TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN GIỮA KỲ MÔN XỬ LÝ DỮ LIỆU LỚN**

**THU THẬP VÀ KHAI PHÁ DỮ LIỆU**

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

*Người thực hiện*: **MAI HOÀNG VIỆT – 51900847**

Lớp **: 19050402**

Khoá  **: 23**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2022**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN GIỮA KỲ MÔN XỬ LÝ DỮ LIỆU LỚN**

**THU THẬP VÀ KHAI PHÁ DỮ LIỆU**

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

Người thực hiện: **MAI HOÀNG VIỆT - 51900847**

Lớp **: 19050402**

Khoá  **: 23**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2022**

LỜI CẢM ƠN

Trước tiên, em xin gửi lời cảm ơn chân thành và lòng biết ơn sâu sắc đối với thầyBùi Thanh Hùng. Thầy là người đã luôn hỗ trợ và hướng dẫn tận tình cho em trong suốt quá trình nghiên cứu và hoàn thành bài báo cáo.

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm đồ án của riêng 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 Tôn Đức Thắng 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 04 tháng 04 năm 2022*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Mai Hoàng Việt*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

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

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

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

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

TÓM TẮT

*Những người nổi tiếng trên thế giới đã đóng góp rất nhiều thành tựu cho sự thay đổi của thế giới, những câu nói của họ có thể khiến cho người khác phải thay đổi tư duy, cách sống của rất nhiều người khác.*

Trong Bài tập này, các bạn sẽ tìm hiểu về cách thu thập và khai phá trên bộ dữ liệu này.

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc100003804)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc100003805)

[TÓM TẮT iv](#_Toc100003806)

[MỤC LỤC 1](#_Toc100003807)

[CHƯƠNG 1 – THU THẬP DỮ LIỆU 3](#_Toc100003808)

[1.1 Bạn hãy viết code cào dữ liệu từ trang web trên, lưu kết quả vào 1 file tương ứng (kq.txt) và mô tả ngắn gọn về cấu trúc của trang web trên 3](#_Toc100003809)

[1.1.1 Viết code cào dữ liệu từ trang web trên và lưu kết quả vào 1 file tương ứng (kq.txt) 3](#_Toc100003810)

[1.1.1.1 Code 3](#_Toc100003811)

[1.1.1.2 Kết quả 3](#_Toc100003812)

[1.1.2 Mô tả ngắn gọn về cấu trúc của trang web trên 4](#_Toc100003813)

[1.2 Với dữ liệu bạn vừa cào về, bạn hãy thực hiện các yêu cầu sau: 4](#_Toc100003814)

[1.2.1 Hãy đọc tất cả các thẻ html (div) với lớp là "quote" và lưu nó trong biến 'result’, hiển thị giá trị biến 'result’ ra màn hình? 4](#_Toc100003815)

[1.2.1.1 Code 4](#_Toc100003816)

[1.2.1.2 Kết quả 5](#_Toc100003817)

[1.2.2 Hãy tìm trong biến 'result’ vừa rồi các dữ liệu có chứa nhãn "small" với class là "author" và in kết quả ra màn hình? 5](#_Toc100003818)

[1.2.2.1 Code 5](#_Toc100003819)

[1.2.2.2 Kết quả 5](#_Toc100003820)

[1.2.3 Hãy 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 6](#_Toc100003821)

[1.2.3.1 Code 6](#_Toc100003822)

[1.2.3.2 Kết quả 7](#_Toc100003823)

[1.2.4 Hãy lưu kết quả ở câu c vào file Quote.csv tương ứng, với mỗi tác giả là 1 dòng dữ liệu. Bạn được yêu cầu thu thập ít nhất 40 câu nói nổi tiếng từ trang web trên một cách tự động theo code của các ý trên? 7](#_Toc100003824)

[1.2.4.1 Code 7](#_Toc100003825)

[1.2.4.2 Kết quả 7](#_Toc100003826)

[CHƯƠNG 2 – KHAI PHÁ DỮ LIỆU 9](#_Toc100003827)

[2.1 Xử lý dữ liệu- Data Imputation 9](#_Toc100003828)

[2.1.1 Giới thiệu 9](#_Toc100003829)

[2.1.2 Cách tiếp cận 9](#_Toc100003830)

[2.1.3 Đánh giá 10](#_Toc100003831)

[2.1.4 Thảo luận 10](#_Toc100003832)

[2.2 Khám phá dữ liệu- Data Exploration 10](#_Toc100003833)

[2.2.1 Giới thiệu 10](#_Toc100003834)

[2.2.2 Cách tiếp cận 10](#_Toc100003835)

[2.2.3 Đánh giá 15](#_Toc100003836)

[2.2.4 Thảo luận 15](#_Toc100003837)

[2.3 Trích xuất đặc trưng- Feature Extraction 15](#_Toc100003838)

[2.3.1 Giới thiệu 15](#_Toc100003839)

[2.3.2 Cách tiếp cận 15](#_Toc100003840)

[2.3.3 Đánh giá 16](#_Toc100003841)

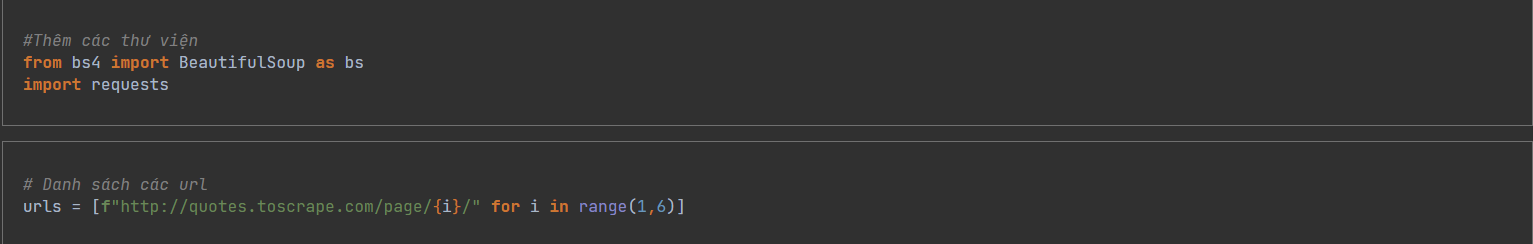
[2.3.4 Thảo luận 16](#_Toc100003842)

[TÀI LIỆU THAM KHẢO 17](#_Toc100003843)

CHƯƠNG 1 – THU THẬP DỮ LIỆU

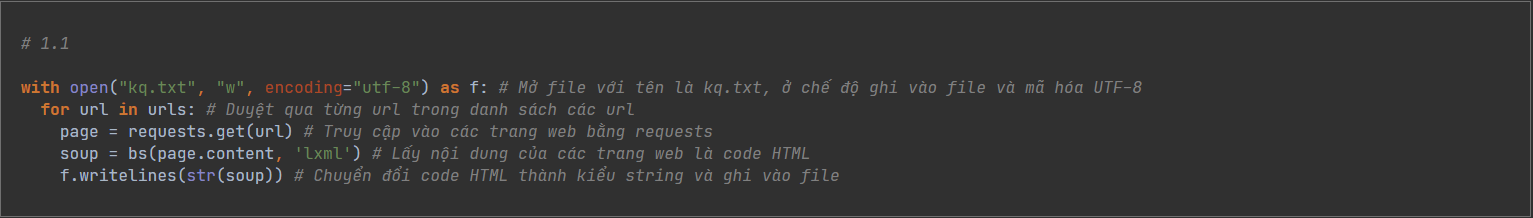
1.1 Bạn hãy viết code cào dữ liệu từ trang web trên, lưu kết quả vào 1 file tương ứng (kq.txt) và mô tả ngắn gọn về cấu trúc của trang web trên

Thêm các thư viện và danh sách các url:

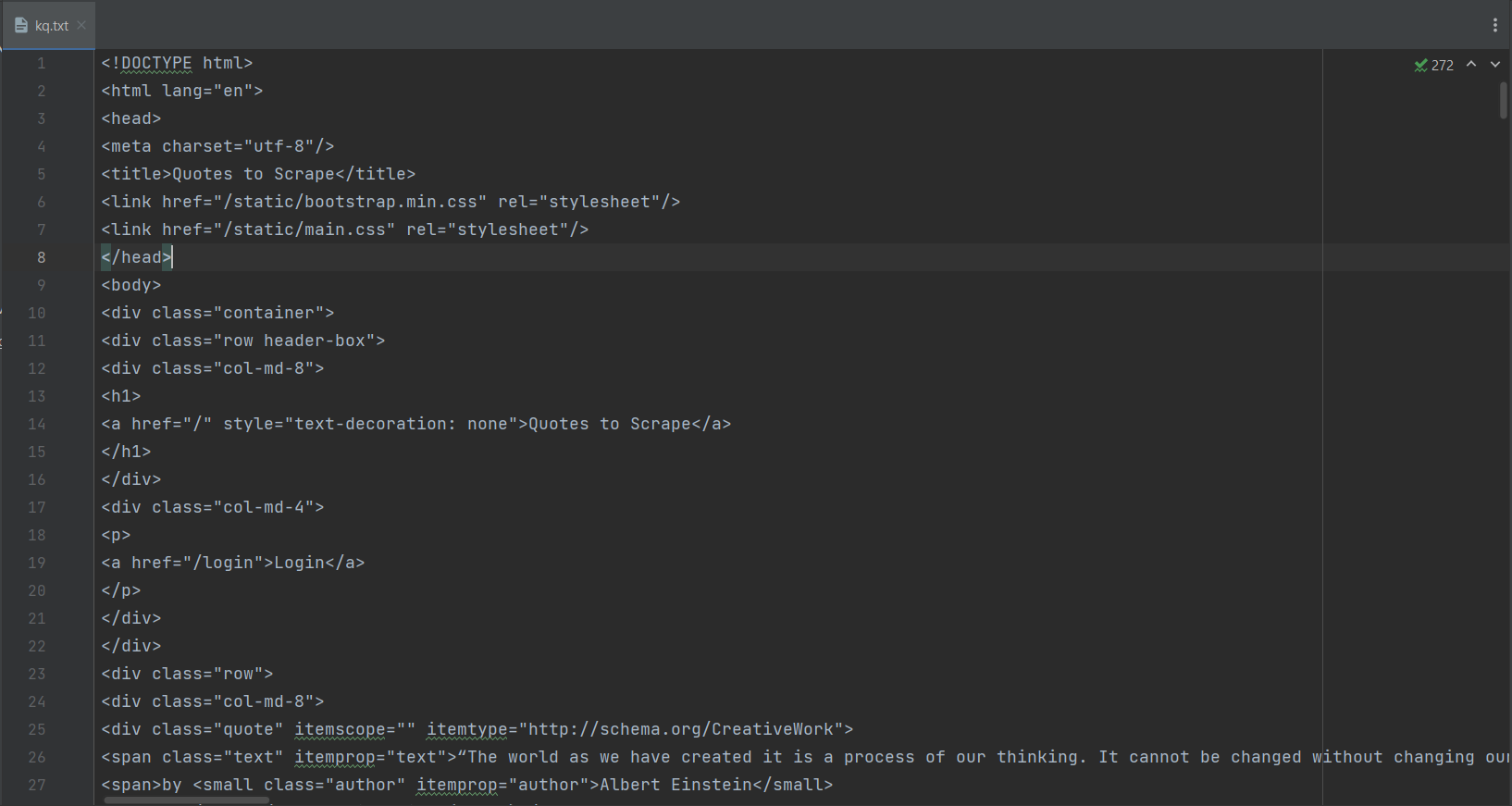


1.1.1 Viết code cào dữ liệu từ trang web trên và lưu kết quả vào 1 file tương ứng (kq.txt)

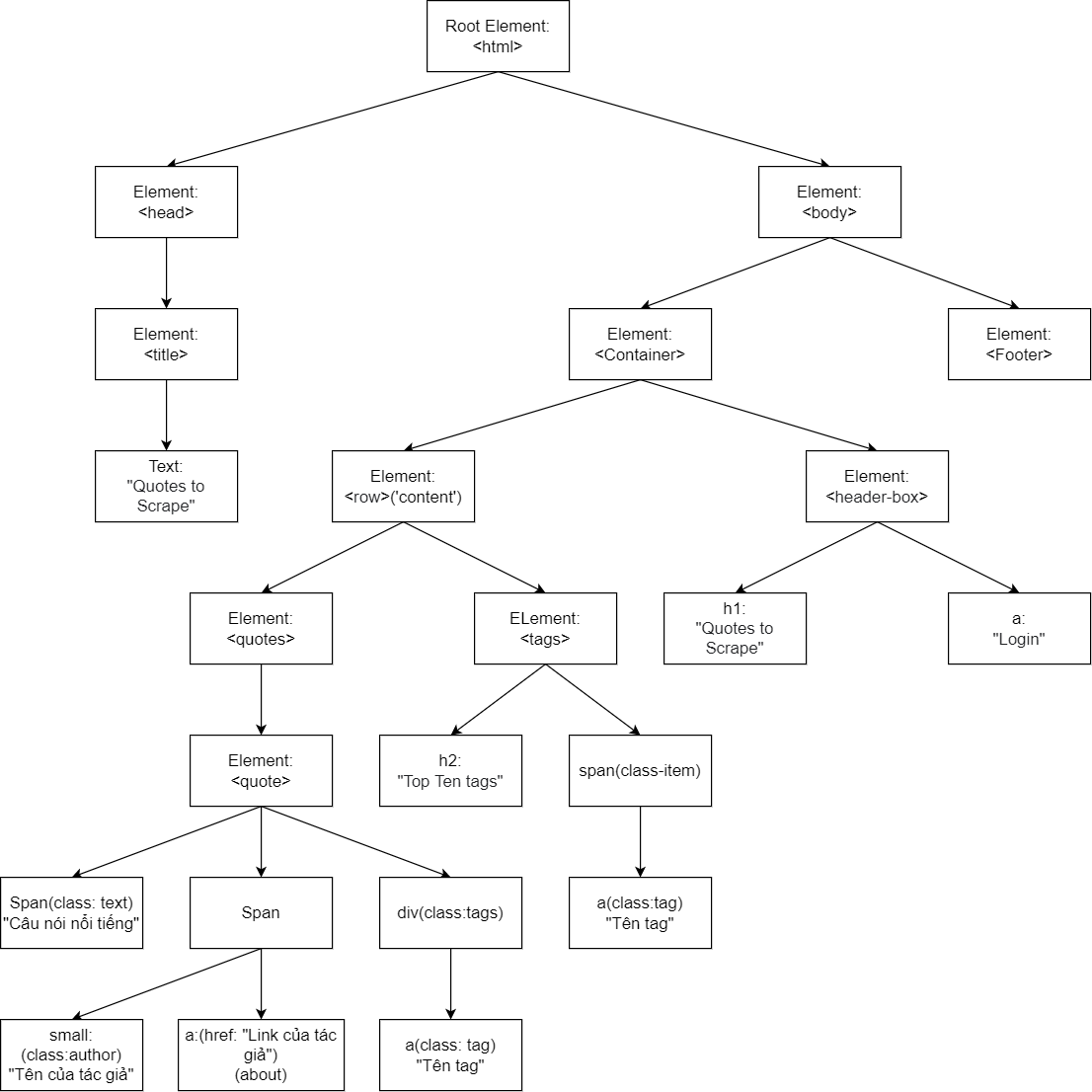
1.1.1.1 Code



1.1.1.2 Kết quả



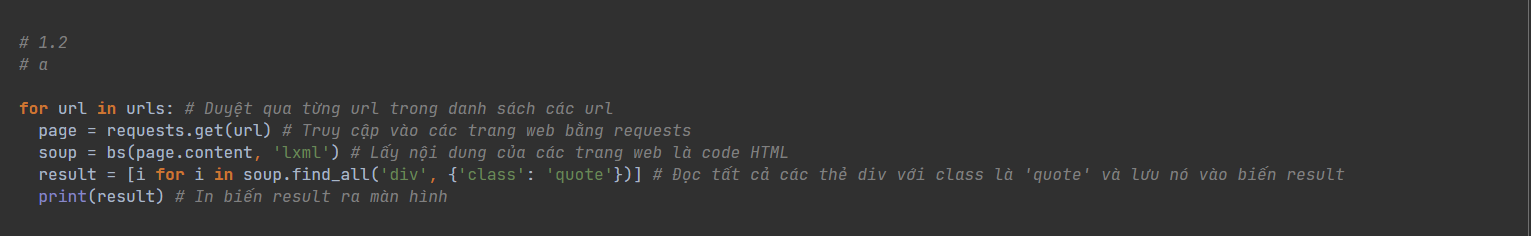
1.1.2 Mô tả ngắn gọn về cấu trúc của trang web trên



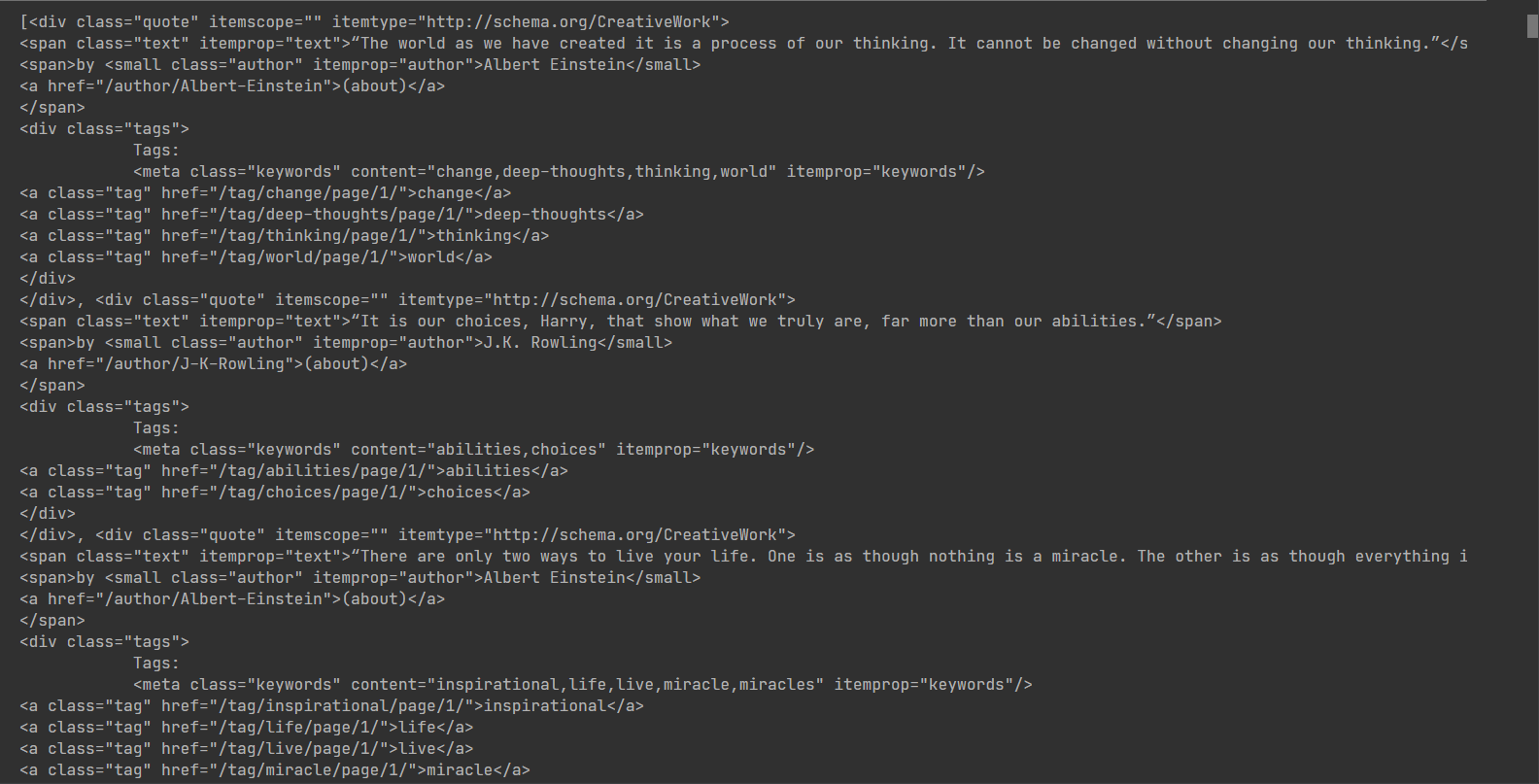
1.2 Với dữ liệu bạn vừa cào về, bạn hãy thực hiện các yêu cầu sau:

1.2.1 Hãy đọc tất cả các thẻ html (div) với lớp là "quote" và lưu nó trong biến 'result’, hiển thị giá trị biến 'result’ ra màn hình?

1.2.1.1 Code

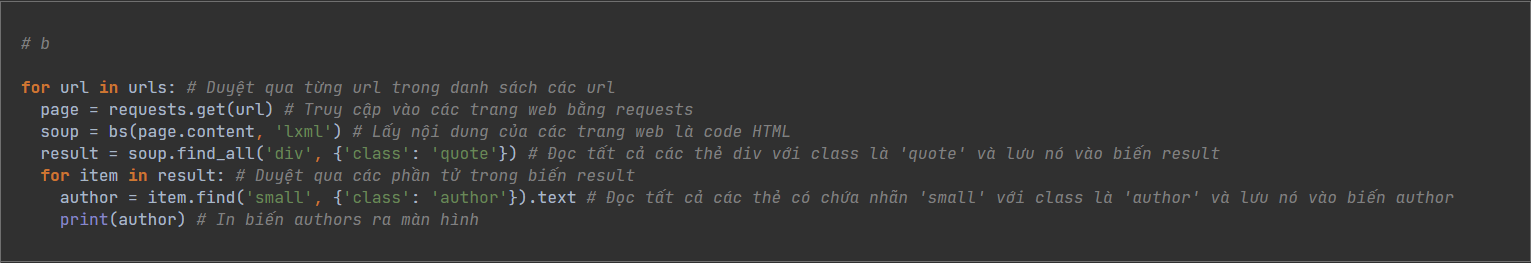


1.2.1.2 Kết quả



1.2.2 Hãy tìm trong biến 'result’ vừa rồi các dữ liệu có chứa nhãn "small" với class là "author" và in kết quả ra màn hình?

1.2.2.1 Code

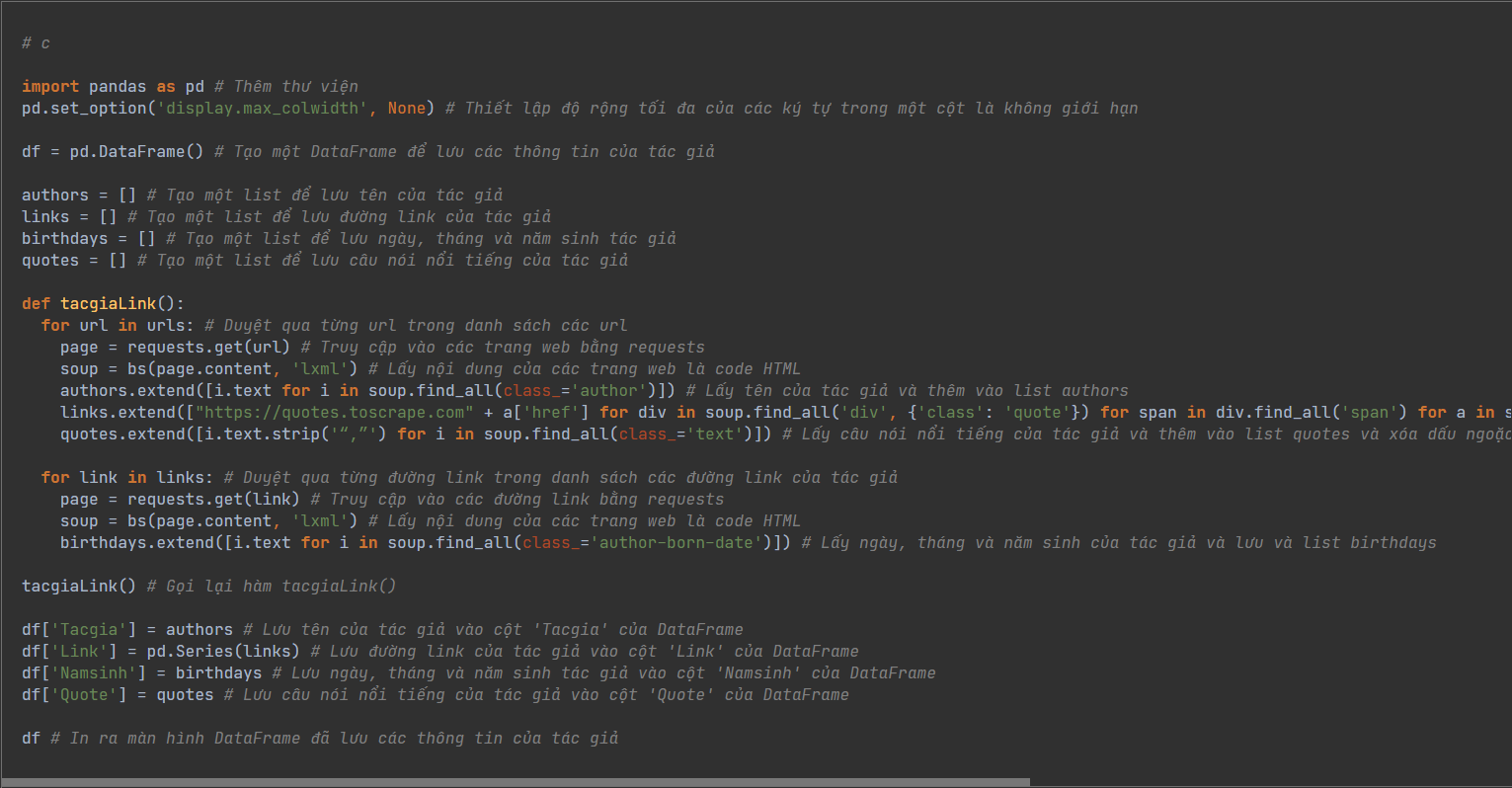


1.2.2.2 Kết quả

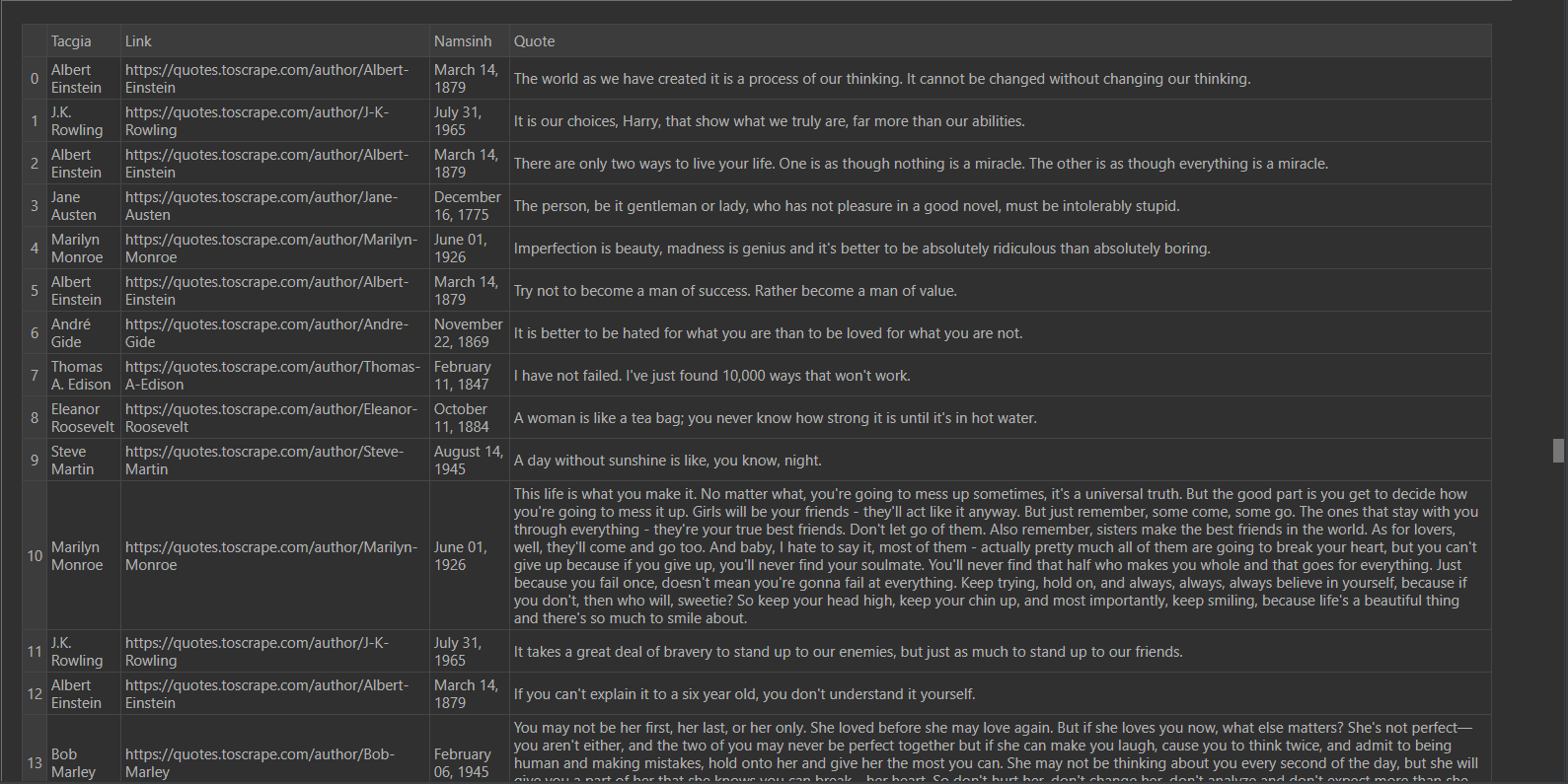


1.2.3 Hãy viết hàm tacgiaLink() để lấy nội dung của mỗi tác giả. Với mỗi tác giả in ra màn hình các nội dung: Tên tác giả, đường link của tác giả, ngày tháng năm sinh và câu nói nổi tiếng của tác giả

1.2.3.1 Code

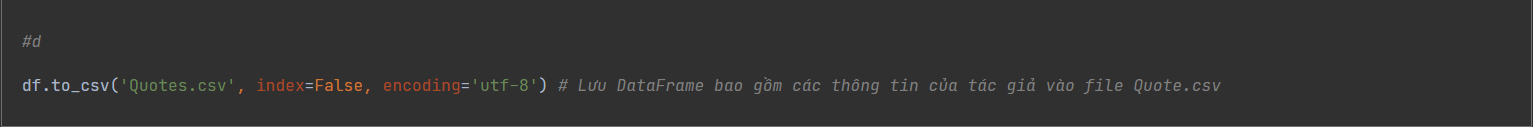


1.2.3.2 Kết quả

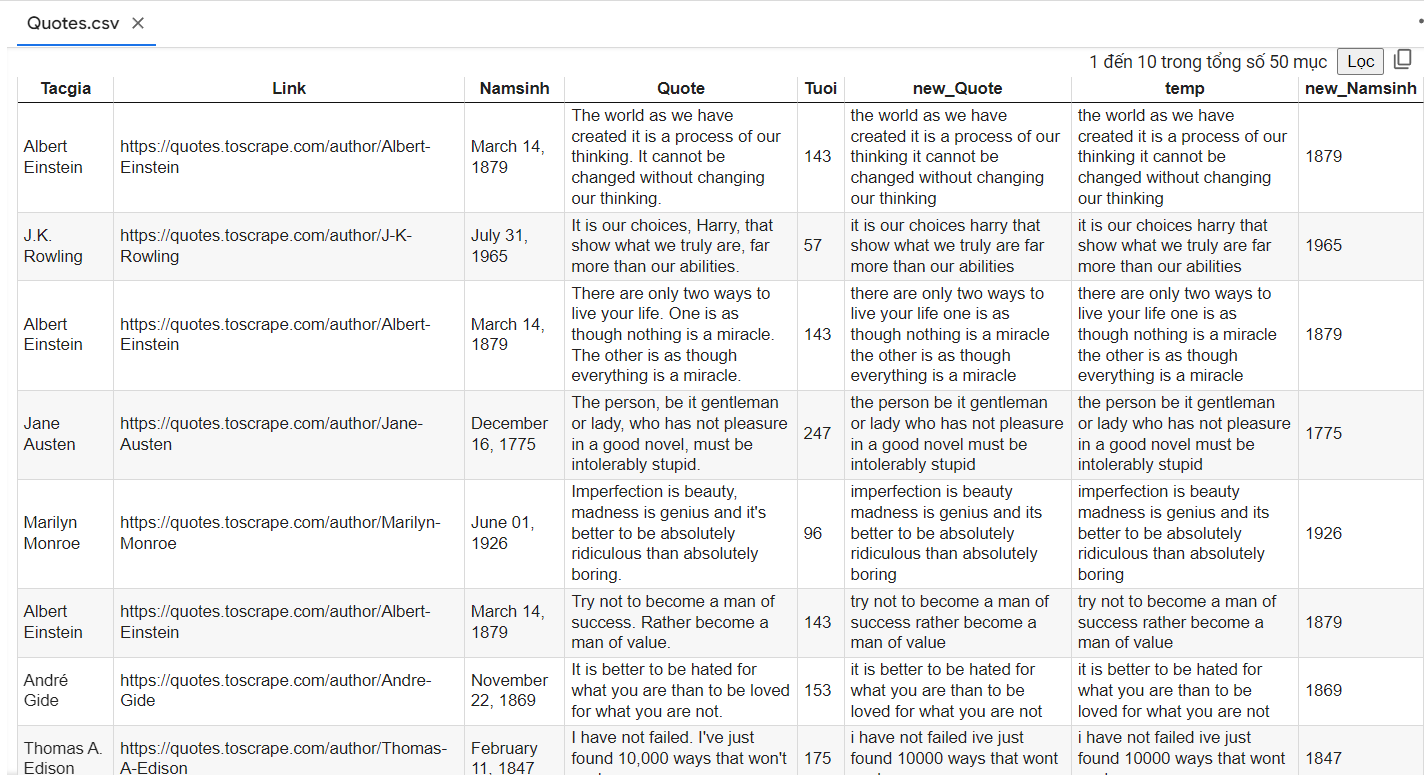


1.2.4 Hãy lưu kết quả ở câu c vào file Quote.csv tương ứng, với mỗi tác giả là 1 dòng dữ liệu. Bạn được yêu cầu thu thập ít nhất 40 câu nói nổi tiếng từ trang web trên một cách tự động theo code của các ý trên?

1.2.4.1 Code



1.2.4.2 Kết quả



CHƯƠNG 2 – KHAI PHÁ DỮ LIỆU

2.1 Xử lý dữ liệu- Data Imputation

2.1.1 Giới thiệu

Trong phần này, cần phải giải quyết hai vấn đề:

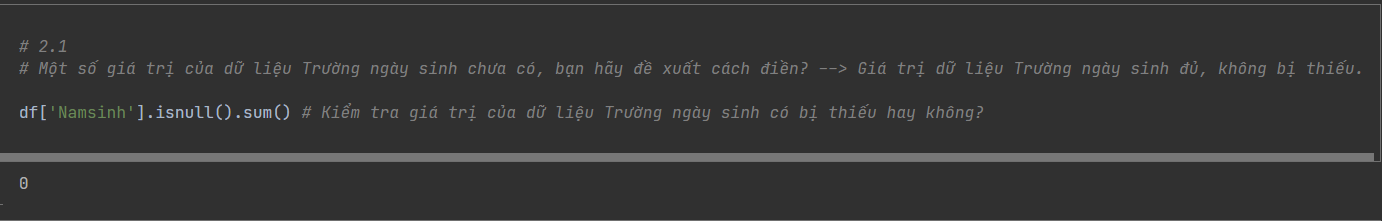
- Vấn đề thứ nhất: một số giá trị của dữ liệu trường Ngaysinh chưa có, chúng ta cần phải đề xuất cách điền cho nó.

- Vấn đề thứ hai: thêm vào trường Tuoi (Tuổi) và đề xuất cách điền tuổi của các tác giả.

2.1.2 Cách tiếp cận

- Đối với vấn đề thứ nhất: cần phải kiểm tra giá trị của dữ liệu trường Ngaysinh có null hay không.

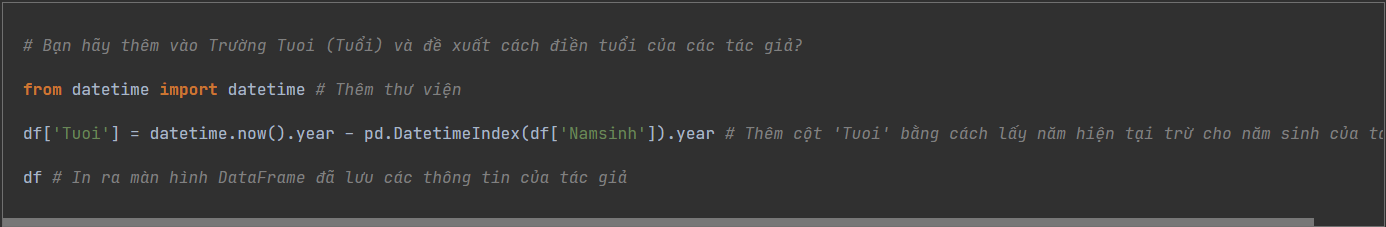
* Cách kiểm tra: sử dụng phương thức isnull() của thư viện pandas



* Nếu kết quả trả về là 0 thì chứng tỏ giá trị của dữ liệu trường Ngaysinh không bị null (đối với trường hợp bộ dữ liệu này).
* Nếu kết quả trả về là khác 0 thì chứng tỏ một số giá trị của dữ liệu trường Ngaysinh bị null và chúng ta cần phải điền cho nó bằng cách sử dụng phương thức loc() của thư viện pandas.

Ví dụ: df.loc[1, ‘Namsinh’] = ‘July 12, 2001’

- Đối với vấn đề thứ hai: cần tạo thêm cột Tuoi. Sau đó dùng phương thức datetime.now().year của thư viện datetime để lấy được năm hiện tại và dựa vào trường Namsinh sử dụng phương thức DatetimeIndex().year của thư viện pandas để lấy được năm sinh của từng tác giả. Chúng ta cần phải xác định được tác giả đó đã mất hay chưa, nếu chưa thì tuổi của tác giả được tính bằng cách lấy năm hiện tại trừ đi năm sinh của tác giả, nếu tác giả đã mất thì tuổi của tác giả được tính bằng cách lấy năm mất trừ đi năm sinh của tác giả.



2.1.3 Đánh giá

- Vấn đề thứ nhất: đã giải quyết.

- Vấn đề thứ hai: chưa được giải quyết.

2.1.4 Thảo luận

Vấn đề thứ hai chưa được giải quyết là vì bộ dữ liệu đã cho bị thiếu. Vì bộ dữ liệu chưa thể hiện được rằng tác giả đó đã mất hay chưa, nếu mất thì cần phải có thêm dữ liệu năm mất của tác giả. Vì thế không thể xác định được tuổi của tác giả nếu tác giả đó đã mất.

2.2 Khám phá dữ liệu- Data Exploration

2.2.1 Giới thiệu

Chúng ta cần khám phá dữ liệu để hiển thị một số thông tin thống kê và phân tích của tập dữ liệu đã cho. Chẳng hạn như:

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

- Thống kê về tác giả và số lượng câu nói của mỗi tác giả

- Thống kê về năm sinh của các tác giả

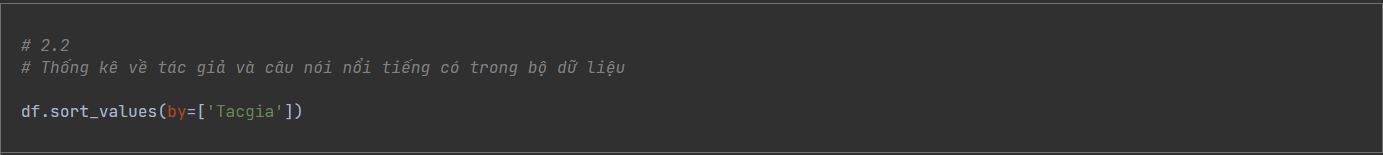
- Thống kê về các từ được sử dụng trong các câu nói

- Thống kê về các câu nói nổi tiếng như: câu dài nhất, ngắn nhất, số từ, ...

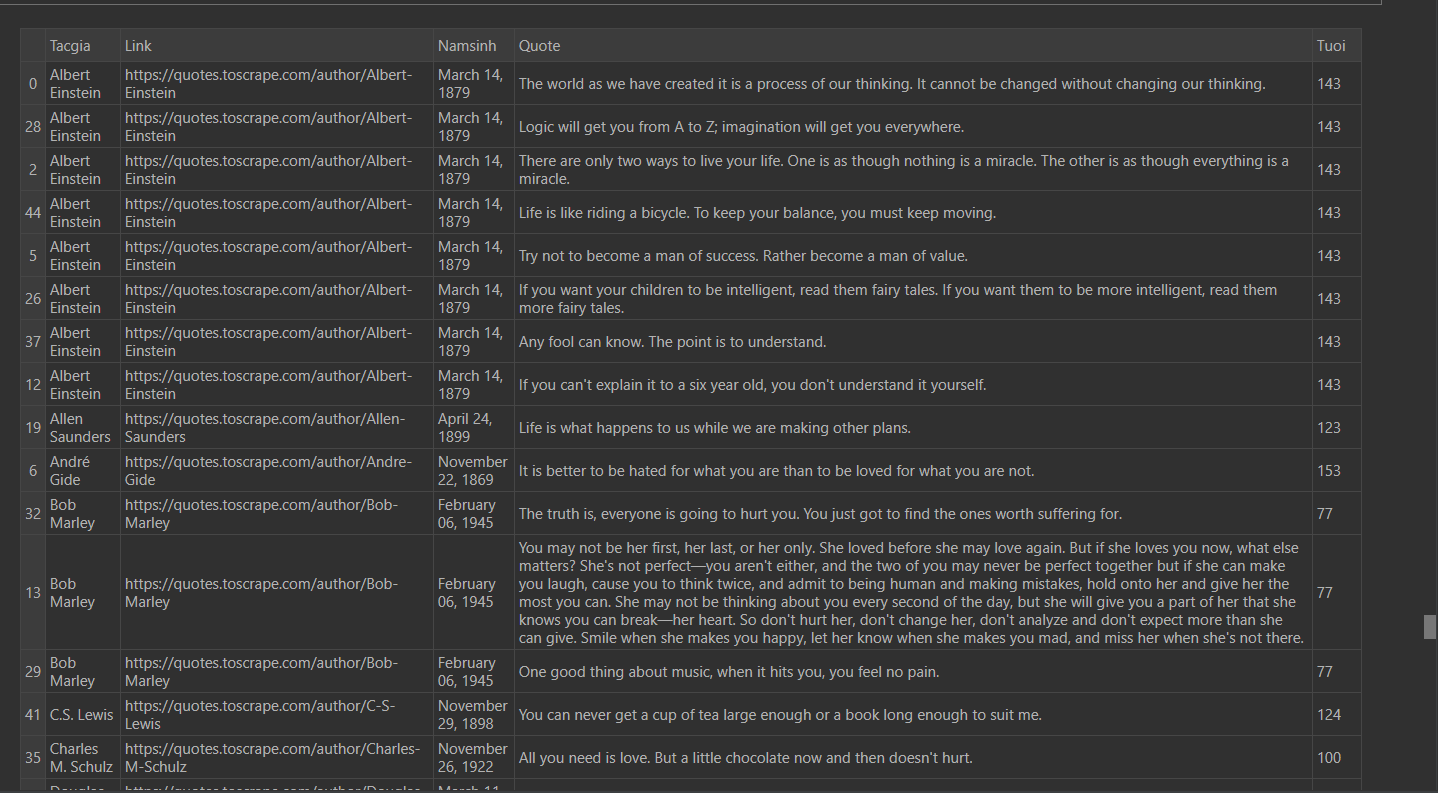
2.2.2 Cách tiếp cận

- Thống kê về tác giả và câu nói nổi tiếng có trong bộ dữ liệu: sử dụng phương thức sort\_values() của thư viện pandas để sắp xếp theo trường Tacgia, mặc định sắp xếp tên của các tác giả theo bảng chữ cái.

* Code:

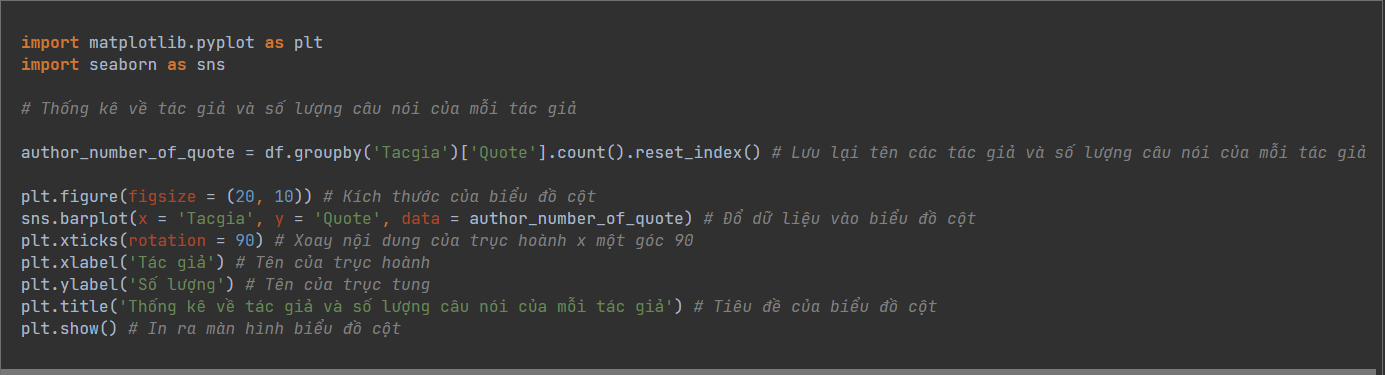


* Kết quả:

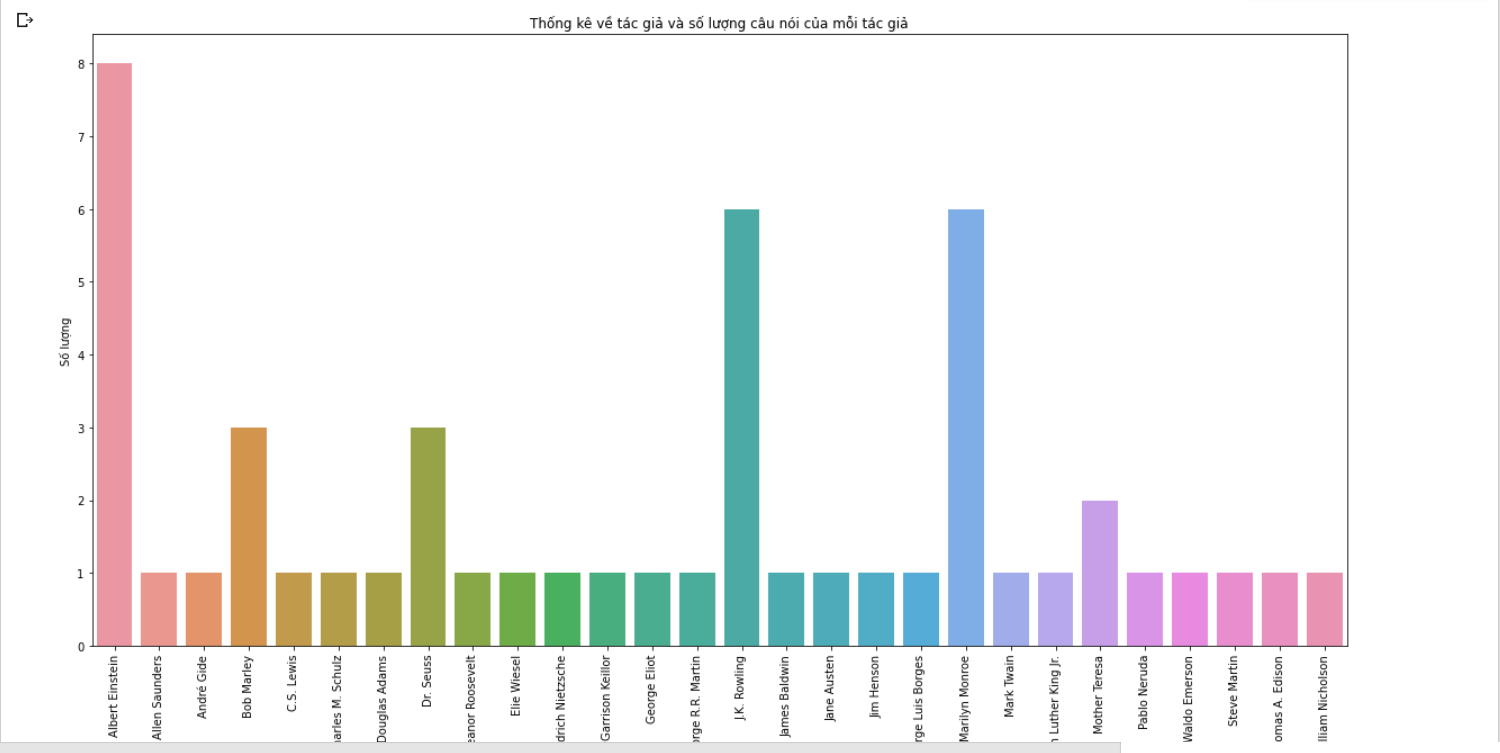


- Thống kê về tác giả và số lượng câu nói của mỗi tác giả: Đầu tiên, sử dụng phương thức groupby() của thư viện pandas để gộp lại tên các tác giả nếu bị trùng, sử dụng phương thức count() để đếm các các câu nói nổi tiếng của tác giả và gán vào biến author\_number\_of\_quote. Sau đó dùng các phương thức của thư viện matplotlib, seaborn để vẽ biểu đồ cột.

* Code:

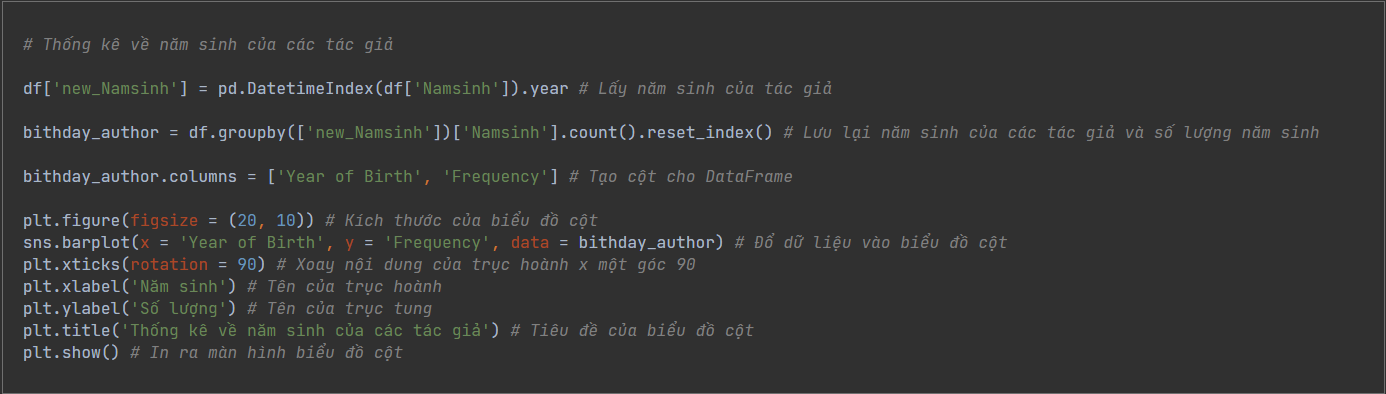


* Kết quả:

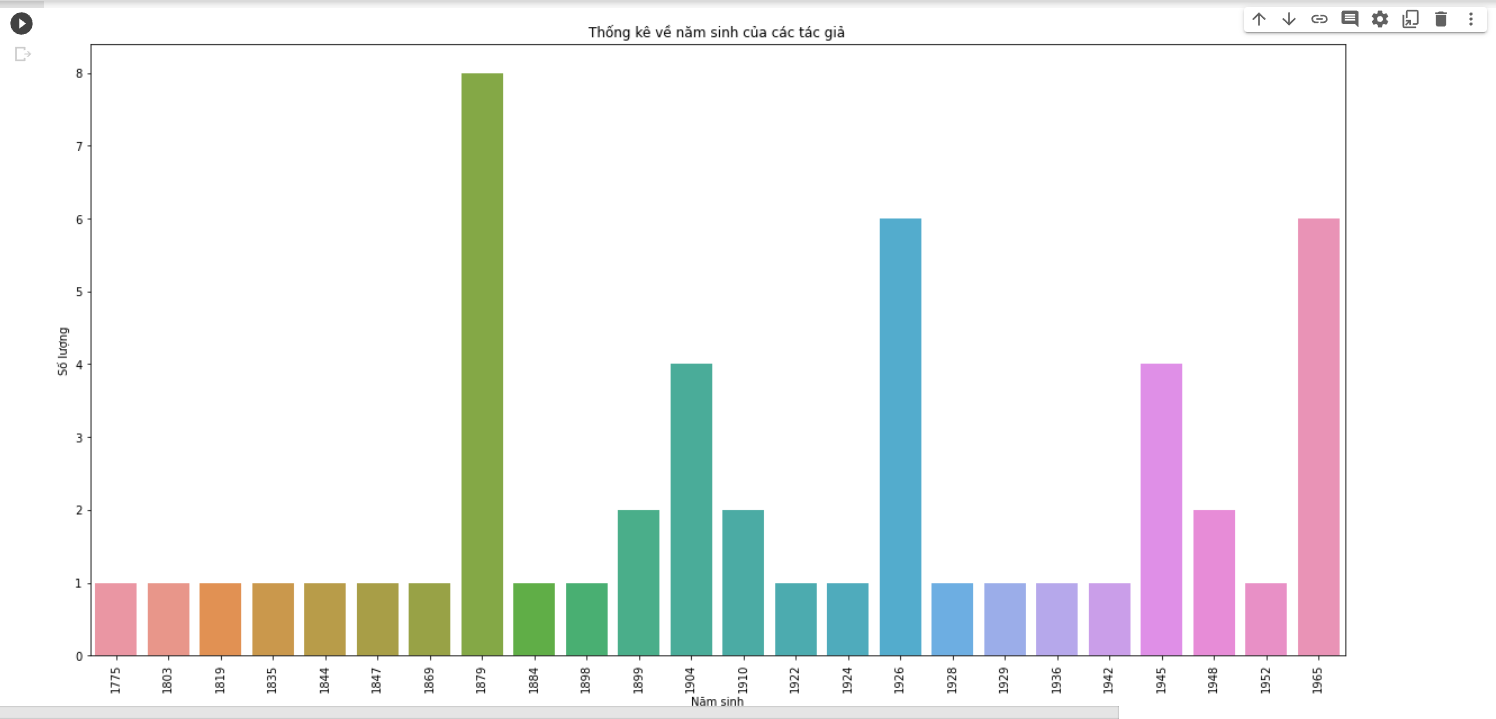


- Thống kê về năm sinh của các tác giả: Đầu tiên, sử dụng DatetimeIndex().year của thư viện pandas để lấy ra năm sinh của các tác giả và lưu vào trường mới được tạo new\_Namsinh. Tiếp theo, sử dụng phương thức groupby() của thư viện pandas để gộp lại năm sinh của các tác giả nếu bị trùng, dùng phương thức count() của thư viện pandas để đếm năm sinh của các tác giả và gán vào biến bithday\_author. Sau đó, dùng phương thức columns để tạo ra hai cột cho DataFrame có tên là Year of Birth và Frequency. Cuối cùng là dùng các phương thức của thư viện matplotlib, seaborn để vẽ biểu đồ cột.

* Code:

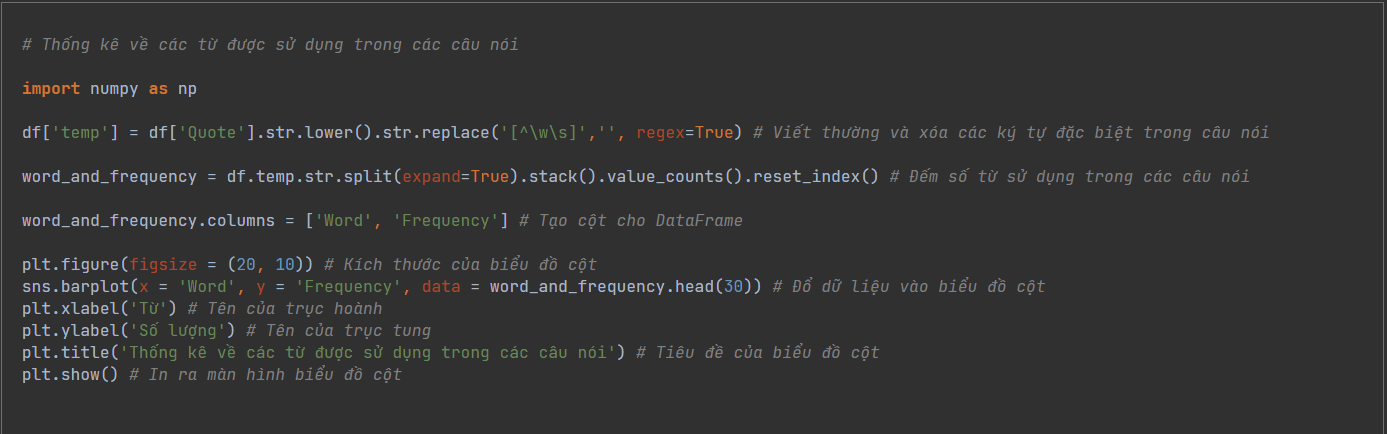


* Kết quả:

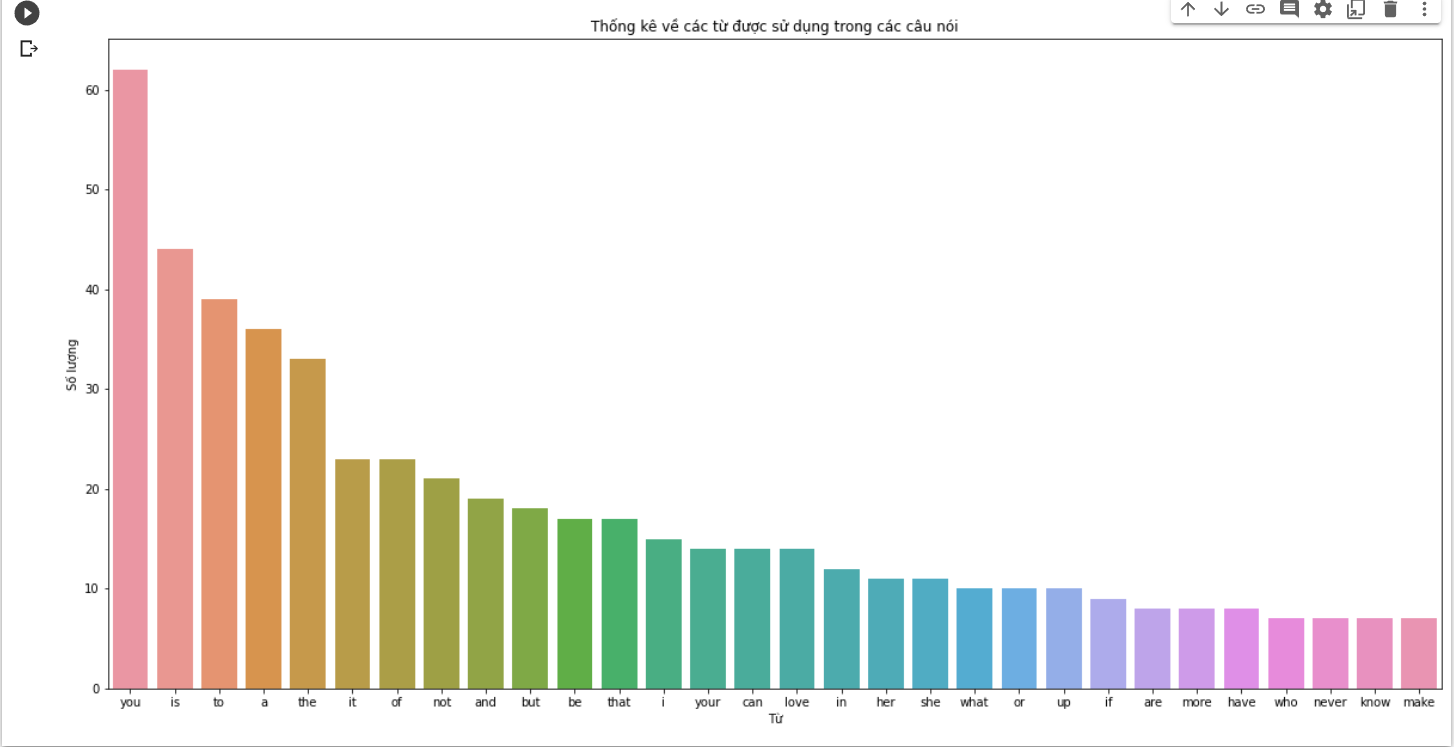


- Thống kê về các từ được sử dụng trong các câu nói: Đầu tiên, dùng phương thức str của pandas để cột Quote có thể sử dụng được các phương thức xử lý chuỗi. Tiếp theo dùng phương thức lower() để viết thường hết các từ trong cột Quote nếu từ đó viết hoa, dùng phương thức replace() để xóa các từ đặc biệt trong cột Quote nếu có ký tự đặc biệt để dễ dàng xử lý đếm các từ và gán vào cột mới được tạo temp. Tiếp tục dùng phương thức split() để chia các câu nói thành các từ riêng biệt, dùng phương thức stack() để sắp xếp số lần xuất hiện của các từ theo chiều giảm dần, dùng hàm values\_counts() để đếm số lần xuất hiện của các từ và gán vào biến word\_frequency. Sau đó, dùng phương thức columns để tạo ra hai cột cho DataFrame có tên là Word và Frequency. Cuối cùng là dùng các phương thức của thư viện matplotlib, seaborn để vẽ biểu đồ cột. Vì do có quá nhiều dẫn đến sơ đồ cột trở nên khó xem nên em đã giới hạn lại 30 từ có số lần xuất hiện nhiều nhất bằng phương thức head().

* Code:

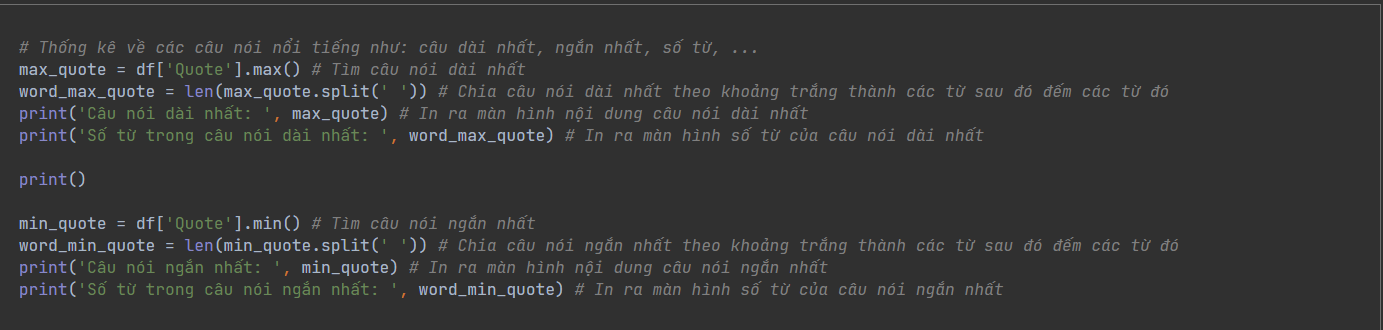


* Kết quả:

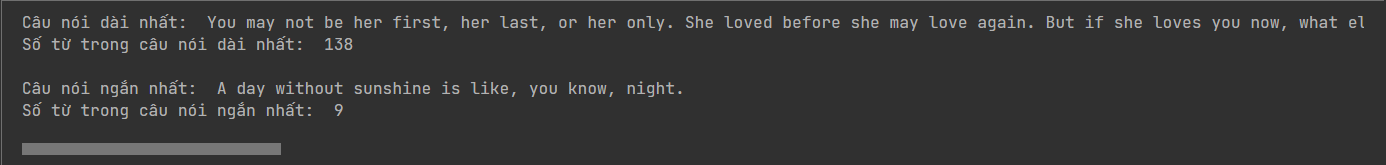


- Thống kê về các câu nói nổi tiếng như: câu dài nhất, ngắn nhất, số từ, ...: dùng phương thức max() và min() của pandas để tìm các câu nói dài nhất và ngắn nhất và lưu lần lượt vào biến max\_quote và min\_quote. Dùng phương thức split để chia các câu thành các từ dựa vào khoảng trắng và dùng phương thức len để đếm số từ trong câu dài nhất, ngắn nhất và lưu lần lượt vào biến word\_max\_quote và word\_min\_quote.

* Code:



* Kết quả:



2.2.3 Đánh giá

- Thống kê về tác giả và câu nói nổi tiếng có trong bộ dữ liệu đã được giải quyết

- Thống kê về tác giả và số lượng câu nói của mỗi tác giả đã được giải quyết

- Thống kê về năm sinh của các tác giả đã được giải quyết

- Thống kê về các từ được sử dụng trong các câu nói đã được giải quyết

- Thống kê về các câu nói nổi tiếng như: câu dài nhất, ngắn nhất, số từ, ... đã được giải quyết

2.2.4 Thảo luận

2.3 Trích xuất đặc trưng- Feature Extraction

2.3.1 Giới thiệu

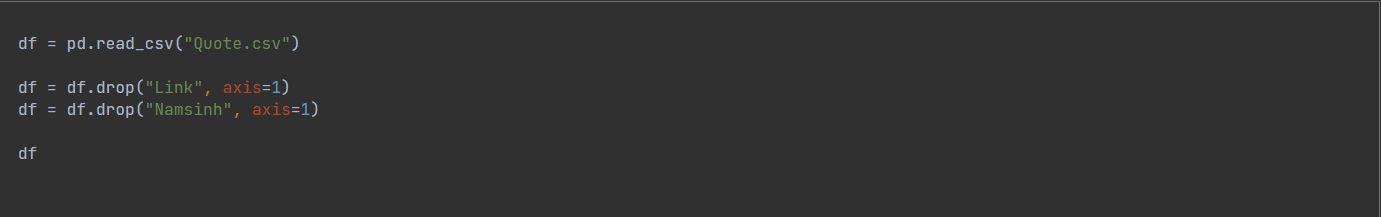
Đề xuất cách trích xuất đặc trưng từ bộ dữ liệu đã cho.

2.3.2 Cách tiếp cận

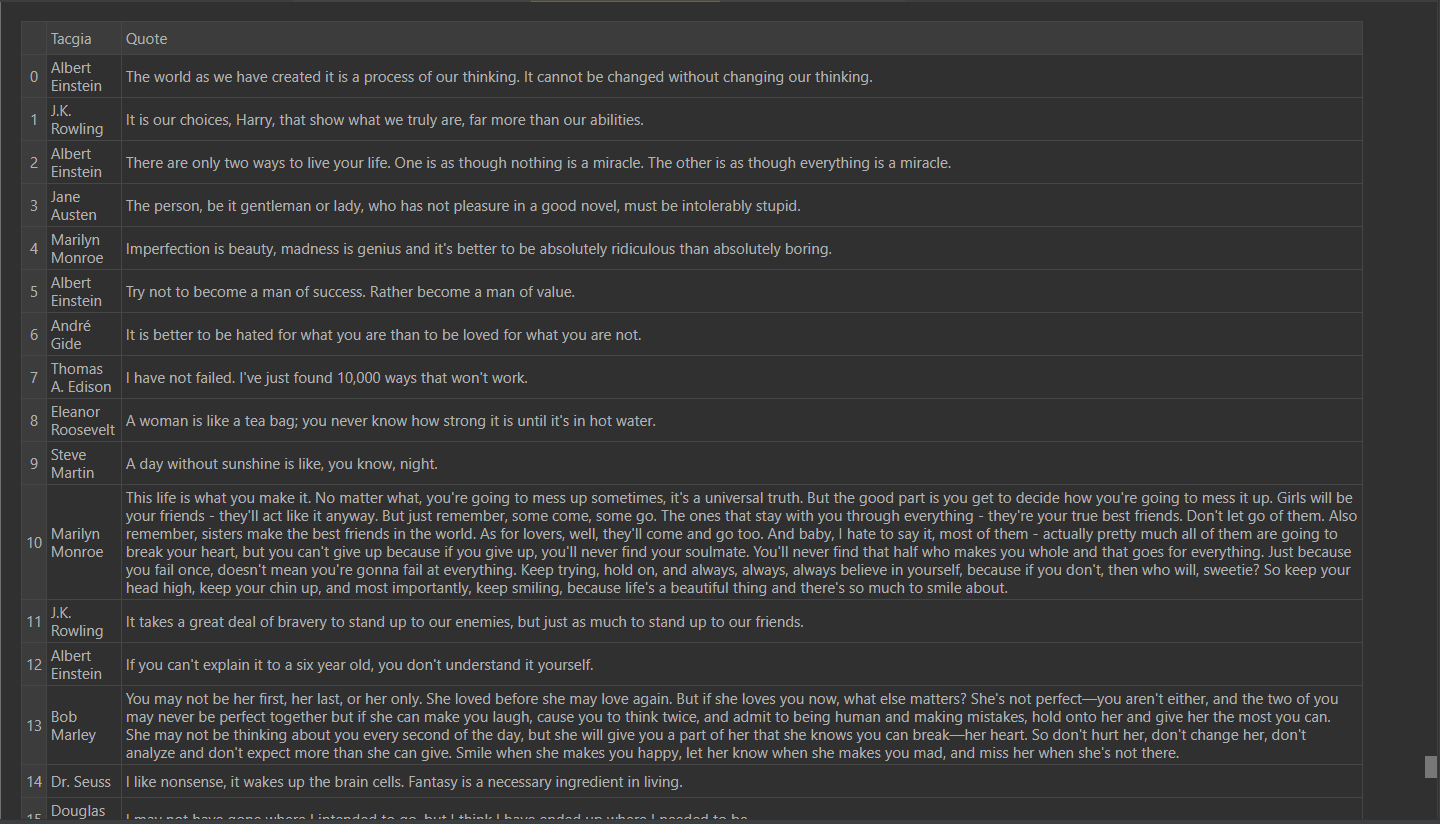
- Lý do: vì ở mục 2.4 có yêu cầu dự đoán tên của người nổi tiếng theo câu nói dựa trên các đặc trưng bạn trích xuất ở trên nên chúng ta sẽ xây dựng mô hình để dự đoán tác giả nào sẽ nói câu nói đó dựa vào câu nói.

- Giải thích cách làm: đọc lại file Quote.csv ban đầu (2.1d) và tiến hành xóa hai trường Link và Namsinh bằng phương thức drop() của thư viện pandas

* Code:



* Kết quả:



2.3.3 Đánh giá

Trích xuất đặc trưng từ bộ dữ liệu đã cho đã giải quyết.

2.3.4 Thảo luận

TÀI LIỆU THAM KHẢO

**Tiếng Anh**

1. Development - Pandas 1.4.2 documentation. (n.d.). Pandas Docs. from <https://docs.scrapy.org/en/latest/intro/tutorial.html>.
2. Beautiful Soup Documentation - Beautiful Soup version 4.8.1 from <https://beautiful-soup-4.readthedocs.io/en/latest/>
3. Marco Santos (Jun 27, 2020). How I used Python to Collect Data from any Website from <https://towardsdatascience.com/how-to-collect-data-from-any-website-cb8fad9e9ec5>