BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH
KHOA CÔNG NGHỆ THÔNG TIN

TIỂU LUẬN MÔN HỌC

Tìm hiểu và báo cáo Mô hình Word2Vec

Sinh viên thực hiện: PHẠM MẠNH TRUNG NGUYÊN
MSSV: 2100000168
Chuyên ngành: Kỹ thuật phần mềm
Môn học: Xử lý ngôn ngữ tự nhiên
Khóa: 2021
Giảng viên hướng dẫn: Ts. Nguyễn Tuấn Đăng

Tp.HCM, Thứ bảy ngày 11 tháng 09 năm 2021

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Điểm tiểu luận:

.....

.....

.....

.....

TP.HCM, Ngày tháng 09 năm 2021

Giáo viên hướng dẫn

(Ký tên, đóng dấu)

LỜI CẢM ƠN

Em xin gửi lời cảm ơn chân thành đến Thầy Nguyễn Tuấn Đăng, giảng viên Bộ môn Xử lý ngôn ngữ tự nhiên - Trường ĐH Nguyễn Tất Thành. Người đã tận tình hướng dẫn, chỉ bảo em trong suốt quá trình học tập cũng như làm tiểu luận, báo cáo.

Tuy báo cáo đã được hoàn thành nhưng không thể tránh khỏi những sự thiếu sót, vì vậy rất mong nhận được sự góp ý của thầy để em có thể bổ sung thêm kiến thức và hoàn thiện hơn cho những dự án sau này.

Em xin chân thành cảm ơn và chúc thầy nhiều sức khỏe!

TP.HCM, ngày 10 tháng 09 năm 2021

Sinh viên thực hiện

Phạm Mạnh Trung Nguyên

LỜI MỞ ĐẦU

Xử lý ngôn ngữ tự nhiên là đang là một chủ đề gây nhức nhối trong ngành công nghệ thông tin. Một phần vì sự đa dạng của ngôn ngữ, phần vì sự tiện ích của nó.

Các nhà phát triển đã và đang tìm cách xử lý triệt để vấn để này để có thể áp dụng vào các bài toán lớn nhưng vơi sự phong phú của ngôn ngữ cũng như đa ngôn ngữ thì việc đó chưa bao giờ là dễ.

Qua tìm hiểu em và nhờ sự hướng dẫn của thầy em đã tìm hiểu và biết được cách xử lý sơ lược các văn bản. Vậy nên hôm nay em xin trình bày báo cáo em đã tìm hiểu đó là “Mô hình Word2Vec” của Gensim.

MỤC LỤC

DANH MỤC CÁC BẢNG HÌNH

DANH MỤC CÁC BẢNG BIỂU

I. TỔNG QUAN

1. Xử lý ngôn ngữ tự nhiên:

Là một nhánh của trí tuệ nhân tạo tập trung vào các ứng dụng trên ngôn ngữ của con người. Trong trí tuệ nhân tạo thì xử lý ngôn ngữ tự nhiên là một trong những phần khó nhất vì nó liên quan đến việc phải hiểu ý nghĩa ngôn ngữ - công cụ hoàn hảo nhất của tư duy và giao tiếp.

2. Python

Python là ngôn ngữ lập trình hướng đối tượng, cấp cao, mạnh mẽ, được tạo ra bởi Guido van Rossum. Nó dễ dàng để tìm hiểu và đang nổi lên như một trong những ngôn ngữ lập trình nhập môn tốt nhất cho người lần đầu tiếp xúc với ngôn ngữ lập trình. Python hoàn toàn tạo kiểu động và sử dụng cơ chế cấp phát bộ nhớ tự động. Python có cấu trúc dữ liệu cấp cao mạnh mẽ và cách tiếp cận đơn giản nhưng hiệu quả đối với lập trình hướng đối tượng. Cú pháp lệnh của Python là điểm cộng vô cùng lớn vì sự rõ ràng, dễ hiểu và cách gõ linh động làm cho nó nhanh chóng trở thành một ngôn ngữ lý tưởng để viết script và phát triển ứng dụng trong nhiều lĩnh vực, ở hầu hết các nền tảng.

II. GENSIM

1. Khái quát:

Gensim là một thư viện dùng để mô hình hóa, lập chỉ mục tài liệu và rút trích tính tương tự với ngữ liệu (corpora) lớn. Đối tượng sử dụng Gensim chủ yếu là cộng đồng xử lý ngôn ngữ tự nhiên và rút trích thông tin.

Gensim bao gồm các thực thi song song các thuật toán như fastText, word2vec và doc2vec, cũng như các phân tích ngữ nghĩa (LSA, LSI, SVD), phân rã trận không âm (NMF), phân bổ Dirichlet tiềm ẩn (LDA), các phép chiếu ngẫu nhiên và tf-idf.

Các khái niệm cốt lõi của Gensim là:
- **Document**: Một văn bản (thường là một chuỗi [string]).
- **Corpus**: Tập hợp các **Docment**.
- **Vecter**: Biểu diễn toán học của một **Docment**.
- **Model**: Một thuật toán để biến đổi cách biểu diễn của các **Vector**.

2. Corpus:

2.1. Khái niệm:

Corpus là 1 dữ liệu tập hợp các văn bản, ngôn ngữ đã được số hoá. Cách dịch thông thường ở VN là "kho ngữ liệu". Ví dụ về corpus như "tuyển tập các tác phẩm của Nam Cao", hay "tuyển tập ca từ của Trịnh Công Sơn",…

Các corpus là 1 tài nguyên quan trọng trong NLP. Từ các corpus, ta có thể rút ra những dữ liệu quan trọng sau :
- Chiết suất 1 cách tự động các qui tắc ngữ pháp "văn mạch tự do".
- Tính toán được xác suất, tần suất xuất hiện của các từ.

Để đảm bảo tính chính xác cho 2 kết luận trên, corpus phải đảm bảo 1 số nguyên tắc nhất định :
- Tính đại diện : các thành phần trong corpus phải có tính phổ quát, đa dạng và phong phú.
- Kích thước : kích thước của corpus càng lớn thì càng được đánh giá cao.

Dựa vào mục đích, cách xây dựng corpus, người ta chia corpus thành các loại sau :
- Corpus thô (raw corpus): đơn giản chỉ là tập hợp các dữ liệu mà không có xử lý gì thêm.
- Corpus được gắn nhãn (tagged corpus) : các dữ liệu trong corpus đã được xử lý như phân tích từ, phân tích cú pháp, gắn nhãn từ loại, …
- Parallel Corpus : được sử dụng nhiều trong ứng dụng máy dịch.

Ngoài cách chia trên, Có thể chia corpus theo cấu tạo của nó:
- Corpus biệt lập : dữ liệu lấy vào 1 cách ngẫu nhiên, biệt lập và không phân biệt với nhau.
- Corpus theo danh mục : dựa vào các danh mục để chia dữ liệu trong corpus thành các nhóm.
- Corpus trùng lặp : các dữ liệu trong corpus có thể ở nhiều nhóm cùng lúc.
- Corpus theo thời gian : các dữ liệu sắp xếp theo thời gian thu thập và thời gian xuất hiện.

2.2. Thống kê trong corpus:

Khái niệm về n-gram : là tần suất xuất hiện của n kí tự (hoặc từ) liên tiếp nhau có trong dữ liệu của corpus.

Với n = 1 và tính trên kí tự, ta có thông tin về tần suất xuất hiện nhiều nhất của các chữ cái. Điều này ứng dụng để làm keyboard : các phím hay xuất hiện nhất sẽ ở những vị trí dễ sử dụng nhất.

Với n = 2, ta có khái niệm bigram. Ví dụ với các chữ cái tiếng Anh, ‘th’,’he’,’in’,’an’,’er’ là các cặp kí tự hay xuất hiện nhất. Ngoài ra, ta có thể biết thêm rằng sau kí tự ‘q’ thì phần lớn đều là kí tự ‘u’.

Với n = 3, ta có trigram. Nhưng vì n càng lớn thì số trường hợp càng lớn nên thường người ta chỉ sử dụng với n = 1,2 hoặc đôi lúc là 3. Ví dụ với các kí tự tiếng Anh, tiếng Anh sử dụng 26 kí tự, vậy với n = 1 thì số trường hợp là 26, n = 2 thì số trường hợp là 26^2 = 676 trường hợp, n = 3 có 17576 trường hợp.

Bigram được sử dụng nhiều trong việc phân tích hình thái (từ, cụm từ, từ loại) cho các ngôn ngữ khó phân tích như tiếng Việt, tiếng Nhật, tiếng Trung, … Dựa vào tần suất xuất hiện cạnh nhau của các từ, người ta sẽ tính cách chia 1 câu thành các từ sao cho tổng bigram là cao nhất có thể. Với thuật giải phân tích hình thái dựa vào trọng số nhỏ nhất, người ta sử dụng n = 1 để xác định tuần suất xuất hiện của các từ và tính trọng số.

Để đảm bảo tính thống kê chính xác đòi hỏi các corpus phải lớn và có tính đại diện cao.

3. Vector:

Để suy ra cấu trúc tiềm ẩn trong **corpus**, chúng ta cần một cách để biểu diễn các tài liệu mà chúng ta có thể thao tác bằng toán học. Một cách tiếp cận là biểu diễn mỗi tài liệu dưới dạng vectơ của các đối tượng.

Ví dụ: một “đối tượng” có thể được coi là một cặp câu hỏi - câu trả lời, ví dụ:

- Từ “chơi” xuất hiện bao nhiêu lần trong tài liệu? 0.
- Tài liệu gồm bao nhiêu đoạn văn? 2.
- Văn bản sử dụng bao nhiêu phông chữ? 5.

Câu hỏi thường chỉ được biểu diễn bằng id số nguyên của nó (như 1 , 2 và 3). Biểu diễn của tài liệu này sau đó trở thành một loạt các cặp. Đây được gọi là vectơ dày đặc , vì nó chứa câu trả lời rõ ràng cho mỗi câu hỏi trên, như câu trên có thể đưa về vectơ:
- (1, 0.0), (2, 2.0), (3, 5.0)

Nếu biết được tất cả các câu hỏi (thông qua các id), Có thể ẩn đi. Như câu trên có thể rút gọn thành:
- (0, 2, 5)

Trong thực tế, vectơ thường bao gồm nhiều giá trị bằng 0. Để tiết kiệm bộ nhớ, Gensim bỏ qua tất cả các phần tử vectơ có giá trị 0.0. Đây được gọi là vectơ thưa thớt hoặc vectơ nhiều từ. Giá trị của tất cả các tính năng bị thiếu trong biểu diễn thưa thớt này có thể được giải quyết rõ ràng thành 0, Như:
- (2, 2.0), (3, 5.0) | 0.0

Giả sử các câu hỏi giống nhau, chúng ta có thể so sánh các vectơ của hai tài liệu khác nhau với nhau. Có thể kết luận rằng các tài liệu tương ứng với các vectơ đó cũng tương tự. Tất nhiên, tính đúng đắn của kết luận đó phụ thuộc vào mức độ chọn các câu hỏi ngay từ đầu, ví dụ:
- (0.0, 2.0, 5.0) (0.1, 1.9, 4.9)

Một cách tiếp cận khác để biểu diễn tài liệu dưới dạng vectơ là mô hình bag-of-words . Trong mô hình bag-of-words, mỗi tài liệu được biểu thị bằng một vectơ có chứa tần số đếm của mỗi từ trong từ điển. Ví dụ:
- Ta có các từ: ['trà', 'sữa', 'đường', 'thìa']
- Và câu "trà sữa thìa trà"
=> Vectơ: [2, 1, 1, 0]

* Lưu ý: các từ không có trong trong ngữ liệu gốc sẽ không được đưa vào vectơ hóa. Cũng lưu ý rằng vectơ này chỉ chứa các mục nhập cho các từ thực sự xuất hiện trong tài liệu. Bởi vì bất kỳ tài liệu nhất định nào sẽ chỉ chứa một vài từ trong số rất nhiều từ trong từ điển, các từ không xuất hiện trong vector hóa được biểu thị bằng 0 như một biện pháp nhằm tiết kiệm không gian.

III. MÔ HÌNH WORD2VEC

1. Khái quát:

Là một thuật toán được sử dụng rộng rãi dựa trên mạng nơ-ron, thường được gọi là "học sâu" (mặc dù bản thân word2vec khá nông). Sử dụng một lượng lớn văn bản thuần túy không có chú thích, word2vec tự động học các mối quan hệ giữa các từ. Đầu ra là các vectơ, một vectơ trên mỗi từ, với các mối quan hệ tuyến tính đáng chú ý cho phép chúng ta thực hiện những việc như:

vec ("vua") - vec ("đàn ông") + vec ("phụ nữ")
= ~ vec ("hoàng hậu")
vec ("Montreal Canadaiens") - vec ("Montreal") + vec ("Toronto")
= ~ vec ("Toronto Maple Leafs").

Word2vec rất hữu ích trong việc gắn thẻ văn bản tự động , hệ thống giới thiệu và dịch máy.

2. Ưu điểm:

Như đã đề cập ở trên, mô hình bag-of-words có hiệu quả rất hiệu nghiệm. Tuy nhiên lại có vài điểm yếu, ví dụ ta có 2 câu như sau:

"Nam thích xem phim, Lan cũng thích xem phim"

"Nam thích bóng đá, Lan thì không"

Từ hai câu xuất thành các Vectơ với thứ tự các phần tử như sau:

1	2	2	2	1	1	0	0	0	0
1	1	0	0	1	0	1	1	1	1
Nam	thích	xem	phim	Lan	cũng	bóng	đá	thì	không

Bảng 1. Vectơ

Điểm yếu thứ nhất là mô hình bag-of-words không lưu thông tin về thứ tự từ vựng vậy nên hai câu "Nam thích Lan" và "Lan thích Nam" tương ứng với các vectơ như nhau. Có một giải pháp cho việc này là sử sụng mô hình túi n-gram để xét các cụm từ có độ dài n để đại diện cho tài liệu có dạng vectơ có độ dài cố định để nắm bắt trật tự. Tuy nhiên sẽ gây nên tính thưa thớt dữ liệu và làm tăng dung lượng lưu trữ.

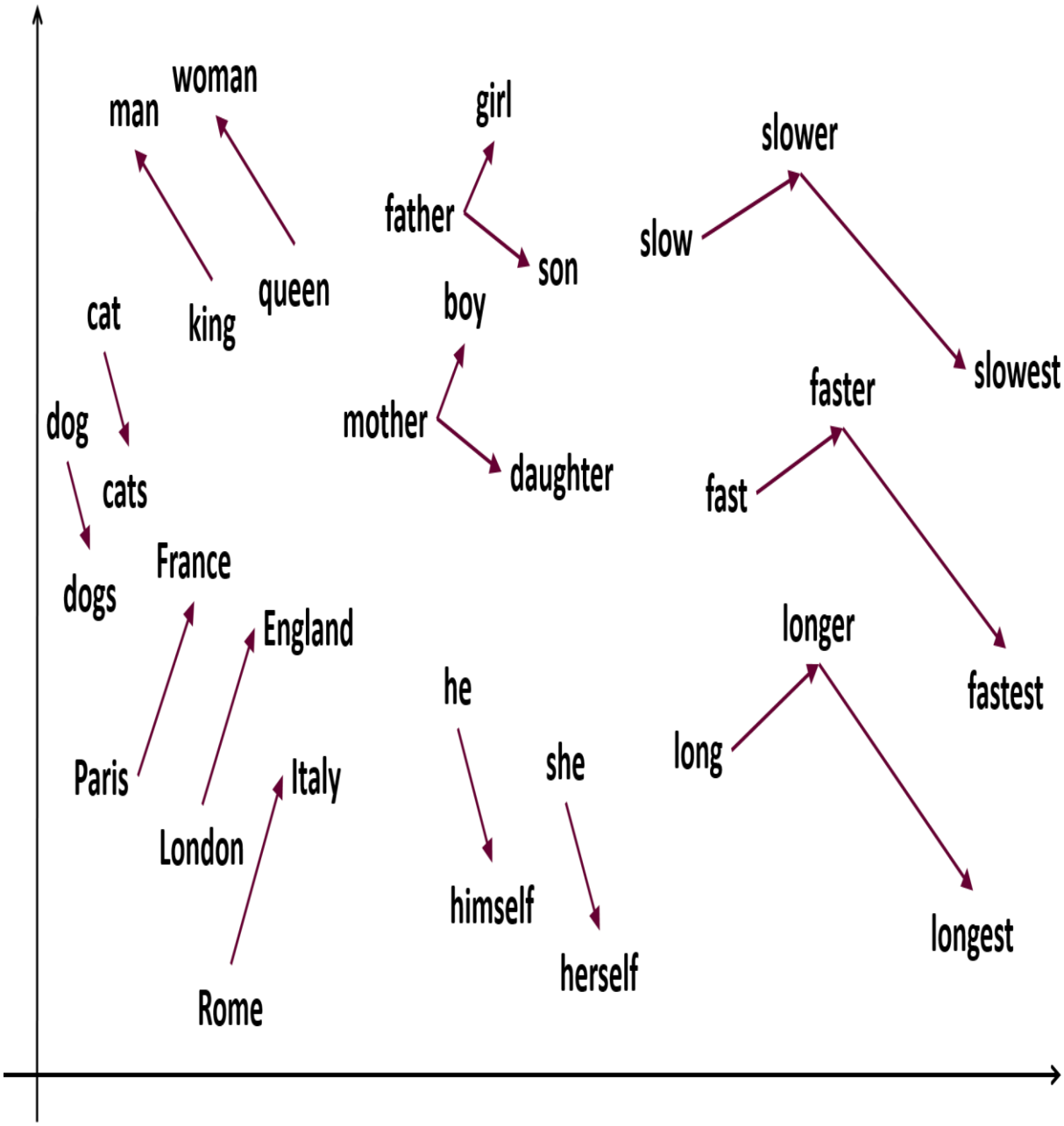
Điểm yếu thứ hai là mô hình bag-of-words không tìm hiểu ý nghĩa của các từ, do đó khoảng cách của các vectơ không phải lúc nào cũng phản ánh sự khác biệt sự khác biệt về nghĩa được.

=> Mô hình Word2Vec ra đời để giải quyết vấn đề thứ hai này!

3. Mô hình Word2Vec:

Là một kỹ thuật xử lý ngôn ngữ tự nhiên, sử dụng một mô hình mạng thần kinh để học các liên kết từ (sự liên quan của từ) từ một **Corpus** có dung lượng lớn. Sau khi được huấn luyện, mô hình có thể phát hiện các từ đồng nghĩa hoặc gợi ý các từ bổ sung cho một phần của câu. Với cái tên nói lên tất cả, word2vec thể hiện cho mỗi từ riêng biệt với một danh sách cụ thể của các số được gọi là vectơ. Các vectơ được lựa chọn cẩn thận sao cho một hàm toán học đơn giản sẽ (độ tương tự cosin giữa các vectơ) cho biết mức độ của độ tương tự ngữ nghĩa giữa các từ được biểu diễn bằng các vectơ đó. Ví dụ:

Từ "mạnh" và từ "khỏe" sẽ gần nhau hơn so với "mạnh" với "Thành phố".



Hình 1.Mô hình Word2Vec

* Mô hình có 2 phiên bản:

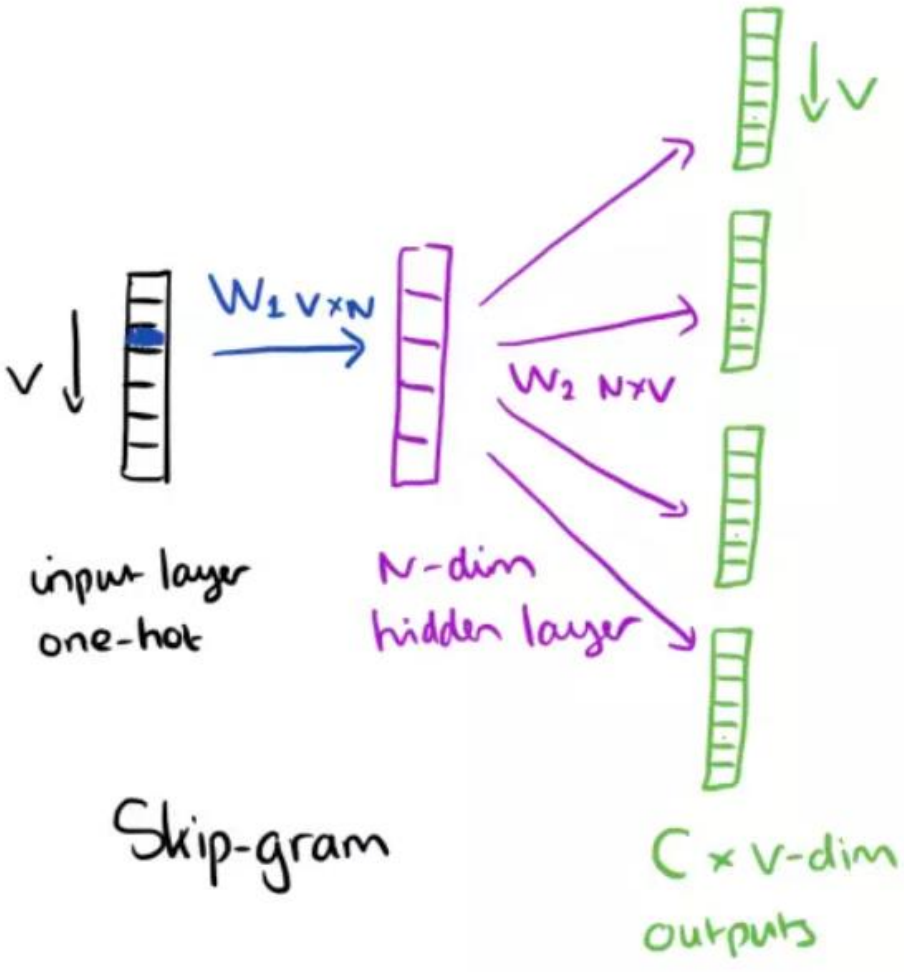
- **Skip-grams** (SG): Được tạo theo cặp (word1, word2) được tạo ra bằng cách di chuyển một cửa sổ trên dữ liệu văn bản, đào tạo nên một mạng lưới thần kinh **1 lớp ẩn** dựa trên các tổng hợp cho một từ ở đầu vào, cho chúng ta phân phối xác suất dự đoán của các từ lân cận với đầu vào. Một mã hóa ảo một nóng của các từ đi qua 'lớp chiếu' đến lớp ẩn; các trọng số chiếu này sau này được hiểu là từ nhúng. Vì vậy, nếu

lớp ẩn có 300 nơ-ron, mạng này sẽ cung cấp cho chúng ta các nhúng từ 300 chiều.

* Input là từ cần tìm mối quan hệ, output là từ các từ có quan hệ gần nhất với từ đó. Ví dụ câu:

"I have a cute dog"

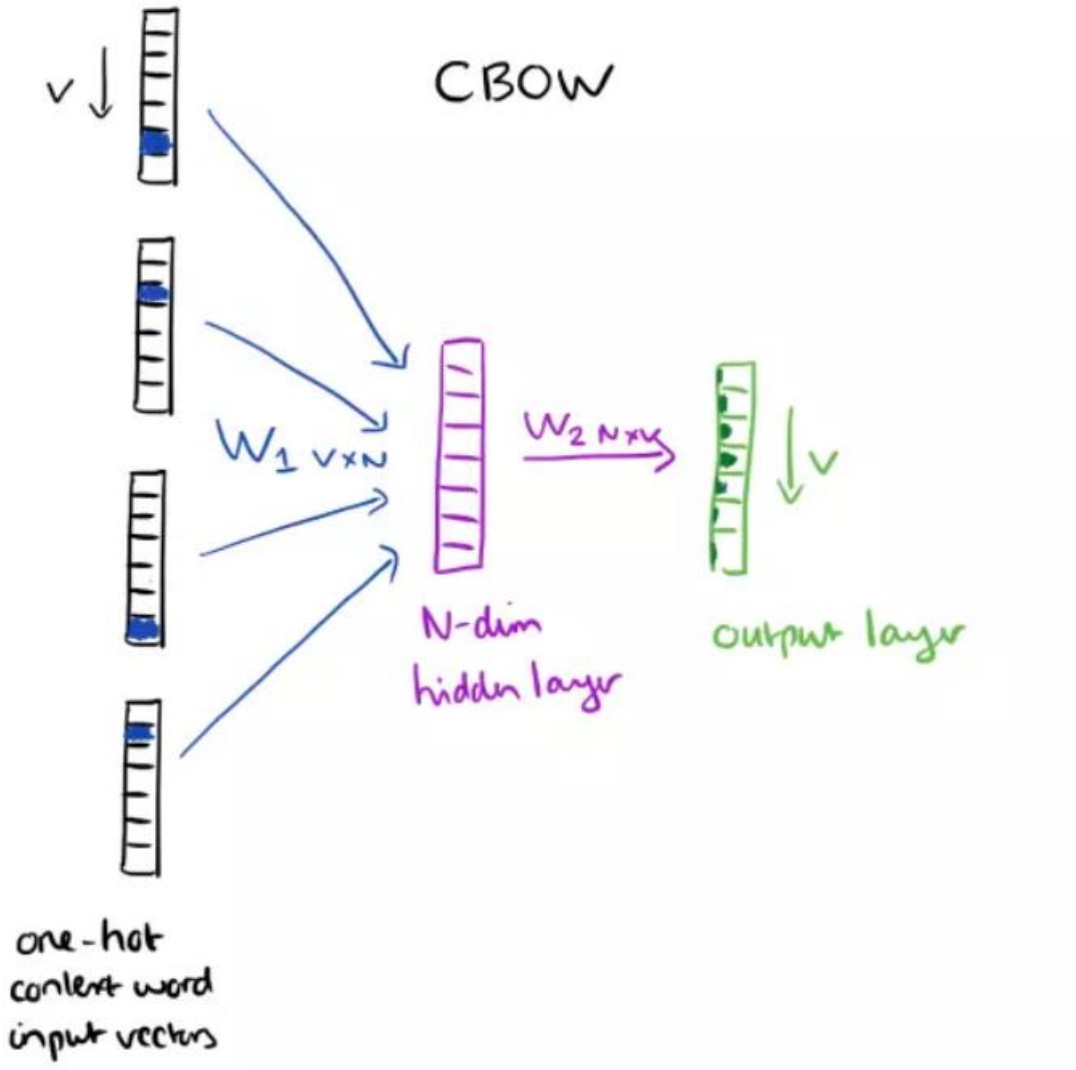
- input từ "a"
- output là “I”, “have”, “cute”, “dog”.



Hình 2. Mô hình Skip-gram

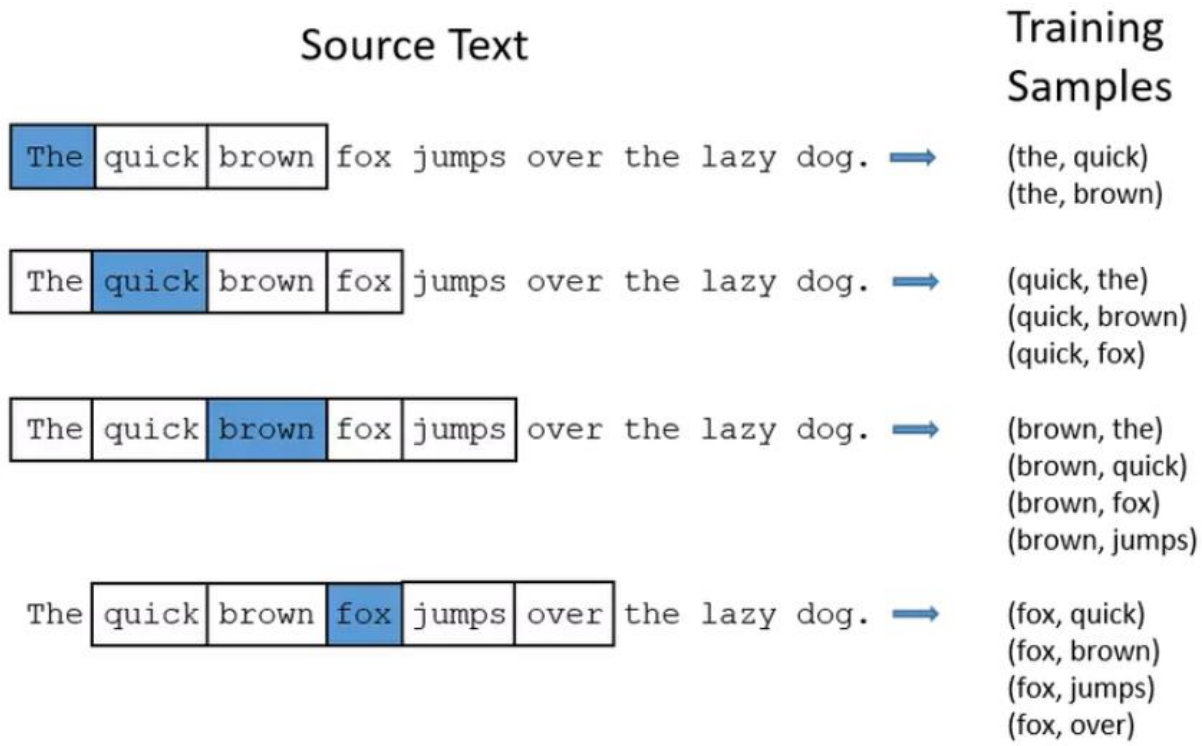
- **Continuous-bag-of-words (CBOW):** Word2vec liên tục giống với mô hình **Skip-grams**. Nó cũng là mạng nơ-ron **1 lớp ẩn**. Nhiệm vụ đào tạo và tổng hợp sử dụng giá trị trung bình của nhiều từ ngữ cảnh đầu

vào, thay vì một từ đơn lẻ như trong **Skip-grams** để dự đoán từ trung tâm. Thì trọng số phép chiếu biến các từ nóng thành vectơ trung bình, có cùng độ rộng với lớp ẩn, được hiểu là từ nhúng.
* input context, output là từ gần nhất với contenxt đó.

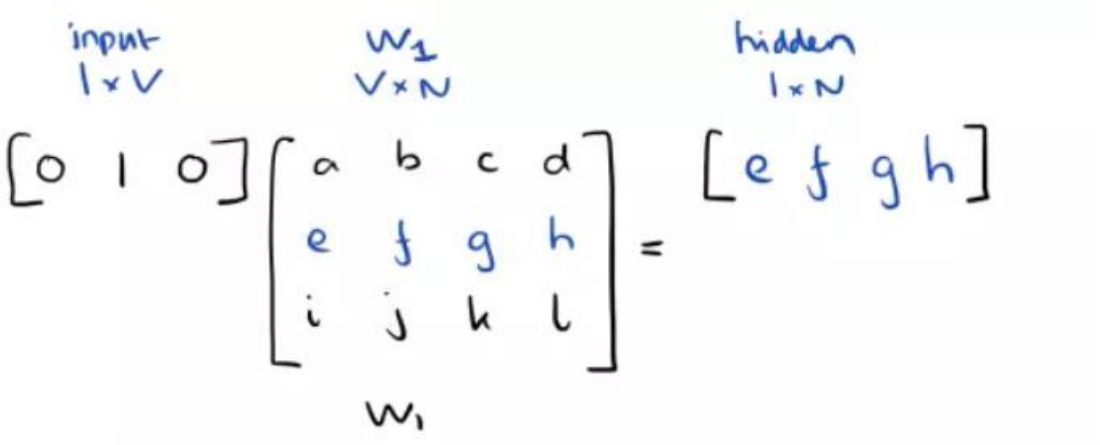


Hình 3. Mô hình Continuous-Bag-of-Words

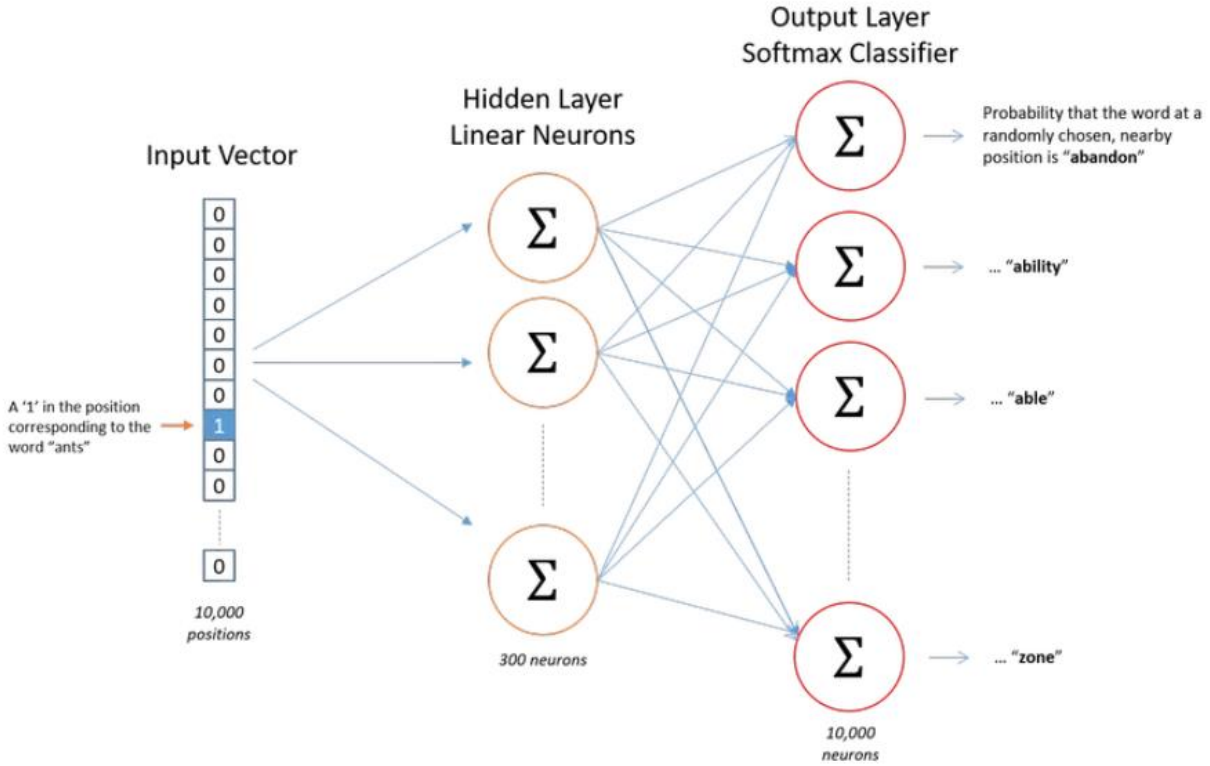
4. Mô hình đào tạo:



Hình 4. Mô hình đào tạo (1)



Hình 5. Mô hình đào tạo (2)



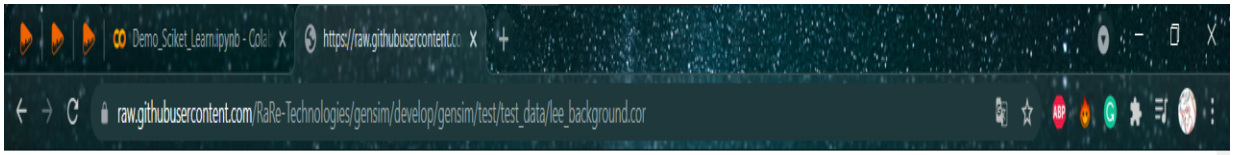
$$y_k = \Pr(word_k \mid word_{context}) = \frac{\exp(activation(k))}{\sum_{n=1}^{V} \exp(activation(n))}$$

Hình 6. Mô hình đào tạo (3)

V. DEMO MÔ HÌNH WORD2VEC

Để Demo mô hình, trước hết ta cần một mô hình đã được đào tạo trước: Mô hình Word2Vec được đào tạo trên một phần của tập dữ liệu Google Tin tức, bao gồm khoảng 3 triệu từ và cụm từ. Một mô hình như vậy có thể mất hàng giờ để đào tạo, nhưng vì nó đã có sẵn nên việc tải xuống và tải nó bằng Gensim chỉ mất vài phút.

Link corpus: https://github.com/RaRe-Technologies/gensim/blob/develop/gensim/test/test_data/lee_background.cor



Hundreds of people have been forced to vacate their homes in the Southern Highlands of New South Wales as strong winds today pushed a huge bushfire towards the town of Hill Top. A new blaze near Goulburn, south-west of Sydney, has forced the closure of the Hume Highway. At about 4:00pm AEDT, a marked deterioration in the weather as a storm cell moved east across the Blue Mountains forced authorities to make a decision to evacuate people from homes in outlying streets at Hill Top in the New South Wales southern highlands. An estimated 500 residents have left their homes for nearby Mittagong. The New South Wales Rural Fire Service says the weather conditions which caused the fire to burn in a finger formation have now eased and about 60 fire units in and around Hill Top are optimistic of defending all properties. As more than 100 blazes burn on New Year's Eve in New South Wales, fire crews have been called to new fire at Gunning, south of Goulburn. While few details are available at this stage, fire authorities says it has closed the Hume Highway in both directions. Meanwhile, a new fire in Sydney's west is no longer threatening properties in the Cranebrook area. Rain has fallen in some parts of the Illawarra, Sydney, the Hunter Valley and the north coast. But the Bureau of Meteorology's Claire Richards says the rain has done little to ease any of the hundred fires still burning across the state. "The falls have been quite isolated in those areas and generally the falls have been less than about five millimetres," she said. "In some places really not significant at all, less than a millimetre, so there hasn't been much relief as far as rain is concerned. "In fact, they've probably hampered the efforts of the firefighters more because of the wind gusts that are associated with those thunderstorms."

Indian security forces have shot dead eight suspected militants in a night-long encounter in southern Kashmir. The shootout took place at Dora village some 50 kilometers south of the Kashmiri summer capital Srinagar. The deaths came as Pakistani police arrested more than two dozen militants from extremist groups accused of staging an attack on India's parliament. India has accused Pakistan-based Lashkar-e-Taiba and Jaish-e-Mohammad of carrying out the attack on December 13 at the behest of Pakistani military intelligence. Military tensions have soared since the raid, with both sides massing troops along their border and trading tit-for-tat diplomatic sanctions. Yesterday, Pakistan announced it had arrested Lashkar-e-Taiba chief Hafiz Mohammed Saeed. Police in Karachi say it is likely more raids will be launched against the two groups as well as other militant organisations accused of targetting India. Military tensions between India and Pakistan have escalated to a level not seen since their 1971 war.

The national road toll for the Christmas-New Year holiday period stands at 45, eight fewer than for the same time last year. 20 people have died on New South Wales roads, with eight fatalities in both Queensland and Victoria. Western Australia, the Northern Territory and South Australia have each recorded three deaths, while the ACT and Tasmania remain fatality free.

Argentina's political and economic crisis has deepened with the resignation of its interim President who took office just a week ago. Aldolfo Rodregiuez Saa told a stunned nation that he could not rescue Argentina because key fellow Peronists would not support his default on massive foreign debt repayment or his plan for a new currency. It was only a week ago that he was promising a million new jobs to end four years of recession, days after his predecessor resigned following a series of failed rescue packages. After announcing that the senate leader, Ramon Puerta, would assume the presidency until congress appoints a new caretaker president, the government said he too had quit and another senior lawmaker would act in the role. Fresh elections are not scheduled until March leaving whoever assumes the presidency with the daunting task of tackling Argentina's worst crisis in 12 years, but this time, isolated by international lending agencies.

Six midwives have been suspended at Wollongong Hospital, south of Sydney, for inappropriate use of nitrous oxide during work hours, on some occasions while women were in labour. The Illawarra Area Health Service says that following an investigation of unprofessional conduct, a further four midwives have been relocated to other areas within the hospital. The service's chief executive officer, Tony Sherbon, says no one was put at risk, because other staff not involved in the use of nitrous oxide were able to take over caring for women in labour. "Well we're very concerned and the body of midwives to the hospital - there are over 70 midwives that work in our service - are very annoyed and angry at the inappropriate behaviour of these very senior people who should know better," he said. "And that's why we've take the action of suspending them and we'll consider further action next week."

The Federal Government says it should be safe for Afghani asylum seekers in Australia to return home when the environment becomes secure. The Government has suspended their applications while the interim government is established in Kabul. The Foreign Affairs Minister Alexander Downer has refused to say for how long the claims process has been put on hold. But he says the major threat to most people seeking asylum is no longer there. "Many Afghans who have tried to get into Australia or for that matter into Britain and other countries in north-west Europe have claimed that they are fleeing the Taliban," he said. "Well, the Taliban is no longer in power in Afghanistan, the Taliban is finished." Meanwhile, there has been a mass airlift of detainees from Christmas Island to the Pacific Island of Nauru. In total, more than 300 people have been flown from the island in two operations using chartered aircraft. The second airlift today delivered 180 asylum seekers to Nauru where they will await processing of their claims for temporary visas. The Department of Immigration says there are now 211 detainees remaining on Christmas Island. A spokesman says a decision regarding their future is yet to be made.

The United States team of Monica Seles and Jan-Michael Gambill scored a decisive victory over unseeded France in their first Hopman Cup match at Burswood Dome in Perth. The pair, runners-up in the $1 million dollar mixed teams event last year, both won their singles encounters to give the US an unbeatable 2-0 lead. The 28-year-old Seles, currently ranked eighth, recovered from a shaky start to overpower Virginie Razzano, 18, who is ranked 72nd. Seles had to fight hard to get home in straight sets, winning 6-3, 6-4 in 62 minutes. Then the 24-year-old Gambill (ranked 21st) wore down a determined Arnaud Clement, 24, (18th) to win 6-4, 6-4 in 91 minutes. The Americans are aiming to go one better than last year when they were beaten by Swiss pair Martina Hingis and Roger Federer in the final of the eight-nation contest. Gambill said the win was a great way to start the tennis year. "I got a little tentative at the end, but it was a great start to my year," he said. "Arnaud is a great scrapper and I am delighted to beat him, even though I am frankly a bit out of shape. "That is one of the reasons I am here. I will be in shape by the end of the tournament. "I just aim to keep improving in the new year, and if I do, I think I have a chance to beat anyone when I am playing well." Gambill was pressed hard by Clement before taking the first set in 47 minutes. But the American gained the ascendancy in the second set, breaking in the third and fifth games. Seles said she had expected her clash with Razzano to be tough. "She was a top junior player in the world, so it was no surprise that she fought so well," she said. Seles said she still had the hunger to strive to regain her position at the top of her sport. "This is why you play," she said. "But I want to try not to peak too early this season. "Seles, slow into her stride, slipped to 2-3 in her opening set against Razzano but recovered quickly, claiming the set after snatching four games in a row. In the second set, Seles broke her opponent in the opening game and completed victory with relative ease despite Razzano's tenacious efforts.

Hundreds of canoeists are enjoying hard-earned New Years Eve celebrations following five days paddling in the Murray River Marathon. After more than 400 exhausting kilometres battling hot and dusty conditions, a Melbourne team took out the strongly contested K-one Cup this afternoon. Two hundred and thirty canoes left Yarrawonga last Thursday, with just over 200 paddling across the finishing line at Swan Hill. The others fell victim to the gruelling conditions. The event raised about $170,000 for the Red Cross.

There has been welcome relief for firefighters in New South Wales overnight with milder weather allowing them to strengthen containment lines around the most severe fires. But fire authorities are not getting overly optimistic as dry and hot weather is forecast to continue. The weather bureau is forecasting temperatures in the high 30s and westerly winds until at least Friday, which means fire authorities are reluctant to get too excited about last night's favourable conditions. Marks Sullivan from the Rural Fire Service says fire fighters are remaining on guard. "A lot of fires that have been burning in the areas around Sydney and the north coast and further south have been burning within areas that are known and are contained," he said. "However, that's not to say that these fires won't pose a threat given the weather conditions that are coming up over the next few days." Despite the caution, the Rural Fire Service says most of the state's fires that threaten property are burning within containment lines. Greater Sydney is

Hình 7. Thư viện mẫu corpus

Nạp dữ liệu:

```
import gensim.downloader as api
wv = api.load('word2vec-google-news-300')
```

Bảng 2. Nạp dữ liệu

Lấy dữ liệu (20 từ đầu tiên):

```
for index, word in enumerate(wv.index2word):
    if index == 20:
        break
    print(f"từ vựng #{index}/{len(wv.index2word)} là {word}")
```

```
từ vựng #0/3000000 là </s>
từ vựng #1/3000000 là in
từ vựng #2/3000000 là for
từ vựng #3/3000000 là that
từ vựng #4/3000000 là is
từ vựng #5/3000000 là on
từ vựng #6/3000000 là ##
từ vựng #7/3000000 là The
từ vựng #8/3000000 là with
từ vựng #9/3000000 là said
từ vựng #10/3000000 là was
từ vựng #11/3000000 là the
từ vựng #12/3000000 là at
từ vựng #13/3000000 là not
từ vựng #14/3000000 là as
từ vựng #15/3000000 là it
từ vựng #16/3000000 là be
từ vựng #17/3000000 là from
từ vựng #18/3000000 là by
từ vựng #19/3000000 là are
```

Bảng 3. Lấy dữ liệu

Có thể lấy được vectơ cho các thuật ngữ mà mô hình đã đào tạo

```
vec_queen = wv['queen']
```

Bảng 4. Lấy Vectơ các từ quen thuộc

Điểm yếu của Word2Vec là không thể lấy vectơ từ không quen thuộc

```
try:
    vec_cameroon = wv['cameroon']
except KeyError:
    print("Từ 'cameroon' không tồn tại bên trong mô hình!")
```

```
Từ 'cameroon' không tồn tại bên trong mô hình!
```

Bảng 5. Lấy vectơ các từ không quen thuộc

Kiểm tra: Các từ có mức độ giống nhau giảm dần

```
pairs = [
    ('car', 'minivan'),    # a minivan is a kind of car
    ('car', 'bicycle'),    # still a wheeled vehicle
    ('car', 'airplane'),   # ok, no wheels, but still a vehicle
    ('car', 'cereal'),     # ... and so on
    ('car', 'communism'),]
for w1, w2 in pairs:
    print('%r\t%r\t%.2f' % (w1, w2, wv.similarity(w1, w2)))
```

```
'car'     'minivan' 0.69
'car'     'bicycle' 0.54
'car'     'airplane'     0.42
'car'     'cereal'  0.14
'car'     'communism'    0.06
```

Bảng 6. Lấy mức độ giống nhau của các cặp từ với từ "car"

Lấy top 10 từ gần với từ “car”:

```
print(wv.most_similar(positive=['car'], topn = 10))
```

```
[('vehicle', 0.7821096181869507),

('cars', 0.7423830032348633),

('SUV', 0.7160962820053101),

('minivan', 0.6907036304473877),

('truck', 0.6735789775848389),

('Car', 0.6677608489990234),

('Ford_Focus', 0.667320191860199),

('Honda_Civic', 0.662684977054596),

('Jeep', 0.6511331796646118),

('pickup_truck', 0.64414381980896)]
```

Bảng 7. Lấy 10 từ gần với từ "car" nhất

Kiểm tra và in ra từ khác biệt

```
print(wv.doesnt_match(['fire', 'water', 'land', 'sea', 'air',
'car']))
```

```
car
```

Bảng 8. Kiểm tra và chọn ra từ khác biệt nhất trong nhóm từ

NHẬN XÉT VÀ KẾT LUẬN

Kết quả đạt được:

Hiểu về thư viện gensim cũng như mô hình Word2Vec.

Hướng phát triển:

Tạo bản demo

Hoàn thiện bài báo cáo.

TÀI LIỆU THAM KHẢO

[1]. Ts. Nguyễn Tuấn Đăng, Đại học Nguyễn Tất Thành:
Các Slide bài giảng.

[2]. Gensim:
https://radimrehurek.com/gensim/auto_examples/index.html#documentation

[3]. Geeks for geeks:
https://www.geeksforgeeks.org/

[4]. Scikit Learn:
https://scikit-learn.org/