TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÀI TẬP GIỮA KỲ**

**MÔN XỬ LÍ DỮ LIỆU LỚN**

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

*Người thực hiện*: **Hoàng Gia Huy – 518H0626**

Lớp **: 18H50302**

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**



**BÀI TẬP GIỮA KỲ**

**MÔN XỬ LÍ DỮ LIỆU LỚN**

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

*Người thực hiện*: **Hoàng Gia Huy – 518H0626**

Lớp **: 18H50302**

Khoá  **: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

LỜI CẢM ƠN

Em xin gửi lời cảm ơn chân thành và sự tri ân sâu sắc đối với các thầy cô của trường Đại học Tôn Đức Thắng, đặc biệt là các thầy cô khoa Công Nghệ Thông Tin của trường đã tạo điều kiện cho em thực tập ở khoa để có nhiều thời gian cho khóa luận tốt nghiệp. Và em cũng xin chân thành cảm ơn thầy Bùi Thanh Hùng đã nhiệt tình hướng dẫn hướng dẫn em hoàn thành tốt dự án giữa kì này.

Trong quá thực hiện dự án, cũng như là trong quá trình làm bài báo cáo, khó tránh khỏi sai sót, rất mong các thầy, cô bỏ qua. Đồng thời do trình độ lý luận cũng như kinh nghiệm thực tiễn còn hạn chế nên bài báo cáo không thể tránh khỏi những thiếu sót, em rất mong nhận được ý kiến đóng góp thầy, cô để em học thêm được nhiều kinh nghiệm và sẽ hoàn thành tốt hơn trong thời gian sắp tới sắp tới.

Em xin chân thành cảm ơn!

**DỰ 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 thầy 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 11 tháng 11 năm 2021*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Hoàng Gia Huy*

PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN

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

Tp. Hồ Chí Minh, ngày tháng năm 2021

(kí và ghi họ tên)

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

**DANH MỤC HÌNH**

**DANH MỤC BẢNG**

MỤC LỤC

[LỜI CẢM ƠN 3](#_Toc87678793)

[PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN 1](#_Toc87678794)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ i](#_Toc87678795)

[MỤC LỤC 1](#_Toc87678796)

[CHƯƠNG 1 – THU THẬP DỮ LIỆU 3](#_Toc87678797)

[1.1 Cào dữ liệu từ trang web 3](#_Toc87678798)

[1.1.1 Cách thức thực hiện 3](#_Toc87678799)

[1.1.1.1 Tiểu mục cấp 3 3](#_Toc87678800)

[1.1.1.2 Tiểu mục cấp 3 tiếp theo. 3](#_Toc87678801)

[1.1.2 Sơ đồ cấu trúc của trang web 3](#_Toc87678802)

[1.2 Thực hiện các yêu cầu 4](#_Toc87678803)

[1.2.1 Đọc tất cả các thẻ div với lớp là "quote" và lưu vào biến result 4](#_Toc87678804)

[1.2.2 Tìm và in ra tên của các tác giả 4](#_Toc87678805)

[1.2.3 Viết hàm lấy nội dung của từng tác giả 4](#_Toc87678806)

[1.2.4 Viết hàm lấy nội dung của từng tác giả 6](#_Toc87678807)

[CHƯƠNG 2 – KHAI PHÁ DỮ LIỆU 7](#_Toc87678808)

[2.1 Xử lý dữ liệu – data imputation 7](#_Toc87678809)

[2.1.1 Đề xuất cách điền ngày sinh chưa có 7](#_Toc87678810)

[2.1.1.1 Tìm ra giá trị theo quy luật 7](#_Toc87678811)

[2.1.1.2 Xoá các giá trị còn thiếu 7](#_Toc87678812)

[2.1.2 Thêm trường tuổi và để xuất cách tính tuổi của tác giả 7](#_Toc87678813)

[2.2 Khám phá dữ liệu – Data exploration 8](#_Toc87678814)

[2.2.1 Thống kê về độ tuổi của từng tác giả 8](#_Toc87678815)

[2.2.2 Thống kê về top 10 tác giả xuất hiện nhiều nhất 8](#_Toc87678816)

[2.2.3 Vẽ đám mây từ khoá 9](#_Toc87678817)

[2.2.4 Thống kê câu dài nhất, ngắn nhất và độ dài trung bình 9](#_Toc87678818)

[2.3 Trích xuất đặc trưng – Feature extraction 9](#_Toc87678819)

[2.3.1 Giới thiệu 9](#_Toc87678820)

[2.3.1.1 Frequency-based embedding 10](#_Toc87678821)

[2.3.1.2 Prediction-based embedding 10](#_Toc87678822)

[2.3.2 Cách tiếp cận 10](#_Toc87678823)

[2.3.2.1 Tiền xử lý dữ liệu văn bản 10](#_Toc87678824)

[2.3.2.2 Term Frequency – Inverse Document Frequency 10](#_Toc87678825)

[2.3.2.3 Word2vec 12](#_Toc87678826)

[2.4 Suy luận 12](#_Toc87678827)

[2.4.1 Chia bộ dữ liệu thành tập train/test 13](#_Toc87678828)

[2.4.2 Huấn luyện mô hình 13](#_Toc87678829)

[2.4.3 Tìm tác giả cách phong cách viết tương đồng 13](#_Toc87678830)

CHƯƠNG 1 – THU THẬP DỮ LIỆU

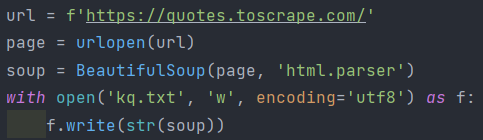
1.1 Cào dữ liệu từ trang web

Dữ liệu về những câu nói của những người nổi tiếng trên thế giới có ở đường link: http://quotes.toscrape.com/. Mỗi trang web hiển thị 10 câu nói của người nỗi tiếng tương ứng này.

1.1.1 Cách thức thực hiện

Để cào dữ liệu từ trang web em sử dụng những thư viện như urllib để gửi 1 request đến trang web và dung thư viện BeautifulSoup để lấy ra thông tin từ trang web.

Tất cả dữ liệu vừa cào về sẽ được lưu vào file kq.txt



1.1.1.1 Tiểu mục cấp 3

- Đây là cấp tiểu mục nhỏ nhất, không thể tiếp tục phân chia.

- Các ý trong tiểu mục được trình bày gạch đầu dòng “-“.

- Các ý nhỏ hơn sử dụng bullet như sau:

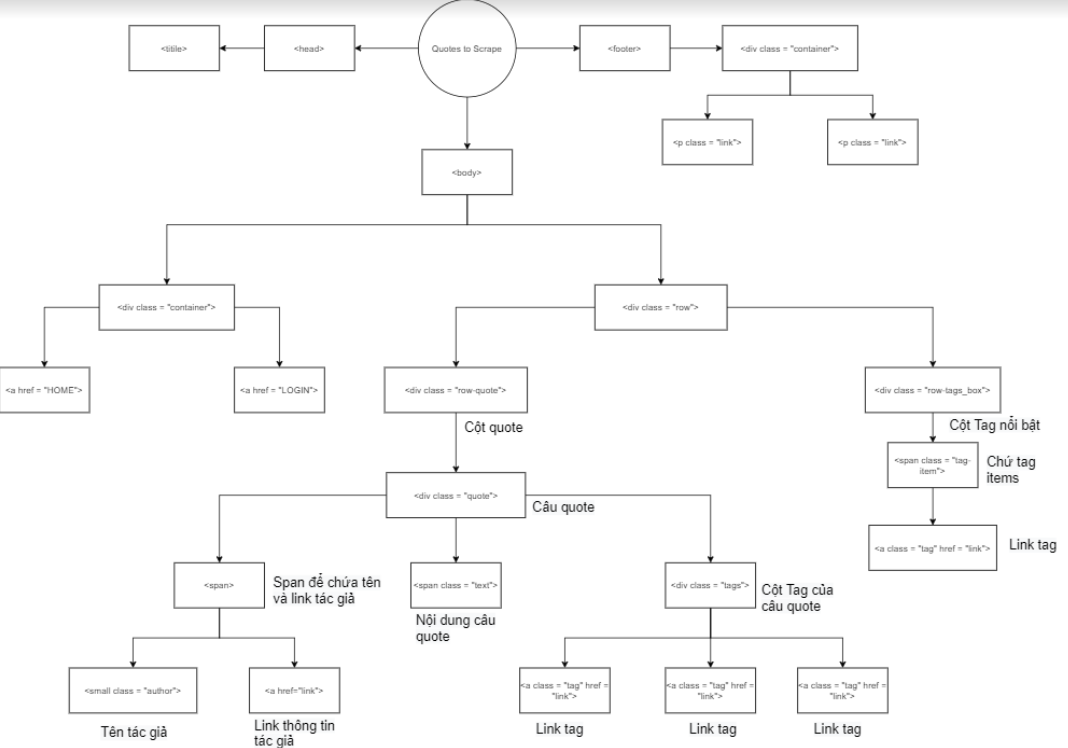
* Ý nhỏ 1.
* Ý nhỏ 2.

- Cần lưu ý rằng đây là cấp sâu nhất, không được phép chia thành 1.1.1.1.1 .

1.1.1.2 Tiểu mục cấp 3 tiếp theo.

Nội dung của tiểu mục thứ ba, khi soạn thảo hãy dùng Styles có sẵn, để khi tạo mục lục sẽ tự động và đồng nhất mỗi khi chúng ta thay đổi format.

1.1.2 Sơ đồ cấu trúc của trang web



1.2 Thực hiện các yêu cầu

Chương này trình bày lý do chọn đề tài, mục đích, đối tượng và phạm vi nghiên cứu, ý nghĩa khoa học và thực tiễn của đề tài; cơ sở khoa học của việc chọn đề tài...;

1.2.1 Đọc tất cả các thẻ div với lớp là "quote" và lưu vào biến result

****

1.2.2 Tìm và in ra tên của các tác giả



1.2.3 Viết hàm lấy nội dung của từng tác giả

Một ‘div :quote’ sẽ chứa tất cả các thông tin của một tác giả như:

* Tên tác giả
* Link
* Câu nói nỗi tiếng của tác giả

Nên việc để lấy các thông tin em chỉ việc dùng find để tìm các thông tin mong muốn trong các thẻ và các thẻ đó chứa các class như ‘author’…..

Riêng ngày tháng năm sinh chúng ta cần truy cập lại trang web chứa đường link cụ thể của 1 tác giả (vd: http://quotes.toscrape.com/author/Albert-Enstein/ ) sau đó chỉ việc đi đến các thẻ span và chọn class : author-bor -date để lấy được ngày tháng năm sinh

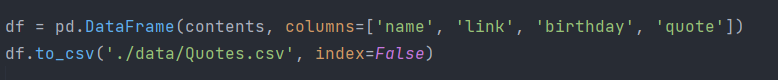


1.2.4 Viết hàm lấy nội dung của từng tác giả

Em thực hiện cào 100 câu từ trang 1 đến trang 10 và lưu nó vào 1 danh sách contents.



Sau đó em sử dụng pandas để chuyển nó về dạng dataframe và lưu toàn bộ nội dung em vừa cào được vào file Quotes.csv

****

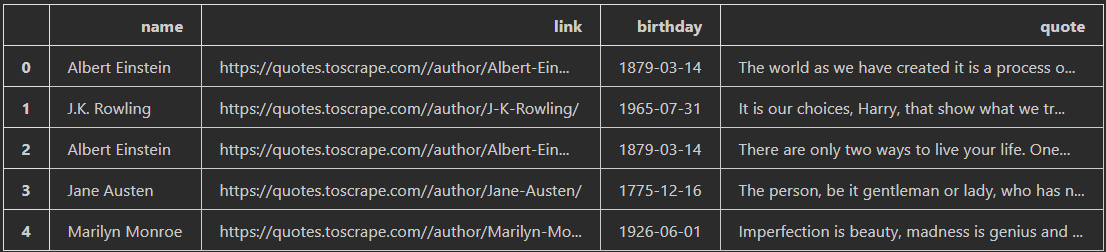
CHƯƠNG 2 – KHAI PHÁ DỮ LIỆU

*Tổng quan*: phân tích, đánh giá các công trình nghiên cứu đã có của tác giả, các tác giả khác trong và ngoài nước liên quan mật thiết đến đề tài; nêu những vấn đề còn tồn tại; chỉ ra những vấn đề mà đề tài cần tập trung nghiên cứu, giải quyết;

2.1 Xử lý dữ liệu – data imputation

2.1.1 Đề xuất cách điền ngày sinh chưa có

Sau khi xem qua bộ dữ liệu thì em thấy không có trường hợp ngày sinh bị thiếu.



Trong trường hợp bộ dữ liệu bị thiếu ta có thể sử dụng những cách như:

2.1.1.1 Tìm ra giá trị theo quy luật

- Sử dụng fillna() có thể điền vào các giá trị bị thiếu bằng những cách như:

* Điền vào giá trị trung bình trên bộ dữ liệu
* Điền voà giá trị xuất hiện nhiều nhất từ bộ dữ liệu

2.1.1.2 Xoá các giá trị còn thiếu

Sử dụng dropna() để loại bỏ toàn bộ các giá trị bị thiếu.

2.1.2 Thêm trường tuổi và để xuất cách tính tuổi của tác giả

Để tính được trường tuổi em thì em cào dữ liệu trực tiếp từ wiki bằng cách:

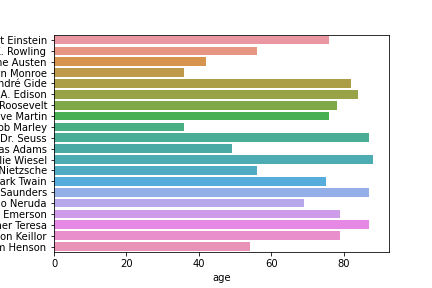
- Tìm các các cái tag id có giá trị là ‘P590’ có nghĩa là ngày mất của tác giả

- Lấy ra ngày mất của tác giả.

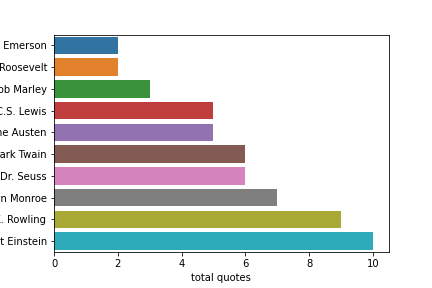
* Nếu giá trị là None đồng nghĩ với việc tác giả còn sống hoặc không biết ngày mất thì tuổi sẽ bằng năm hiện tại trừ cho năm sinh
* Nếu giá trị có thể lấy được thì tuổi sẽ bằng năm sinh trừ năm mất

2.2 Khám phá dữ liệu – Data exploration

2.2.1 Thống kê về độ tuổi của từng tác giả



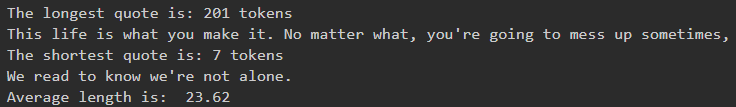
2.2.2 Thống kê về top 10 tác giả xuất hiện nhiều nhất



2.2.3 Vẽ đám mây từ khoá

****

2.2.4 Thống kê câu dài nhất, ngắn nhất và độ dài trung bình

******

2.3 Trích xuất đặc trưng – Feature extraction

2.3.1 Giới thiệu

Trích xuất đặc trưng là việc đưa dữ liệu thô như đoạn văn bản không có cấu trúc thành dữ liệu có cấu trúc mà thuật toán có thể học được. Với việc dữ liệu là văn bản ta thường sử dụng phương pháp đó là Word Embedding.

Word Embedding là tên gọi chung của các mô hình ngôn ngữ và các phương pháp học theo đặc trưng trong Xử lý ngôn ngữ tự nhiên(NLP), ở đó các từ hoặc cụm từ được ánh xạ sang các vector số (thường là số thực). Đây là một công cụ đóng vai trò quan trọng đối với hầu hết các thuật toán, kiến trúc Machine Learning, Deep Learning trong việc xử lý Input ở dạng text, do chúng chỉ có thể hiểu được Input ở dạng là số, từ đó mới thực hiện các công việc phân loại, hồi quy,vv…

Word Embedding được phân chủ yếu thành 2 loại:

* Frequency-based embedding.
* Prediction-based embedding.

2.3.1.1 Frequency-based embedding

- Frequency-based Embedding: dựa vào tần số xuất hiện của các từ để tạo ra các vector từ, trong đó có 2 loại phổ biến nhất:

* Bag of words
* Term Frequency-Inverse Document Frequency

2.3.1.2 Prediction-based embedding

Prediction-based embedding: là xây dựng các vector từ dựa vào các mô hình dự đoán. Tiêu biểu nhất chính là Word2vec, nó là sự kết hợp của 2 mô hình: CBOW (Continous Bag Of Words) và Skip-gram.

2.3.2 Cách tiếp cận

2.3.2.1 Tiền xử lý dữ liệu văn bản

- **Tiền xử lý dữ liệu văn bản** là cách làm sạch bộ dữ liệu bằng cách loại bỏ nhưng dấu câu, khoảng trắng dư thừa, thay đổi từ về 1 dạng nào đó, xoá các từ không cần thiết,…

- Ở trong bài làm em đã sử dụng những phương pháp đó là:

+ Loại bỏ dấu câu

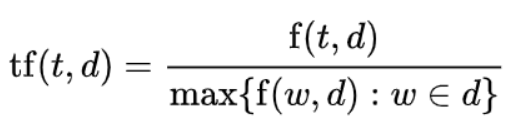
+ Chuyển từ viết hoa thành viết thường

+ Chuyển các từ về từ loại của nó

2.3.2.2 Term Frequency – Inverse Document Frequency

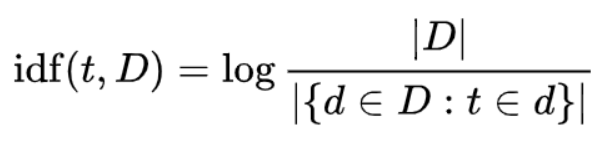
- **Term Frequency – Inverse Document Frequency(TFIDF)** là 1 kĩ thuật sử dụng trong khai phá dữ liệu văn bản. Trọng số này được sử dụng để đánh giá tầm quan trọng của một từ trong một văn bản. Giá trị cao thể hiện độ quan trọng cao và nó phụ thuộc vào số lần từ xuất hiện trong văn bản nhưng bù lại bởi tần suất của từ đó trong tập dữ liệu. Một vài biến thể của tf-idf thường được sử dụng trong các hệ thống tìm kiếm như một công cụ chính để đánh giá và sắp xếp văn bản dựa vào truy vấn của người dùng, Nó gồm có 2 phần đó là:

* **TF** (Tần suất xuất hiện của từ): là số lần từ xuất hiện trong văn bản. Vì các văn bản có thể có độ dài ngắn khác nhau nên một số từ có thể xuất hiện nhiều lần trong một văn bản dài hơn là một văn bản ngắn



Trong đó:

* tf(t, d): tần suất xuất hiện của từ t trong văn bản d
* f(t, d): Số lần xuất hiện của từ t trong văn bản d
* max({f(w, d) : w ∈ d}): Số lần xuất hiện của từ có số lần xuất hiện nhiều nhất trong văn bản d
* **IDF** (Nghịch đảo tần suất của văn bản): giúp đánh giá tầm quan trọng của một từ . Khi tính toán TF , tất cả các từ được coi như có độ quan trọng bằng nhau.



Trong đó:

* idf(t, D): giá trị idf của từ t trong tập văn bản
* |D|: Tổng số văn bản trong tập D
* |{d ∈ D : t ∈ d}|: thể hiện số văn bản trong tập D có chứa từ t.

2.3.2.3 Word2vec

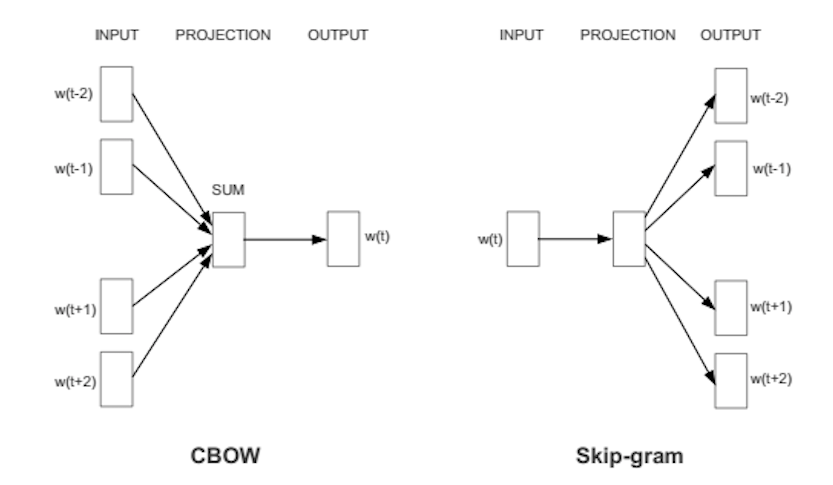
Word2vec là mô hình học trực tiếp word vector có số chiều thấp trong quá trình dự đoán các từ xung quanh mỗi từ. Đặc điểm của phương pháp này là nhanh hơn và có thể dễ dàng kết hợp một câu một văn bản mới hoặc thêm vào từ vựng.

Word2vec sử dụng một mạng neural 2 lớp với duy nhất 1 tầng ẩn, lấy đầu vào là một corpus lớn và sinh ra không gian vector (với số chiều khoảng vài trăm), với mỗi từ duy nhất trong corpus được gắn với một vector tương ứng trong không gian.

Các word vectors được xác định trong không gian vector sao cho những từ có chung ngữ cảnh trong corpus được đặt gần nhau trong không gian. Dự đoán chính xác cao về ý nghĩa của một từ dựa trên những lần xuất hiện trước đây.

Có 2 cách xây dựng word2vec:

* Sử dụng ngữ cảnh để dự đoán mục tiêu(CBOW).
* Sử dụng một từ để dự đoán ngữ cảnh mục tiêu(skip-gram).



2.4 Suy luận

2.4.1 Chia bộ dữ liệu thành tập train/test

Ở phần này em chia bộ dữ dữ liệu với tỷ lệ 80% train và 20% để test

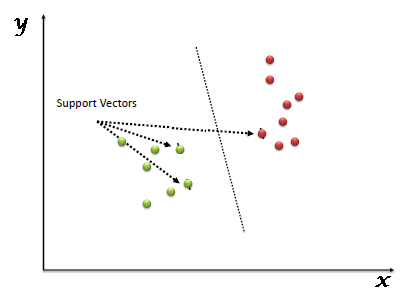
Với X danh sách cách câu nói đã được chuyển về dạnh vector và y là danh sách các tác giả đã chuyển về dạng mã hoá 0, 1, 3… bằng Label encoder



2.4.2 Huấn luyện mô hình

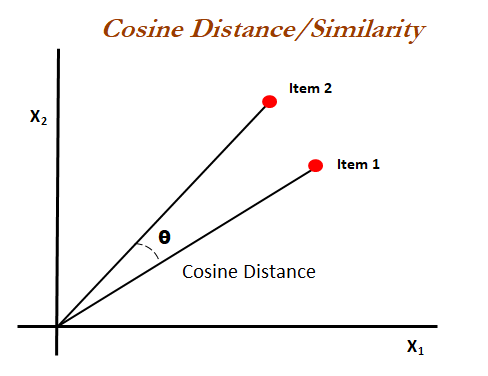
Mô hình mà em sử dụng để huấn luyện đó là mô hình Support Vector Machine (SVM). Mô hình SVM là một mô hình có giám sát, nó có thể sử dụng cho cả việc phân loại hoặc đệ quy. Tuy nhiên nó được sử dụng chủ yếu cho việc phân loại.

Trong mô hình này, chúng ta vẽ đồi thị dữ liệu là các điểm trong n chiều ( ở đây n là số lượng các tính năng bạn có) với giá trị của mỗi tính năng sẽ là một phần liên kết. Sau đó chúng ta thực hiện tìm “siêu phăng" (hyper-plane) phân chia các lớp. Hyper-plane nó chỉ hiểu đơn giản là 1 đường thẳng có thể phân chia các lớp ra thành hai phần riêng biệt.

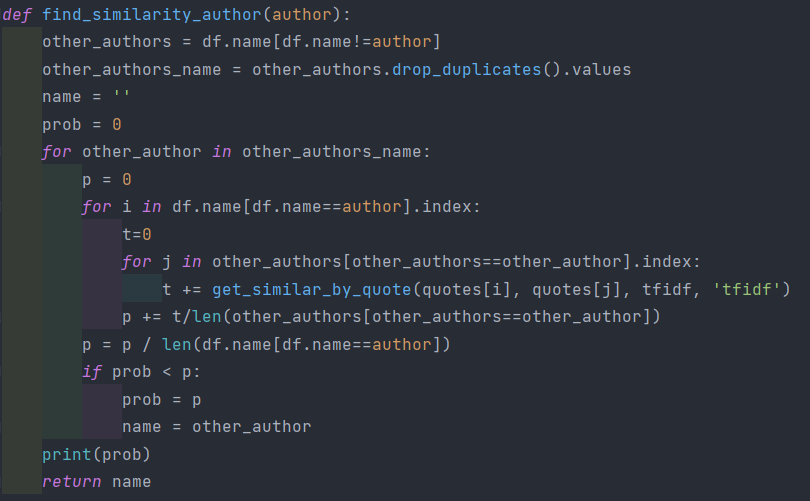


2.4.3 Tìm tác giả cách phong cách viết tương đồng

Để tính độ tương đồng trong phong cách viết của các tác giả ta sử dụng độ đo cosine similarity. Nếu như góc tạo thành giữa 2 vector câu nói càng gần nhau thì độ tương đồng trong phong cách viết của 2 tác giả càng lớn.



Thuật toán để tìm tác giả có phong cách viết tương đồng



**TÀI LIỆU THAM KHẢO**

**Tiếng Việt**

1. [Machine Learning cơ bản (machinelearningcoban.com)](https://machinelearningcoban.com/2017/04/09/smv/)
2. [SƠ LƯỢC WORD EMBEDDING (viblo.asia)](https://viblo.asia/p/so-luoc-word-embedding-gDVK2RAeKLj)
3. [Feature Engineering (Phần 4): Phương pháp xử lý truyền thống với dữ liệu dạng văn bản (Text Data) (viblo.asia)](https://viblo.asia/p/feature-engineering-phan-4-phuong-phap-xu-ly-truyen-thong-voi-du-lieu-dang-van-ban-text-data-1Je5EvWYKnL)
4. [TF-IDF là gì? Code demo thuật toán TF-IDF với dữ liệu tiếng Việt (nguyenvanhieu.vn)](https://nguyenvanhieu.vn/tf-idf-la-gi/)

**Tiếng Anh**

1. [models.word2vec – Word2vec embeddings — gensim (radimrehurek.com)](https://radimrehurek.com/gensim/models/word2vec.html)
2. [python - Parse birth and death dates from Wikipedia? - Stack Overflow](https://stackoverflow.com/questions/12250580/parse-birth-and-death-dates-from-wikipedia?fbclid=IwAR0zhy8WoN7sA7i8nLzbQxN_-ipOyd9GGfo4xHEVte98FFPfjN-kypuX8cc)
3. [Linguistic Features · spaCy Usage Documentation](https://spacy.io/usage/linguistic-features)

**PHỤ LỤC**

Phần này bao gồm những nội dung cần thiết nhằm minh họa hoặc hỗ trợ cho nội dung luận văn như số liệu, biểu mẫu, tranh ảnh. . . . nếu sử dụng những câu trả lời cho một *bảng câu hỏi thì bảng câu hỏi mẫu này phải được đưa vào phần Phụ lục ở dạng nguyên bản* đã dùng để điều tra, thăm dò ý kiến; **không được tóm tắt hoặc sửa đổi**. Các tính toán mẫu trình bày tóm tắt trong các biểu mẫu cũng cần nêu trong Phụ lục của luận văn. Phụ lục không được dày hơn phần chính của luận văn