Đại học quốc gia TP.HCM
Trường đại học công nghệ thông tin

Khoa khoa học máy tính

Lớp: CS221.O11

BÁO CÁO ĐỒ ÁN CUỐI KÌ

MULTI-STAGE DOCUMENT RANKING FOR VIETNAMESE NEWS RETRIEVAL

GV hướng dẫn: TS. Nguyễn Trọng Chỉnh, ThS. Nguyễn Đức Vũ

Sinh viên thực hiện
21521024 - Nguyễn Trung Kiên
21522792 - Phạm Quốc Việt
21522121 - Nguyễn Văn Hùng

Mục lục

1 Lý do lựa chọn đề tài

Ngày nay, với sự phát triển của công nghệ thông tin, các công cụ tìm kiếm như Google, Bing,... đã trở thành một phần không thể thiếu trong cuộc sống của con người. Chỉ với vài cú click chuột, chúng ta có thể tìm thấy mọi thông tin mình cần chỉ trong tích tắc. Điều mà các thành viên trong nhóm thắc mắc, đó là **làm thế nào mà những công cụ này có thể hiểu được những câu truy vấn của chúng ta và trả về các kết quả liên quan đến những gì chúng ta muốn?** Việc tìm hiểu và xây dựng một hệ thống có thể làm điều tương tự như vậy là **lý do chính** mà nhóm thực hiện đề tài.

Hơn nữa, Internet lúc này đã trở thành một nguồn tài nguyên thông tin khổng lồ. Tuy nhiên, việc tìm kiếm thông tin trên Internet, đặc biệt là đối với các **bài báo tiếng Việt**, vẫn còn gặp nhiều khó khăn. Việc hướng đến truy vấn các bài báo Việt Nam (Vietnamese News Retrieval) cho đề tài xuất phát từ việc chúng em quan sát được các trang web tổng hợp tin tức từ các báo khác như baomoi.com - một nguồn thông tin phong phú và đa dạng, lại có hạn chế trong việc xếp hạng các bài báo theo độ liên quan với truy vấn tìm kiếm. Vì thế, nhóm mong muốn thực hiện một cách tiếp cận nhằm xử lý vấn đề đã nêu.

Ngoài ra, tài liệu và nguồn tài nguyên cũng là một lý do quan trọng. Đề tài mà nhóm thực hiện thuộc lĩnh vực Information Retrieval (Truy vấn thông tin) và Natural Language Processing (Xử lý ngôn ngữ tự nhiên), là một lĩnh vực phổ biến trong cộng đồng nghiên cứu khoa học máy tính nói chung và nghiên cứu khoa học nói riêng. Do đó, có rất nhiều tài liệu và nguồn tài nguyên hữu ích để chúng em có thể tham khảo và học tập như các bài báo khoa học, các mã nguồn,... để phục vụ cho đồ án.

Nhóm chúng em mong muốn rằng những gì mà nhóm đã học được từ lớp **CS221.O11** được thể hiện qua đồ án môn học có thể phần nào đóng góp vào thực tế. Chúng em vô cùng cảm ơn **TS. Nguyễn Trọng Chỉnh, ThS. Nguyễn Đức Vũ** đã nhiệt tình giảng dạy, giải đáp cho chúng em những vấn đề còn thắc mắc trong quá trình học tập, hỗ trợ nhóm trong quá trình thực hiện đồ án cuối kỳ. Đồ án chắc chắn không tránh khỏi những sai sót, chúng em hi vọng nhận được những góp ý của các thầy về đồ án của mình.

2 Giới thiệu bài toán

2.1 Đặt vấn đề

Cho một truy vấn q và một ngữ liệu D là tập hợp nhiều bài báo khác nhau, trong đó mỗi bài báo bao gồm tiêu đề (title) và tổng quan (abstract). Hãy xây dựng một hệ thống truy vấn giúp trả về các bài báo từ D thỏa mãn các tính chất sau:

- Chứa các ngữ cảnh quan trọng liên quan đến q.
- **Ranking:** Xếp hạng các bài báo từ trên xuống dưới dựa vào độ đo nào đó nhằm thể hiện mức độ liên quan với truy vấn.

2.2 Mô tả ngữ liệu

Ngữ liệu được thu thập từ các bài báo trên 5 trang báo điện tử gồm: Lao Động, Dân Trí, VnExpresses, VTC và báo Đảng Cộng Sản. Với mỗi website được craft dữ liệu, ta sẽ thu thập 7 features bao gồm: title (tiêu đề), abstract (tổng quan), source (nguồn bài báo), link (đường dẫn bài báo), topic (chủ đề), time (thời gian đăng bài) và imglink (đường dẫn chứa ảnh tượng trưng cho bài báo). Số lượng các bài báo được thu thập là 49542.

Hình 1: Ví dụ về một bài báo trên báo điện tử

Chẳng hạn, ở hình 1 là một bài báo trên báo Lao Động, ta thu được các trường dữ liệu sau:

- title: Những nhân tố mới của tuyển Việt Nam tại Asian Cup 2023

- abstract: Danh sách tuyển Việt Nam chuẩn bị cho vòng chung kết Asian Cup 2023 xuất hiện nhiều nhân tố đáng chú ý như: Filip Nguyễn, Hai Long và Văn Việt.
- source: Báo Lao Động
- link: https://laodong.vn/the-thao/nhung-nhan-to-moi-cua-tuyen-viet-nam-tai-asian-cup-2023-1288638.ldo
- topic: Thể thao
- time: 05/01/2024 06:54
- imglink: https://media-cdn-v2.laodong.vn/storage/newsportal/2024/1/4/1288638/Filip-Nguyen.jpg

Dưới đây là phần đầu trong tập dataset mà nhóm thực hiện:

```
data.head()
```

	title	abstract	source	link	topic	time	imglink
0	Thành nhà Hồ miễn phí tham quan nhân Ngày di s...	Thanh Hóa - Nhân kỷ niệm 18 năm ngày Di sản vă...	Báo Lao Động	https://laodong.vn/xa-hoi/thanh-nha-ho-mien-ph...	Xã hội	21/11/2023 17:20	https://media-cdn-v2.laodong.vn/storage/newspo...
1	Ngang nhiên thu phí tại cầu phao đã bị cấm hoạ...	Thái Nguyên - Mặc dù đã bị cấm hoạt động hơn m...	Báo Lao Động	https://laodong.vn/xa-hoi/ngang-nhien-thu-phi-...	Xã hội	21/11/2023 17:08	https://media-cdn-v2.laodong.vn/storage/newspo...
2	Trong 10 tháng, Cần Thơ xảy ra 39 vụ sạt lở gâ...	Cần Thơ - So với cùng kỳ năm 2022, trong 10 th...	Báo Lao Động	https://laodong.vn/xa-hoi/trong-10-thang-can-t...	Xã hội	21/11/2023 16:48	https://media-cdn-v2.laodong.vn/storage/newspo...
3	Hiện trạng nút giao đang thí điểm bỏ đèn giao ...	Nút giao Châu Văn Liêm - Lê Quang Đạo - Mễ Trì...	Báo Lao Động	https://laodong.vn/video-xa-hoi/hien-trang-nut...	Xã hội	21/11/2023 16:44	https://media-cdn-v2.laodong.vn/storage/newspo...
4	Dự án chống ùn tắc hơn 800 tỉ đồng ở Hà Nội sa...	Hà Nội - Khởi công cuối năm 2019, Dự án mở rộn...	Báo Lao Động	https://laodong.vn/photo/du-an-chong-un-tac-ho...	Xã hội	21/11/2023 16:38	https://media-cdn-v2.laodong.vn/storage/newspo...

Hình 2: Phần đầu tiên trong tập data tìm kiếm

Thời điểm đăng bài của các bài báo là từ 11 giờ 27 phút, ngày 1 tháng 4 năm 2017 đến 23 giờ 59 phút, ngày 11 tháng 12 năm 2023.

2.3 Mô tả Input/Output

a) Input:

- Ngữ liệu: tập hợp các bài báo trên các báo điện tử.
- Truy vấn mà người dùng muốn tìm kiếm ở dạng text.

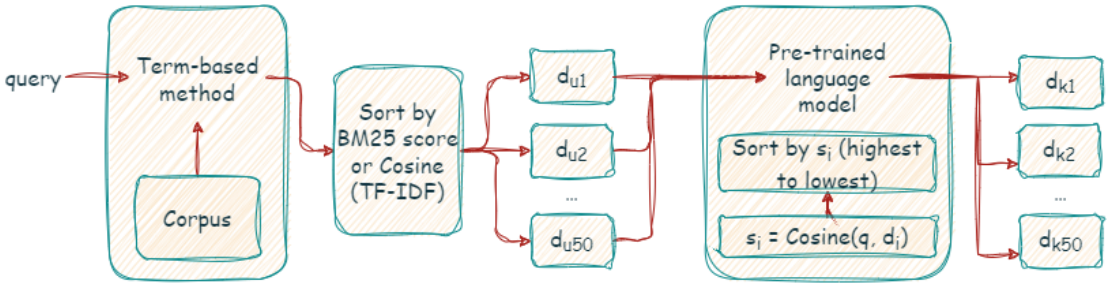
b) Output: Danh sách các bài báo trong ngữ liệu sao cho các bài báo này có liên quan đến truy vấn và được sắp xếp từ trên xuống dưới ứng với mức độ liên quan (nhiều hay ít với truy vấn).

3 Giới thiệu phương pháp chính

Phương pháp **Multi-stage Document Ranking**: Thay vì sử dụng một mô hình đơn lẻ để xếp hạng tất cả các tài liệu, multi-stage document ranking áp dụng một chuỗi các mô hình khác nhau, mỗi mô hình thực hiện một chức năng cụ thể. Hướng thực hiện của nhóm sẽ là sử dụng một **phương pháp term-based** như TF-IDF hay BM25 để lọc các bài báo có nội dung không liên quan, các kết quả được lọc ra từ các phương pháp term-based sẽ được đưa vào một mô hình ngôn ngữ được đào tạo trước (pre-trained language model) để xét về mặt ngữ nghĩa văn bản.

3.1 Pineline của Multi-stage Document Ranking

Ngữ liệu và truy vấn sau khi được tiền xử lý dữ liệu sẽ đi qua pineline như hình 3.



Hình 3: Pineline xử lý bài toán đặt ra

Sử dụng phương pháp term-based, ta có được 50 bài báo "khớp"nhất so với query, sau đó 50 bài báo này sẽ được sắp xếp lại **(reranking)** qua pre-trained language model để sắp xếp được những bài báo khớp với query nhất, cả về mặt ngữ cảnh và ngữ nghĩa.

Ý nghĩa của việc sử dụng pre-trained language model: nắm được ngữ nghĩa của văn bản, do đó khắc phục được nhược điểm không nắm bắt ngữ nghĩa của văn bản của các phương pháp term-based như TF-IDF hay BM25 vì chúng có thể dẫn đến sai sót nếu các từ có nghĩa tương tự được coi là khác nhau. Các phương pháp term-based được sử dụng trước là để giảm bớt lượng dữ liệu được đưa vào pre-trained language model để tính toán, ngoài ra là còn để lọc bớt các bài báo chứa nội dung không liên quan đến truy vấn.

Phần **3.2** và **3.3** sẽ giới thiệu về các phương pháp term-based và các pre-trained language model được sử dụng cho đồ án.

3.2 Các phương pháp term-based

3.2.1 BM25

BM25 là một hàm xếp hạng giúp xếp hạng một tập các tài liệu dựa vào việc tính xác xuất xuất hiện của các từ khóa trong truy vấn của user.

Định nghĩa: Cho một truy vấn q chứa các từ khóa $q_1, q_2, ..., q_n$, BM25 score của một tài liệu D là:

$$score(q, D) = \sum_{i=1}^{n} IDF(q_i) \frac{f(q_i, D)(k_1 + 1)}{f(q_i, D) + k_1(1 - b + b\frac{|D|}{avgdl})}$$

- $f(q_i, D)$ là tần số q_i xuất hiện trong tài liệu D.
- $|D|$ là số lượng các từ trong tài liệu D.
- avgdl là độ dài trung bình số lượng từ của các tài liệu trong ngữ liệu.
- k_1 và b là các hệ số tự do. Thông thường, k_1 nằm trong khoảng $[1.2; 2.5]$ và b được gán bằng 0.75.

Các bài báo có BM25 score cao có nghĩa là chúng chứa những từ khóa có liên quan đến truy vấn. Ngược lại, các bài báo có BM25 score thấp chứa nội dung

không liên quan đến truy vấn và sẽ không được xử lý trong pre-trained language model.

3.2.2 TF-IDF

TF-IDF (term frequency–inverse document frequency) là một thước đo quan trọng để đánh giá mức độ quan trọng của một từ trong một văn bản. TF-IDF được tính bằng cách nhân tần suất xuất hiện của một từ trong văn bản với tần suất nghịch đảo của từ đó trong toàn bộ tập dữ liệu. Tuy nhiên, giá trị TF thường còn được chuẩn hóa bằng cách lấy số lần xuất hiện của từ trong văn bản, chia cho tổng số từ trong văn bản đó.

Cụ thể, với từ khóa t, văn bản d và ngữ liệu D, ta có các công thức được sử dụng để tính các giá trị tf(t, d), idf(t, D), tf-idf(t, d, D) như sau:

$$\text{TF}(t, d) = \frac{\text{Số lần từ } t \text{ xuất hiện trong văn bản } d}{\text{Tổng số từ trong văn bản } d}$$

$$\text{IDF}(t, D) = 1 + \log\left(\frac{\text{Tổng số văn bản trong toàn bộ ngữ liệu } N}{\text{Số văn bản chứa từ } t}\right)$$

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

Thực chất, việc dùng TF-IDF là để biểu diễn từng văn bản và truy vấn dưới dạng một vector, sau đó mới sử dụng một phương thức để so sánh độ tương đồng như L2 hay Cosine similarity để tính toán mức độ tương đồng. Trong đồ án, **cosine similarity** là độ đo được sử dụng.

3.3 Pre-trained language model

3.3.1 PhoBERT

PhoBERT là một mô hình ngôn ngữ được đào tạo trước (pre-trained language model) cho ngôn ngữ Việt. Việc thực thi của PhoBERT cũng tương tự như BERT.

Ta sẽ sử dụng pooled_output của token [CLS] như feature vector cho văn bản. Sau khi đưa qua mô hình PhoBERT, truy vấn và nội dung các bài báo đều đã được chuyển về các feature vector.

3.3.2 Vietnamese-SBERT

Vietnamese-SBERT là một mô hình ngôn ngữ được đào tạo trước được sử dụng để biểu diễn văn bản tiếng Việt dưới dạng feature vector. Mô hình này được xây dựng dựa trên mô hình SentenceBERT, một mô hình học máy phổ biến được sử dụng để biểu diễn văn bản tiếng Anh.

Vietnamese-SBERT được đào tạo trên tập dữ liệu Vietnamese NLI and STSb. Mục đích chính của mô hình này nhận ra là để nhận ra được các văn bản đồng nghĩa nhau, thông qua các feature vector.

4 Cài đặt phương pháp

4.1 Tiền xử lý dữ liệu

Với tác vụ truy vấn, nhóm sử dụng 2 features: title và abstract. Với từng bài báo trong ngữ liệu, dữ liệu dạng text tương ứng với bài báo đó sẽ là kết hợp của title và abstract, sau đó đi qua 4 bước bao gồm:

- Lowercasing (Di sản → di sản)
- Word segmentation (di sản → di_sản)
- Loại bỏ stopwords
- Loại bỏ các ký tự đặc biệt (tp.hcm → tphcm)

Câu truy vấn được nhập vào cũng được đi qua 4 bước tiền xử lý dữ liệu trên.

4.2 Đánh giá mức độ tương đồng giữa truy vấn và văn bản

Để đánh giá và so sánh độ tương đồng giữa truy vấn và nội dung bài báo, nhóm sử dụng phép đo **cosine similarity** cho tác vụ này. Cụ thể như sau:

- Xét 2 vector u và v là các [CLS] đã được padding về kích thước 768x1 sau khi được xử lý qua multi-stage document ranking, độ tương đồng giữa q và u được tính bằng cách tính cosine của 2 vector này.

$$CosSimilarity(u, v) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$$

- Trong đó: $\mathbf{u} \cdot \mathbf{v}$ là tích vô hướng của u và v, $\|\mathbf{u}\|$ và $\|\mathbf{v}\|$ lần lượt là chuẩn 2 của vector u và v. Nếu giá trị CosSimilarity của 2 vector tương ứng với 2 văn bản càng gần 1 thì 2 văn bản này càng đồng nghĩa với nhau và ngược lại, nếu càng gần -1 thì 2 văn bản này càng có ý nghĩa ngược nhau.

5 Đánh giá mô hình

Trong phần đánh giá, nhóm sẽ thực hiện so sánh độ hiệu quả của hướng tiếp cận Multi-stage document ranking cho truy vấn các bài báo tiếng Việt bằng cách kết hợp một trong các phương pháp term-based (BM25, TF-IDF) cùng với một trong các pre-trained language model (PhoBERT, Vietnamese-SBERT), so sánh tất cả các cặp có thể tạo ra, và so sánh hiệu quả của từng phương pháp term-based và pre-trained language model.

Với truy vấn q, ta gọi D_{pred} là danh sách các bài báo đã được sắp xếp từ trên xuống dưới là liên quan với q sau khi sử dụng multi-stage document ranking, D_{truth} là danh sách các bài báo đã được sắp xếp từ trên xuống dưới mà ta thật sự mong muốn. Ta so sánh kết quả giữa mô hình và thực tế dựa trên 10 bài báo đầu tiên bằng $nDCG@10$.

5.1 nDCG@10

Công thức của $nDCG@10$ được tính như sau:

$$nDCG@10 = \frac{DCG@10}{IDCG@10}$$

trong đó:

- $DCG@10 = \sum_{i=1}^{10} \frac{rel_i}{\log_2(i+1)}$, rel_i là độ liên quan của bài báo thứ i trong D_{pred}.
- $IDCG@10 = \sum_{i=1}^{10} \frac{rel_i}{\log_2(i+1)}$, rel_i là độ liên quan của bài báo thứ i trong D_{truth}.
- Giá trị $nDCG@10$ nằm trong đoạn $[0; 1]$, càng gần 1 thì mô hình của ta càng tốt.

5.2 Tập đánh giá

Tập đánh giá gồm 352 bài báo đã được xếp hạng và gán nhãn mức độ liên quan (dành cho việc tính $nDCG@10$) với truy vấn: "Ronaldo giàu cỡ nào?

	title	abstract	link	score
0	Ronaldo giàu cỡ nào? So sánh tài sản của Ronal...	Cristiano Ronaldo là một trong những cầu thủ g...	https://vebongdaonline.vn/ronaldo-giau-co-nao/...	4
1	Tài sản của Ronaldo so với cầu thủ giàu có nhấ...	Siêu sao Cristiano Ronaldo sẽ kiếm được mức lư...	https://bongdaplus.vn/bong-da-the-gioi/tai-san...	4
2	Cristiano Ronaldo giàu cỡ nào: Mỗi tuần kiếm 2...	Với mức thu nhập 117 triệu USD trong năm 2020,...	https://cafef.vn/cristiano-ronaldo-giau-co-nao...	4
3	Ronaldo sở hữu khối tài sản trị giá gần nửa tỷ...	Theo Goal, ước tính tổng giá trị khối tài sản ...	https://cafeland.vn/doanh-nhan/doanh-nhan/rona...	4
4	Tài sản 1 tỉ USD và thương hiệu vượt xa phạm v...	Tạp chí Forbes công bố những ngôi sao thể thao...	https://laodong.vn/bong-da-quoc-te/tai-san-1-t...	3
...	...	...	...	...
347	U não chèn ép suýt chết sau một tháng đau đầu	Chàng trai 21 tuổi nhức đầu, lơ mơ, ngủ lì bì ...	https://vnexpress.net/u-nao-chen-ep-suyt-chet-...	0
348	Thuyên tắc ối sau sinh, sản phụ hai lần chạm c...	Sản phụ 34 tuổi ngưng tim ngưng thở khi vừa si...	https://vnexpress.net/thuyen-tac-oi-sau-sinh-s...	0
349	Bì lợn bị tẩy trắng - món ăn tiềm ẩn độc tố	Bì lợn giàu collagen, là thực phẩm được ưa chu...	https://vnexpress.net/bi-lon-bi-tay-trang-mon-...	0
350	Cách ăn giúp mẹ hai con lấy lại vóc dáng thời ...	Bảo Ly chọn cách tiêu thụ thực phẩm theo "quy ...	https://vnexpress.net/cach-an-giup-me-hai-con-...	0
351	Tỷ lệ tiêm chủng thấp, bệnh sởi bủa vây nước Anh	Số ca mắc sởi ở toàn nước Anh tăng chóng mặt, ...	https://vnexpress.net/ty-le-tiem-chung-thap-be...	0

352 rows × 4 columns

Hình 4: Dữ liệu tập đánh giá

Quy tắc gán nhãn mức độ liên quan được quy định như sau:

- 0: Bài báo không chứa nội dung liên quan đến truy vấn.
- 1: Bài báo chứa thông tin giúp ta có được các nội dung bên lề (một phần các chủ đề) liên quan đến truy vấn.
- 2: Bài báo chứa thông tin hữu ích giúp ta có được các nội dung chứa khía cạnh liên quan đến các chủ đề của truy vấn.
- 3: Bài báo chứa các thông tin mang tính cụ thể hơn so với mức 2 (tức là hơn cả về mặt chủ đề, nó mang đến các thông tin quan trọng về người, sự kiện nào đó...), sao cho giải thích được một phần truy vấn.
- 4: Bài báo chứa các ngữ cảnh quan trọng chứa đựng hoặc giải thích được đầy đủ các thông tin liên quan đến truy vấn.

Mức độ liên quan giữa bài báo ứng với truy vấn *"Ronaldo giàu cỡ nào?"* được thể hiện trên cột "score"của tập đánh giá. Trong tập đánh giá, có 4 bài báo có relevance score là 4, 4 bài báo là 3, 7 bài báo là 2, 6 bài báo là 1 và 331 bài còn lại là 0.

5.3 Kết quả thu được

Giá trị $nDCG$@10 của từng hướng tiếp cận trong bảng dưới (kết quả được làm tròn đến chữ số thập phân thứ 4). Kết quả cao nhất của bảng 1 được **in đậm**, cao thứ nhì được gạch chân.

Bảng 1: Kết quả $nDCG$@10 của một số phương pháp

Hướng tiếp cận	Phương pháp	ndcg@10
Phương pháp term-based	TF-IDF	0.9259
	BM25	0.7433
Pre-trained language model	Vietnamese-SBERT	0.9256
	PhoBERT	0.7383
Multi-stage document ranking	TF-IDF + Vietnamese-SBERT	**0.9582**
	TF-IDF + PhoBERT	0.8019
	BM25 + Vietnamese-SBERT	0.9256
	BM25 + PhoBERT	0.7825

Nhận xét:

- Kết quả đạt được tốt nhất là khi sử dụng TF-IDF + Vietnamese-SBERT với $nDCG$@10 có giá trị **0.9582**, vượt trội hơn hẳn so với việc chỉ sử dụng TF-IDF hay Vietnamese-SBERT. Điều này cũng đã cho thấy được sự hiệu quả của phương pháp **Multi-stage Document Ranking**.

- Giữa Vietnamese-SBERT và PhoBERT, nhóm cho rằng pooled_output của token [CLS] trong mô hình PhoBERT sẽ hiệu quả với tác vụ classification hơn là khi so sánh với Vietnamese-SBERT, một mô hình ngôn ngữ đào tạo trước nhằm phát hiện văn bản đồng nghĩa với nhau.

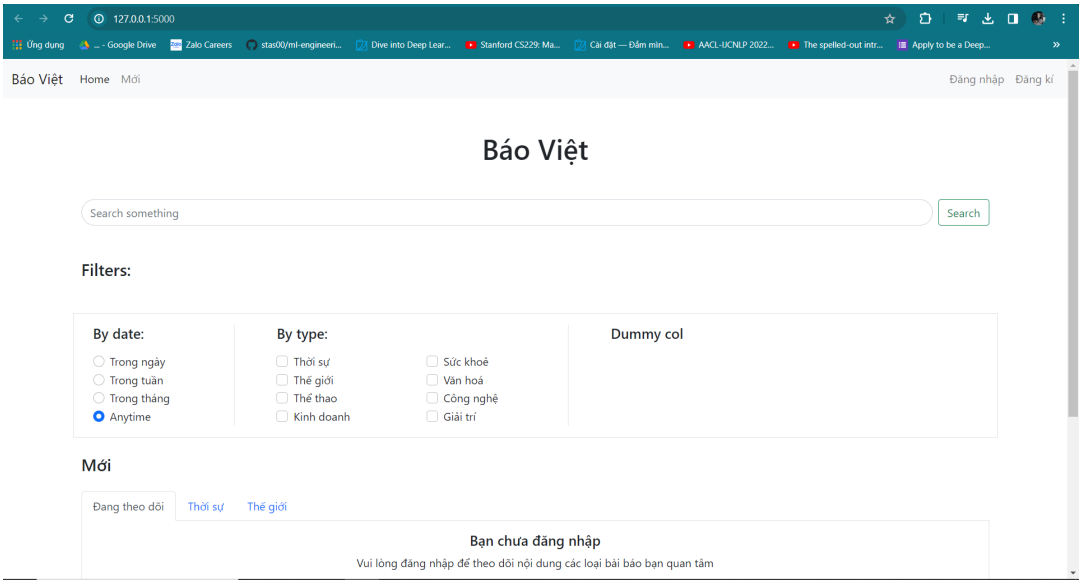
- Về thời gian thực thi của từng phương pháp, vì mỗi lần chạy thời gian luôn có sự sai lệch, không cố định nên nhóm không lập bảng so sánh. Tuy nhiên, sau nhiều lần chạy khác nhau, nhóm nhận thấy thời gian chạy TF-IDF và BM25 đều khoảng dưới 1s và đặc biệt với TF-IDF cho kết quả $nDCG$@10 rất cao. Thời gian chạy các phương pháp khác đều từ 10s trở lên. Do đó, có sự **"trade-off"** giữa thời gian tính toán và độ chính xác của phương pháp. Nếu ta quan tâm đến thời gian người dùng phải chờ, thì TF-IDF chắc chắn

vẫn sẽ là một lựa chọn tốt, còn nếu ta cần ưu tiên độ chính xác nhưng phải chờ lâu hơn một chút, sử dụng TF-IDF kết hợp Vietnamese-SBERT sẽ là lựa chọn ổn nhất.

6 Hướng mở rộng và phát triển

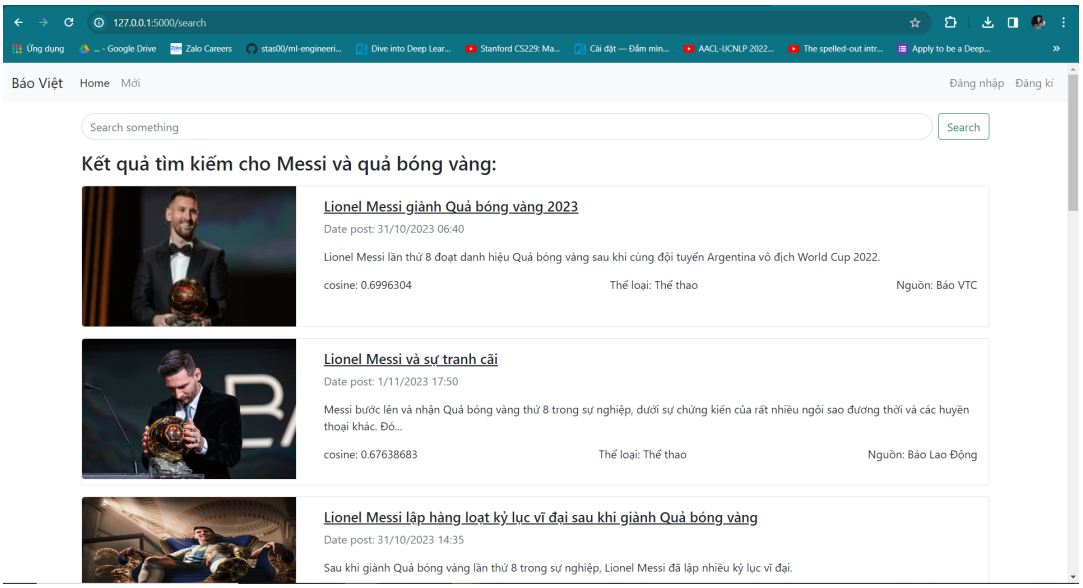
Tập dataset mà nhóm tự thu thập có thể được sử dụng để huấn luyện mô hình cho bài toán Multi-label classification, nhằm mục đích phân biệt thể loại báo điện tử (chính trị, thể thao...).

Hơn nữa, nhóm đã hoàn thành một trang web (localhost) sử dụng Flask trên framework bootstrap5 nhằm đưa mô hình vào ứng dụng thực tế.



Hình 5: Giao diện web giúp tìm kiếm các báo điện tử liên quan đến truy vấn

Khi nhập xong câu truy vấn, sau khi bấm nút "Search"hoặc nhấn phím Enter, các kết quả tương ứng sẽ được trả về theo thứ tự từ trên xuống dưới.



Hình 6: Kết quả cho truy vấn: ”Messi và quả bóng vàng”

Hướng tiếp cận của nhóm có thể được sử dụng cho việc tìm kiếm sách. Chẳng hạn khi một ai đó đã từng đọc một cuốn sách rất hay, nhưng sau này lại không nhớ ra mà chỉ nhớ một chút. Truy vấn lúc này sẽ là mô tả mà người đó còn nhớ về cuốn sách. Dựa vào thông tin đó, mô hình **Multi-stage Document Ranking** sử dụng kết hợp TF-IDF và Vietnamese-SBERT sẽ giúp tìm ra các cuốn sách có thể là cuốn mà người đó tìm kiếm.

7 Tài liệu tham khảo

1. PhoBERT: Pre-trained language models for Vietnamese - https://aclanthology.org/2020.findings-emnlp.92.pdf

2. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding -https://arxiv.org/pdf/1810.04805.pdf

3. Improvements to BM25 and Language Models Examined - https://www.cs.otago.ac.nz/homepages/andrew/papers/2014-2.pdf

4. Rank-BM25: A two line search engine - https://pypi.org/project/rank-bm25/

5. IR-BERT: Leveraging BERT for Semantic Search in Background Linking for News Articles -https://arxiv.org/abs/2007.12603

6. Ad-hoc retrieval with BERT -https://arxiv.org/abs/2007.12603

7. Vietnamese-SBERT: https://huggingface.co/keepitreal/vietnamese-sbert

8. TF-IDF là gì?: https://vi.wikipedia.org/wiki/Tf%E2%80%93idf

9. Thực thi TF-IDF: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html