INTRODUCTION TO DATA SCIENCE

FINAL PROJECT

Giảng viên: TS. Bùi Thanh Hùng
Bộ môn Khoa học dữ liệu, Khoa Công nghệ thông tin
Đại học Công nghiệp thành phố Hồ Chí Minh
Email: buithanhhung@iuh.edu.vn
Website: https://sites.google.com/site/hungthanhbui1980/

I. ĐỀ BÀI

Bài 1 (30 điểm):

Xây dựng bộ dữ liệu theo như hướng dẫn mô tả sau:

1- **(6đ)** Lựa chọn 1 chủ đề và chọn các câu hỏi/câu trả lời survey tiếng Anh về chủ đề đó (tối thiểu 100 câu hỏi/trả lời)
2- **(6đ)** Translate các câu hỏi và câu trả lời đó qua tiếng Việt, đánh giá các câu dịch và nộp lại bản đánh giá các câu dịch đó trong báo cáo.
3- **(10đ)** Sử dụng các luật (ít nhất 5 luật, nộp các luật này trong báo cáo) nhờ ChatGPT sinh ra các câu hỏi/câu trả lời tương ứng (mỗi câu hỏi gốc sinh ra 5 câu hỏi mới) (tối thiểu 500 câu hỏi/trả lời từ ChatGPT)
4- **(6đ)** Bạn hãy tạo ra phiếu chấm đánh giá các câu hỏi và câu trả lời đó, giải thích cách làm của bạn, nộp lại phiếu đánh giá trong báo cáo.
5- **(2đ)** Lưu bộ dữ liệu đã sinh ra ở trên thành 2 bộ dữ liệu tương ứng: Data, DChatGPT. Mỗi bộ dữ liệu đều có câu hỏi/trả lời tương ứng.

Bài 2 (40 điểm):

Thực hiện các yêu cầu sau:

2.1 **(5đ)** Hãy lựa chọn một chủ đề liên quan đến vấn đề cần giải quyết ở IUH và đặt ra ít nhất 2 câu hỏi có liên quan đến chủ đề ở trên

2.2 **(10đ)** Xây dựng ít nhất 10 câu hỏi khảo sát cho việc trả lời 2 câu hỏi ở câu hỏi ở 2.1, trong đó có 07 câu hỏi với câu trả lời là lựa chọn phương án trả lời từ 1 tới 5 và 03 câu hỏi với câu trả lời là text

2.3 **(10đ)** Thực hiện khảo sát ít nhất 50 bạn trong trường bằng phiếu khảo sát in ra ở 2.2, nộp lại các phiếu này)

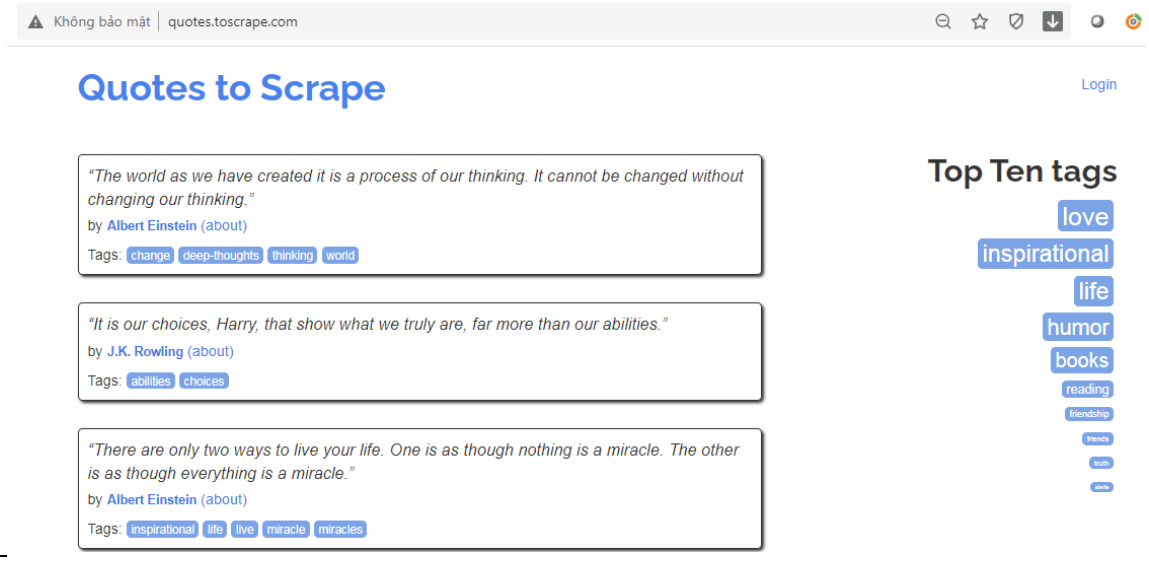
2.4 **(15đ)** Số hóa dữ liệu thu được và Phân tích dữ liệu trên bằng Python**.**
Mỗi phân tích đều bao gồm các nội dung theo yêu cầu sau:

- Mô tả về dữ liệu
- Thực hiện các thống kê căn bản
- Tìm mối tương quan giữa các câu hỏi khảo sát và kết quả
- Xác định các yếu tố quan trọng ảnh hưởng đến kết quả
- Trực quan hóa dữ liệu và kết quả

Bài 3 (30 điểm):

3.1. Thu thập dữ liệu (5đ)

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:



3.1.1 - Bạn hãy viết code cào dữ liệu từ trang web trên, lưu kết quả vào 1 file tương ứng (kq.txt) và vẽ sơ đồ mô tả ngắn gọn về cấu trúc của trang Web trên? **(1đ)**

3.1.2 - 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? **(1đ)**

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? **(1đ)**

c. Hãy viết hàm tacgiaLink() để lấy nội dung của mỗi tác giả. Với mỗi tác giả in ra màn hình các nội dung **(1đ)**

- ✓ 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? **(1đ)**

3.2. Khai phá dữ liệu (25 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 2: Mô tả về bộ dữ liệu Quote.csv

3.2.1. Xử lý dữ liệu- Data Imputation (2 điểm):

- Bạn hãy thêm vào Trường STT và điền tự động dữ liệu của trường này?
- 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?

3.2.2. Khám phá dữ liệu- Data Exploration (10 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.

3.2.3. Trích xuất đặc trưng- Feature Extraction (3 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.

3.2.4. Suy luận (10 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? **(5 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? **(5 điểm)**

II. CÁCH THỨC NỘP BÀI

Đóng gói 3 bài và Báo cáo thành 1 file nén duy nhất đặt tên tiếng Việt không dấu theo cú pháp: Số nhóm-Ho va ten SV1- Ho va ten SV2 và nộp lên LMS theo thời gian quy định.

Ví dụ: NHOM 1-Bui Thanh Hung-Nguyen Phuong Lan.zip

Nội dung của file nén gồm:

- Bài 1: File dữ liệu csv
- Bài 2: File code Python + File Dữ liệu

  50 phiếu khảo sát nộp trực tiếp cho thầy
- Bài 3: File code Python + File Dữ liệu
- Báo cáo file doc với mẫu như gửi đính kèm.

  Trong báo cáo trình bày nội dung Bài 1, Bài 2 và Bài 3 (chỉ báo cáo Phần 3.2)

  Mỗi bài trình bày theo nội dung câu hỏi của mỗi bài.

  Chú ý mỗi câu hỏi nên bao gồm: *giới thiệu, cách tiếp cận, đánh giá, thảo luận, v.v.* phù hợp để bạn hoàn thành các yêu cầu ở mỗi câu hỏi một cách hiệu quả.

  Cuối báo cáo phải bao gồm phần tự chấm điểm như trong mẫu gửi kèm.

III. CHẤM BÀI

Các nhóm sẽ tham gia buổi chấm bài theo thời gian thông báo.