

TRƯỜNG ĐẠI HỌC THỦY LỢI
KHOA CÔNG NGHỆ THÔNG TIN



BÀI TẬP LỚN

Môn: Xử Lý Ngôn Ngữ Tự Nhiên

ĐỀ TÀI

Ứng Dụng Mô Hình Vector Space Model Và BERT Trong Việc Truy Vấn Thông Tin Câu Hỏi

Giảng viên hướng dẫn: PGS.TS Nguyễn Quang Hoan

Sinh viên thực hiện: Đồng Anh Quân
Mã Sinh Viên: 2251262626

Hà Nội - 2025

MỤC LỤC

Chương I. Giới Thiệu	2
1.1 Bối cảnh và lý do chọn đề tài	2
1.2 Mục tiêu của đề tài	2
1.3 Phạm vi thực hiện	2
Chương II. Cơ Sở Lý Thuyết	3
2.1 Truy xuất thông tin (Information Retrieval)	3
2.2. Mô hình không gian vector (Vector Space Model - VSM)	4
2.3. Kỹ thuật Bag of Words (BoW)	5
2.4. Mô hình BERT	5
2.5. Độ đo tương đồng giữa câu truy vấn và văn bản	7
Chương 3: Dữ Liệu Và Tiền Xử Lý Dữ Liệu	7
3.1. Giới thiệu về bộ dữ liệu MS MARCO	7
3.2 Tiền xử lý dữ liệu	8
Chương 4: Xây Dựng Mô Hình Và Thực Nghiệm	11
4.1 Tổng quan Pipeline mô hình sử dụng Vector Space Model	11
4.2 Tổng quan Pipeline của mô hình sử dụng BERT	13
4.3 Kết quả thực nghiệm	15
4.3.1. Mô hình Vector Space Model	15
4.3.2. Mô Hình BERT	16
Chương 5: Kết Luận Và Hướng Phát Triển	17
5.1 Kết luận	17
5.2 Hướng phát triển	17
Tổng Kết Đề Tài	18
Tài Liệu Tham Khảo	19

Chương I. Giới Thiệu

1.1 Bối cảnh và lý do chọn đề tài

Trong thời đại bùng nổ thông tin hiện nay, người dùng thường xuyên phải tiếp cận và tìm kiếm thông tin từ các nguồn dữ liệu khổng lồ như internet, tài liệu học thuật, cơ sở tri thức nội bộ,... Việc truy vấn thông tin một cách hiệu quả và chính xác đã trở thành một yêu cầu thiết yếu trong nhiều lĩnh vực như giáo dục, chăm sóc khách hàng, trợ lý ảo và hệ thống hỏi–đáp.

Một trong những mô hình nền tảng được áp dụng phổ biến trong các hệ thống truy vấn văn bản là Vector Space Model (VSM). Đây là mô hình biểu diễn tài liệu và truy vấn dưới dạng vector trong không gian nhiều chiều, cho phép đánh giá mức độ tương đồng giữa chúng dựa trên các phép đo toán học như cosine similarity. Ưu điểm nổi bật của VSM là tính đơn giản, khả năng mở rộng và hiệu quả trong việc xử lý dữ liệu văn bản phi cấu trúc.

Trong khuôn khổ bài tập lớn này, em nghiên cứu tập trung vào bài toán truy vấn thông tin câu hỏi – tức là xây dựng một hệ thống có thể nhận vào một câu hỏi và tìm ra các câu trả lời liên quan nhất từ một tập dữ liệu văn bản đã cho. Việc áp dụng mô hình VSM trong ngữ cảnh này cho phép hệ thống đánh giá mức độ liên quan giữa câu hỏi và các đoạn văn bản trả lời tiềm năng, từ đó sắp xếp và hiển thị kết quả phù hợp.

1.2 Mục tiêu của đề tài

- Tìm hiểu lý thuyết và cách triển khai mô hình Vector Space Model.
- Áp dụng VSM để giải quyết bài toán truy vấn thông tin dựa trên câu hỏi.
- Thực nghiệm với bộ dữ liệu có sẵn để đánh giá tính hiệu quả của mô hình.

1.3 Phạm vi thực hiện

- Hệ thống truy vấn sử dụng mô hình Vector Space Model và đo độ tương đồng cosine.
- Dữ liệu đầu vào là tập hợp các câu hỏi và văn bản trả lời bằng tiếng Anh.
- Không áp dụng các mô hình học sâu hay học máy nâng cao trong phạm vi đề tài.

Bằng việc xây dựng và thử nghiệm hệ thống này, đề tài không chỉ giúp làm rõ vai trò thực tiễn của Vector Space Model trong truy vấn thông tin, mà còn tạo nền tảng cho các hướng phát triển tiếp theo với các mô hình hiện đại hơn như BM25, word embeddings hoặc BERT.

Chương II. Cơ Sở Lý Thuyết

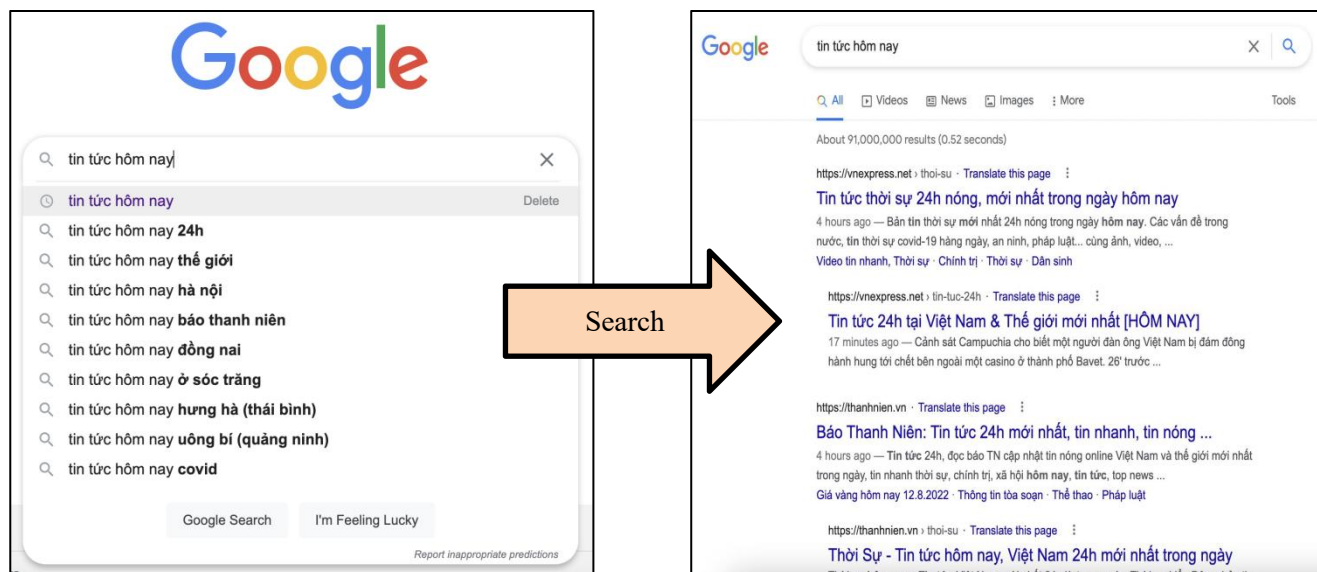
Trong lĩnh vực Truy xuất Thông tin (Information Retrieval – IR), việc tìm kiếm nội dung liên quan từ một tập hợp lớn văn bản dựa trên một truy vấn đầu vào là một nhiệm vụ trọng tâm. Để thực hiện điều này, các mô hình biểu diễn và đánh giá mối quan hệ giữa truy vấn và tài liệu đã được phát triển. Một trong những mô hình kinh điển, nền tảng và hiệu quả là Mô hình Không gian Vector (Vector Space Model – VSM).

2.1 Truy xuất thông tin (Information Retrieval)

Truy xuất thông tin là quá trình tìm kiếm các tài liệu phù hợp trong một tập dữ liệu lớn dựa trên nhu cầu thông tin (truy vấn) của người dùng. Khác với hệ thống cơ sở dữ liệu truyền thống (yêu cầu dữ liệu có cấu trúc), truy xuất thông tin thường xử lý các văn bản phi cấu trúc như email, bài báo, đoạn văn,...

Các ứng dụng thực tế của IR bao gồm:

- Công cụ tìm kiếm (Google, Bing,...)
- Trợ lý ảo (Siri, Google Assistant,...)
- Hệ thống hỏi – đáp tự động
- Tìm kiếm trong cơ sở tri thức doanh nghiệp



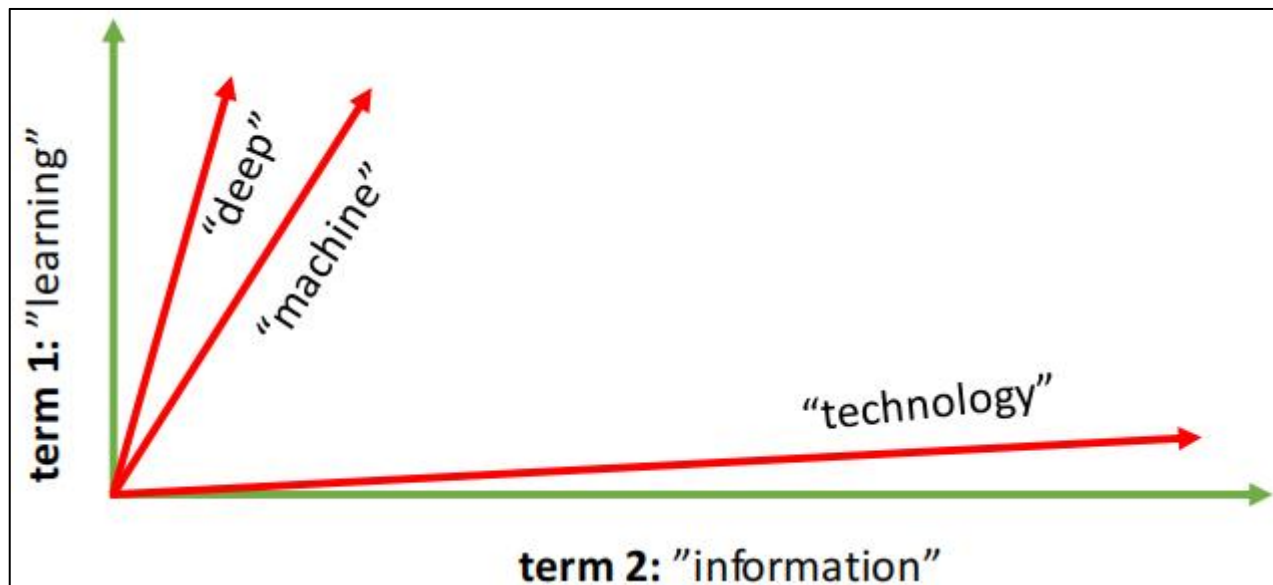
Hình 2.1.1: Truy xuất thông tin theo câu hỏi

2.2. Mô hình không gian vector (Vector Space Model - VSM)

Vector Space Model (VSM) là một mô hình toán học dùng để biểu diễn các văn bản và truy vấn dưới dạng các vector trong một không gian nhiều chiều. Mỗi chiều trong không gian này tương ứng với một từ (term) trong tập từ vựng (vocabulary) trích xuất từ toàn bộ tập văn bản.

Ý tưởng chính của VSM:

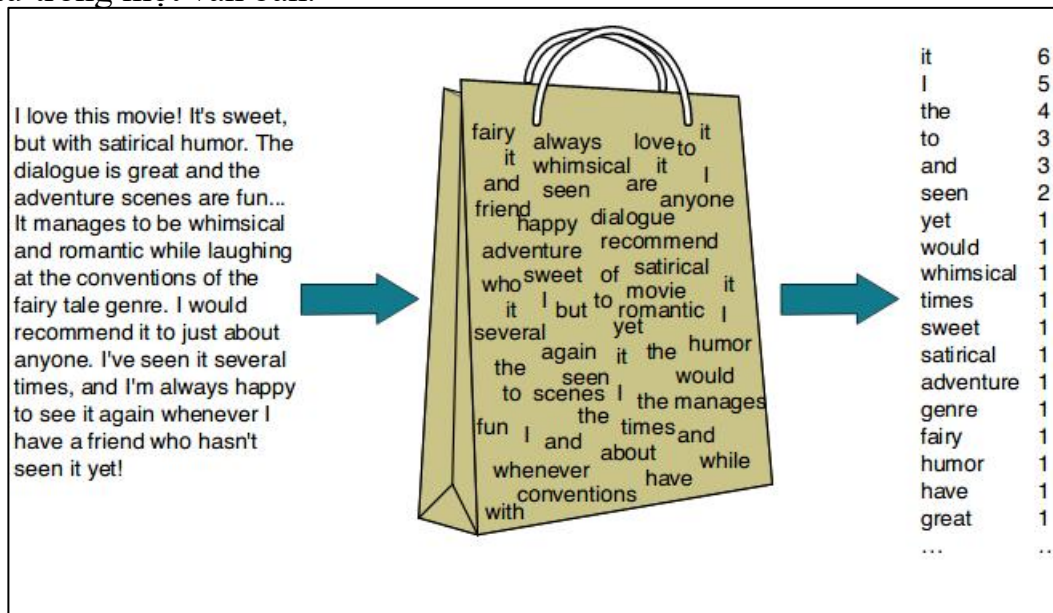
- Văn bản và truy vấn đều được biểu diễn dưới dạng vector số.
- Độ tương đồng giữa truy vấn và văn bản được tính bằng các phép đo vector, điển hình là độ tương đồng cosine (cosine similarity).
- Các văn bản có độ tương đồng cao với truy vấn sẽ được xem là phù hợp và được xếp hạng cao hơn.



Hình 2.2.1: Minh họa mô hình không gian vector

2.3. Kỹ thuật Bag of Words (BoW)

Bag of Words (BoW) là một kỹ thuật đơn giản nhưng hiệu quả để biểu diễn văn bản dưới dạng vector. Thay vì quan tâm đến thứ tự từ, BoW chỉ xem xét tần suất xuất hiện của các từ trong một văn bản.



Hình 2.3.1: Kỹ thuật Bag of Words

Các bước chính trong BoW:

- Xây dựng tập từ vựng: Gộp tất cả các từ xuất hiện trong tập dữ liệu, loại bỏ từ trùng.
- Mã hóa văn bản: Với mỗi văn bản (hoặc truy vấn), đếm số lần xuất hiện của từng từ trong từ vựng.
- Biểu diễn vector: Mỗi văn bản được biểu diễn dưới dạng một vector số nguyên có chiều dài bằng kích thước từ vựng.

Ví dụ: Ta có tập từ vựng ["trí", "tuệ", "nhân", "tạo", "học", "máy"]

Câu hỏi: "trí tuệ nhân tạo là gì?"

Biểu diễn Vector BoW: [1, 1, 1, 1, 0, 0]

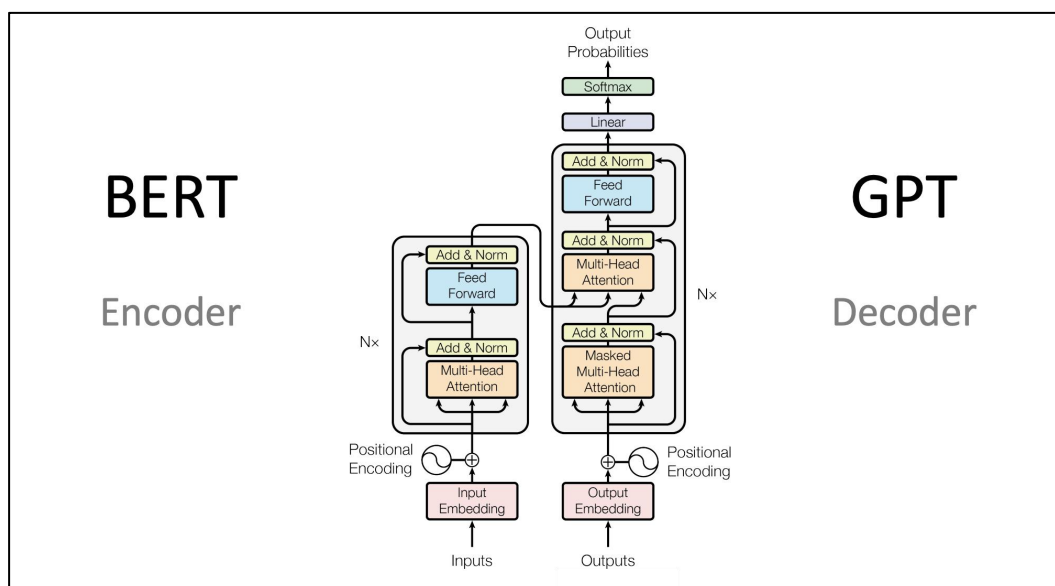
2.4. Mô hình BERT

BERT (Bidirectional Encoder Representations from Transformers) là một mô hình ngôn ngữ sâu được phát triển bởi nhóm nghiên cứu của Google vào năm 2018, nhằm mục đích cải thiện khả năng hiểu ngữ cảnh trong các tác vụ xử lý ngôn ngữ tự nhiên (NLP). Không giống như các mô hình trước đây chỉ xử lý văn bản theo chiều từ trái sang phải hoặc phải sang trái, BERT sử dụng cơ chế attention hai chiều (bidirectional attention),

cho phép mô hình học được ngữ cảnh của một từ dựa trên cả hai phía (trái và phải) trong câu.

Đặc điểm nổi bật của BERT:

- Tiền huấn luyện hai nhiệm vụ (pretraining):
- Masked Language Model (MLM): BERT che một số từ ngẫu nhiên trong văn bản đầu vào và học cách dự đoán những từ bị ẩn này.
- Next Sentence Prediction (NSP): Mô hình học mối quan hệ giữa hai câu để hiểu được dòng chảy ngữ nghĩa giữa các câu.
- Kiến trúc Transformer: BERT được xây dựng dựa trên kiến trúc Transformer Encoder, giúp xử lý song song các từ và hiểu mối quan hệ giữa các từ trong chuỗi một cách hiệu quả.



Hình 2.4.1: BERT sử dụng Encoder của Transformer

- BERT đã đạt được hiệu suất vượt trội trên nhiều tác vụ chuẩn trong NLP, bao gồm:
- Trả lời câu hỏi (QA)
- Phân loại cảm xúc
- Tìm kiếm văn bản (Information Retrieval)
- Trích xuất thực thể có tên (NER)

2.5. Độ đo tương đồng giữa câu truy vấn và văn bản

Sau khi truy vấn và các văn bản đều đã được biểu diễn dưới dạng vector BoW, ta có thể đánh giá mức độ liên quan giữa chúng thông qua độ tương đồng cosine, được tính như sau:

$$\text{cosine_similarity}(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \times \|\vec{B}\|}$$

$\vec{A} \cdot \vec{B}$: là vector biểu diễn của câu hỏi và câu trả lời

$\|\vec{A}\| \cdot \|\vec{B}\|$: Độ dài của vector A và B

Chương 3: Dữ Liệu Và Tiền Xử Lý Dữ Liệu

3.1. Giới thiệu về bộ dữ liệu MS MARCO

Trong bài này ta sử dụng bộ dữ liệu [MS MARCO \(Microsoft Machine Reading Comprehension\)](#) là một trong những tập dữ liệu phổ biến trong lĩnh vực truy vấn thông tin (Information Retrieval). Bộ dữ liệu gồm các cặp câu hỏi và đoạn văn bản được trích xuất từ các truy vấn thực tế trên công cụ tìm kiếm Bing, kèm theo câu trả lời ngắn gọn do con người gán nhãn.

answers (sequence)	passages (sequence)	query (string)	query_id (int32)	query_type (string)	wellFormedAnswers (sequence)
["Approximately \$15,000 per...	{ "is_selected": [1, 0, 0, 0, 0, 0], "passage_text": ["The average Walgreens salary...	"walgreens store sales average"	9,652	"numeric"	[]
["\$21,550 per year", "The...	{ "is_selected": [0, 1, 0, 0, 0, 0, 0, 0], "passage_text": ["A bartender's income is...	"how much do bartenders make"	9,653	"numeric"	[]
["A boil, also called a...	{ "is_selected": [0, 0, 0, 0, 0, 0, 1, 0], "passage_text": ["Knowledge center. A boil, also...	"what is a furuncle boil"	9,654	"description"	[]
["Detect and assess a wide...	{ "is_selected": [0, 0, 0, 0, 1, 0, 0, 0, 0], "passage_text": ["Urinalysis: One way to test fo...	"what can urinalysis...	9,655	"description"	[]
["Shigellosis, diseases of the...	{ "is_selected": [0, 0, 0, 0, 1, 0, 0, 0, 0], "passage_text": ["Since vitamin A is fat-soluble...	"what is vitamin a used for"	9,656	"description"	[]
["The initiation of cell...	{ "is_selected": [0, 0, 0, 0, 0, 1, 0, 0, 0], "passage_text": ["1 Fail to stop uncontrolled...	"what causes genetic...	9,657	"description"	[]
["\$2.51 - \$3.17 per square foot"...	{ "is_selected": [0, 0, 0, 0, 0, 1], "passage_text": ["1 A lot depends on how much is...	"cost to frame basement"	9,658	"numeric"	[]

Hình 3.1: Bộ dữ liệu MS MARCO

Cấu trúc bộ dữ liệu MS MARCO bao gồm:

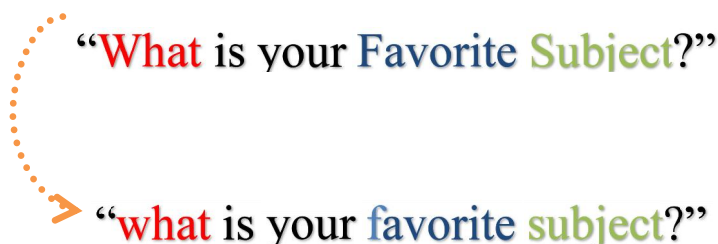
- answers: Danh sách các câu trả lời ngắn do con người gán nhãn.
- passages: Danh sách các đoạn văn. Mỗi đoạn có is_selected và passage_text.
- query: Câu hỏi/truy vấn người dùng nhập (ngôn ngữ tự nhiên).
- query_id: ID duy nhất cho mỗi truy vấn.
- query_type: Loại truy vấn: "numeric", "description", "yesno", ...
- wellFormedAnswers: Câu trả lời ngắn gọn, đầy đủ ngữ pháp (trống).

3.2 Tiền xử lý dữ liệu

Để tăng hiệu quả trong quá trình truy vấn và giảm nhiễu cho mô hình, toàn bộ văn bản (bao gồm cả câu hỏi và văn bản) được đưa qua phần tiền xử lý bao gồm các bước sau:

Các bước tiền xử lý:

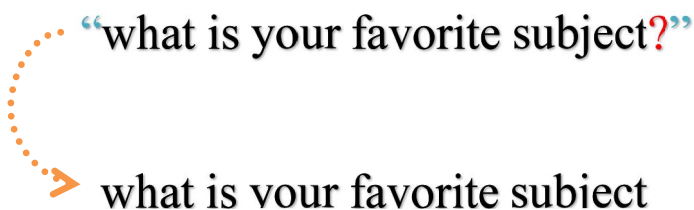
- Chuyển về chữ thường (lowercase): Tất cả các ký tự trong văn bản được chuyển thành chữ thường nhằm đồng nhất dữ liệu, tránh phân biệt giữa các từ



“What is your Favorite Subject?”
→ “what is your favorite subject?”

Hình 3.2.1: Chuyển các chữ viết hoa về viết thường

- Loại bỏ dấu câu (punctuation): Các dấu câu như .,?!:;'" được loại bỏ để chỉ giữ lại nội dung văn bản quan trọng cho việc phân tích.



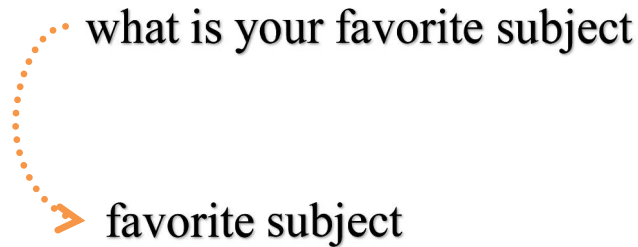
“what is your favorite subject?”
→ what is your favorite subject

Hình 3.2.2: Loại bỏ các dấu câu

- Loại bỏ từ dừng (stopwords): Các từ không mang nhiều thông tin như “the”, “is”, “and”, “in”, v.v. được loại bỏ để giảm độ nhiễu trong biểu diễn vector.

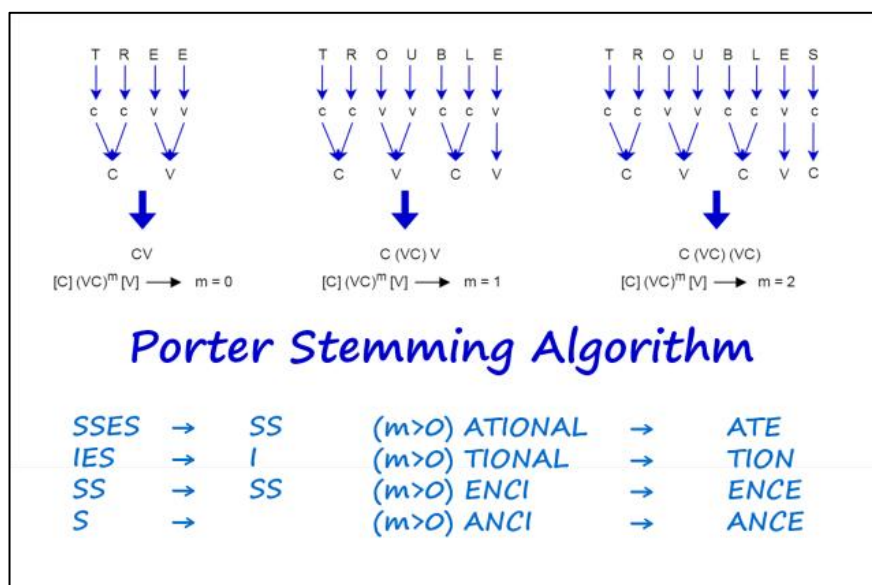
a	did	herself	not	the	we've
about	didn't	him	of	their	were
above	do	himself	off	theirs	weren't
after	does	his	on	them	what
again	doesn't	how	once	themselves	what's
against	doing	how's	only	then	when
all	don't	i	or	there	when's
am	down	i'd	other	there's	where
an	during	i'll	ought	these	where's
and	each	i'm	our	they	which
any	few	i've	ours	they'd	while
are	for	if	ourselves	they'll	who
aren't	from	in	out	they're	who's
as	further	into	over	they've	whom
at	had	is	own	this	why
be	hadn't	isn't	same	those	why's
because	has	it	shan't	through	with
been	hasn't	it's	she	to	won't
before	have	its	she'd	too	would
being	haven't	itself	she'll	under	wouldn't
below	having	let's	she's	until	you
between	he	me	should	up	you'd
both	he'd	more	shouldn't	very	you'll
but	he'll	most	so	was	you're
by	he's	mustn't	some	wasn't	you've
can't	her	my	such	we	your
cannot	here	myself	than	we'd	yours
could	here's	no	that	we'll	yourself

Hình 3.2.3: Các Stopwords phổ biến

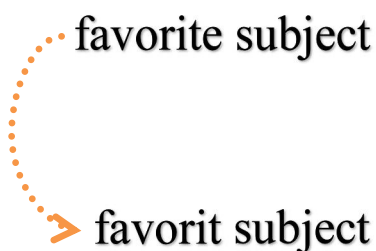


Hình 3.2.4: Loại bỏ các stopwords

- Stemming: Các từ được rút gọn về gốc bằng các thuật toán như *Porter Steaming*, giúp giảm số lượng từ vựng và tăng khả năng khái quát hoá (ví dụ: “running”, “runner”, “ran” → “run”).



Hình 3.2.5: Thuật toán Porter Steaming

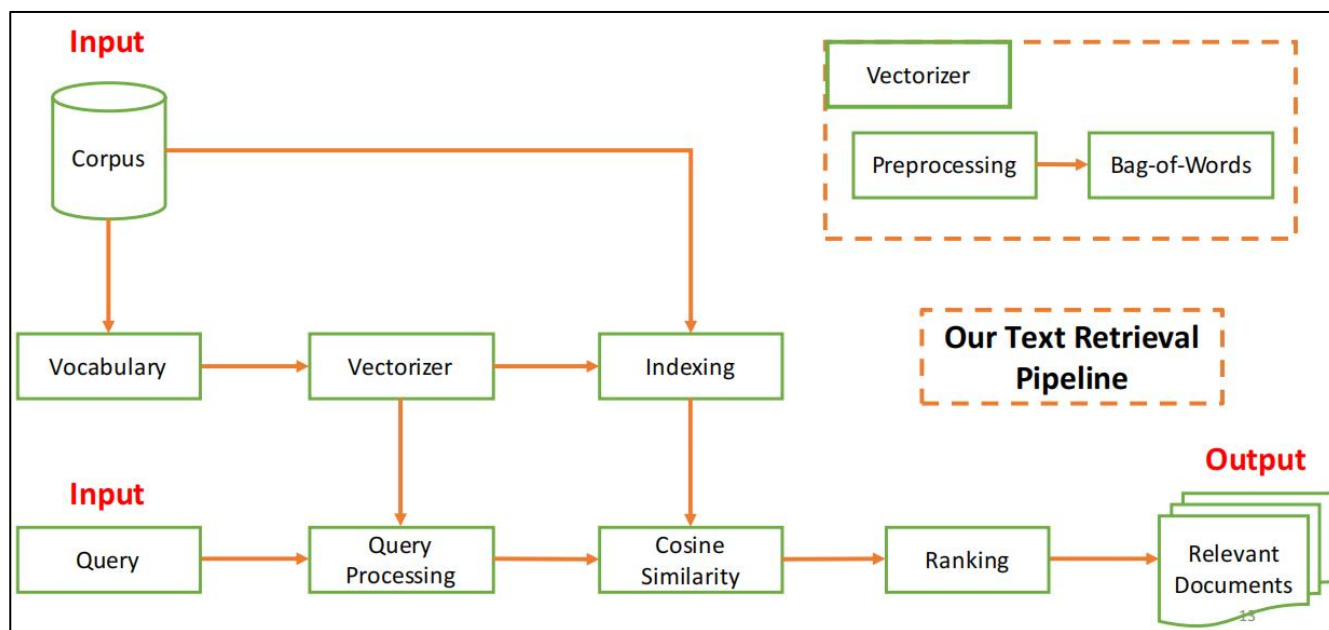


Hình 3.2.5: Từ sau khi đi qua thuật toán Porter Steaming

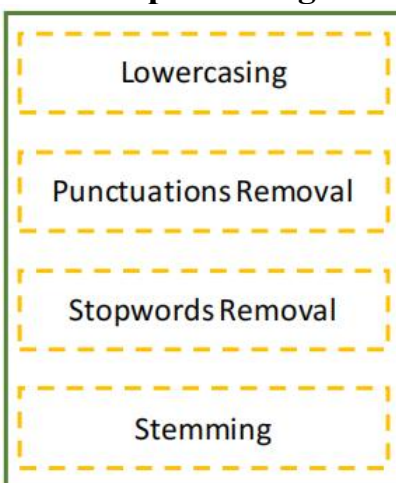
Chương 4: Xây Dựng Mô Hình Và Thực Nghiệm

4.1 Tổng quan Pipeline mô hình sử dụng Vector Space Model

Mô hình truy vấn được xây dựng dựa trên Vector Space Model (VSM) sử dụng phương pháp Bag-of-Words (BoW) để biểu diễn văn bản. Mô hình gồm các bước chính: tiền xử lý dữ liệu, tạo bảng chỉ mục, biểu diễn vector, tính độ tương đồng cosine và xếp hạng kết quả truy vấn.



Preprocessing



Hình 4.1.1: Pipeline của mô hình

- **Input:** Bao gồm tập dữ liệu văn bản (Corpus) và câu truy vấn (Query).
- **Preprocessing:** Chuẩn hóa văn bản như chuyển về chữ thường, loại bỏ dấu câu, stopwords, và sử dụng Porter Stemmer như đã mô tả ở phần tiền xử lý.
- **Vectorizer:** Sử dụng Bag-of-Words để chuyển văn bản thành vector.
- **Indexing:** Tạo bảng chỉ mục cho các văn bản để truy xuất nhanh.

	words1	words2	words3	words4	words5
doc1	0	0	1	0	0
doc2	2	0	1	1	0
doc3	0	0	1	1	0
doc4	0	0	1	1	1

Hình 4.1.2: Phần Indexing

- **Query Processing:** Xử lý câu truy vấn tương tự như văn bản để đưa về vector cùng không gian.
- **Similarity Measurement:** Tính toán độ tương đồng giữa truy vấn và các văn bản bằng công thức Cosine Similarity.
- **Ranking:** Sắp xếp các văn bản theo mức độ phù hợp với truy vấn.

	read	book	ai	machine	learn	how
doc1	1	1	1	0	0	0
doc2	0	1	0	1	1	0
doc3	0	0	1	0	2	1
query	0	0	0	1	1	0

Cosine

Ranked List

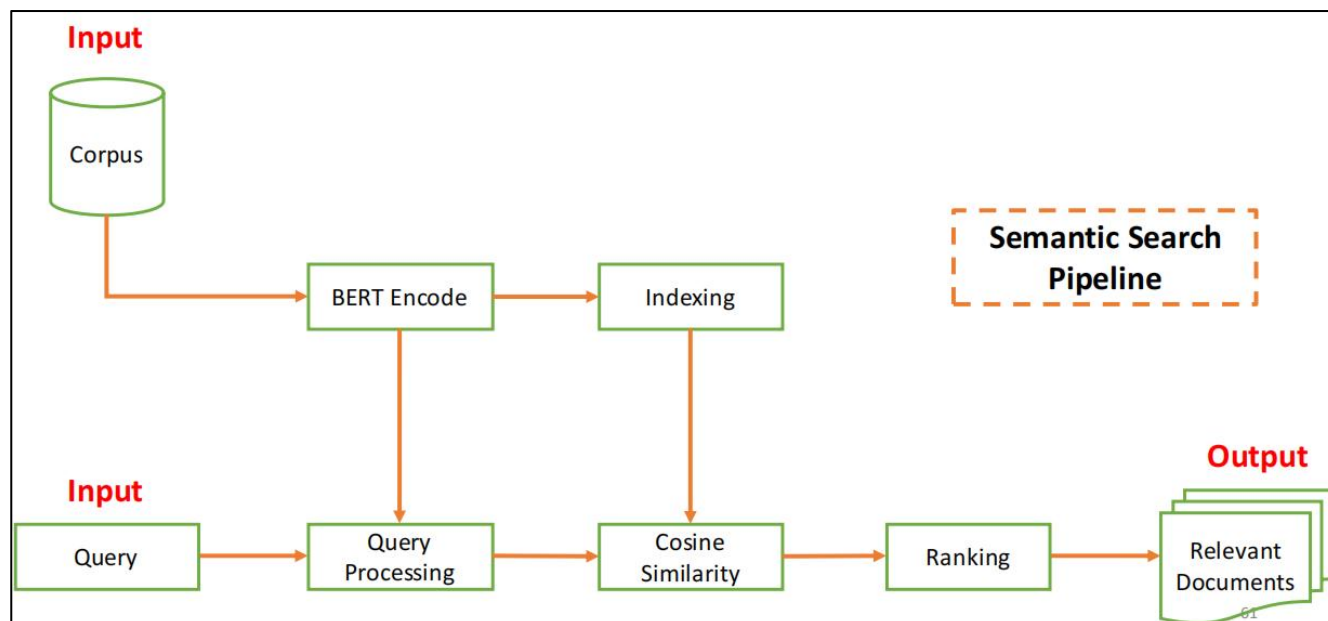
DocID	Similarity
d2	0.8165
d3	0.5774
d1	0.0000

Hình 4.1.3: Phần Ranking

- **Output:** Trả về danh sách các văn bản phù hợp nhất.

4.2 Tổng quan Pipeline của mô hình sử dụng BERT

Mô hình Semantic Search sử dụng BERT nhằm tìm kiếm các tài liệu có ý nghĩa tương đồng với truy vấn đầu vào, không chỉ dựa trên từ khóa mà còn dựa trên ngữ nghĩa sâu của văn bản

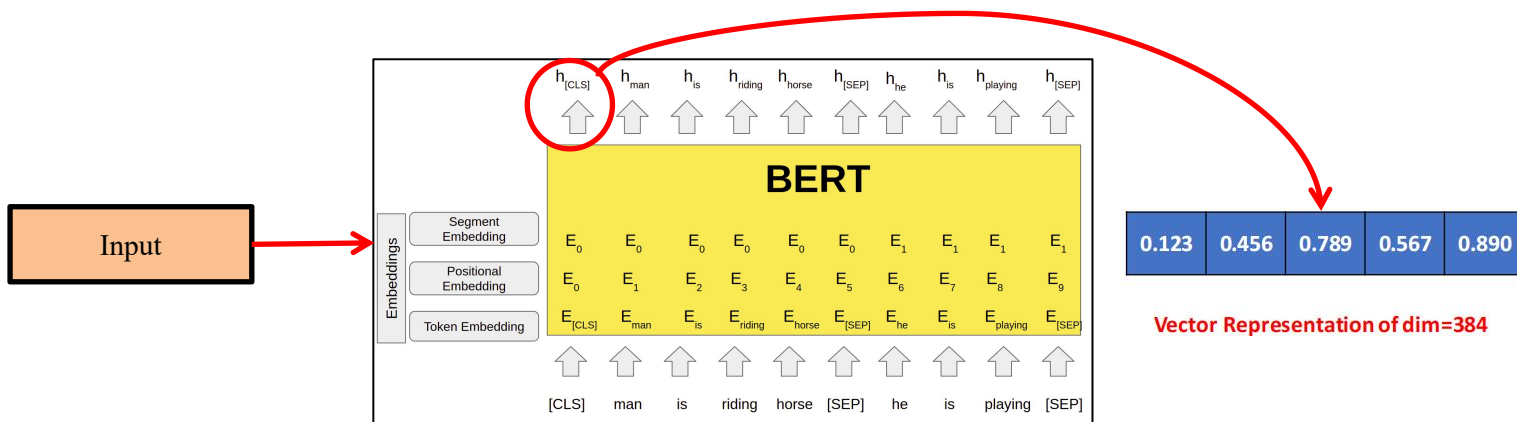


Hình 4.2.1: Pipeline của mô hình

Các bước của mô hình bao gồm:

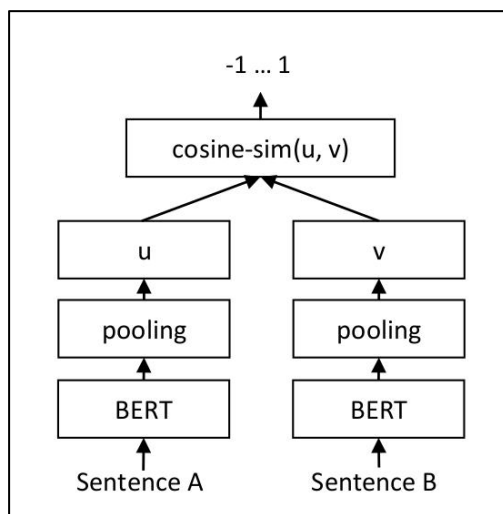
- **Input:** Corpus và Query

- **BERT Encode:** Tất cả văn bản trong Corpus được mã hóa bằng mô hình BERT để tạo ra các vector ngữ nghĩa (semantic embeddings). Những vector này thể hiện ý nghĩa của từng đoạn văn thay vì chỉ biểu diễn bằng từ khóa. Tương tự, truy vấn đầu vào cũng được đưa qua mô hình BERT để mã hóa thành vector ngữ nghĩa tương ứng.



Hình 4.2.1: Quá trình BERT Encoder

- **Query Processing:** Truy vấn sau khi được mã hóa có thể trải qua bước xử lý bổ sung như chuẩn hóa độ dài vector hoặc lọc nhiễu để cải thiện độ chính xác khi so sánh với các đoạn văn trong tập dữ liệu.
- **Indexing:** Các vector biểu diễn ngữ nghĩa của toàn bộ corpus được lưu trữ và tổ chức lại thành chỉ mục (index) để truy xuất nhanh chóng. Thay vì so khớp từng đoạn một, chỉ mục giúp tối ưu hóa quá trình tìm kiếm.
- **Cosine Similarity:** Khi nhận được một truy vấn, mô hình tính độ tương đồng cosine giữa vector của truy vấn và tất cả các vector trong corpus. Cosine similarity đo lường góc giữa hai vector — càng gần 1 thì hai văn bản càng giống nhau về mặt ngữ nghĩa.



Hình 4.2.2: Tính Cosine từ Vector sau khi đi qua BERT Encoder

- **Ranking:** Các đoạn văn bản trong corpus được xếp hạng dựa trên điểm tương đồng với truy vấn. Đoạn văn nào có điểm cao hơn sẽ được coi là phù hợp hơn.
- **Output:** Cuối cùng, hệ thống trả về danh sách các tài liệu liên quan có điểm tương đồng cao nhất, được sắp xếp theo thứ tự từ cao xuống thấp.

4.3 Kết quả thực nghiệm

- Dữ liệu sử dụng: Bộ dữ liệu MS MARCO (Microsoft Machine Reading Comprehension) với các cặp query - passage.

- Tiêu chí đánh giá: Dựa trên độ chính xác của việc xếp hạng câu trả lời phù hợp với truy vấn thông qua việc tính toán Cosine Similarity.

4.3.1. Mô hình Vector Space Model

Query: what is the official language in Canada
Top: 1, Score: 0.7071 aboriginal languages of canada communities where aboriginal languages are spoken are found in all regions of canada there are around 60 distinct indigenous languages in canada falling into 10 separate language families languages generally have many varieties or dialects and the aboriginal languages in canada are no exception many of these languages have several more or less mutually intelligible dialects particularly when the language is distributed over a large area
Top: 2, Score: 0.5252 Most Alaskans continue to accept the name Eskimo, particularly because Inuit refers only to the Inupiat of northern Alaska, the Inuit of Canada, and the Kalaallit of Greenland, and it is not a word in the Yupik languages of Alaska and Siberia. 1 Comparative Yupik and Inuit. However, the people of Canada and Greenland prefer other names. Inuit, meaning people, is used in most of Canada, and the language is called Inuktitut in eastern Canada although other local designations are used also.
Top: 3, Score: 0.5108 there are approximately 100 different languages spoken in canada not including languages spoken by non native immigrants and residents the 2 official languages of canada are english and french here is a list of most of the languages spoken in canada 1 french and english have equal status as canada's official languages english is the language spoken by a majority of canadians french is the language spoken by the overwhelming majority of quebecers quebecers new brunswick is canada's only officially bilingual province french is widely spoken in eastern and northeastern ontario manitoba was canada's first bilingual province and still is federally there are only two official languages english and french
Top: 4, Score: 0.5080 Canola is the most toxic oil on the face of the earth. No such plant as the "canola plant" canola oil comes from Canada. That is how it got its name canola (Canada oil) the people of Canada think that us Americans are out of our minds to use this toxic oil for cooking, for in Canada it is used to grease machinery.
Top: 5, Score: 0.4170 It is referred to in the currency markets as the Canadian Dollar. The currency of Canada is the Canadian Dollar. It is abbreviated CAD. Canada's currency is Dollar. It is not the same as US\$. Its current value is almost at par with us\$.

Hình 4.3.1.1: Kết quả thực nghiệm 1

Nhận xét

Ưu điểm:

- Mô hình VSM sử dụng kỹ thuật Bag-of-Words cùng với vector hóa và so sánh Cosine Similarity, cho phép truy xuất nhanh chóng và đơn giản.

- Tìm được các văn bản chứa nhiều từ khóa giống với câu truy vấn. Ví dụ: kết quả Top 1, 2 đều có chứa nhiều từ liên quan như "aboriginal languages", "canada", "languages spoken" v.v.

Hạn chế:

- Mô hình thiếu khả năng hiểu ngữ nghĩa sâu sắc: các câu được xếp hạng cao chủ yếu dựa trên số lượng từ trùng lặp, không xét tới ngữ cảnh.

4.3.2. Mô Hình BERT

```
Query: What is the official language in Canada

=====
Top 5 most similar sentences in corpus:

Document rank 1
the two official languages every country in the world has official languages that are spoken amongst the population these languages are generally what
(Score: 0.7396)

Document rank 2
improving your english and french canada has two official languages english and french english is the most commonly spoken language in most provinces a
(Score: 0.7207)

Document rank 3
there are approximately 100 different languages spoken in canada not including languages spoken by non native immigrants and residents the 2 official l
(Score: 0.7135)

Document rank 4
a multitude of languages are used in canada according to the 2011 census english and french are the mother tongues of 56 9 % and 21 3 % of canadians re
(Score: 0.7117)

Document rank 5
other languages the prominence of languages other than english and french varies across the country mostly influenced by immigration in western canada
(Score: 0.6740)
```

Hình 4.3.2.1: Kết quả thực nghiệm 2

Nhận xét

Ưu điểm:

- Hiểu ngữ cảnh tốt hơn nhờ kiến trúc Transformer hai chiều, giúp nắm bắt mối quan hệ giữa các từ trong câu, cả phía trước và sau.
- Kết quả trả về sát với nội dung truy vấn hơn. Các đoạn Top 1 đến 5 đều trả lời đúng trọng tâm câu hỏi: "What is the official language in Canada".
- Điểm số giữa các kết quả cũng đồng đều và cao hơn, thể hiện khả năng mô hình hóa ngữ nghĩa vượt trội.

Hạn chế:

- Tốc độ xử lý chậm hơn so với VSM do cấu trúc mạng nơ-ron phức tạp.
- Cần tài nguyên tính toán lớn hơn.

Chương 5: Kết Luận Và Hướng Phát Triển

5.1 Kết luận

Trong đề tài này, em đã xây dựng và so sánh hai hệ thống truy xuất thông tin: một dựa trên mô hình truyền thống Vector Space Model (VSM) và một dựa trên mô hình học sâu hiện đại BERT (Bidirectional Encoder Representations from Transformers).

Thông qua các thực nghiệm với tập truy vấn cụ thể, chúng tôi thu được các kết luận như sau:

- Về hiệu quả truy xuất: Mô hình VSM có khả năng xử lý và tìm kiếm nhanh chóng với chi phí tính toán thấp. Tuy nhiên, do chỉ dựa vào biểu diễn từ theo mô hình túi từ (bag-of-words), mô hình này không nắm bắt được ngữ nghĩa sâu và thường trả về các văn bản chứa từ khóa nhưng không đúng ngữ cảnh.

- Về độ chính xác ngữ nghĩa: Mô hình BERT cho kết quả vượt trội nhờ khả năng hiểu ngữ cảnh và mối quan hệ giữa các từ trong câu. Các câu trả lời được xếp hạng cao bởi BERT đều sát với nội dung truy vấn, cho thấy sức mạnh của các mô hình ngôn ngữ hiện đại trong việc giải quyết bài toán tìm kiếm ngữ nghĩa.

- Về độ phù hợp kết quả: BERT đã loại bỏ các kết quả gây nhiễu thường thấy trong VSM. Ví dụ, truy vấn "What is the official language in Canada" với VSM trả về văn bản nói về dầu canola, trong khi BERT trả về các câu liên quan trực tiếp đến tiếng Anh và tiếng Pháp – hai ngôn ngữ chính thức tại Canada.

Từ đó, có thể khẳng định rằng mô hình ngôn ngữ BERT phù hợp hơn trong các hệ thống truy xuất thông tin hiện đại, đặc biệt là trong môi trường dữ liệu phi cấu trúc và ngữ nghĩa phức tạp.

5.2 Hướng phát triển

Để nâng cao hiệu quả và khả năng ứng dụng của hệ thống truy xuất thông tin, các hướng phát triển trong tương lai có thể bao gồm:

- Tối ưu hóa tốc độ và chi phí tính toán của BERT: Do BERT có chi phí tính toán cao, việc áp dụng các phiên bản rút gọn như DistilBERT, TinyBERT hoặc các kỹ thuật như vector indexing (FAISS, Annoy) sẽ giúp cải thiện thời gian phản hồi mà vẫn giữ được chất lượng kết quả.

- Tích hợp đánh giá kết quả bằng người dùng thực tế: Để đánh giá chính xác hơn mức độ hài lòng, có thể tích hợp phản hồi người dùng (relevance feedback) vào hệ thống để tự động cải thiện khả năng xếp hạng truy vấn.

- Mở rộng hệ thống cho truy vấn tiếng Việt: Áp dụng các mô hình BERT tiền huấn luyện cho tiếng Việt như PhoBERT hoặc viBERT sẽ giúp hệ thống hoạt động hiệu quả hơn với ngôn ngữ bản địa.

Tổng Kết Đề Tài

Trong bối cảnh lượng dữ liệu văn bản ngày càng gia tăng mạnh mẽ, việc xây dựng các hệ thống truy xuất thông tin hiệu quả, chính xác và thông minh đóng vai trò quan trọng trong nhiều lĩnh vực như giáo dục, nghiên cứu, dịch vụ khách hàng và tìm kiếm tài liệu. Đề tài đã thực hiện việc khảo sát, xây dựng và đánh giá hai mô hình truy xuất thông tin là Mô hình Vector Space Model (VSM) sử dụng kỹ thuật tiền xử lý truyền thống kết hợp biểu diễn văn bản bằng mô hình túi từ (Bag-of-Words) và mô hình Semantic Search sử dụng BERT để mã hóa ngữ nghĩa văn bản và truy vấn.

Kết quả thực nghiệm cho thấy rằng mô hình BERT mang lại hiệu quả vượt trội về mặt ngữ nghĩa và độ phù hợp của kết quả so với VSM. Mô hình này có khả năng hiểu truy vấn tốt hơn, loại bỏ được các kết quả chứa từ khóa nhưng không liên quan ngữ cảnh – một điểm yếu của mô hình truyền thống.

Đề tài không chỉ cung cấp cái nhìn tổng quan về hai phương pháp truy xuất thông tin, mà còn góp phần minh chứng cho xu hướng chuyển dịch từ các mô hình thống kê đơn giản sang các mô hình học sâu ngữ nghĩa trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP). Kết quả từ đề tài có thể là nền tảng cho các nghiên cứu mở rộng tiếp theo, như tích hợp học tăng cường, cải tiến hiệu năng hệ thống, hoặc áp dụng trên các dữ liệu đa ngôn ngữ và dữ liệu tiếng Việt.

Tài Liệu Tham Khảo

1. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.
2. Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130–137.
2. Jurafsky, D., & Martin, J. H. (2021). Speech and Language Processing (3rd ed., draft). Stanford University.
3. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*
5. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of EMNLP-IJCNLP*.
6. Salton, G., Wong, A., & Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11), 613–620.
7. Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
8. Wolf, T., et al. (2020). Transformers: State-of-the-art Natural Language Processing. In *Proceedings of the 2020 EMNLP: System Demonstrations*.
9. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.
10. Suresh, K. (2020). Text Preprocessing Techniques in NLP. *Towards Data Science*.