TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO MÔN XỬ LÍ DỮ LIỆU LỚN**

**BÁO CÁO GIỮA KÌ**

*Instructor:* **Gv.BÙI THANH HÙNG**

*Student*: **Nguyễn Thị Lan Anh - 518H0470**

Class **: 18H50204**

Course  **: 22**

**Ho Chi Minh City, 2020**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



THANKS FOR

A difficult semester due to the influence of Covid-19, we would like to express our sincere thanks to Mr. Đỗ Trí Nhật’s efforts for creating lectures and guiding us wholeheartedly to complete. into this subject.

**REPORT IS COMPLETED AT TON DUC THANG UNIVERSITY**

We hereby guarantee that this is our product created under the guidance of Mr.Bui Thanh Hung. The references and results in this report are truthful and have not been published in any way before. The data in the tables for analysis, comments and evaluation collected by the author from different sources are clearly stated in the references section.

In addition, the report also uses a number of comments, assessments as well as data of other authors, other organizations and organizations with citations and origin notes.

**If I detect any fraud I take full responsibility for the content of my project. Ton Duc Thang University is not related to any copyright or copyright infringement caused by me during the implementation process (if any)**.

*Ho Chi Minh City, month day year*

*author*

*(sign and write full name)*

*Nguyễn Thị Lan Anh*

TEACHER'S CONFIRMATION AND EVALUATION SECTION

**The certified part of the instructor**

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

*Ho Chi Minh City, month day year*

(sign and write full name)

**The teacher evaluation section marks**

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

*Ho Chi Minh City, month day year*

(sign and write full nam

1. **Thu thập dữ liệu**

1.1

**Cấu trúc trang web quotetoscrape**:

Trang web có 3 phần chính:

**-header:**

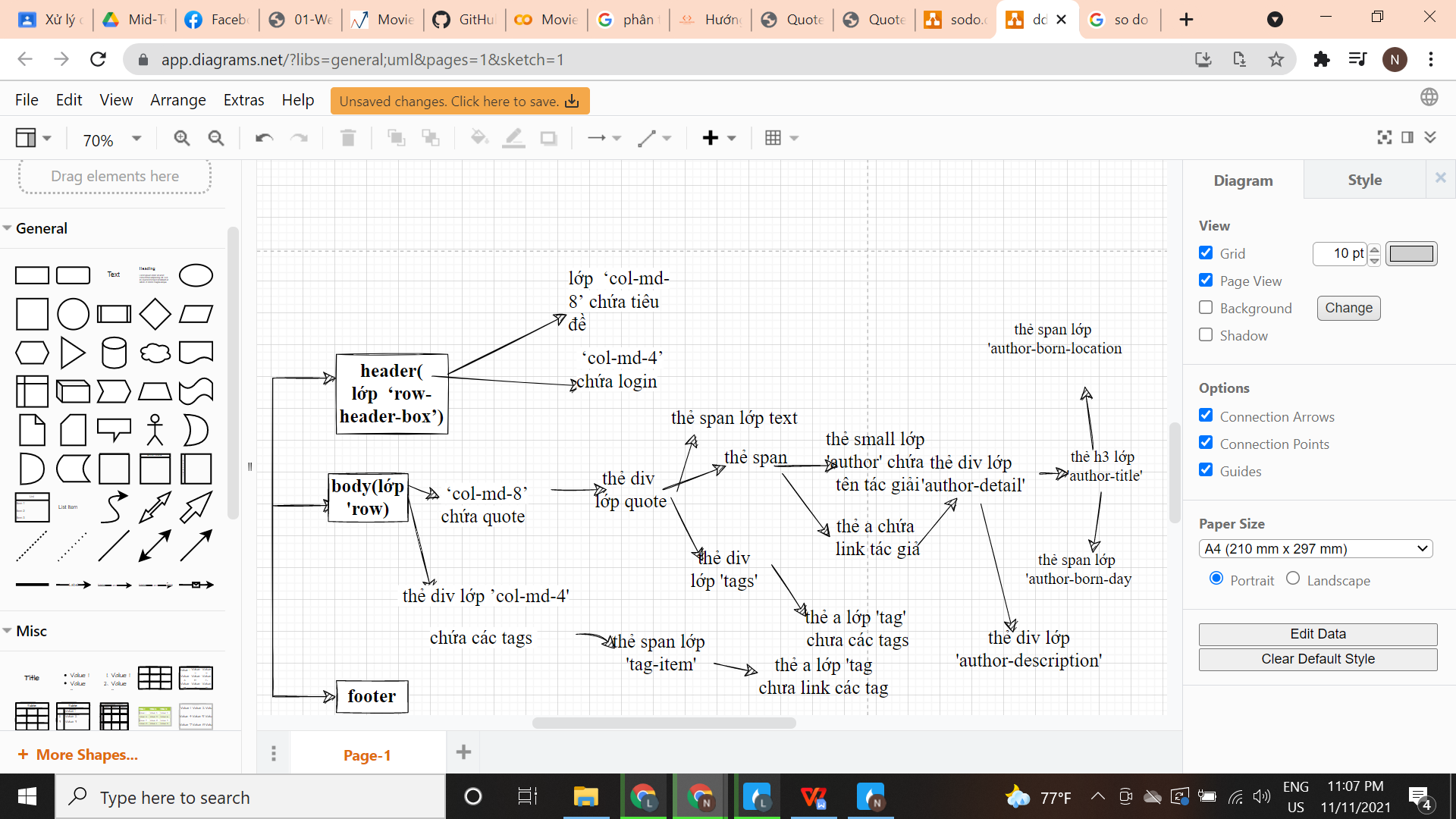
lớp chính là ‘row-header-box’, các lớp con gồm ‘col-md-8’ chứa tiêu đề trang web, lớp ‘col-md-4’ chưa phần login

**-body:**

tất cả các thẻ đều chứa trong thẻ div lớp ‘row’. tiếp theo lớp ‘row’ là 2 lớp con ‘col-md-8’ chưa nội dung thông tin các quote ,’col-md-4 tags-box’ chưa các thẻ tag chủ đề chính cho người dùng lựa chọn.bên trong lớp ‘col-md-8’ chứa tập hợp các lớp ‘quote’ mang thẻ div chứa tất cả thông tin như quote, tác giả, ngày tháng năm sinh,…. mỗi thẻ div lớp ‘quote’ sẽ chứa 1 dữ liệu quote, tác giả, các tags. Bên trong mỗi lớp quote chứa các lớp con như: thẻ span lớp ‘text’ chưa nội dung của quote, tiếp theo là thẻ span chứa thẻ small lớp ‘author’ chứa tên tác giả và link tác giải chứa trong thẻ a. thẻ div lớp ‘tags’ chứa các tags nằm bên trong thẻ a lớp ‘tag’. bên trong thẻ div lớp ’col-md-4 tags-box’ chứa các tên thẻ và liên kết nằm trong thẻ a lớp ‘tag-item’.

**-footer**

**Sơ đồ cấu trúc**



Lấy dữ liệu của tất cả các thẻ div trong 6 trang đầu, sau đo chạy vòng lặp for in ra quote từ thẻ span lớp text, author từ thẻ small lớp author, tags từ thẻ div lớp tags, link từ thẻ a của thẻ div lớp tags. Đưa các dữ liệu vào dataframe và in ra.

**1.2**

1. Dùng lệnh find\_all thẻ div và lớp quote
2. Chạy vòng lặp for, dùng hàm find để tìm thẻ small lớp author và in.
3. D Tương tự câu 1.1. sau đó xuất file bằng hàm to\_csv.
4. **Khai phá dữ liệu**

**2.1 xử lí dữ liệu**

**A.**

- Dùng hàm df.loc để thêm thông tin ngày sinh bằng tên tác giải của cột ‘author’ và cột ‘dob’

df.loc[df['author'] =='Albert Einstein', 'dob'] = 'November 30, 1835'

-Dùng hàm df.loc để thêm thông tin ngày sinh bằng chỉ mục của dòng và tên cột

df.loc[1].at['dob']= 'November 30, 1835'

-Dùng hàm df.iat để thêm thông tin ngày sinh bằng chỉ mục của dòng và cột

df.iat[2, 3]='November 30, 1835'

-Chuyển cột dữ liệu ‘dob; thành dạng dữ liệu datetime

df['dob'] = pd.to\_datetime(df['dob'])

-Sau đó tính tuổi bằng cách lấy ngày tháng năm hiện tại bằng hàm pd.to\_datime(‘today’) trừ đi dữ liệu ngày tháng năm của cột ‘dob’ sau đó gán vào cột ‘age’ của dataframe df để tạo thêm cột ‘age’

df['age'] = today.year - df['dob'].dt.year

**2.2 khám phá dữ liệu**

-thống kê tác giải và câu nói nổi tiếng trong bộ dữ liệu bằng cách tạo dataframe mới tên df\_thongke với dữ liệu láy từ cột ‘quote’ và ‘author’ của dataframe cũa df. Hiển thị 20 tên đầu tiên

df\_thongke =df[['quote', 'author']]

df\_thongke[0:20]

- thống kê độ dài của tên tác giả bằng hàm str.split để tách từ trong tên, sau đó đếm bằng hàm .value\_count()

count = df['author'].str.split().apply(len).value\_counts()

count.index = count.index.astype(str) + ' words:'

count.sort\_index(inplace=True)

Count

-thống kê độ dài của mỗi câu quote tương tự như câu trên

count = df['quote'].str.split().apply(len).value\_counts()

count.index = count.index.astype(str) + ' words:'

count.sort\_index(inplace=True)

Count

Dùng hàm groupby để thống kê số quote dựa theo tên của tác giải

Dùng hàm groupby và get\_group để thống kê số quote dựa theo tên của tác giả được chọn

**2.3 trích xuất dữ liệu**

1. **Giai đoạn "làm sạch" và tiền xử lý dữ liệu dạng văn bản**

-Xóa các ký tự đặc biệt, Các ký tự đặc biệt thường là các ký tự không phải là chữ và số, thường gây "nhiễu" cho dữ liệu. dùng regular expressions ****(regexes)**** để xử lý vấn đề này

-Xóa các stopwords: stopwords là các từ có ít hoặc không có ý nghĩa gì đặc biệt khi xây dựng các đặc trưng. Đây thường là những từ giới từ, trợ từ có tần suất xuất hiện tương đối cao trong một văn bản thông thường ví dụ như: a, an, the... dù không có một danh sách chung tác stopwords tuy nhiên có thể sử dụng thư viện nltk.

-Bên cạnh đó cũng có thể thực hiện một số biện pháp tiêu chuẩn khác như tokenization, xóa bỏ các khoảng trắng thừa, chuẩn hóa chữ cái viết hoa, các biện pháp sửa lỗi chính tả, sửa lỗi ngữ pháp, xóa các ký tự lặp lại

- Thống kê độ phổ biến của từ bằng hàm FreqDist

1. **Giai đoạn trích xuất dữ liệu**

**Sử dụng mô hình Bag of Words**

Đây là mô hình biểu diễn không gian vecto đơn giản nhất cho các văn bản phi cấu trúc. Mô hình không gian vecto chỉ đơn giản là một mô hình toán học để biểu diễn, đại diện cho văn bản (hoặc bất kỳ dạng dữ liệu nào khác) dưới dạng vecto số sao cho mỗi chiều của vecto là một đặc trưng của dữ liệu. Mô hình **Bag of words** biểu diễn cho mỗi mẫu dữ liệu văn bản dưới dạng một vecto số trong đó mỗi chiều là một từ cụ thể trong kho dữ liệu và giá trị có thể là tần số của nó xuất hiện trong đoạn văn bản (giá trị có thể là 0 hoặc 1) hoặc thậm chí là các giá trị có trọng số. Tên mô hình này là **Bag of words** thể hiện theo đúng nghĩa đen của nó nghĩa là một túi các từ, không quan tâm đến trật tự, trình tự, ngữ pháp.

**xác định độ quan trọng của một từ trong các câu quote bằng TF-IDF**

 thuật ngữ TF-IDF ([term frequency - inverse document frequency](http://en.wikipedia.org/wiki/Tf%E2%80%93idf" \t "https://viblo.asia/p/_blank)) là một phương thức thống kê được biết đến rộng rãi nhất để xác định độ quan trọng của một từ trong đoạn văn bản trong một tập nhiều đoạn văn bản khác nhau. Nó thường được sử dụng như một trọng số trong việc khai phá dữ liệu văn bản. TF-IDF chuyển đổi dạng biểu diễn văn bản thành dạng không gian vector (VSM), hoặc thành những vector thưa thớt

1. **Giai đoạn chia dữ liệu thành các tập train, test**
2. **Giai đoạn áp dụng thuật toán để dự đoán tên tác giải**

**Thuật toán Random Forests**

Random Forests là thuật toán học có giám sát (supervised learning). Nó có thể được sử dụng cho cả phân lớp và hồi quy. Nó cũng là thuật toán linh hoạt và dễ sử dụng nhất. Một khu rừng bao gồm cây cối. Người ta nói rằng càng có nhiều cây thì rừng càng mạnh. Random forests tạo ra cây quyết định trên các mẫu dữ liệu được chọn ngẫu nhiên, được dự đoán từ mỗi cây và chọn giải pháp tốt nhất bằng cách bỏ phiếu.

nó là một phương pháp tổng hợp (dựa trên cách tiếp cận phân chia và chinh phục) của các cây quyết định được tạo ra trên một tập dữ liệu được chia ngẫu nhiên. Bộ sưu tập phân loại cây quyết định này còn được gọi là rừng. Cây quyết định riêng lẻ được tạo ra bằng cách sử dụng chỉ báo chọn thuộc tính như tăng thông tin, tỷ lệ tăng và chỉ số Gini cho từng thuộc tính. Mỗi cây phụ thuộc vào một mẫu ngẫu nhiên độc lập. Trong bài toán phân loại, mỗi phiếu bầu chọn và lớp phổ biến nhất được chọn là kết quả cuối cùng. Trong trường hợp hồi quy, mức trung bình của tất cả các kết quả đầu ra của cây được coi là kết quả cuối cùng. Nó đơn giản và mạnh mẽ hơn so với các thuật toán phân loại phi tuyến tính khác.

**tính toán các độ đo accuracy, f1-score,..**

**Giai đoạn tính độ tương đồng của văn bản**

### **Document Similarity**

Document Similarity (hay độ tương tự của văn bản) là quá trình sử dụng số liệu dựa trên khoảng cách hoặc độ tương tự có thể sử dụng để xác định mức độ tương đương của một văn bản với bất kỳ văn bản nào khác dựa trên các đặc trưng được trích xuất ra từ **bag of words** hoặc **tf-idf**.

**Phân cụm tài liệu sử dụng đặc trưng tương tự**

Phân cụm **(clustering)** tận dụng phương pháp học không giám sát để nhóm các điểm dữ liệu (trong trường hợp này là các mẫu văn bản) thành các nhóm hoặc các cụm. Chúng ta sẽ sử dụng thuật toán phân cụm không giám sát để thử nhóm các mẫu văn bản tương tự từ kho văn bản đã có trước đó. Có hai phương pháp phân cụm chính là **agglomerative** và **divisive methods**

Bài này em sẽ dùng phương pháp **agglomerative**. Đây là phương pháp phâm cụm từ dưới lên trên, tức là mỗi mẫu hoặc tài liệu ban đầu được xếp vào các cụm riêng biệt và được nối với nhau bằng cách sử dụng các thang đo khoảng cách và các tiêu chuẩn để hợp nhất liên kết

Việc lựa chọn tiêu chí để hợp nhất các liên kết quyết định rất lớn đến kết quả của bài toán. Một vài tiêu chuẩn liên kết như **Ward**, liên kết hoàn chỉnh **(Complete linkage)**, liên kết trung bình **(Average linkage)...** Các tiêu chí này nhằm mục đích chọn các cặp cụm dữ liệu ở tầng thấp nhất và ở các cụm cao hơn để hợp nhất dựa trên giá trị tối ưu của hàm mục tiêu. Trong ví dụ sau đây, chúng ta sẽ chọn phương pháp phương sai tối thiểu **Ward** để làm tiêu chí liên kết các cặp cụm với mục đích giảm thiểu tổng phương sai trong cụm đó.