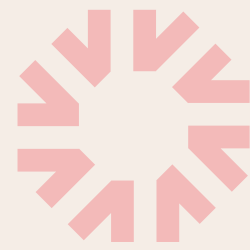






Khai phá dữ liệu hướng lĩnh vực

Giảng viên: TS. Trần Mai Vũ



Nhóm 3 : Lê Minh Tâm - 22024500
 Phạm Thu Trang - 22024528
 Phạm Tùng Chi - 22024525

Đóng góp

Lê Minh Tâm

- Phân tích dữ liệu để tìm ra thiếu sót của các mô hình được sử dụng
- Xây dựng mô hình LDA và trích xuất từ khóa
- Sử dụng vncorenlp để tokenize
- Theo dõi tiến độ của nhóm
- Hỗ trợ làm slides

40%

Phạm Thu Trang

- Phân tích dữ liệu để tìm ra thiếu sót của các mô hình được sử dụng
- Lọc nhiễu (ký tự đặc biệt, thông số đặc biệt)
- Hỗ trợ chạy code
- Hỗ trợ diễn giải topic
- Làm slides chính

30%

Phạm Tùng Chi

- Phân tích dữ liệu để tìm ra thiếu sót của các mô hình được sử dụng
- Xử lý stopwords
- Lọc tags
- Hỗ trợ chạy code
- Hỗ trợ diễn giải topic
- Làm slides chính

30%

Nội dung

**Phân
tích dữ
liệu**

**Tiền xử
lý dữ
liệu**

**Xây dựng
topic ẩn**

**Phân
cụm dữ
liệu**

**Mô hình
học máy
đa nhãn**



Phân tích & quan sát thực tế

Trên thực tế, các trang web thường được phân cụm theo 2 mức



Kinh doanh Bất động sản

NetZero

Quốc tế

Doanh nghiệp

Chứng khoán

Ebank

Vĩ mô

Tiền của tôi

VnExpress

Kinh doanh Thể thao Thể

Net Zero

Tài chính

Đầu tư

Thị trường

Doanh nhân

Tư vấn tài chính

Vietnamnet

KINH TẾ

Lao động - Việc làm

Tài chính

Chứng khoán

Kinh doanh

Baomoi

Mục đích

- Nhiều lý do liên quan đến quản lý nội dung, trải nghiệm người dùng và tối ưu hóa công cụ tìm kiếm (SEO).
- Hỗ trợ cải thiện độ chính xác, ngữ cảnh hóa, và hiệu suất của quá trình trích xuất.

Ngữ cảnh hóa nội dung

Cách tổ chức phản ánh cấu trúc thông tin trong đời sống

Tạo ra từ khóa cụ thể hơn, đúng trọng tâm, phù hợp với những gì người dùng thực sự cần tìm.

Phân tích cách từ khoá được trích ra

Topic : Tuyên dương HSG

Nội dung bài báo

Trận chung kết “Đường lên đỉnh Olympia” năm 2024, Võ Quang Phú Đức đã dẫn đầu tất cả các vòng thi để giành ngôi quán quân với số điểm 220. Đây là vòng nguyệt quế thứ 3 của trường THPT Chuyên Quốc Học Huế nói riêng và của tỉnh Thừa Thiên Huế nói chung. Trước đó, Võ Quang Phú Đức có những thành tích đáng nể như: Học sinh giỏi quốc gia, Huy chương Bạc quốc tế... Phát biểu tại buổi lễ, Chủ tịch UBND tỉnh Thừa Thiên Huế Nguyễn Văn Phương nhấn mạnh, "đạt được vòng nguyệt quế là một thành quả vô cùng xứng đáng với những cố gắng của Phú Đức, chúng ta tự hào về em. Những thành công của em là nguồn cảm hứng mạnh mẽ cho các bạn trẻ khác trong tỉnh và cho chúng ta". Chủ tịch UBND tỉnh Thừa Thiên Huế Nguyễn Văn Phương cảm ơn bố mẹ của em, cảm ơn thầy cô giáo ở những ngôi trường em từng theo học; đặc biệt là thầy cô giáo của Trường THPT chuyên Quốc Học - Huế đã luôn quan tâm, hỗ trợ và tạo mọi điều kiện để em phát triển và tỏa sáng tài năng của mình...

Từ khóa

Võ Quang Phú Đức
Đường lên đỉnh Olympia
Phú Đức
Thừa Thiên Huế
UBND tỉnh Thừa Thiên Huế
Nguyễn Văn Phương
quán quân
THPT chuyên Quốc học
vòng nguyệt quế
THPT chuyên Quốc học Huế
Bằng khen
nhà vô địch
vinh danh
tuyên dương
học sinh giỏi quốc gia
trao tặng

Phân tích cách từ khoá được trích ra

Topic : Tuyên dương HSG



Từ khoá chung theo cụm

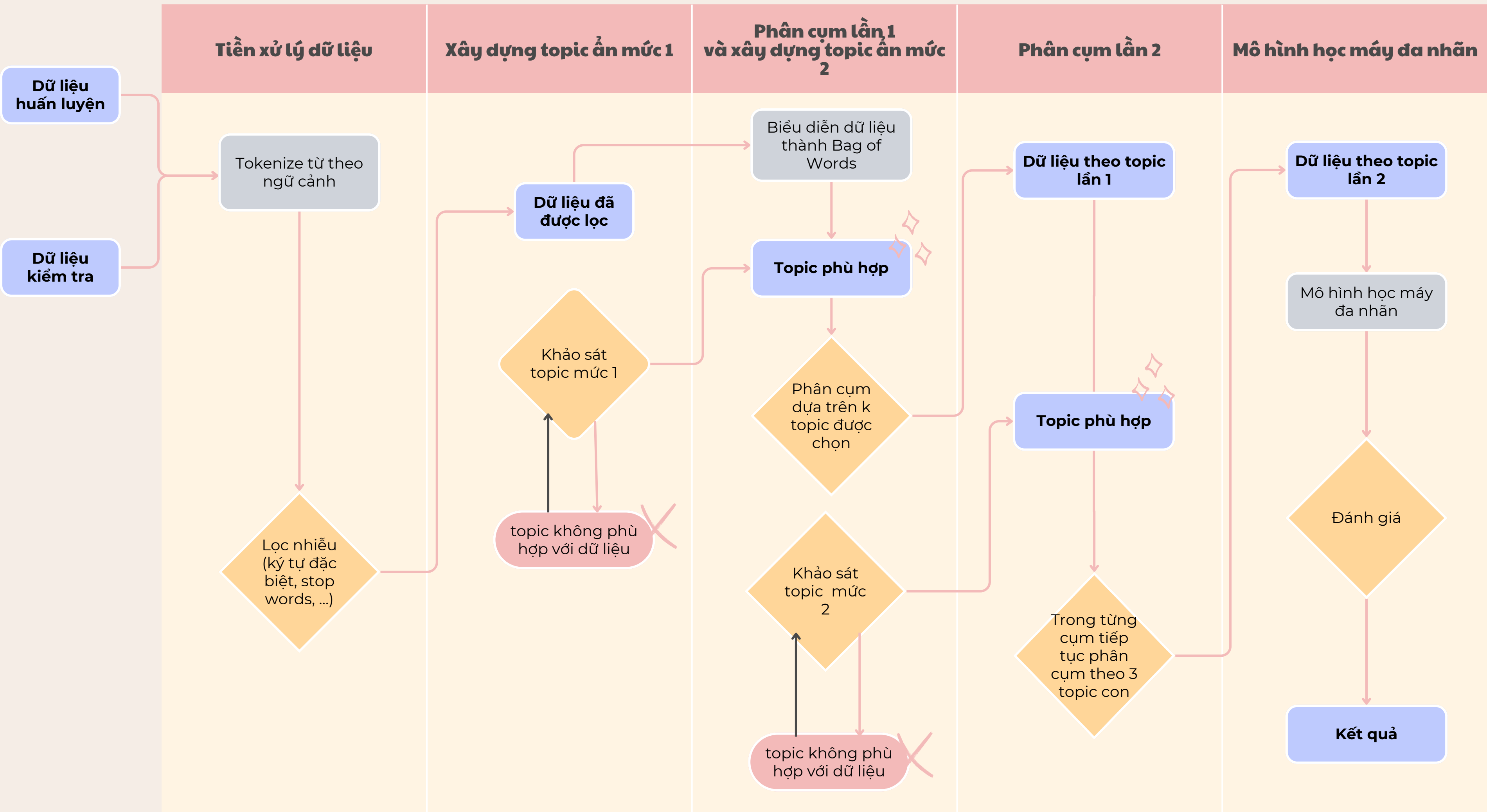
quán quân
Bằng khen
nhà vô địch
vinh danh
tuyên dương
học sinh giỏi quốc gia
trao tặng

Từ khoá riêng của bài báo



Võ Quang Phú Đức
Đường lên đỉnh Olympia
Phú Đức
Thừa Thiên Huế
UBND tỉnh Thừa Thiên Huế
Nguyễn Văn Phương
THPT chuyên Quốc học
vòng nguyệt quế
THPT chuyên Quốc học
Huế

Loại từ trong tags: danh từ, động từ, tên riêng chiếm tỉ lệ cao



Tiền xử lý dữ liệu

Thống kê content không hữu ích

- Những content có ký tự < 100 thì sẽ bị coi là không hữu ích:
 - Không cung cấp đủ ngữ cảnh hoặc thông tin quan trọng để phân tích.
 - Ảnh hưởng đến phân phối topic.

➡ Loại bỏ các hàng này.

```
Số lượng văn bản có độ dài < 100 ký tự: 688
Nội dung các văn bản ngắn trong cột 'content':
- Vietnam+
- Mời quý vị và các bạn xem video dưới đây:
- Theo baokiemtoan.vn
- Theo Vũng Tàu Review, Vie Limo, Cùng Đi
- Mời quý vị và các bạn xem video dưới đây:
- Theo AP
- Theo Vietnam plus
- Dương Giang/TTXVN
- Ảnh
- Xuống hạng
- Đá play-off trụ hạngXuống hạng
- Vietnam+
- Thực hiện: Nhóm phóng viên
- Hà Mi (thực hiện)
- Thiết kế: Mỹ Trang
- Mời quý vị và các bạn xem video dưới đây:
- Video: Toàn cảnh mở đá Nà Cà thuộc huyện Bạch Thông, Bắc Kạn
- '
```

Tiền xử lý dữ liệu

Thống kê số tags bị xoá

Số hàng thay đổi	86,415
Số tags bị xoá	287,713
Tổng số tags bị xoá	1.988.499
Phần trăm	14,47%

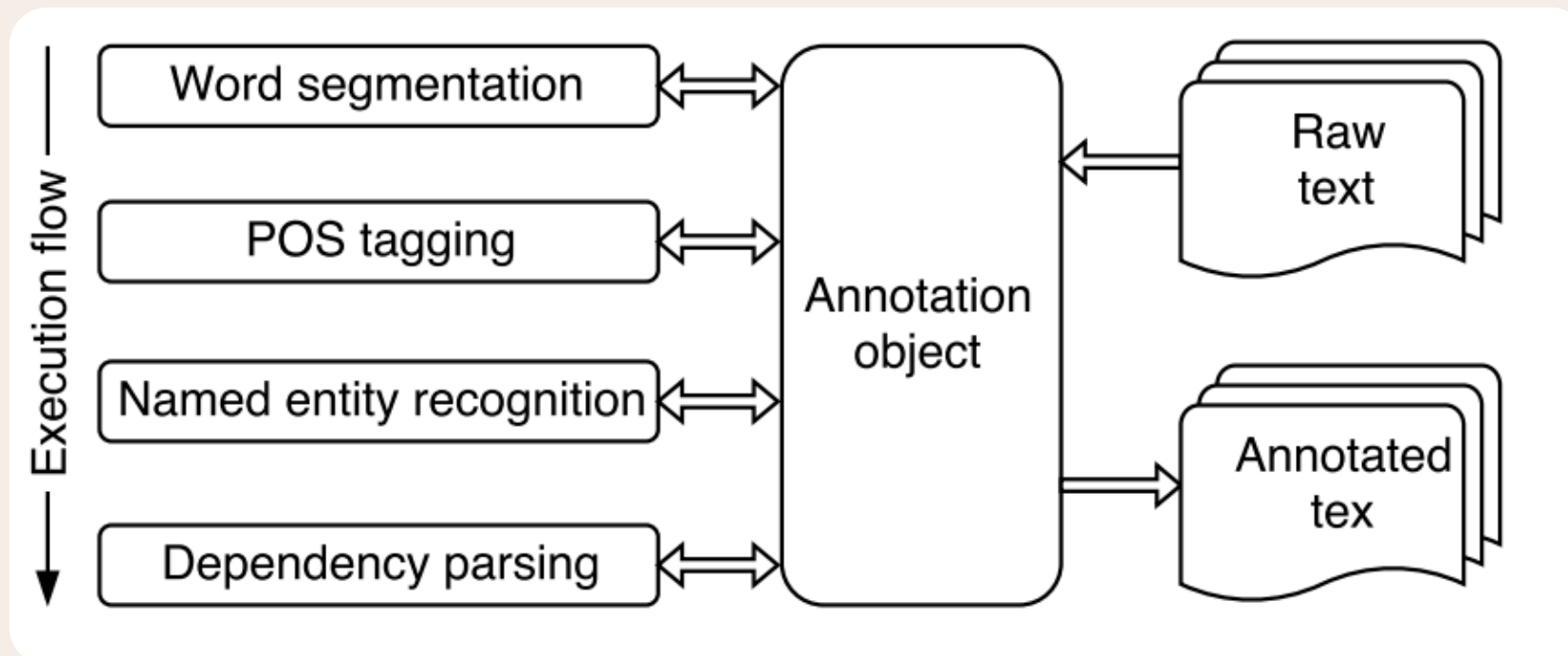
Các ký tự đặc biệt

@, #, \$, ^, &, *, (,), !, :, +, -, =, ...

Tách từ và tách token

Sử dụng **vncorenlp**, một thư viện được công bố bởi chính tác giả của **PhoBERT**.

- Java 1.8+ (Yêu cầu)
- File VnCoreNLP-1.2.jar (27MB) và thư mục models (115MB) cần được đặt trong cùng một thư mục làm việc.



Mô hình vẫn còn gặp nhiều vấn đề liên quan đến nhập nhằng ngữ nghĩa

Nhập nhằng ngữ nghĩa

Một cụm từ có thể tạo ra nhiều cách ghép nghĩa, ví dụ trong “từ trường phái” thì “từ trường” và “trường phái” đều có nghĩa.

→ **Mô hình tách thành “từ trường” và “phái” là sai.**

Nhập nhằng kết hợp

Một từ ghép đẳng lập có các thành phần đều mang nghĩa riêng, ví dụ “hoa cỏ” gồm “hoa” và “cỏ” đều là từ đơn có nghĩa.

→ **Mô hình chưa tách token thành công với từ ghép đẳng lập.**

Thành ngữ, tục ngữ

Thành ngữ, tục ngữ như “nước sôi lửa bỏng” hay “đầu voi đuôi chuột” cần được tách đúng thành cụm thay vì tách riêng từng từ.

→ **Mô hình chưa tách token thành công với từ ghép thành ngữ, tục ngữ.**

Ngộ nhận tên riêng

Cụm từ “Olympic Tin học quốc tế” bị phân tách sai thành “Olympic_Tin”, “học” và “quốc tế”

→ **Mô hình nhận diện ngữ cảnh nhầm lẫn rằng “Olympic Tin” là một tên riêng, chỉ vì cả hai từ đều viết hoa.**

Tiền xử lý dữ liệu

Xử lý dữ liệu nhiều

Stop words

Thường xuất hiện với tần suất cao nhưng không đóng góp nhiều vào việc phân tích ngữ nghĩa của văn bản.

Xây dựng thủ công bổ sung danh sách stopwords dựa trên bộ dữ liệu.

Bao gồm: đại từ, mạo từ, trạng từ, giới từ, liên từ, câu hỏi,...

Loại từ	Ví dụ
Mạo từ	Cái, con, nó, chiếc, ấy, ...
Đại từ	Tôi, chúng tôi, bạn, họ, ...
Trạng từ	hầu như, thỉnh thoảng, ...
Liên từ	nhưng, tuy nhiên, mặc dù, hoặc,...

Tiền xử lý dữ liệu

Thông số đặc biệt

	Date	Long numbers	ISBN	IP	Percentage	Time	Comments	Mail	URL
Tần suất trong Tag	0	4	0	0	0	1	0	0	0
Phần trăm trong Tag	0%	0%	0%	0%	0%	0%	0%	0%	0%
Tần suất trong Content	3896	269	68	168	5	4419	466	91	295
Phần trăm trong Content	0.03%	0%	0%	0%	0%	0.04%	0%	0%	0%

Phân cụm theo topic ẩn

LDA

(Latent Dirichlet Allocation)

Nguyên lý hoạt động

LDA giả định rằng:

- Content là một tập hợp các từ được tạo ra từ một hoặc nhiều topic.
- Topic là một phân phối xác suất trên các từ.

Dựa trên phân phối Dirichlet

- Phát hiện các topic tiềm ẩn trong tập content.
- Gán từ trong content vào các topic với xác suất cao nhất.

Biểu diễn dưới dạng công thức toán học

- Ma trận phân phối topic-content:
 - Content nào liên quan đến topic nào (với tỷ lệ bao nhiêu).
- Ma trận phân phối từ-topic:
 - Từ nào phổ biến trong topic nào.

Kết quả

- Phân phối topic-content: Ví dụ, content 1 là 70% về "Thể thao" và 30% về "Công nghệ".
- Phân phối từ-topic: Ví dụ, topic "Thể thao" có 40% từ "bóng đá", 30% từ "cầu thủ", 20% từ "trận đấu".

Tìm phân phối topic-content (θ):

- Mỗi content là sự pha trộn của các topic.
 - **Document 1:** [0.7 (topic 1), 0.3 (topic 2)]
 - **Document 2:** [0.2 (topic 1), 0.8 (topic 2)]

Tìm phân phối từ-topic (ϕ):

- Mỗi topic là sự kết hợp của các từ.
 - **Topic 1:** ["ai" (0.5), "tech" (0.3), "industry" (0.2)]
 - **Topic 2:** ["cooking" (0.4), "recipes" (0.4), "food" (0.2)]



Xây dựng topic ẩn

Xây dựng Document-Term Matrix

- Ma trận $M \times N$, trong đó M là số content và N là số từ trong từ điển.
- Mỗi phần tử $A[i,j]$ là số lần từ j xuất hiện trong content i .

Gán từng từ trong content vào các topic

- Khởi tạo, mỗi từ trong content được gán **ngẫu nhiên** vào một topic. Quá trình cập nhật diễn ra bằng cách tính toán **xác suất từ đó thuộc về topic nào**, dựa trên hai yếu tố:
 - Xác suất topic xuất hiện trong content.
 - Xác suất từ đó thuộc về topic.

Tối ưu hóa phân phối bằng Gibbs Sampling hoặc Variational Inference đến khi hội tụ

- **Hội tụ bằng Gibbs Sampling:** Khi số lần thay đổi gán topic cho các từ không còn nhiều.
- **Hội tụ bằng Variational Inference:** Khi các tham số θ và ϕ không còn thay đổi đáng kể sau mỗi vòng cập nhật.

Phân tích và diễn giải topic

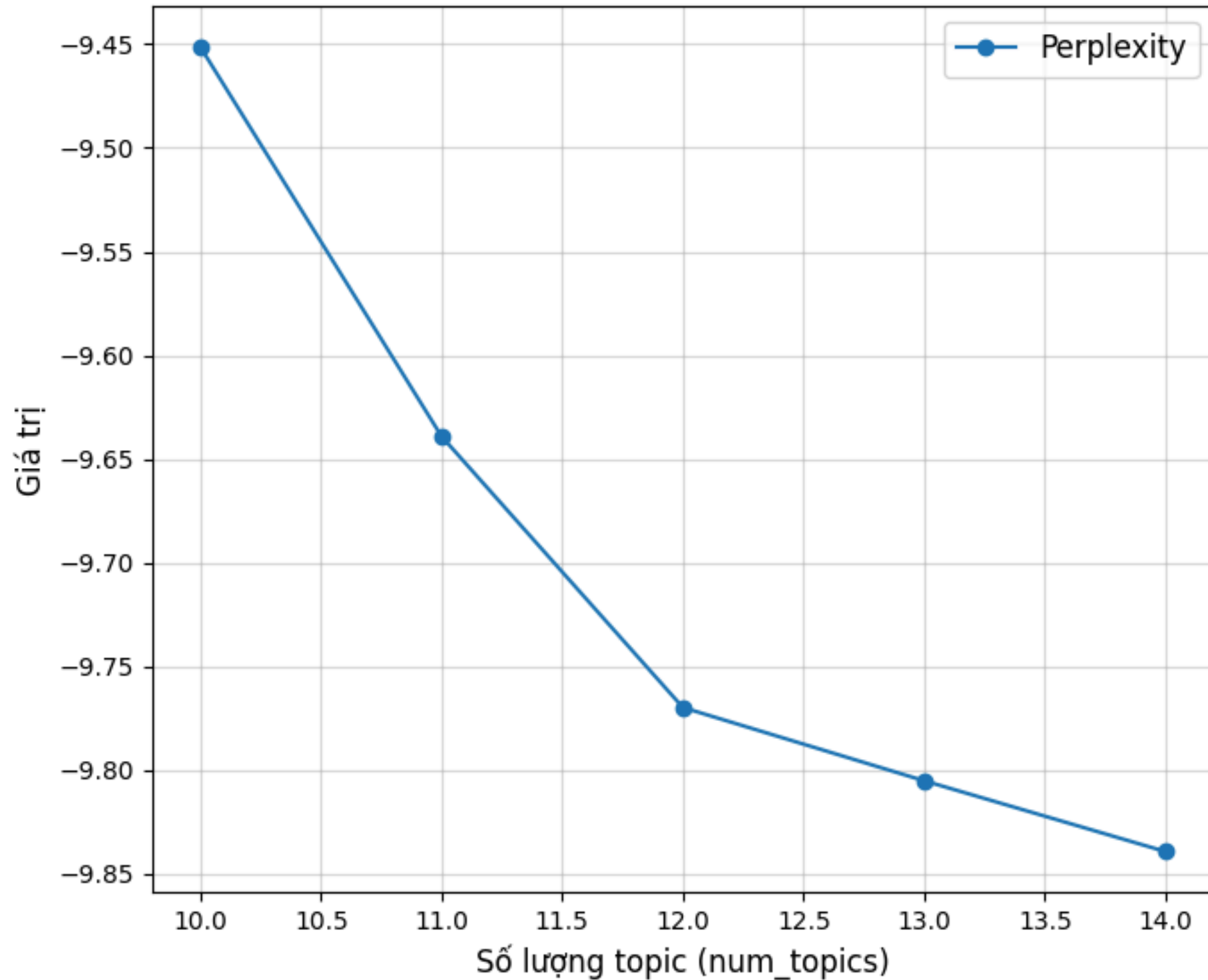
- Diễn giải các topic dựa trên các từ xuất hiện nhiều nhất trong mỗi topic.

Phương pháp đánh giá

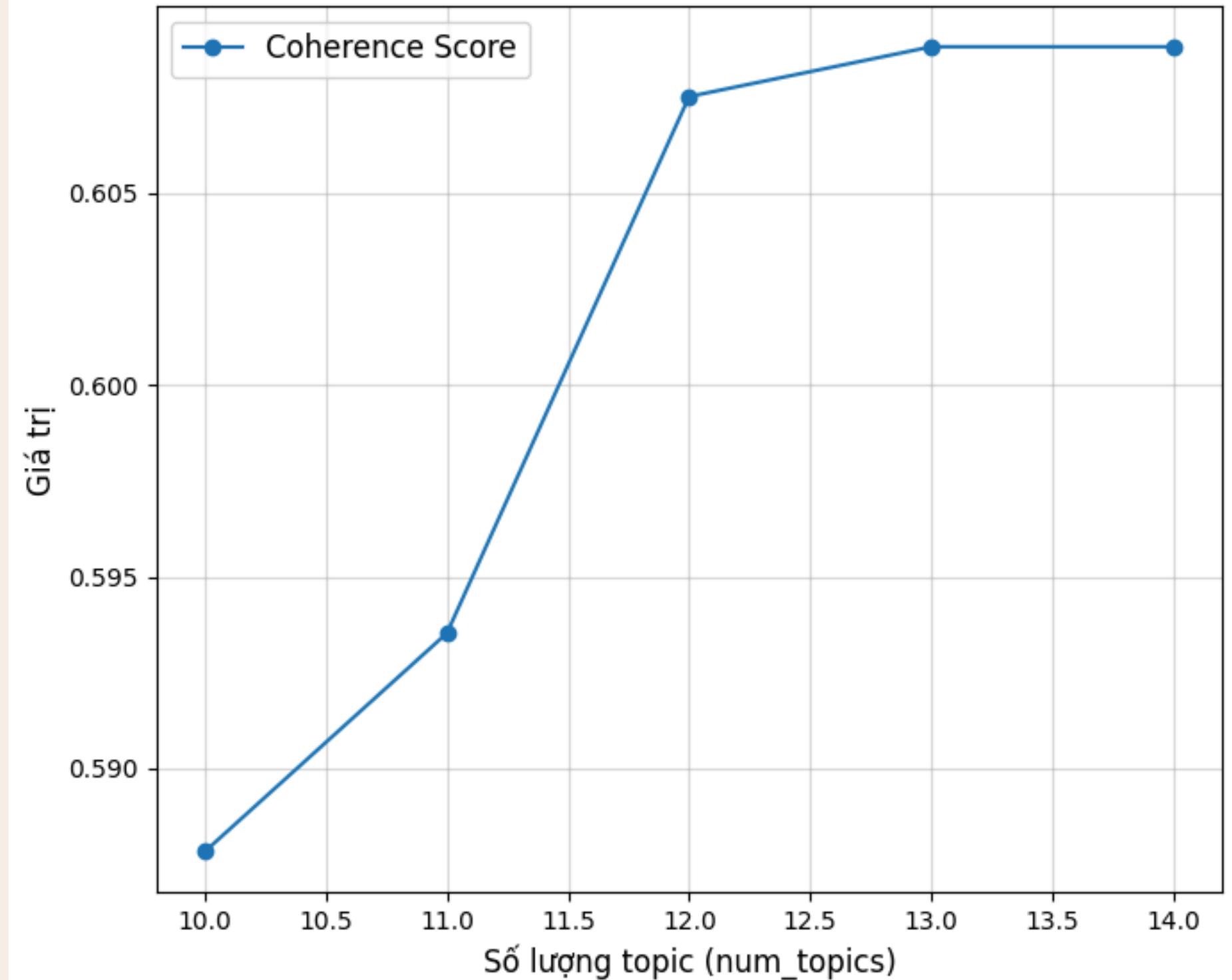
Chỉ số	Coherence Score	Perplexity
Khái niệm	Đo lường mức độ liên quan và tính hợp lý của các từ trong mỗi topic. Topic có độ đồng nhất cao sẽ có các từ có mối quan hệ chặt chẽ với nhau.	Đo độ "khó khăn" trong việc dự đoán một content từ mô hình LDA. Perplexity càng thấp, mô hình càng tốt trong việc nắm bắt cấu trúc của dữ liệu.
Ứng dụng	Đánh giá tính hợp lý của các topic.	Đánh giá khả năng dự đoán và tổng quát của mô hình.
Cách tính	<ul style="list-style-type: none">LDA chọn các từ có xác suất cao nhất cho mỗi topic.Dùng công thức UMass hoặc c_v để tính độ tương quan giữa các từ.	$\text{Perplexity}(D) = \exp \left(-\frac{1}{M} \sum_{i=1}^M \log p(w_i \theta) \right)$
Mối quan hệ với chất lượng topic	Phản ánh trực tiếp chất lượng ngữ nghĩa của topic.	Không trực tiếp phản ánh chất lượng ngữ nghĩa.

Lựa chọn số lượng Topic

So sánh Perplexity theo số lượng topic



So sánh Coherence score theo số lượng topic



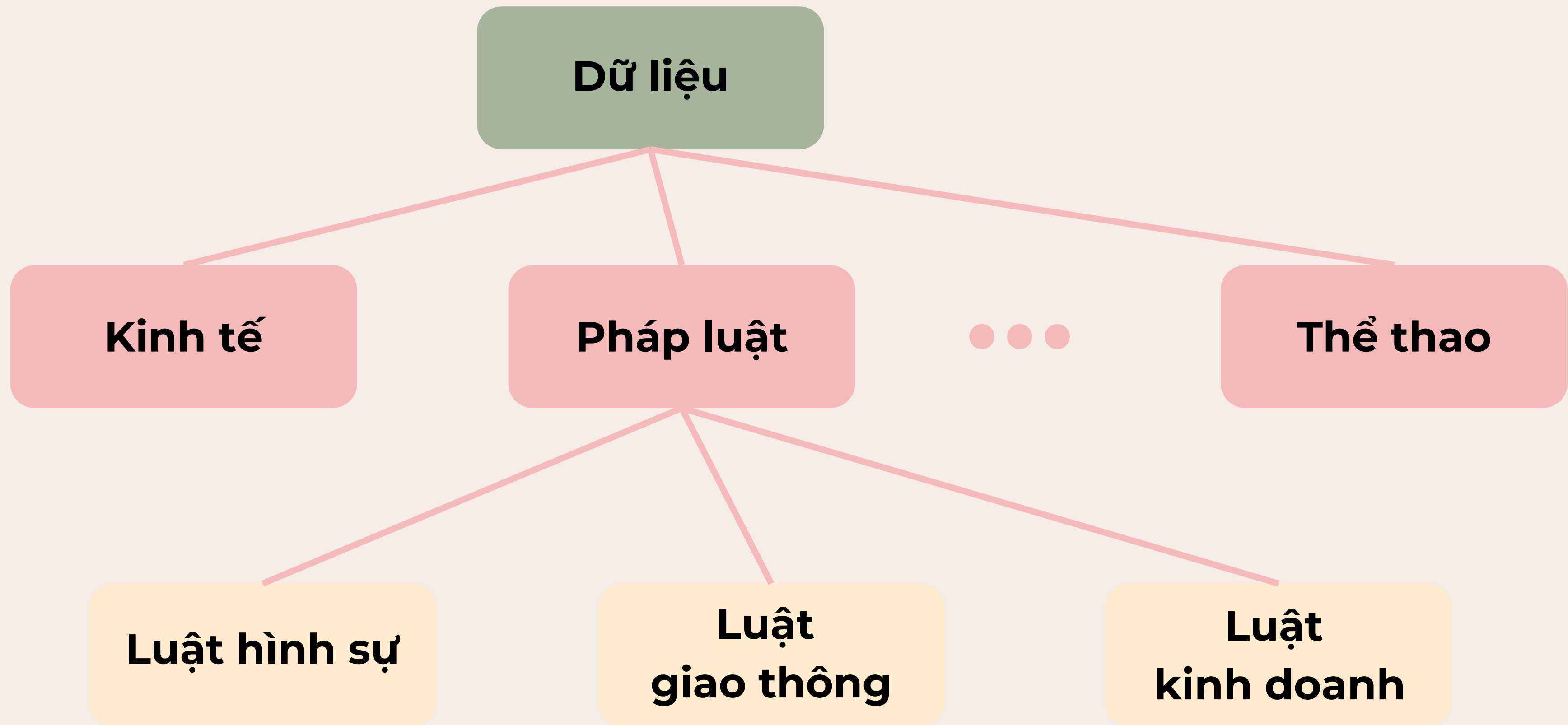
Một số từ khóa

Pháp luật	Quốc tế	Đất đai	Thời sự	Thể thao	Giải trí
công_an	israel	quy_định	phát_triển	trận	gia_đình
người	trung_quốc	luật	việt_nam	giải	diễn_viên
tỉnh	mỹ	dự_án	đầu_tư	đội	thiết_kế
vụ	tấn_công	triển_khai	xây_dựng	đấu	phụ_nữ
vi_phạm	nga	đất	doanh_nghiệp	bóng	gia_đình
đối_tượng	nước	đề_nghị	sản_phẩm	việt_nam	khán_giả
thông_tin	bầu_cử	đầu_tư	dịch_vụ	HLV	ca_sĩ
xe	năm	địa_phương	kinh_tế	cầu_thủ	giúp
số	ukraine	quy_hoạch	ngành	CLB	khán_giả
hành_vi	tổng_thống	cơ_sở	công_nghê	thi_đấu	đẹp

quan trọng

Công nghệ	Sức khỏe	Giáo dục	Xã hội	Kinh doanh	Khí tượng
tỷ	bệnh	thi	việt_nam	đấu_giá	mưa
apple	bác_sĩ	điểm	tổ_chức	đồng	gió
xe	bệnh_viện	ngành	chủ_tịch	thị_trường	sạt_lở
triệu	món	tốt_nghiệp	công_tác	vàng	đất
iphone	đường	trường	tỉnh	nhà_đầu_tư	khu_vực
điện	điều_trị	kết_quả	đại_hội	bán	biển
thuế	thịt	đại_học	lãnh_đạo	usd	bão
khách_hàng	bệnh_nhân	xét	nhân_dân	mua	thiên_tai
tiền	thuốc	học	đoàn_kết	phiên	tuyến
công_ty	người	học_sinh	trung_ương	lượng	cấp

theo từng topic



Mô hình trích xuất đa nhãn

- Chỉ tính TF-IDF trên các từ ghép

```
Số từ đơn trong cột 'tags': 256659
Tổng số từ trong cột 'tags': 1613859
Tỉ lệ từ đơn: 15.90%
```

→ TF-IDF

Mục đích

- Đánh giá mức độ quan trọng của một từ trong một content so với toàn bộ tập content.

TF (Tần suất từ)

- Tần suất xuất hiện của từ trong một content

$$TF(t) = \frac{\text{số lần từ } t \text{ xuất hiện}}{\text{tổng số từ trong tài liệu}}$$

IDF (Nghịch đảo tần suất content)

- Giảm ảnh hưởng của các từ phổ biến (như "và", "là")

$$IDF(t) = \log \frac{\text{số tài liệu}}{\text{số tài liệu chứa từ } t}$$

Công thức TF-IDF

$$TF-IDF(t) = TF(t) \times IDF(t)$$

Thư viện hỗ trợ

- Summa

Ưu điểm

- Nhanh chóng và dễ triển khai
- Không cần dữ liệu huấn luyện (unsupervised).
- Kết quả tương đối tốt cho các văn bản ngắn và rõ ràng.
- Có thể thay đổi các tham số như số lượng từ/câu hoặc tỷ lệ tóm tắt.

Khuyết điểm

- Phụ thuộc vào chất lượng tokenization và POS tagging.
- Không xử lý ngữ cảnh ngữ nghĩa sâu như các mô hình NLP hiện đại (GPT, BERT).



TextRank

Mục đích

- Thuật toán xếp hạng dựa trên đồ thị, dùng để trích xuất từ khóa và tóm tắt văn bản tự động.

Gán nhãn từ loại

- Các từ trong văn bản sẽ được phân loại theo từ loại (danh từ, động từ, tính từ, v.v.). Các từ quan trọng thường là danh từ, động từ, và tính từ, vì chúng mang nhiều ý nghĩa ngữ nghĩa.

Xây dựng đồ thị

- Mỗi từ hoặc câu là một nút.
- Kết nối các nút nếu chúng có mối quan hệ ngữ nghĩa (ví dụ: từ xuất hiện gần nhau, câu liên quan).

Xếp hạng nút

- Sử dụng thuật toán PageRank để tính mức độ quan trọng của mỗi nút dựa trên các kết nối.

$$R(V_i) = (1 - d) + d \sum_{V_j \in In(V_i)} \frac{R(V_j)}{Out(V_j)}$$

Trích xuất từ khóa

- 16 từ quan trọng được xếp hạng cao nhất.

Điểm mạnh của TextRank

Xây dựng cụm từ (N-grams)

Khi trích xuất từ khóa, các từ thường được nhóm thành **cụm từ (n-grams)**, và POS tagging giúp xác định xem cụm từ nào là hợp lý và có ý nghĩa.

Ví dụ:

- “an_toàn giao_thông” là một cụm từ do hay xuất hiện cùng nhau, và là một danh từ (noun + noun).

Lựa chọn từ quan trọng dựa trên POS tagging

Các từ thường xuất hiện trong từ khóa bao gồm:

- **Danh từ (Nouns)**: Mang thông tin cụ thể, định danh đối tượng hoặc khái niệm (ví dụ: "machine learning", "data").
- **Động từ (Verbs)**: Miêu tả hành động hoặc trạng thái, giúp xác định mối quan hệ giữa các khái niệm (ví dụ: "automates", "understanding").

Điểm yếu của TextRank

Bỏ sót một số từ khóa quan trọng mang tính đại diện nếu từ đó không nằm trong các cụm liên kết chặt chẽ.

Trận chung kết “Đường lên đỉnh Olympia” năm 2024, Võ Quang Phú Đức đã dẫn đầu tất cả các vòng thi để giành ngôi **quán quân** với số điểm 220. Đây là vòng nguyệt quế thứ 3 của trường THPT Chuyên Quốc Học Huế nói riêng và của tỉnh Thừa Thiên Huế nói chung. Trước đó, Võ Quang Phú Đức có những thành tích đáng nể như: **Học sinh giỏi quốc gia, Huy chương Bạc quốc tế**... Phát biểu tại buổi lễ, Chủ tịch UBND tỉnh Thừa Thiên Huế Nguyễn Văn Phương nhấn mạnh, "đạt được vòng nguyệt quế là một thành quả vô cùng xứng đáng với những cố gắng của Phú Đức, chúng ta tự hào về em. Những thành công của em là nguồn cảm hứng mạnh mẽ cho các bạn trẻ khác trong tỉnh và cho chúng ta"...



Gộp TextRank & TF-IDF

Mục đích

- Vừa lấy được từ khóa theo từng cụm, vừa lấy được từ khóa riêng của bài

Khai thác điểm mạnh của TextRank

- Tìm ra những từ quan trọng, xuất hiện nhiều trong chính bài báo đấy nhưng không phải từ khóa chính của cụm.

Khai thác điểm mạnh của TF-IDF

- Xác định được những từ khóa mang tính đại diện cho topic chính nhưng không xuất hiện nhiều trong cụm từ khóa của TextRank.
- Bổ sung những từ khóa quan trọng nhưng không được đánh giá cao bởi TextRank.

Kết hợp

- Sử dụng TextRank để tìm ra từ khóa, sau đó gán thêm các từ khóa mà TextRank bỏ sót bằng TF-IDF



Sử dụng mô hình phân loại nhị phân để trích xuất đa nhãn

Bản chất bài toán trích xuất từ khóa

Mỗi token trong văn bản là từ khóa hay không phải từ khóa?

Đặc trưng cần thiết

- Word embedding
- TF-IDF của token đó trong toàn cụm
- TF-IDF của token đó trong văn bản
- POS Tag (N, V, A)
- isName (token có phải tên riêng hay không)
- Label (token có phải tag hay không)

	idx	Từ	TF-IDF Value	POS	Tag_N	POS	Tag_V	Labels
0	0	ngưỡng	0.010582		1.0		0.0	0
1	0	u40	0.005998		0.0		1.0	1
2	0	thành_viên	0.091738		1.0		0.0	0
3	0	snsd	0.001853		0.0		1.0	0
4	0	duy_trì	0.112928		1.0		0.0	0
...
1069811	4107	lan_toà	0.023215		0.0		1.0	0
1069812	4107	ưa_chuộng	0.073891		1.0		0.0	0
1069813	4107	thế_giới	0.182701		0.0		0.0	0
1069814	4107	loài	0.076173		1.0		0.0	0
1069815	4107	sinh_vật	0.016872		1.0		0.0	0

[1066142 rows x 6 columns]



Micro Precision: 0.21
Micro Recall: 0.32
Micro F1-Score: 0.25

Điểm yếu

- Thời gian chạy lâu ~1 tiếng/cụm con
- Dữ liệu quá lớn, Kaggle bị quá tải
- Phụ thuộc rất lớn vào khả năng tokenize của vncorenlp - thứ mà nhóm không kiểm soát được
- Không thể dự đoán từ khóa được ghép từ ≥ 2 token

- Không biết trước label của thầy
- Không xử lý tốt từ khóa hiếm hoặc từ mới
- Không tự tin về khả năng tổng quát hóa của mô hình



TEAM 3

Thank You
for your time and attention