Xây dựng hệ thống đề xuất tin tức dựa vào bình luận của người dùng

Vũ Hữu Tùng, Nguyễn Văn Hữu Nghĩa and Mai Đức Thuận

Trường Đại Học Công Nghệ Thông Tin - Đại Học Quốc Gia TP.HCM, Việt Nam.

Contributing authors: 19522497@gm.uit.edu.vn; 19521900@gm.uit.edu.vn; 19522316@gm.uit.edu.vn;

Tóm tắt nội dung

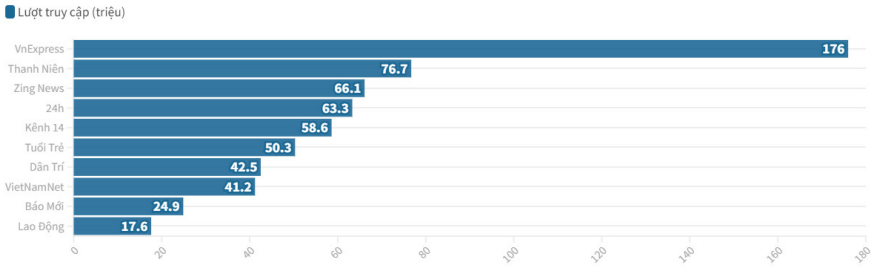
Nhóm chúng em thực hiện một hệ khuyến khị tin tức cho người dùng trên Vnexpress. Nhóm thu thập dữ liệu bình luận của người dùng và các tin tức trên trang báo Vnexpress. Từ đó nhóm chúng em xây dựng hệ khuyến nghị dựa trên bộ dữ liệu. Phương pháp nhóm chúng em sử dụng là TF-IDF và LDA model để xử lý bình luận của người dùng. Để xử lý tin tức chúng em tính các sentiment core, topic modeling score và metric score dựa vào 2 thành phần trên. Từ metric score hệ thống sẽ đề xuất các tin tức cho người dùng.

Keywords: Recommendation system - Hệ khuyến nghị - Tin tức - Vnexpress

1 Giới thiệu

1.1 Thống kê nhu cầu đọc báo điện tử tại Việt Nam

Đọc báo điện tử ở Việt Nam đang ngày càng phát triển. Mọi người có thể tìm được rất nhiều tin tức trên Internet. Nó rất dễ tiếp cận, giúp mọi người cập nhật tin tức nhanh hơn và đặc biệt phù hợp với mọi lứa tuổi.



Hình 1 TOP 10 báo điện tử và trang thông tin điện tử tổng hợp có lượt truy cập lớn nhất Việt Nam tháng 07/2022

Qua hình trên cho chúng ta thấy nhu cầu đọc báo điện tử là rất lớn. Đặc biệt là trang báo Vnexpress cao vượt trội so với các trang còn lại - cao hơn gấp đôi so với trang thứ 2 là "Thanh Niên". Do đó nhóm chúng em quyết định thu thập dữ liệu trên trang báo Vnexpress.

1.2 Mục tiêu của đề tài

Đề tài khuyến nghị tin tức hỗ trợ rất nhiều cho người dùng khi họ muốn tìm kiếm tin tức nào đó. Hệ thống giúp người dùng nhanh chóng tìm được tin tức mình mong muốn góp phần nâng cao sự trải nghiệm của ho. Đồng thời hệ khuyến nghị giúp tăng khả năng quay trở lại sử dụng dịch vụ của người dùng. Các trang báo không những giữ chân được người dùng củ mà còn thu hút được nhiều người dùng mới,lợi nhuận gia tăng, từ đó hỗ trợ rất nhiều cho việc kinh doanh.

Nhóm chúng em đặt mục tiêu xây dựng một hệ thống khuyến nghị tin tức trên VnExpress. Chúng em thu thập dữ diệu bình luận của người dùng và các tin tức thực tế trên Vnexpress. Từ những bình luận thực tế của người dùng về một tin tức nào đó để nhóm áp dụng những thuật toán khuyến nghị để đưa ra những tin tức phù hợp với nu cầu của người dùng.

2 Bộ dữ liệu

Nhóm chúng em thu thập 2 bộ dữ liệu: Bộ dữ diệu về bình luận của người dùng và bộ dữ liệu về các tin tức

2.1 Dữ liệu bình luận của người dùng

2.1.1 Crawl dữ liệu

Nguồn crawl dữ liệu: Chúng em thực hiện crawl trực tiếp từ trang báo điện tử Vnexpress. Theo đường link: https://vnexpress.net/

	userid	comments	comments_id	article_id	title	url
0	1002611535	[Ủa, iOS có tính năng khóa App từ lâu rồi mà. ...	[47968853, 47800135, 47703811, 47645657, 47584...	[4529373, 4527283, 4524455, 4522989, 4521000, ...	[Ý tưởng iOS 17 có khóa ứng dụng - VnExpress S...	[https://vnexpress.net/y-tuong-ios-17-co-khoa-...
1	1002611542	[Lâu nay phạt nguội xe máy vẫn là vấn đề nan g...	[48165879, 48157811, 48157769, 48132386, 48088...	[4537874, 4537422, 4537406, 4535990, 4535329, ...	[Môtô CSGT tông liên hoàn, ba người bị thương ...	[https://vnexpress.net/moto-csgt-tong-lien-hoa...
2	1002611620	[Messi 0d nhưng lại được định giá gần gấp 3 là...	[48163846, 48163447, 48162509, 48159323, 48156...	[4537829, 4537929, 4537576, 4537847, 4537527, ...	[10 đội tuyển đắt giá nhất tại World Cup 2022 ...	[https://vnexpress.net/10-doi-tuyen-dat-gia-nh...
3	1002611661	[Hy vọng 30 năm nữa hoàn thành để tôi có thể đ...	[47983237, 38548809, 37953883, 37023659, 36523...	[4528558, 4226031, 4201350, 4167048, 4144167, ...	[Nghiên cứu đường sắt tốc độ cao 250 km/h - Vn...	[https://vnexpress.net/nghien-cuu-duong-sat-to...
		[Năm rồi vì 20	[46444497, 46311714,	[4492456, 4488831,	[Man Utd có thể nhờ	[https://vnexpress.net/man-utd-co-

Hình 2 Bộ dữ liệu bình luận của người dùng

Các bước thực hiện:
Bước 1: Tìm người dùng. Thực thi file get-user.py. Định kỳ 5 kết quả mới nhất sẽ lưu ở mục Checkpoint.
Bước 2: Tìm bình luận người dùng.Thực thi file get-user-comments.py. Sửa đường dẫn đến file mới nhất ở thư mục checkpoints trước khi chạy. Chạy xong comment của mỗi người dùng sẽ ở thư mục user-comments.

2.1.2 Bộ dữ liệu

a. **Thông tin bộ dữ liệu:** Nguồn dữ liệu từ trang báo điện tử Vnexpress, gồm 1172 dòng, 6 cột

Số thứ tự	Tên thuộc tính	Kiểu dữ liệu	Mô tả
1	userid	object	Mã người dùng
2	comments	object	Bình luận của người dùng
3	comments_id	object	Mã bình luận của người dùng
4	article_id	object	Mã tin tức được người dùng bình luận
5	title	object	Tên tin tức được người dùng bình luận
6	url	object	Đường link tin tức có bình luận của người dùng

Bảng 1 Dữ liệu bình luận của người dùng

b. **Tiền xử lý:** Trước khi chạy mô hình chúng em tiền xử lý dữ liệu theo các bước sau đây.

Bước	Cách xử lý	Ví dụ
1	Xóa những người dùng không có bình luận nào	1183 sample -> 1172 sample
2	Chuẩn hóa unicode	rat hay -> rất hay
3	Xóa thẻ html, URL	youtube.com đa dạng -> đa dạng
4	Chuyển đổi chữ thường	Phân Tích Dữ Liệu -> phân tích dữ liệu
5	Loại bỏ stop word	bánh mì rất ngon -> bánh mì ngon
6	Tách từ bằng VNcoreNLP	bạn là cầu thủ -> bạn là cầu_thủ
7	Xóa khoảng trắng thừa	Tin tức hay -> tin tức hay

Bảng 2 Các bước tiền xử lý dữ liệu

2.2 Dữ liệu tin tức

2.2.1 Crawl dữ liệu

Nguồn crawl dữ liệu: Chúng em thực hiện crawl trực tiếp từ trang báo điện tử Vnexpress.Theo đường link: https://vnexpress.net/

	URL	article	news_content
0	https://vnexpress.net/co-gai-met-moi-vi-qua-xi...	Cô gái mệt mỏi vì quá xinh đẹp	MỹMái tóc bồng bềnh, khuôn mặt xinh đẹp cùng v...
1	https://vnexpress.net/dau-bep-doi-tuyen-nhat-t...	Đầu bếp đội tuyển Nhật tiếp tục nấu món may mắ...	Hy vọng món lươn nướng sẽ mang lại chiến thắng...
2	https://vnexpress.net/bao-nhieu-lau-nen-goi-da...	Bao nhiêu lâu nên gội đầu một lần?	Tần suất gội đầu của mỗi người tùy thuộc vào n...
3	https://vnexpress.net/lam-the-nao-giup-cha-me-...	Làm thế nào giúp cha mẹ già dùng smartphone?	Hầu hết người già đều hiểu biết hạn chế về côn...
4	https://vnexpress.net/tai-sao-khong-nen-ham-no...	Tại sao không nên ham nổi tiếng?	Nổi tiếng là thứ rất nhiều người khao khát như...
5	https://vnexpress.net/biet-thu-pho-theo-kien-t...	Biệt thự phố theo kiến trúc Á Đông	Q Villa được xây dựng trên mảnh đất diện tích ...
6	https://vnexpress.net/nhung-gian-qua-cua-ba-me...	Những giàn quả của bà mẹ Hải Dương	Khác với các loại bầu bí, trồng các loại dưa y...
7	https://vnexpress.net/th-ra-mat-sua-trai-cay-b...	TH ra mắt sữa trái cây bổ sung vi chất cho bé	TH ra mắt TH true Juice milk Topkid, nước uống...
8	https://vnexpress.net/cach-chon-bon-tam-trong-...	Cách chọn bồn tắm trong không gian nội thất nh...	Phòng tắm nhà tôi có diện tích nhỏ nhưng lại m...
9	https://vnexpress.net/dau-hieu-ban-la-ung-vien...	Dấu hiệu bạn là ứng viên trúng tuyển	Những dấu hiệu nhà tuyển dụng xem bạn là ứng v...

Hình 3 Bộ dữ liệu tin tức

Các bước thực hiện:
Bước 1: Tìm tên tin tức. Thực thi file article.py
Bước 2: Tìm nội dung của tin tức. Thực thi file content.py

2.2.2 Bộ dữ liệu

a. **Thông tin bộ dữ liệu:** Nguồn dữ liệu từ trang báo điện tử Vnexpress, gồm 548 dòng, 3 cột

Số thứ tự	Tên thuộc tính	Kiểu dữ liệu	Mô tả
1	URL	object	Đường link tin tức
2	article	object	Tên tin tức
3	news_content	object	Nội dung tin tức

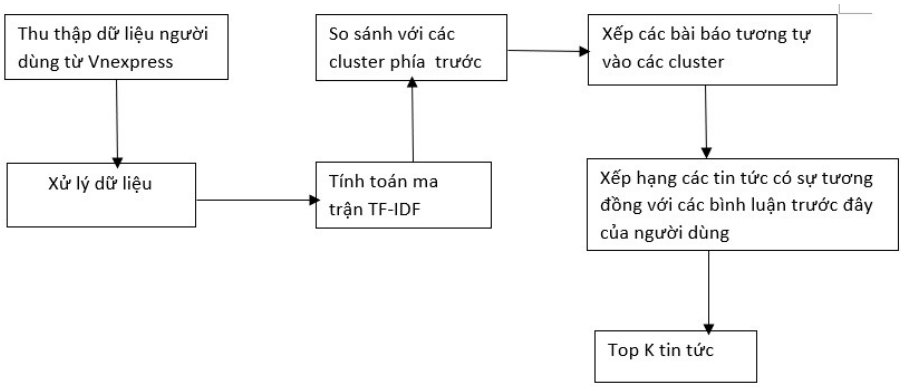
Bảng 3 Dữ liệu tin tức

b. **Tiền xử lý:** Trước khi chạy mô hình chúng em tiền xử lý dữ liệu theo các bước sau đây.

Bước	Cách xử lý	Ví dụ
1	Chuẩn hóa unicode	rat hay -> rất hay
2	Xóa thẻ html, URL	youtube.com đa dạng -> đa dạng
3	Chuyển đổi chữ thường	Phân Tích Dữ Liệu -> phân tích dữ liệu
4	Loại bỏ stop word	bánh mì rất ngon -> bánh mì ngon
5	Tách từ bằng VNcoreNLP	bạn là cầu thủ -> bạn là cầu_thủ
6	Xóa khoảng trắng thừa	Tin tức hay -> tin tức hay

Bảng 4 Các bước tiền xử lý dữ liệu

2.3 Mô tả bài toán



Hình 4 Mô hình bài toán

Input: Mã người dùng
Output: Các tin tức được khuyến nghị cho người dùng

3 Phương pháp

Như chúng ta đã biết, có hai phương pháp hay thường gặp ở một hệ khuyến nghị thường có đó là Content-Based filtering và Collaborative filtering. Với hai phương pháp này áp dụng trong việc đề xuất suất một bài báo dựa vào comment có những đặc điểm như sau:

- Content-Based filtering [1]: Dựa vào comment của của user, ta có thể xác định sự quan tâm của người dùng với một số chủ đề nào đó. Hay dựa vào sự tương đồng giữa sự quan tâm của người dùng với nội dung của bài báo đó.
- Collaborative filtering [2]: Dựa vào nội dung comment chúng ta có thể xác định phân nhóm những người dùng cụ thể, và từ nhóm đó chúng ta có thể recommend những bài báo mà những người trong nhóm đó đọc hoặc tương tác cho users khác.

Tuy vậy, cả 2 phương pháp trên đều có những bất cập như:

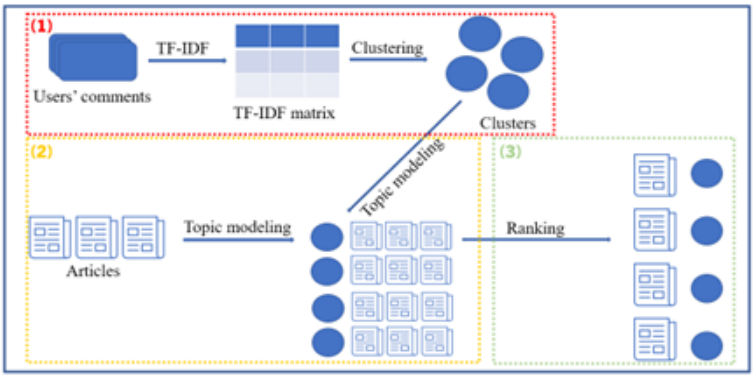
- Content-Based filtering: Các từ hiếm có trọng số lớn trong thuật toán nên đôi khi làm giảm giá trị hiệu suất. Ví dụ, một người nào đó comment từ "bầu cử" thì có thể có những bài báo có nội dung về bầu cử sẽ được đề xuất cho anh ấy
- Collaborative filtering: một số bài báo mà chưa ai đọc hoặc tương tác cho bất cứ người nào trong một nhóm tuy nhiên nội dung bài báo lại phù hợp với sự quan tâm của nhóm đó cũng sẽ ko được recommend.

Để kết hợp cả hai thuật toán để tận dụng điểm lợi thế và hạn chế bất lợi từ hai thuật toán trên nhóm chúng em đã sử dụng Hybrid-filtering [3][4]. Cụ thể hơn, nhóm chúng em sẽ sử dụng phương pháp feature combination[5]. Đặc điểm là sẽ kết hợp những nguồn dữ liệu đề xuất khác nhau để vào một hệ thống gợi ý duy nhất. Với bài toán lần này sẽ là kết hợp dựa trên lịch sử comment của người dùng trên vnexpress đề cập chung về những topic cụ thể để xây dựng các nhóm người dùng (collaborative filtering)[6] và sự tương đồng giữa nội dung bài báo và comment của các user (Content-based filtering)

3.1 Xây dựng những nhóm người dùng dựa trên comments

(1) Trước tiên, các comments được tiền xử lý. Sau đó, chúng sẽ được mã hóa bằng phương pháp TF-IDF. Kết quả thu được là một ma trận mxn với m là số lượng user và n là số từ. Từ ma trận vừa thu được ta sử dụng một thuật toán để phân cụm người dùng. Thuật toán phân cụm được nhóm chúng em chọn là K-Means. Các cụm sẽ được hình thành dựa vào sự quan tâm của họ thông qua các comments của người đó.
(2) Quá trình tiếp theo là phân loại các bài báo phù hợp với sự quan tâm vào các cụm đã tạo trước đó. Để phân loại được các bài báo chúng em sử dụng các mô hình topic modeling mà cụ thể hơn là mô hình LDA

- Topic modeling là gì: là một dạng mô hình thống kê dùng để khám phá các tóm tắt "chủ đề" xảy ra trong một tập dữ liệu. Mô hình hóa chủ đề là một



Hình 5 Pipeline xây dựng cụm người dùng

công cụ thường xuyên sử dụng để khám phá các cấu trúc ngữ nghĩa tiềm ẩn trong văn bản.

- LDA: là lớp mô hình sinh xác suất (generative probabilistic model) cho phép xác định một tập hợp các chủ đề tưởng tượng (imaginary topics) mà mỗi topic sẽ được biểu diễn bởi tập hợp các từ. Mục tiêu của LDA là mapping toàn bộ các văn bản sang các topics tương ứng sao cho các từ trong mỗi một văn bản sẽ thể hiện những topic tưởng tượng ấy. LDA sử dụng lý thuyết thống kế bayes về xác xuất tiền nghiệm và hậu nghiệm của các topic đối với các văn bản và các từ.

Cụ thể mục định sử dụng LDA trong bài toán của nhóm em như sau:

- Sử dụng LDA để khai phá số lượng topic ẩn có trong mỗi cluster từ word, document, corpus.
- Sau khi tìm ra số lượng topic ẩn phù hợp nhất có trong mỗi cluster, nhóm chúng em sẽ xây dựng một bộ gồm word, document, corpus từ các bài báo đã được thu thập trước đó. Đây là một bộ unseen corpus của của mô hình, LDA sẽ xác định topic phù hợp cho mỗi bài báo cho trong mỗi cụm dựa vào topic percentage contribution (tcp - range [0,1]) cao nhất mà LDA tính toán được.

(3) Quá trình cuối cùng mà chúng ta phải làm để kết thúc quá trình đầu tiên là ranking các bài báo để một bài báo chỉ có thể được phần vào một nhóm duy nhất. Quá trình ranking này diễn ra như sau:

- Sau khi thu được topic percentage contribution (tcp) của mỗi bài báo trong mỗi cụm nó sẽ được chuẩn hóa mean centering (tcpcent):

$$tcpcent_{i,j} = tcp_{i,j} - \frac{\sum_{n}^{k=0} tcp_{k,j}}{n} \tag{1}$$

Trong đó: i : bài báo thứ i; j : cụm thứ j; $\frac{\sum_{n}^{k=0} tcp_{k,j}}{n}$: mean tcp các bài báo của cụm j; n: số lượng bài báo.

- Tiến hành tính sentiment score (sts) trên mỗi cụm và mỗi bài báo. Phương pháp mà nhóm em áp dụng là sử dụng một từ điển từ vựng cảm xúc - VnEmolex, đếm số từ tích cực và tiêu cực có trong một đoạn văn bản - với sự chuẩn hóa để tính toán với công thức:

$$ss = \frac{pos\_count - neg\_count}{total\_word} \tag{2}$$

Trong đó: ss ∈ [-1,1]; pos_count: số lượng từ tích cực; neg_count: số lượng từ tiêu cực; $total\_count$: số lượng từ một văn bản.
Sau đó, chúng ta tính toán sự tương đồng về mặt cảm xúc giữa các cụm và bài báo (sentiment similarity metric - ssm). Việc này bảo đảm là nếu nội dung của bài báo không phù hợp về mặt cảm xúc và gây khó chịu cho người đọc sẽ được giảm đi khả năng được khuyến nghị

$$ssm_{i,j} = CosineSimilarity(ssa_i, ssc_j) \tag{3}$$

Trong đó: $ssm_{i,j} \in$ {-1,0,1}; ssa_i: sentiment score của bai báo i; ssc_j: sentiment score của cụm j.
Với 1 là khi sentiment score của cụm và bài báo cùng trên một khoảng [-1,0) hoặc (0,1]. Với -1 ngược lại nếu 2 điểm ngược khoảng. Giá trị 0 nhận được khi bất cứ 1 trong 2 điểm sentiment score có giá trị là 0.
- Kết thúc quá trình ranking, ta sẽ tính toán weight được tổng hợp từ 2 điểm là topic percentage contribution và sentiment similarity metric. Dựa vào weight bài báo trên mỗi cụm, weight của một bài báo ở cụm nào cao nhất thì bài báo đó sẽ được đưa vào nhóm người dùng đó để đề xuất.

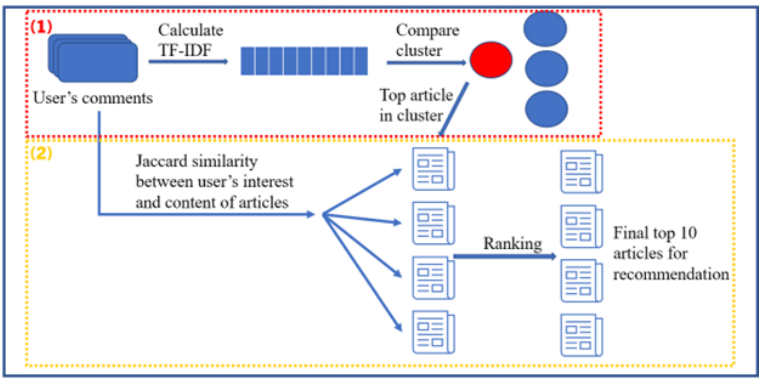
$$Weight_{i,j} = 0.8 * tcpcent_{i,j} + 0.2 * ssm_{i,j} \tag{4}$$

$$i \in j \quad IF \quad Weight_{i,j} = argmax(Weight_{i,k}) \quad WITH \quad k \in [0, n] \tag{5}$$

Trong đó: $Weight_{i,j} \in$ [-1,1]; i: bài báo thứ i; j: cụm thứ j; n: số lượng cụm.
Việc căn chỉnh 2 hệ số 0.8 và 0.2 là tùy vào mục đích của hệ khuyến nghị. Trong khi, tcpcent thể hiện về mặt nội dụng bài báo phù hợp với sự quan tâm hay không, thì ssm là thể hiện về mặt cảm xúc bài báo có phù hợp hay không. Như vậy các bài báo phù hợp với sự quan tâm của mỗi cụm đã được tìm ra.

3.2 Xây dựng hệ thống recommend kết hợp

Hệ khuyến nghị kết hợp sẽ có 2 quá trình trước khi tìm ra những bài báo nào phù hợp với người dùng.
(1) Người dùng được đề xuất tin tức sẽ được mã hóa những comments và phân vào cụm phù hợp bằng mô hình tf-idf và k-means được xây dựng trước đó. Sau khi đã xác định được cụm phù hợp ta sẽ có được những bài báo đã được xác định là có weight cao, phù hợp để đề xuất trong cụm đó.



Hình 6 Hệ khuyến nghị kết hợp

(2) Tuy vậy, để tìm ra những bài báo phù hợp với những người dùng đó chúng ta cần thêm Jacard similarity để so sánh những sự quan tâm thông qua comments với nội dung của từng bài báo. Jaccard similarity được sử dụng để xác định sự giống nhau giữa hai tài liệu văn bản, có nghĩa là hai tài liệu văn bản gần nhau như thế nào về ngữ cảnh của chúng, tức là có bao nhiêu từ phổ biến tồn tại trên tổng số từ.

$$J_{u,i} = \frac{len(w_u) \cap len(w_i)}{len(w_u) + len(w_i) + len(w_u \cap w_i)} \tag{6}$$

Trong đó: $J_{u,i}$: jacard score; w_u: tập từ được tạo bởi các comments user cần đề xuất; w_i: tập từ của bài báo.

Cuối cùng, ta sẽ ranking một lần nữa bằng cách sử dụng công thức để tình final score recommend

$$final\_score = 0.5 * weight_{i,j} + 0.5 * J_{u,i} \tag{7}$$

Trong đó: $J_{u,i}$: jacard score; $weight_{i,j}$: weight bài báo thứ i được phân vào cụm j trước đó.

Việc căn chỉnh 2 hệ số 0.5 và 0.5 tùy thuộc vào mong muốn của hệ khuyến nghị. Trong khi weight thể hiện rằng nội dung bài báo có thể phù hợp với những người dùng trong nhóm, nhờ đó một người dung có thể tiếp cận những nội dung có thể thú vị với mình. Còn nếu ta tăng hệ số J lên thì hệ khuyến nghị sẽ ưu tiên những bài báo có chứa những từ khóa mà người dùng đã comments.

4 Thực nghiệm

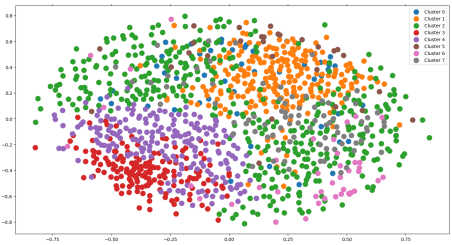
4.1 Kết quả phân cụm dựa trên comments

Sau khi tiền xử lý và mã hóa các comment bằng tf-idf và thu được một ma trận (1172,1136) (users x word). Để xác định số lượng cụm cụ thể nhóm chúng em đã sử dụng phương pháp Elbow và tìm ra số lượng là 8. Sau khi phân cụm,

nhóm chúng em xác định những từ có tần suất xuất hiện nhiều nhất trong cụm như sau:

- Cụm 0: iphone apple mua máy samsung xài tiền android giá màn_hình
- Cụm 1: bóng trận đội cầu_thủ việt_nam thắng hlv thua giải tiền
- Cụm 2: tiền mua giá học lương đất đầu lao_động thuế đóng
- Cụm 3: tui mua mấy tiền chạy học giá đường chả lắm
- Cụm 4: thi gia việt_nam pha chi mua tha chúc_mừng tiền chúc
- Cụm 5: mua tiền giá đường chạy máy nga mấy điện lái
- Cụm 6: messi bóng trận fan cr7 barca đội cầu_thủ real ghi_bàn
- Cụm 7: chồng tiền học sống gia_đình mua thể gái mấy đừng

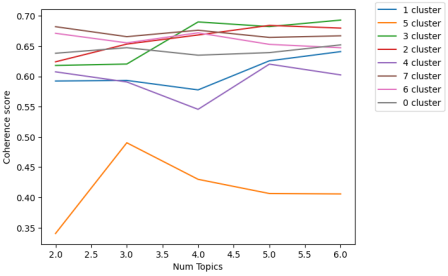
Có thể nhận thấy các cụm phân chia khá rõ sự quan tâm của các người dùng trong các cụm. Ví dụ cụm 0 là nhóm người dùng quan tâm nhiều tin tức về công nghệ, trong khi cụm 1 là quan tâm đến thể thao Việt Nam.



Hình 7 Plot các cụm sau khi giảm chiều

4.2 Kết quả topic modeling và sentiment score trên mỗi cụm

Để thực hiện topic modeling, chúng em tiến hành xây dựng dictionary và corpus. Với dictionary là 6868 token và 1172 documents trong corpus. Sau đó với mỗi cụm chúng em tiến hành tìm ra số lượng topic phù hợp dựa trên độ đo C_v measure (combining normalized pointwise similarity and cosine similarity). Mô hình LDA sẽ được áp dụng với các tham số đầu vào là random_state=100, update_every=1, chunksize=100, passes=5, alpha='auto', per_word_topics=True. Từ hình 8 ta có số lượng topic ẩn phù hợp từ cụm 0 đến 7 lần lượt là 4, 6, 5, 3, 6, 4, 4, 4, 5.
Kết quả sentiment score trên các cụm ở bảng 5. Từ kết quả này, ta thấy rằng đa phần các sự quan tâm của người dùng đều thể hiện cảm xúc gần trung tính, việc này là do các comment trên VNexpress đa phần ngôn từ có kiểm duyệt nên đa phần mọi người thường thể hiện quan điểm đóng góp ít sử dụng ngôn từ quá tiêu cực. Việc này sẽ ảnh hưởng đến việc recommend kết quả cuối cùng.



Hình 8 C_v measure trên từng cụm

Bảng 5 Kết quả sentiment score trên các cụm

Cụm	Sentiment score
Cụm 0	0.043645
Cụm 1	0.053209
Cụm 2	0.072294
Cụm 3	0.050132
Cụm 4	0.039421
Cụm 5	0.05307
Cụm 6	0.048361
Cụm 7	0.073401

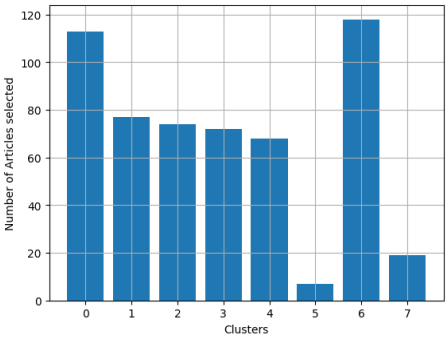
4.3 Kết quả phân chia các bài báo vào các cụm người dùng đã được xây dựng sẵn

Với tổng cộng 548 bài báo chúng em đã xây dựng dictionary và corpus cho tập dữ liệu này. Với 1400 token và 548 documents được tạo ra trong dictionary và corpus đây là cơ sở để chúng em phân chia các bài báo vào topic cụ thể trong mỗi cụm. Kết quả sau khi tình toán ta thu được 1 dataframe recommend các bài báo như hình 9.

	cluster	metric	url	article_text
0	0	0.656170	https://vnexpress.net/hang-nao-o-viet-nam-lon-...	hang việt_nam giới hang_động nhiên giới nằm qu...
1	0	0.638090	https://vnexpress.net/co-can-lam-sach-ruot-gia...	sạch ruột_giả trắc_nghiệm giúp chức_năng bệnh ...
2	0	0.536330	https://vnexpress.net/hang-trieu-sam-bien-bo-l...	hàng triệu sam biển trứng mỹvào mùa sinh_sản s...
3	0	0.531290	https://vnexpress.net/ung-dung-cua-hieu-ung-la...	ứng_dụng hiệu_ứng sen khoa_học vật_liệu khả_nă...
4	1	0.517469	https://vnexpress.net/xa-doi-ten-toi-co-can-la...	giấy khai_sinh 92019 song sáp_nhập đối địa_dan...
5	1	0.514989	https://video.vnexpress.net/tin-tuc/nhip-song/...	chàng trai rêu tươi tranh nói hoàng tuấn rêu m...
6	1	0.511869	https://vnexpress.net/erina-mano-vo-nguoi-mau-...	erina mano hoạt_động nổi_bật làng giải_trí vai...
7	1	0.508109	https://vnexpress.net/loai-ong-xay-lau-dai-cat...	loài ong xây đài cát bài biển mỹmỗi ong đào cả...
8	1	0.507629	https://vnexpress.net/cau-thu-dortmund-dao-pho...	thành_viên hội đồng_viên clb dortmund nội hòi ...
9	1	0.497069	https://vnexpress.net/nhung-khu-cho-giang-sinh...	đức chợ giáng_sinh quy khu chợ ravenna gorge n...
10	6	0.491746	https://vnexpress.net/thanh-tra-chinh-phu-manh...	nộithanh tra phủ quy_định kết tra báo_cáo thủ_...
11	1	0.484269	https://vnexpress.net/ve-dep-truong-ba-chi-tuo...	chi trải biển cảm may_mắn thần_yêu_thương đẹp ...
12	0	0.465610	https://vnexpress.net/o-my-co-nganh-ngon-ngu-a...	thông ngành ngôn_ngữ phân_vân ngành ngôn_ngữ t...
13	6	0.465506	https://vnexpress.net/messi-toa-sang-trong-tra...	messi ghi_bàn sút trận tình_huống bàn thắng vọ...
14	7	0.463366	https://vnexpress.net/lam-gan-het-nam-thi-bi-c...	hợp_đồng thời_hạn công hiện công cắt nhân_công...
15	7	0.463366	https://vnexpress.net/pho-chu-tich-tinh-quang-...	quảng ninhsau chức phó chủ_tịch ubnd tỉnh quản...
16	7	0.463046	https://vnexpress.net/cau-do-dong-nghiep-nao-g...	robert sắt_hai phòng cảnh_sát_tra phát_hiện gh...
17	7	0.462566	https://vnexpress.net/ben-trong-phong-vip-gioi...	trận đấu world_cup 2022 chủ qatar cung dịch vi...
18	7	0.462486	https://vnexpress.net/duong-day-lo-de-giao-dic...	tinhđăng thị văn đồng_phạm_chức mang_lưới tiền...
19	7	0.462086	https://vnexpress.net/mau-thuy-toi-so-dan-ong-...	máu thuỷ trai chính_phục lợm máu thuỷ trai sin...

Hình 9 Phân phối các bài bào vào mỗi topic phù hợp trong mỗi cụm

Lưu ý, metric: $Weight_{i,j}$
Tỷ lệ phân chia các bài bào vào các cụm ở hình 8.



Hình 10 Tỷ lệ phân chia các bài bào vào các cụm

4.4 Demo kết quả cuối cùng

Sau khi xử lý các comment của người được khuyến nghị ta phân vào các các cụm đã có sẵn, tính toán điểm số đề xuất cuối cùng dựa trên weight đã tính toán ở trên và jaccard similarity của comment của user đó với nội dung các bài báo Một số recommend ở hình 11.

user id	comment	url	final_score
1002637038	Chuẩn quá. Bạn chắc là giáo viên nên thấm hiểu', 'Nga sắp hết tên lửa rồi mà....', 'Tổng thống và ngoại trưởng Ukraine bác bỏ mà', ' chỉ có đàm phán thôi, mới mong có chút hi vọng hòa bình, chung sống.', 'Loại đội tuyển Nga khỏi trận playoff.', 'Trai lành ngoan thì đổ vỏ, ăn sau', 'Vũ khí Nga ngon	https://vnexpress.net/an-banh-flan-the-nao-de-khong-thua-can-4543563.html	0.563005033
		https://vnexpress.net/du-kien-khong-xet-tuyen-dai-hoc-som-voi-moi-phuong-thuc-4542353.html	0.55846918
		https://vnexpress.net/be-tac-tim-duong-thi-chung-chi-tieng-trung-4542516.html	0.550031704
		https://vnexpress.net/argentina-vs-australia-4543840-tong-thuat.html	0.547885026
		https://vnexpress.net/che-do-an-kieng-cho-nguoi-ung-thu-vu-4544283.html	0.543590119
		https://vnexpress.net/chung-chi-tieng-anh-noi-lep-ve-tren-san-nha-4540327.html	0.541924642
		https://vnexpress.net/ngap-lut-sat-nui-sau-mua-lon-4543354.html	0.541845042
		https://vnexpress.net/cdv-argentina-an-mung-doi-tuyen-vao-tu-ket-world-cup-2022-4543858.html	0.539801399
		https://vnexpress.net/thu-tuong-phai-tinh-lai-gia-dien-gio-mat-troi-de-hai-hoa-loi-ich-4544048.html	0.537424417
		https://vnexpress.net/tp-hcm-to-chuc-thi-tro-lai-chung-chi-tieng-anh-quoc-te-cho-tre-em-4543190.html	0.536447854
1002638512	Ôi vậy phải gọi là chú rồi. Vì Espana 82 cháu mới chào đời!', 'Người ta đang đá ở Anh nên nói vậy là đúng rồi, còn Bồ mà là ứng viên thì nên tháo tạ ra thưa hlv online!', 'Các anh rất dũng cảm nhưng hãy để các anh công an làm việc đó sẽ ít nguy hiểm hơn.', 'Đằng nào cũng sẽ tóm được thôi,	https://vnexpress.net/phat-hien-hai-loai-thuc-vat-nui-cao-moi-tren-day-andes-4542180.html	0.588126771
		https://vnexpress.net/10-loai-tra-giup-chua-cam-lanh-4543815.html	0.585095763
		https://vnexpress.net/tour-xem-world-cup-vong-trong-hiem-gia-300-600-trieu-dong-4542416.html	0.575325142
		https://vnexpress.net/su-nguy-hai-cua-vet-xuoc-chao-chong-dinh-4543928.html	0.575081071
		https://vnexpress.net/co-hoi-dau-tu-bat-dong-san-thoi-thi-truong-thanh-loc-4542847.html	0.57461576
		https://vnexpress.net/se-co-tau-nghi-dem-tren-song-sai-gon-4543533.html	0.573775755
		https://vnexpress.net/nhung-thao-duoc-va-vitamin-danh-bay-cam-lanh-4543607.html	0.570727671
		https://vnexpress.net/khach-quoc-te-den-viet-nam-thang-11-cao-nhat-tu-khi-mo-cua-4542484.html	0.569388809
		https://vnexpress.net/jimmii-nguyen-toi-khong-phai-nguoi-chong-tot-4543338.html	0.569028675
		https://vnexpress.net/ukraine-mia-mai-ty-phu-elon-musk-4544140.html	0.56813087
1002641874	'Ôi, nhan sắc ngày xưa. . .', '30%, còn ghê hơn xã hội đen cho vay nặng lãi.', 'Ai cũng biết Vietnam và TQ không chơi thì thôi chứ chơi là toàn chọn đồ đắt tiền nhất.', 'Bản thân Thổ Nhĩ Kỳ lạm phát 80% còn nghiêm trọng hơn.', 'Hy vọng sau này NHM bớt đem QH so sánh với Son.', 'Elon dùng Twitter để ông có thể bán CP tesla giá cao	https://vnexpress.net/dam-cuoi-tro-ve-tuoi-tho-cua-cap-doi-ha-noi-4541443.html	0.541071896
		https://vnexpress.net/ha-lan-lap-ky-luc-trong-ban-mo-ty-so-vao-luoi-my-4543942.html	0.524649232
		https://vnexpress.net/pele-toi-van-manh-me-4543945.html	0.522588966
		https://vnexpress.net/dac-quyen-cua-lop-chon-trong-truong-hoc-trung-quoc-4543941.html	0.521298465
		https://vnexpress.net/khach-vip-xem-world-cup-2022-the-nao-4542352.html	0.517268779
		https://vnexpress.net/thuy-dung-khoc-nghe-bo-dan-do-khi-ve-nha-chong-4543831.html	0.516744059
		https://vnexpress.net/gioi-tre-trung-quoc-do-xo-lam-nha-nuoc-vi-kinh-te-di-xuong-4543272.html	0.515800144
		https://vnexpress.net/bo-truong-y-te-phu-cap-y-bac-si-truc-lac-hau-4544216.html	0.515146941
		https://vnexpress.net/startup-so-ban-hang-doat-quan-quan-tai-nang-khoi-nghiep-quoc-gia-2022-4543979.html	0.511206745
		https://vnexpress.net/hang-gia-ban-tran-lan-tren-cac-kenh-thuong-mai-dien-tu-4543333.html	0.507986173

Hình 11 Một số recommend tin tức

4.5 Đánh giá hệ khuyến nghị

Việc đánh giá hệ thống khuyến nghị tin tức là một thách thức với nhóm chúng em. Đa phần các hệ thống khuyến nghị là dự đoán liệu người dùng có thích những gì để đề xuất hay không. Việc đánh giá hệ khuyến nghị tin tức sẽ là sự so sánh những sự đề xuất và dự đoán với những gì người dùng thực sự click, rating hay feedback những bài báo được đề xuất. Do đó, chúng ta phải tiếp cận với ba phương pháp. Đầu tiên là tiếp cận với nguồn dữ liệu user về sự tương tác trên website. Thứ hai, deploy hệ khuyến nghị tin tức để đánh giá hiệu suất thực tế. Cuối cùng, thực hiện khảo sát với tập mẫu lớn. Với phạm trù là một đồ án môn học nhóm chúng em chưa đủ điều kiện và khả năng thực hiện cả ba cách tiếp cận đánh giá mô hình trên để đánh giá chính xác hiệu suất hệ thống khuyến nghị này.

5 Kết luận

Xây dựng được hệ thống khuyến nghị tin tức dựa vào bình luận của người dùng.
Tiếp cận nhiều phương pháp cho bài toán cụ thể là Content-Based filtering và Collaborative filtering.
Khó khăn: nhóm chúng em chưa tìm được phương pháp thích hợp để đánh giá cho bài toán này. Đây cũng là cơ hội và thách thức để bài toán có thể phát triển hơn trong tương lai.

Hướng phát triển: Áp dụng thêm nhiều phương pháp để hệ thống ngày càng tối ưu hơn cho người dùng.

Tài liệu

[1] Pazzani, M., Billsus, D.: Content-based recommendation systems, 325–341 (1999)

[2] Balabanovic, M., Shoham, Y.: Fab: Content-based, collaborative recommendation, 66–72 (1997)

[3] Billsus, D., Pazzani, M.J.: A personal news agent that talks, learns and explains, 268–275 (1999)

[4] A Hybrid Recommendation for Music Based on Reinforcement Learning. Yu Wang (2020)

[5] News Recommender System: a Review of Recent Progress, Challenges, and Opportunities. Shaina Raza, Chen Ding (2021)

[6] Personalized News Recommendation: Methods and Challenges. Chuhan Wu,Fangzhao Wu,Yongfeng Huang,Xing Xe (2020)