TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO MÔN XỬ LÝ DỮ LIỆU LỚN**

**BÀI BÁO CÁO GIỮA KỲ**

*Người hướng dẫn*: **TS.** **Bùi Thanh Hùng**

*Người thực hiện*: **Trần Anh Vũ – 51800517**

*Lớp***: 18050201**

Khoá  **: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

Logo

Description automatically generated

**BÁO CÁO MÔN XỬ LÝ DỮ LIỆU LỚN**

**BÀI BÁO CÁO GIỮA KỲ**

*Người hướng dẫn*: **TS.** **Bùi Thanh Hùng**

*Người thực hiện*: **Trần Anh Vũ – 51800517**

*Lớp***: 18050201**

Khoá  **: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

LỜI CẢM ƠN

Để hoàn thành báo cáo này trước tiên em xin gửi đến cảm ơn thầy Bùi Thanh Hùng – người đã tận tình hướng dẫn, giúp đỡ em hoàn thành đố án cuối kì này lời cảm ơn sâu sắc nhất.

Với điều kiện thời gian cũng như kinh nghiệm còn chế của sinh viên, bài báo cáo này không thế tránh khỏi được những thiếu sót. Em rất mong nhận được sự chỉ bảo, đóng góp ý kiến của thầy để em có điều kiện bổ sung, nâng cao kiến thức của mình, phục vụ tốt hơn công việc thực tế sau này.

**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 / chúng tôi và được sự hướng dẫn của ThS. 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 tháng năm

Tác giả

Trần Anh Vũ

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

Với sự phát triển của khoa học hiện đại rất nhanh chóng hiện nay dẫ đến tình trạng có nhiều dữ liệu khổng lồ cần phải xử lý và phân tích.

Do đó, qua bài giữa kỳ này em xin trình bày về hai phần mà trong đề giữa kỳ yêu cầu:

Phần 1 – Thu thập dữ liệu

Phần 2 – Khai phá dữ liệu

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc99998340)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc99998341)

[TÓM TẮT iv](#_Toc99998342)

[MỤC LỤC 1](#_Toc99998343)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 3](#_Toc99998344)

[PHẦN 1 – THU THẬP DỮ LIỆU 4](#_Toc99998345)

[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? 4](#_Toc99998346)

[1.2 Với dữ liệu bạn vừa cào về, bạn hãy thực hiện các yêu cầu sau: 6](#_Toc99998347)

[PHẦN 2 – KHAI PHÁ DỮ LIỆU 10](#_Toc99998348)

[2.1 Xử lý dữ liệu- Data Imputation 10](#_Toc99998349)

[2.1.1 Phần giới thiệu: 10](#_Toc99998350)

[2.1.2 Cách tiếp cận 10](#_Toc99998351)

[2.1.3 Đánh giá 11](#_Toc99998352)

[2.2 Khám phá dữ liệu- Data Exploration 12](#_Toc99998353)

[2.2.1 Phần giới thiệu 12](#_Toc99998354)

[2.2.2 Cách tiếp cận 12](#_Toc99998355)

[2.2.3 Đánh giá 14](#_Toc99998356)

[TÀI LIỆU THAM KHẢO 15](#_Toc99998357)

**DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT**

**CÁC KÝ HIỆU**

**CÁC CHỮ VIẾT TẮT**

df – DataFrame

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

**DANH MỤC HÌNH**

[Ảnh 1 Khai báo thư viện 5](#_Toc99998191)

[Ảnh 2 Code để lấy toàn phần của web 6](#_Toc99998192)

[Ảnh 3 Cấu trúc sơ bộ của trang web https://quotes.toscrape.com/ 6](#_Toc99998193)

[Ảnh 4 Câu 1.2 ý 1 7](#_Toc99998194)

[Ảnh 5 Kết quả của câu 1.2 ý 1 8](#_Toc99998195)

[Ảnh 6 Code và kết quả của câu 1.2 ý 2 8](#_Toc99998196)

[Ảnh 7 Code để lấy ngày sinh của tác giả 9](#_Toc99998197)

[Ảnh 8 Hàm tacgiaink() 10](#_Toc99998198)

[Ảnh 9 Kết quả hàm tacgiaLink() 10](#_Toc99998199)

[Ảnh 10 code lưu dữ liệu vào file csv 11](#_Toc99998200)

[Ảnh 11 df có missingvalue 12](#_Toc99998201)

[Ảnh 12 dữ liệu missing được thay bằng số 0 12](#_Toc99998202)

[Ảnh 13 code điền tuổi và thêm trường ‘Tuoi’ của tác giả vào df 12](#_Toc99998203)

[Ảnh 14 Thống kê về tác giả và các câu nói 13](#_Toc99998204)

[Ảnh 15 tác giả và số câu nói 13](#_Toc99998205)

[Ảnh 16 Các tác giả có cùng độ tuổi 14](#_Toc99998206)

[Ảnh 17 Thống kê về năm sinh và độ tuổi 14](#_Toc99998207)

[Ảnh 18 Khảo sát về độ dài các câu nói 14](#_Toc99998208)

[Ảnh 19 Số lượng tác giả trong df 15](#_Toc99998209)

**DANH MỤC BẢNG**

PHẦN 1 – THU THẬP DỮ LIỆU

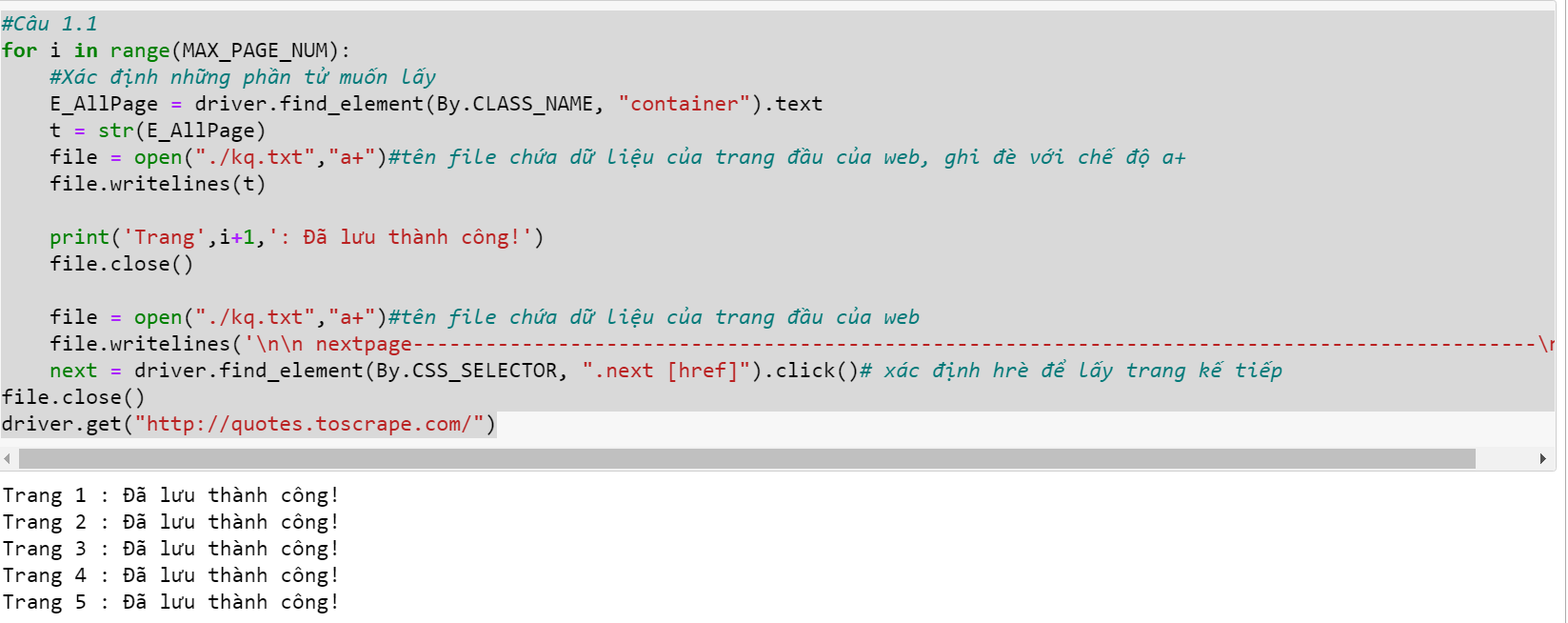
* 1. Bạn hãy viết code cào dữ liệu từ trang web trên, lưu kết quả vào 1 file tương ứng (kq.txt) và mô tả ngắn gọn về cấu trúc của trang Web trên?
* Do phần 1 yêu cầu sử cào dữ liệu từ web với ít nhất 40 dòng nên em đã chọn selenium để thực hiện việc cào dữ liệu tự động ở các trang:
* Đầu tiên là tải driver trên trang chủ cảu selenium(cần tải driver phù hợp với trình duyệt muốn sử dụng), ở đây em sẽ dùng chrome version 100 để làm bài.
* Sau đó import các thư viện để sử dụng.

Graphical user interface, text, application, email

Description automatically generated

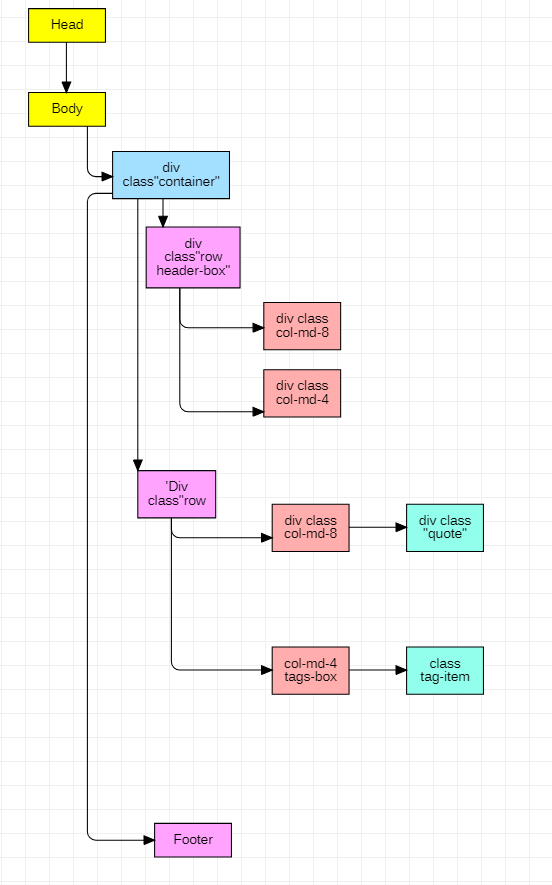
Ảnh 1 Khai báo thư viện

* Tiếp theo em sẽ khai báo driver để chạy selenium và sử dụng phương thức get để chỉ định trang web cần cào:
* Tiếp theo em sẽ thực hiện code về giai đoạn lấy tất cả nội dụng của ý 1.1 và lưu vào file ‘kq.txt’.



Ảnh 2 Code để lấy toàn phần của web

* Mô tả về cấu trúc của trang web(em sẽ mô tả về cấu trúc web bằng sơ đồ sau):



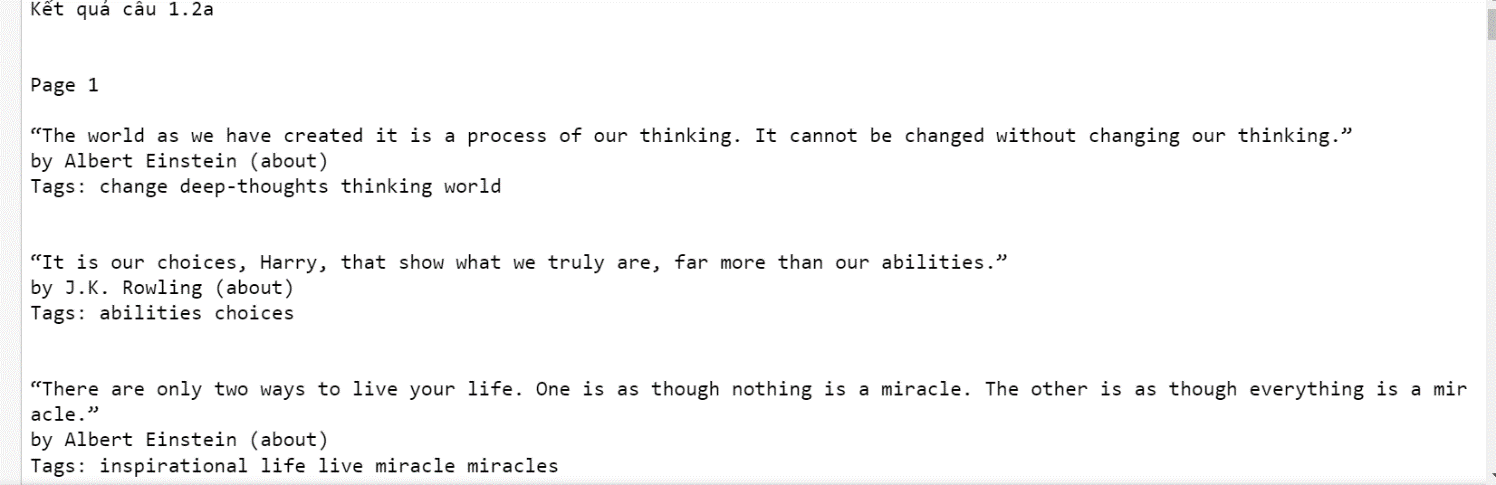
Ảnh 3 Cấu trúc sơ bộ của trang web <https://quotes.toscrape.com/>

* Mô tả về cấu trúc trên trang web: 1. Web chia ra các thẻ và tên class như sau:
  + Body:
    - Thẻ div class="container"
      * Thẻ div class="row header-box" chứa tên của trang web(Quotes to Scrape) và vị trí để login tài khoản
      * Thẻ row chứa hai thẻ có tên class="col-md-8"(chứa phần lớn nội dung chính của trang như các câu nói, tên tác giả, thông tin của tác giả và các tags(liên quan)) và class="col-md-4 tags-box"(chứa những tag được phổ biển trong top 10(Top Ten tags))
      * Thẻ 'ul' chia trang trước và trang sau.
      * Cuối cùng là thẻ div class= "Footer"(chứa các thông tin cơ bản về trang web)
  1. Với dữ liệu bạn vừa cào về, bạn hãy thực hiện các yêu cầu sau:
* Với câu 1.2 ý 1:
  + Em thực hiện như sau em sẽ cào 5 trang của web và in ra màn hình những dòng có class theo yêu cầu.

Graphical user interface, text, application, email

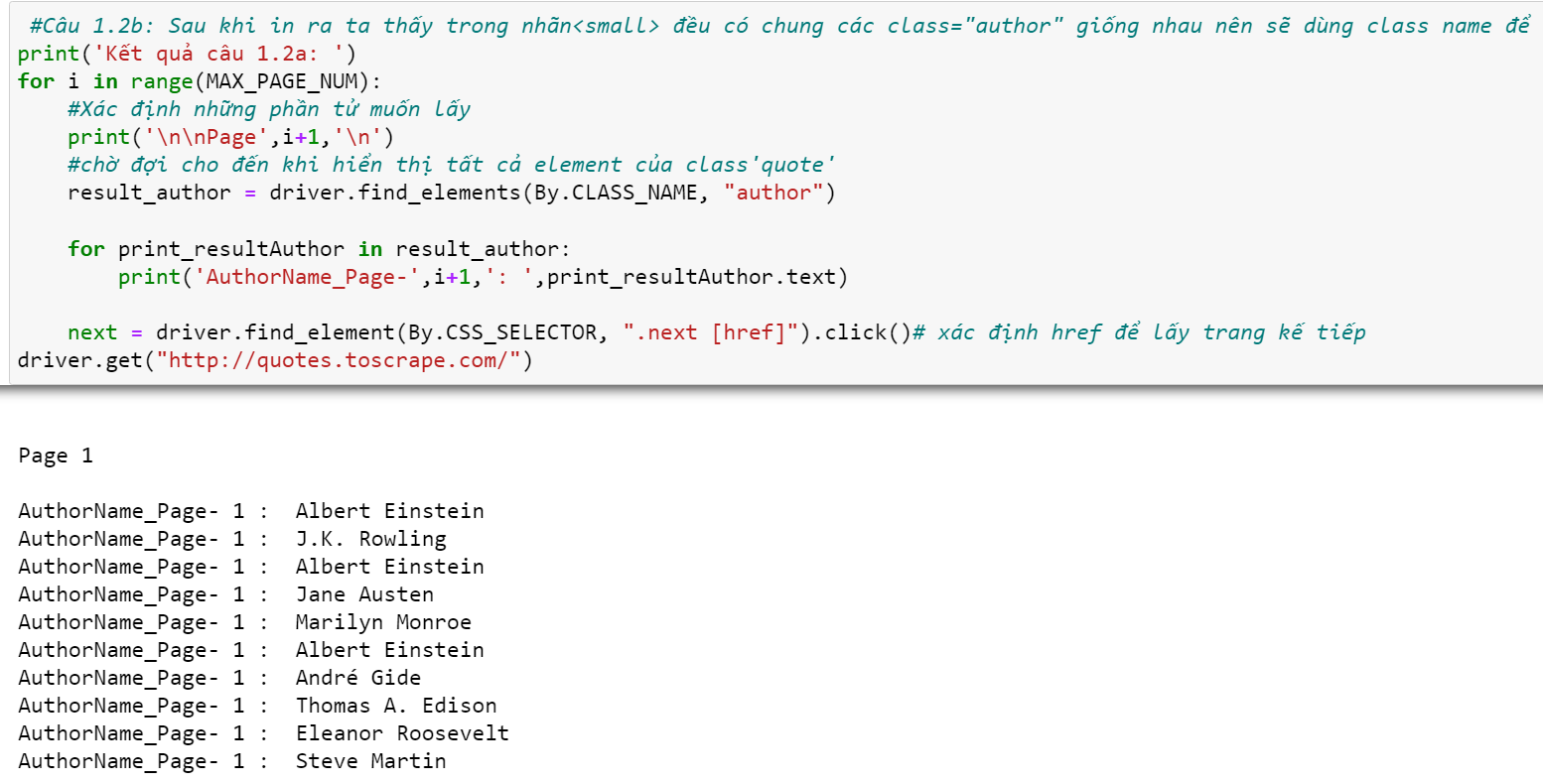
Description automatically generated

Ảnh 4 Câu 1.2 ý 1



Ảnh 5 Kết quả của câu 1.2 ý 1

* Tiếp đến là phần 1.2 ý 2 đề sẽ yêu cầu lấy dữ liệu ở các nhãn <small> và có class”author”, và sau khi xem xét thì kết quả là sẽ phải lấy tên của tác giả.



Ảnh 6 Code và kết quả của câu 1.2 ý 2

* Tiếp theo để thực hiện yêu cầu viết hàm tacgiaLink() để in ra kết quả bao gồm Tên tác giả, Đường link của tác giả, Ngày tháng năm sinh, câu nói nổi tiếng của tác giả. Em sẽ làm như sau:
  + Đầu tiên em sẽ viết một đoạn code để lấy dữ liệu ngày sinh trước vì nếu muốn lấy dữ liệu ngày sinh phải truy cập vào một link khách nên em đã lấy trước và lưu vào biến: data0, data1, data2, data3, data4(tương úng từ trang 1 đến 5).

A picture containing table

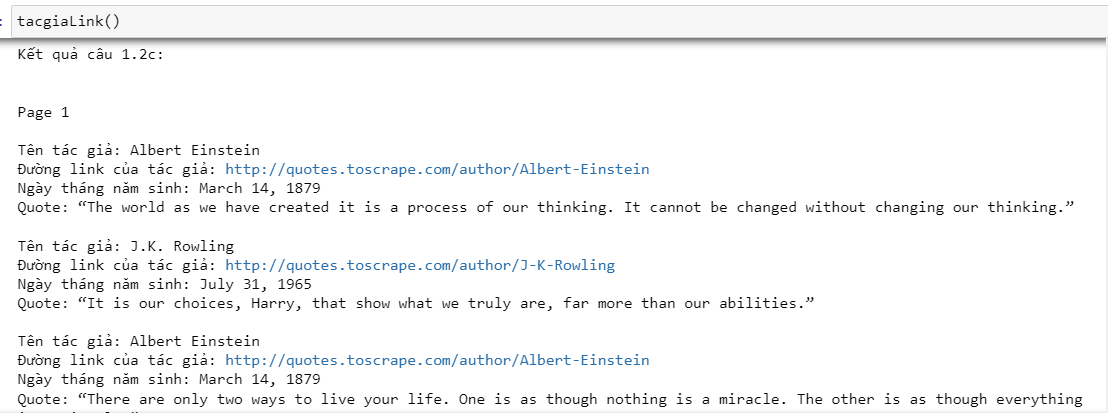
Description automatically generated

Ảnh 7 Code để lấy ngày sinh của tác giả

* Sau đó với những dữ liệu cần lấy còn lại như(tên tác giả, đường link, câu nói) em sẽ lấy theo các phương thức như sau: tên tác giả(class”author”), đường link(xpath), câu nói(xpath). Và sau đó in ra:



Ảnh 8 Hàm tacgiaink()



Ảnh 9 Kết quả hàm tacgiaLink()

* Và yêu cầu cuối cùng là lưu kết quả vào file Quote.csv:



Ảnh 10 code lưu dữ liệu vào file csv

PHẦN 2 – KHAI PHÁ DỮ LIỆU

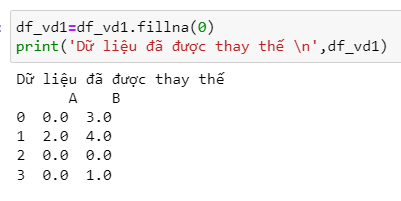
2.1 Xử lý dữ liệu- Data Imputation

* + 1. Phần giới thiệu:
* Với những dữ liệu khi được lấy về thì không phải dữ liệu nào cũng được định dạng sẵn hay có đầy đủ các ô dữ liệu cho nên cần phải vận dụng các kỹ thuật xử lý dữ liệu để xử lý những vấn đề như missing data.
  + 1. Cách tiếp cận
* Sau khi thực hiện yêu cầu 2.1 ý 1với dữ liệu em thu về từ web ở phần 1 sẽ không có ô bị thiếu như đề cần tìm hiểu nên em có lấy 1 ví dụ đơn giản về missing data và đề xuất cách xử lý nó.
* Đầu tiên em sẽ tạo dataframe với hai dòng như ảnh dưới và thực hiện xử lý nó:

Graphical user interface, text, application

Description automatically generated

Ảnh 11 df có missingvalue

* Với trường hợp này em sẽ đề xuất dùng hàm fillna() để có thể thay thế các giá trị missing.
  + - 

Ảnh 12 dữ liệu missing được thay bằng số 0

* Tiếp theo ở ý 2 để thực hiện việc thêm trường tuổi và đề xuất cách điền tuổi của các tác giả
* Đầu tiên em sẽ đề xuất phương án điền số tuổi của các tác giả theo cách em làm: đầu tiên em sẽ chuyển dữ liệu của trường ‘Namsinh’ về int64(dạng datetime) trước
* Tiếp theo, em sẽ dùng ‘relativedelta’ để thực hiện lấy ngày hiện tại để trừ cho trường năm sinh để lấy được tuổi của tác giả cho đến hiện tại. Và thực hiện thêm cột tuổi vào dataframe.
* Graphical user interface, text, application

  Description automatically generated

Ảnh 13 code điền tuổi và thêm trường ‘Tuoi’ của tác giả vào df

* + 1. Đánh giá
* Bước xử lý dữ liêu này rất quan trọng và là bước đầu tiên để xử lý dữ liệu với mục đích làm sạch dữ liệu mới được thu về. Đây là một bước rất quan trọng trong việc xử lý dữ liệu lớn.

2.2 Khám phá dữ liệu- Data Exploration

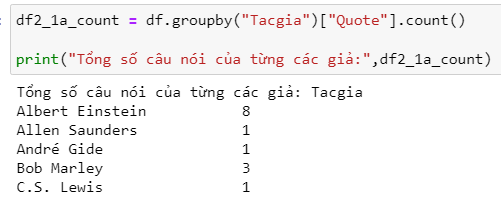
2.2.1 Phần giới thiệu

* Khám phá dữ liệu là bước mà khi đã xử lý dữ liệu xong, sẽ tiến hành dùng dữ liệu đó để tạo ra các phân tích ban đầu về dữ liệu để phục vụ cho các bước phân tích chuyên sâu về sau. Nhìn chung nó sẽ giúp người thu thập hiểu rõ hơn về dữ liệu mới được thu thập.
  + 1. Cách tiếp cận
* Để tiếp cận được phương pháp này em sẽ tiến hành các yêu cầu của đề bài:
  + Thống kê về tác giả và câu nói nổi tiếng có trong bộ dữ liệu: Để thực hiện yêu cầu này em sẽ dùng hàm groupby và thống kê tác giả và những câu nói trong dữ liệu.

Text

Description automatically generated

Ảnh 14 Thống kê về tác giả và các câu nói



Ảnh 15 tác giả và số câu nói

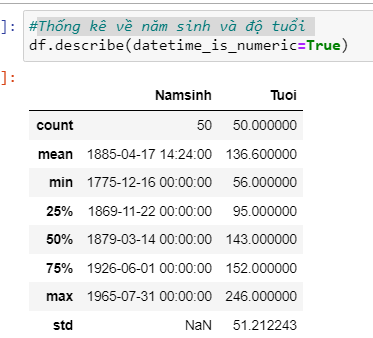
* + Thống kê về năm sinh và độ tuổi của các tác giả: yêu cầu này em sẽ gom nhóm những tác giác có cùng độ tuổi với nhau:

Text

Description automatically generated

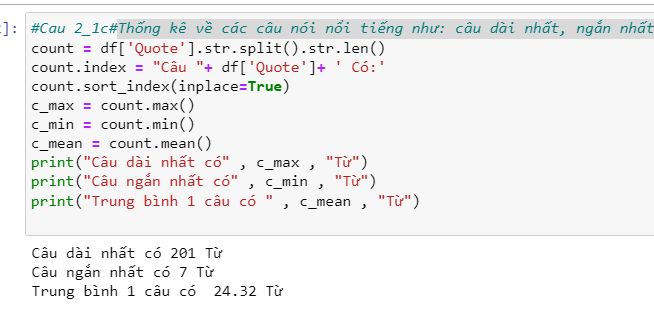
Ảnh 16 Các tác giả có cùng độ tuổi

* + Thống kê về năm sinh và độ tuổi:



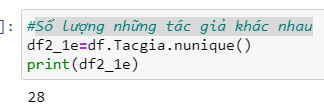
Ảnh 17 Thống kê về năm sinh và độ tuổi

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



Ảnh 18 Khảo sát về độ dài các câu nói

* + Số lượng những tác giả khác nhau: với dữ liệu thu được thì các tác giả sẽ có nhiều câu nói, khảo sát này sẽ giúp chúng ta nhận thấy số lượng tác giả trong 50 câu nói thu được.



Ảnh 19 Số lượng tác giả trong df

2.2.3 Đánh giá

* Nhìn chung thì bước này sẽ giúp người xử lý hiểu hơn về dữ liệu mà thu được sẽ giúp định hướng mục đích dùng dữ liệu vào các ứng dụng khác nhau. Bằng việc vận dụng thư viện pandas sẽ giúp dễ dàng hoàn thành hơn.
* Giúp giảm tải những dữ liệu cần xử lý bằng cách biện pháp gom nhóm.

TÀI LIỆU THAM KHẢO

[1] <https://www.selenium.dev/documentation/>

[2] <https://www.analyticsvidhya.com/blog/2020/08/web-scraping-selenium-with-python/>

[3] <https://bigdatauni.com/tin-tuc/cac-phuong-phap-imputation-don-gian-cho-missing-values.html>

[4] <https://pandas.pydata.org/docs/reference/frame.html>

[5] <https://www.sisense.com/glossary/data-exploration/>