Thủ tục hành chính (học phí, học bổng, …) liên quan đến sinh viên được giải quyết kịp thời, bạn có đồng ý điều đó không? | | | | | |

CẢM ƠN BẠN ĐÃ THAM GIA KHẢO SÁT. CHÚC BẠN 1 NGÀY TRÀN ĐẦY NĂNG LƯỢNG VÀ HỌC TẬP TỐT.

# 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 | 12 | |
| | **3.2.3. Trích xuất đặc trưng** | 5 | 5 | |
| | **3.2.4. Suy luận** | 15 | 12 | Chúng em đưa ra nhận định chưa sâu sắc và không có code kèm theo |
| **Báo cáo** | **(**chú ý các chú ý 2,3,4,6 ở trang trước, nếu sai sẽ bị trừ điểm nặng**)** | 10đ | 10đ | |
| **Làm việc nhóm** | Chú ý trả lời đúng 4 yêu cầu trong phần làm việc nhóm | 10đ | 10đ | |
| **Tổng điểm (120)** | | | 117 | |
| **Đôi qua thang điểm 10 (120=10 điểm)** | | | 9.75 | |