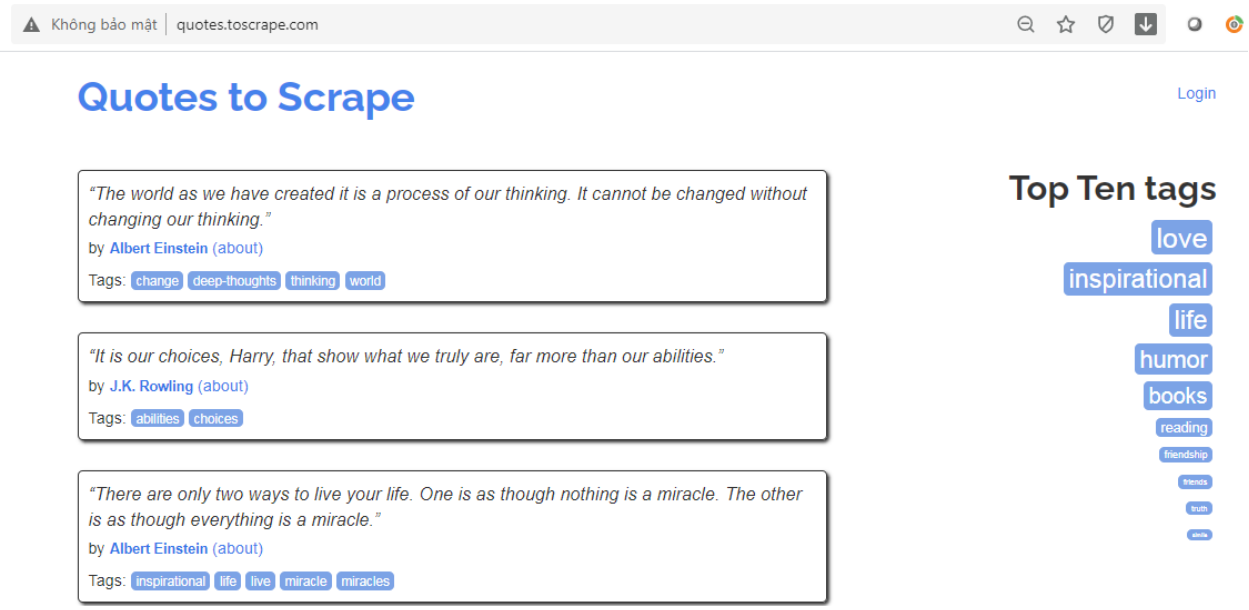
Assignment

Lecturer: Doctor Bui Thanh Hung
Data Science Laboratory
Faculty of Information Technology
Industrial University of Ho Chi Minh city
Email: hung.buithanhcs@gmail.com)
Website: https://sites.google.com/site/hungthanhbui1980/

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.

I. Thu thập dữ liệu (20 điểm)

Dữ liệu về những câu nói của Những người nổi tiếng trên thế giới có ở đường link: http://quotes.toscrape.com/ , trang Web này có giao diện như sau:



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? **(5 điểm)**

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

a. 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? **(2 điểm)**

b. 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? **(2 điểm)**

c. Hãy viết hàm tacgiaLink() để lấy nội dung của mỗi tác giả. Với mỗi tác giả in ra màn hình các nội dung **(10 điểm)**

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

d. 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 điểm)**

II. Khai phá dữ liệu (60 điểm)

Với bộ dữ liệu bạn đã thu thập ở trên có nội dung:

Trường	**Kiểu dữ liệu**	**Mô tả**
Tacgia	Text	Tên tác giả
Link	Text	Đường link của tác giả
Namsinh	Date	Ngày tháng năm sinh
Quote	Text	Câu nói nổi tiếng của tác giả

Bảng 1: Mô tả về bộ dữ liệu Quote.csv

Bạn hãy viết code thực hiện các yêu cầu sau.

2.1. Xử lý dữ liệu- Data Imputation (5 điểm):

- Một số giá trị của dữ liệu Trường ngày sinh chưa có, bạn hãy đề xuất cách điền?
- Bạn hãy thêm vào Trường Tuoi (Tuổi) và đề xuất cách điền tuổi của các tác giả?

2.2. Khám phá dữ liệu- Data Exploration (15 điểm):

Bạn 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 năm sinh và độ tuổi của các tác giả,
- Thống kê về các câu nói nổi tiếng như: câu dài nhất, ngắn nhất, số từ, …

- Thống kê về các từ được sử dụng trong các câu nói,
- Phân tích, trực quan mối quan hệ giữa giữa tác giả và câu nói nổi tiếng,
- Phân tích, trực quan mối quan hệ giữa các tác giả với nhau,…

Trên đây chỉ là một số gợi ý, bạn có thể đề xuất thêm các phân tích, thống kê khác.

2.3. Trích xuất đặc trưng- Feature Extraction (10 điểm):

Hãy đề xuất cách trích xuất đặc trưng từ bộ dữ liệu đã cho, cung cấp lý do và giải thích cách làm của bạn.

2.4. Suy luận (30 điểm):

Bạn được yêu cầu phân loại câu nói theo tên người nổi tiếng và tính độ tương đồng phong cách nói giữa các tác giả theo 2 yêu cầu sau:

- 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? **(15 điểm)**
- Hãy đề xuất cách tính độ tương đồng phong cách nói giữa các tác giả và tìm ra các tác giả có phong cách nói tương đồng nhau nhất? **(15 điểm)**

III. Báo cáo (20 điểm)

Báo cáo của bạn được trình bày theo từng phần ở trên, sử dụng ngôn ngữ tiếng Việt. Báo cáo phải có cấu trúc rõ ràng cho từng yêu cầu ở trên, với phần I- ***bạn chỉ cần đưa ra cách hiện thực (code), riêng câu 1.1 ý 2 bạn vẽ sơ đồ cấu trúc)***; còn Phần 2- ***với mỗi yêu cầu đều bao gồm: phần giới thiệu, cách tiếp cận, đánh giá, thảo luận, v.v. phù hợp*** để bạn hoàn thành các yêu cầu ở trên một cách hiệu quả.

IV. Nộp bài

Bạn hãy nộp bài trên LMS trước 23h ngày 28/11/2024 (Source code + Data + Báo cáo theo hướng dẫn ở mục III).

Chấm bài Slot 13 ngày 29/11/2024.