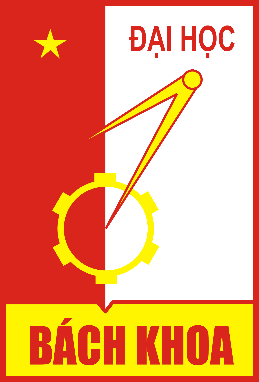
TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

──────── \* ───────

**ĐỒ ÁN 3**

**ĐỊNH HƯỚNG CÔNG NGHỆ**

**TÌM CHỦ ĐỀ TỪ TẬP CÁC VĂN BẢN VỚI**

**LATENT DIRICHLET ALLOCATION**

***Giảng viên hướng dẫn:* Trần Vĩnh Đức**

***Sinh viên thực hiện***: **Nguyễn Thị Thái Hà – 20166033**

***Mã lớp:*** **683459**

***Hà Nội, tháng 5 năm 2019***

**MỤC LỤC**

**[1. GIỚI THIỆU BÀI TOÁN](#_Toc8924873)** [2](#_Toc8924873)

[**2. TỔNG QUAN VỀ MÔ HÌNH CHỦ ĐỀ LATENT DIRICHLET ALLOCATION (LDA)** 3](#_Toc8924874)

[**2.1. Chủ đề** 3](#_Toc8924875)

[**2.2. Mô hình chủ đề** 3](#_Toc8924876)

[**2.3. Mô hình LDA** 3](#_Toc8924877)

[**2.4. Tham số trong mô hình LDA** 4](#_Toc8924878)

[**3. PHÂN TÍCH YÊU CẦU BÀI TOÁN** 4](#_Toc8924879)

[**4. QUY TRÌNH THỰC HIỆN** 5](#_Toc8924880)

[**4.1. Thu thập dữ liệu** 5](#_Toc8924881)

[**4.2. Tiền xử lý dữ liệu** 7](#_Toc8924882)

[**4.2.1. Tách từ** 7](#_Toc8924883)

[**4.2.2. Loại bỏ stop words** 8](#_Toc8924884)

[**4.3. Tìm chủ đề cho tập văn bản bằng LDA với Gensim** 8](#_Toc8924885)

[**5. KẾT QUẢ THỰC NGHIỆM** 8](#_Toc8924886)

[**6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN** 9](#_Toc8924887)

[**TÀI LIỆU THAM KHẢO** 10](#_Toc8924888)

[**PHỤ LỤC** 11](#_Toc8924889)

# **1. GIỚI THIỆU BÀI TOÁN**

Phương pháp tiếp cận theo mô hình chủ đề được nhiều công trình nghiên cứu quan tâm trong lĩnh vực phân tích mạng xã hội và rút trích thông tin. Chủ đề thể hiện động cơ, sở thích của cá nhân khi tạo lập văn bản, do đó khám phá chủ đề là khám phá được thuộc tính quan trọng, sở thích, thói quen, hành vi của cá nhân hay cộng đồng trên mạng xã hội.

Tuy nhiên, đối với tính chất của mạng xã hội, chủ đề được trao đổi trên mạng xã hội là tiềm ẩn. Chính vì vậy, việc khám phá chủ đề và hiểu được nội dung thông điệp trao đổi của người dùng là một thách thức lớn và là bài toán khó, đặc biệt là đối với tập ngữ liệu văn bản tiếng Việt.

Từ nhu cầu thực tiễn, em tiến hành tìm hiểu các bước thực hiện để tìm ra chủ đề cho tập các văn bản tin tức bằng phương pháp Latent Dirichlet Allocation (LDA). Từ đó, em sẽ tiến hành xây dựng chương trình thực nghiệm áp dụng các bước đã tìm hiểu ở trên.

## **2. TỔNG QUAN VỀ MÔ HÌNH CHỦ ĐỀ LATENT DIRICHLET ALLOCATION (LDA)**

## **2.1. Chủ đề**

- Theo từ điển Oxford, chủ đề là một vấn đề được trình bày trong văn bản, bài luận hay trong cuộc hội thoại.

- Chủ đề tiềm ẩn là chủ đề chưa biết (chưa được gán nhãn) trong quá trình tạo lập văn bản của người dùng.

## **2.2. Mô hình chủ đề**

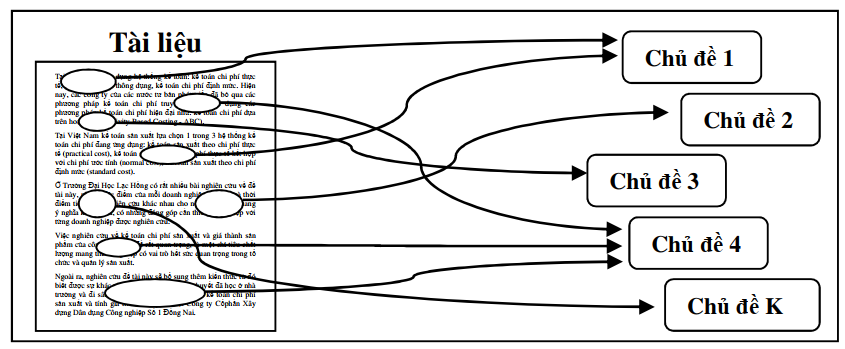
- Mô hình chủ đề được Deerwester cùng cộng sự đề xuất năm 1990, sau đó còn nhiều những nghiên cứu khác như Hofmann, 1999 và Blei et al, 2003.

- Mô hình chủ đề cho phép kiểm tra và khai thác tập tài liệu văn bản dựa trên việc tìm kiếm và thống kê các từ có liên quan đến chủ đề trong mỗi tài liệu, và khám phá ra những chủ đề tiềm ẩn trong tài liệu văn bản đó.

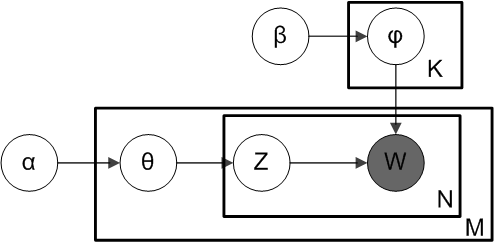
- Một số phương pháp tiếp cận hiện nay trong việc mô hình nội dung tài liệu dựa trên ý tưởng tính phân bố xác suất của mỗi từ đặc trưng trong tài liệu. Phân bố này xem văn bản là hỗn hợp nhiều chủ đề, mỗi chủ đề là sự kết hợp của nhiều từ kèm phân bố xác suất riêng cho từng từ trong chủ đề.

- Trong cách tiếp cận phân tích mạng xã hội theo chủ đề, các nghiên cứu chủ yếu tập trung xây dựng mô hình toán học, ước lượng tham số mô hình dựa trên nền tảng mạng xác suất Bayes.

## **2.3. Mô hình LDA**

Latent Dirichlet Allocation (LDA) là một mô hình sinh xác suất (generative probabilistic model) cho các tập ngữ liệu rời rạc theo nhóm. Về bản chất, LDA là một mô hình mạng Bayes theo 3 cấp, trong đó mỗi tài liệu (document) được mô tả dưới dạng kết hợp ngẫu nhiên của một tập các chủ đề. Mỗi chủ đề (topic) là một phân bố rời rạc của một tập các từ đặc trưng (word). Theo đó, mỗi tài liệu tồn tại nhiều chủ đề tiềm ẩn và mỗi chủ đề có thể sẽ tồn tại trong nhiều tài liệu khác nhau, đây có thể được xem là mối quan hệ nhiều - nhiều giữa tài liệu và chủ đề. 

## **2.4. Tham số trong mô hình LDA**



Trong đó:

- là phân bố xác suất các chủ đề của tài liệu (phân bố này thỏa phân bố )

- là tham số tập trung

- là một chủ đề được rút ra từ phân bố đa thức

- là phân bố xác suất các từ khóa cho chủ đề , phân bố này thỏa phân bố

- là tập từ đặc trưng được rút ra từ phân bố đa thức phát sinh bởi chủ đề.

# **3. PHÂN TÍCH YÊU CẦU BÀI TOÁN**

Bài toán tìm chủ đề cho tập văn bản có hai hướng nghiên cứu chính. Hướng thứ nhất là xây dựng giải thuật mới và cải thiện các giải thuật đã có nhằm nâng cao hiệu suất. Hướng thứ hai là tập trung vào tập dữ liệu cùng với việc xử lý chúng để chạy các giải thuật đã có và đánh giá các giải thuật đó.

Trong đồ án này, em tập trung vào hướng thứ hai, không cải thiện giải thuật đã có hay hiện thực các giải thuật đó mà chỉ sử dụng chúng thông qua thư viện sẵn có.

Về mặt thuận lợi:

- Có nhiều công trình nghiên cứu liên quan đề tìm hiểu và học hỏi.

- Thư viện sẵn có tích hợp nhiều giải thuật có thể sử dụng thuận tiện.

- Ngoài ra, em còn sử dụng thư viện hỗ trợ trong giai đoạn tiền xử lý dữ liệu.

Về mặt khó khăn: Dữ liệu của em ở đây là tin tức trên các trang báo mạng Việt Nam. Tuy nhiên việc lấy các tin tức đó về là khó khăn lớn nhất của bài toán này. Để có thể đưa ra mô hình tốt cần một lượng dữ liệu tương đối lớn. Việc thực hiện thủ công sẽ tốn rất nhiều thời gian cho nên cần xây dựng chương trình tự động lấy tin tức trên các trang báo mạng Việt Nam. Mà trên các trang báo mạng ngoài nội dung tin tức (dữ liệu cần thu thập) còn rất nhiều mục khác nhau như hình ảnh, video, mục lục tin, các trang trí của trang báo, các đường dẫn đến các tin liên quan của tin tức đó, quảng cáo… Em cần phải hiện thực để chương trình loại bỏ những phần đó và giữ lại nội dung, tiêu đề và các thông tin cần thiết khác của tin tức. Ngoài ra, các trang báo mạng khác nhau cũng có cấu trúc trình bày tin tức khác nhau, bản thân mỗi trang cũng có nhiều cách trình bày tin khác nhau (tin ảnh, tin video, tin tường thuật trực tiếp…). Ứng với mỗi cấu trúc trình bày, em sẽ hiện thực một cách lấy tin khác nhau, do đó mất khá nhiều thời gian.

# **4. QUY TRÌNH THỰC HIỆN**

## **4.1. Thu thập dữ liệu**

Thu thập dữ liệu là một trong những bước tiền đề quan trọng để huấn luyện tạo ra mô hình tìm chủ đề cho tập văn bản. Dữ liệu càng nhiều, càng đa dạng thì mô hình càng bao quát và đầy đủ hơn.

Trong giới hạn bài toán này, em chỉ tiến hành tìm chủ đề cho tập văn bản trong 8 chủ đề (thể loại) bao gồm: thế giới, thể thao, pháp luật, kinh doanh, công nghệ, giáo dục, sức khỏe, du lịch. Để tạo sự đa dạng cho tin tức thu thập về, em tiến hành lấy tin từ nhiều trang báo mạng Việt Nam bao gồm: vnexpress.net, vietnamnet.vn, tuoitre.vn.

Mỗi tin tức lấy về sẽ bao gồm các thông tin: thể loại, tiêu đề, nội dung, đường dẫn (URL – Uniform Resourse Locator). Tin tức thu thập được sẽ được lưu trữ trong CSDL - MySQL.

Em thực hiện lấy tin tức tự động bằng cách phân tích RSS (Rich Site Summary) – gồm các bài báo mới được cập nhật gần đây nhất, để lấy các link bài viết. Danh sách các RSS đã sử dụng được trình bày ở phụ lục 1.

Các RSS của các website được trình bày dưới dạng cấu trúc:

<rss …>

<channel> //các thẻ con của chanel

<item>

…

**<link>**

**http://... //url cần lấy**

**</link>**

…

</item> //danh sách các item khác

</chanel>

</rss>

Trong thẻ chanel có chứa nhiều thẻ con item, mỗi item tương ứng với một bài viết. Nó chứa các thông tin tóm tắt cả mỗi bài. Thông tin cần lấy chính là đường dẫn của các bài viết. Việc phân tích RSS để lấy đường dẫn tin tức mang lại nhiều lợi ích: mỗi lần phân tích sẽ lấy được nhiều đường dẫn, nhanh hơn so với việc lấy từng đường dẫn.

Từ các đường dẫn lấy ở trên, em tiến hành phân tích các bài viết html để lấy thông tin cần thiết. Mỗi website có nhiều mô hình đối tượng tài liệu DOM (Document Object Model) khác nhau để trình bày cho các dạng bài viết khác nhau. Một bài viết tin tức bình thường bao gồm tiêu đề, giới thiệu, nội dung; sẽ có mô hình DOM khác với một bài viết đưa tin chỉ có hình ảnh và khác với một bài viết tường thuật trực tiếp một trận bóng đá – đây là dạng bài viết cập nhật liên tục trong suốt trận đấu sẽ có hình ảnh, clip kèm theo.

Em thực hiện phân tích mô hình DOM của các dạng bài viết phổ biến trên các website sau: vnexpress.net, vietnamnet.vn, tuoitre.vn, 24h.com.vn, thanhnien.vn, nld.com.vn.

Mô hình đối tượng tài liệu DOM được dùng nhiều trong website của vnexpress.net:

<html …>

<div class=”title\_news”>…</div>

//thẻ chứa tiêu đề bài viết

…

<p class=”Normal”>…</p> //Thẻ chứa một đoạn nội dung

</html>

Mô hình đối tượng tài liệu DOM được dùng nhiều trong website của vietnamnet.vn:

<html …>

<h1 class=”title”>…</h1>

//thẻ chứa tiêu đề bài viết

…

<div id=”ArticleContent” class=”ArticleContent”>

<p>…</p> //Thẻ chứa một đoạn nội dung

</div>

</html>

Mô hình đối tượng tài liệu DOM được dùng nhiều trong website của tuoitre.vn:

<html …>

<h1 class=”article-title”>…</h1>

//thẻ chứa tiêu đề bài viết

…

<div class=”content fck” id=”main-detail-body”>

<p>…</p> //Thẻ chứa một đoạn nội dung

</div>

</html>

Bảng thống kê tập dữ liệu tin tức

|  |  |
| --- | --- |
| **Tên chủ đề** | **Số lượng** |
| Thế giới | 267 |
| Thể thao | 437 |
| Pháp luật | 292 |
| Kinh doanh | 259 |
| Công nghệ | 476 |
| Giáo dục | 420 |
| Sức khỏe | 394 |
| Du lịch | 220 |
| **Tổng** | **2765** |

## **4.2. Tiền xử lý dữ liệu**

### **4.2.1. Tách từ**

Khác với tiếng Anh, các từ được tách nhau bởi khoảng trắng, trong tiếng Việt, các từ được cấu thành từ một hoặc nhiều tiếng. Có hai loại từ chính là từ đơn và từ ghép. Từ đơn là những từ được cấu thành bằng một tiếng, ví dụ: đi, chạy, cười, đùa, vui, buồn,… Từ ghép là từ được cấu thành từ hai hay nhiều tiếng, mà những tiếng đó có quan hệ ngữ nghĩa với nhau, ví dụ: chợ búa, bếp núc, cơ quan, trường học,… Từ là đơn vị cấu thành nên câu trong tiếng Việt, chứ không phải tiếng, nếu tách từ dựa trên khoảng trắng sẽ không giữa được ngữ nghĩa của từ. Do đó, việc tách từ trong giai đoạn tiền xử lý dữ liệu sẽ cho hiệu quả phân loại cao hơn so với việc tách tiếng.

Hiện tại đã có nhiều công trình nghiên cứu liên quan và nhiều công cụ đã được xây dựng thành công để tách từ tiếng Việt với độ chính xác tương đối cao. Trong bài toán này, em sử dụng công cụ tách từ Python Vietnamese Toolkit (pyvi).

### **4.2.2. Loại bỏ stop words**

Stop word hay còn gọi là từ dừng. Đây là những từ xuất hiện nhiều trong tất cả các văn bản thuộc mọi thể loại trong tập dữ liệu, hay những từ chỉ xuất hiện trong một hoặc một vài văn bản. Nghĩa là stop word là những từ xuất hiện quá nhiêu lần hoặc quá ít lần. Chúng không có ý nghĩa và không chứa thông tin đáng giá để chúng ta sử dụng. Ví dụ như các từ: thì, mà, là, và, hoặc, bởi,…

Trong việc tìm chủ đề cho tập văn bản thì sự xuất hiện của những từ đó không những không giúp gì trong việc tìm chủ đề mà còn là nhiễu và làm giảm độ chính xác của mô hình. Trong bài toán này, em sẽ tiến hành loại bỏ stop word dựa trên file vietnamese-stopwords-dash.txt lưu tập các stop word thường gặp trong tiếng Việt.

## **4.3. Tìm chủ đề cho tập văn bản bằng LDA với Gensim**

- Bước 1: Dữ liệu sau bước tiền xử lý sẽ được lưu vào file clean\_data.txt. Tập từ điển sẽ được tạo từ file này và tiến hành loại bỏ đi các từ xuất hiện dưới 10 lần trên tổng tất cả các văn bản (nghĩa là từ quá hiếm) và loại bỏ từ xuất hiện trên 90% trên tổng tất cả các văn bản (nghĩa là từ này quá phổ biến).

- Bước 2: Tạo tập corpus cho tập dữ liệu bằng cách dùng hàm doc2bow của thư viện Gensim. Mỗi corpus của một văn bản là một danh sách bao gồm id của các từ xuất hiện trong văn bản đó và số lần từ đó xuất hiện.

- Bước 3: Tạo mô hình LDA với thư viện Gensim và chạy.

# **5. KẾT QUẢ THỰC NGHIỆM**

Danh sách các chủ đề được tìm ra và các từ đặc trưng tương ứng với chủ đề đó:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Chủ đề 1** | | **Chủ đề 2** | | **Chủ đề 3** | | **Chủ đề 4** | |
| mỹ | 0.037 | da | 0.014 | du\_lịch | 0.014 | đồng | 0.023 |
| trung quốc | 0.026 | giúp | 0.014 | du\_khách | 0.011 | công\_ty | 0.021 |
| usd | 0.017 | ung\_thư | 0.012 | đường | 0.010 | tiền | 0.017 |
| thương\_mại | 0.011 | bệnh | 0.011 | thành\_phố | 0.009 | tp | 0.013 |
| hai | 0.010 | bác\_sĩ | 0.010 | dân | 0.009 | triệu | 0.010 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Chủ đề 5** | | **Chủ đề 6** | | **Chủ đề 7** | | **Chủ đề 8** | |
| công\_nghệ | 0.012 | ảnh | 0.015 | đội | 0.017 | trường | 0.031 |
| máy | 0.012 | đi | 0.014 | giải | 0.016 | thi | 0.028 |
| ảnh | 0.012 | công\_an | 0.010 | mùa | 0.014 | học | 0.016 |
| màn\_hình | 0.011 | điều\_tra | 0.008 | trận | 0.012 | học\_sinh | 0.015 |
| chụp | 0.009 | hai | 0.007 | bóng | 0.012 | giáo\_dục | 0.010 |

Từ đây có thể thấy:

- Chủ đề 1 tương ứng với chủ đề: thế giới

- Chủ đề 2 tương ứng với chủ đề: sức khỏe

- Chủ đề 3 tương ứng với chủ đề: du lịch

- Chủ đề 4 tương ứng với chủ đề: kinh doanh

- Chủ đề 5 tương ứng với chủ đề: công nghệ

- Chủ đề 6 tương ứng với chủ đề: pháp luật

- Chủ đề 7 tương ứng với chủ đề: thể thao

- Chủ đề 8 tương ứng với chủ đề: giáo dục

Kết quả tìm được các từ đặc trưng khá chính xác với các chủ đề trong tập dữ liệu đầu vào.

# **6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN**

Đồ án này đã bao gồm cơ bản các bước xử lý để xây dựng một mô hình chủ đề. Qua quá trình làm đồ án này, em đã tiếp thu được mô hình chủ đề là như thế nào. Kết quả thực nghiệm cuối cùng cũng đã đưa ra được các chủ đề với các từ đặc trưng tương ứng với chủ đề đó tuy còn nhiều hạn chế như quá trình thu thập dữ liệu, lọc dữ liệu đầu vào còn hạn chế, chưa đưa ra phân loại được đối với các tập văn bản mới.

Trong tương lai, em dự định sẽ áp dụng phương pháp học máy SVM với tập dữ liệu huấn luyện để phân lớp văn bản và gán nhãn chủ đề. Tiếp tục cải tiến và khắc phục những hạn chế đã nêu nhằm tối đa hoá độ chính xác quá trình tìm chủ đề, phân lớp và gán nhãn chủ đề.

# **TÀI LIỆU THAM KHẢO**

[1] Lê Vĩnh Phú, Diệp Minh Hoàng – “Phân loại tin tức tiếng Việt sử dụng các phương pháp học máy” – Luận văn tốt nghiệp, Đại học Bách Khoa Thành phố Hồ Chí Minh, 2014.

[2] Trần Ngọc Phúc, Phạm Trần Vũ, Phạm Công Xuyên, Nguyễn Vũ Duy Quang – “Phân loại nội dung tài liệu web tiếng Việt”, 2013.

[3] “Topic modeling using LDA and Gibbs Sampling explained” - https://medium.com/@tomar.ankur287/topic-modeling-using-lda-and-gibbs-sampling-explained-49d49b3d1045

[4] Hồ Trung Thành , Đỗ Phúc - “Mô hình tích hợp khám phá, phân lớp và gán nhãn chủ đề tiếp cận theo mô hình chủ đề”, 2014.

[5] “The Basics Steps of an LDA Topic Model in Python” - https://medium.com/@jessmo/the-basics-steps-of-an-lda-topic-model-in-python-d8c363a3966e

# **PHỤ LỤC**

Danh sách các trang RSS của các trang web mà em đã sử dụng:

|  |  |  |
| --- | --- | --- |
| **Tên chủ đề** | **Website** | **RSS** |
| Thế giới | vnexpress.net | https://vnexpress.net/rss/the-gioi.rss |
| vietnamnet.vn | https://vietnamnet.vn/rss/the-gioi.rss |
| tuoitre.vn | https://tuoitre.vn/rss/the-gioi.rss |
| thanhnien.vn | https://thanhnien.vn/rss/the-gioi/goc-nhin.rss |
| nld.com.vn | https://nld.com.vn/thoi-su-quoc-te.rss |
| Thể thao | vnexpress.net | https://vnexpress.net/rss/the-thao.rss |
| vietnamnet.vn | https://vietnamnet.vn/rss/the-thao.rss |
| tuoitre.vn | https://tuoitre.vn/rss/the-thao.rss |
| 24h.com.vn | https://www.24h.com.vn/upload/rss/thethao.rss |
| https://www.24h.com.vn/upload/rss/bongda.rss |
| nld.com.vn | https://nld.com.vn/the-thao.rss |
| Pháp luật | vnexpress.net | https://vnexpress.net/rss/phap-luat.rss |
| vietnamnet.vn | https://vietnamnet.vn/rss/phap-luat.rss |
| tuoitre.vn | https://tuoitre.vn/rss/phap-luat.rss |
| 24h.com.vn | https://www.24h.com.vn/upload/rss/anninhhinhsu.rss |
| thanhnien.vn | https://thanhnien.vn/rss/viet-nam/phap-luat.rss |
| nld.com.vn | https://nld.com.vn/phap-luat.rss |
| Kinh doanh | vnexpress.net | https://vnexpress.net/rss/kinh-doanh.rss |
| vietnamnet.vn | https://vietnamnet.vn/rss/kinh-doanh.rss |
| tuoitre.vn | https://tuoitre.vn/rss/kinh-doanh.rss |
| 24h.com.vn | https://www.24h.com.vn/upload/rss/taichinhbatdongsan.rss |
| nld.com.vn | https://nld.com.vn/kinh-te.rss |
| Công nghệ | vnexpress.net | https://vnexpress.net/rss/so-hoa.rss |
| vietnamnet.vn | https://vietnamnet.vn/rss/cong-nghe.rss |
| tuoitre.vn | https://tuoitre.vn/rss/nhip-song-so.rss |
| 24h.com.vn | https://www.24h.com.vn/upload/rss/congnghethongtin.rss |
| thanhnien.vn | https://thanhnien.vn/rss/cong-nghe-thong-tin/san-pham-moi.rss |
| https://thanhnien.vn/rss/cong-nghe/xu-huong.rss |
| https://thanhnien.vn/rss/cong-nghe-thong-tin/y-tuong.rss |
| https://thanhnien.vn/rss/cong-nghe-thong-tin/kinh-nghiem.rss |
| nld.com.vn | https://nld.com.vn/cong-nghe-thong-tin.rss |
| Giáo dục | vnexpress.net | https://vnexpress.net/rss/giao-duc.rss |
| vietnamnet.vn | https://vietnamnet.vn/rss/giao-duc.rss |
| tuoitre.vn | https://tuoitre.vn/rss/giao-duc.rss |
| 24h.com.vn | https://www.24h.com.vn/upload/rss/giaoducduhoc.rss |
| thanhnien.vn | https://thanhnien.vn/rss/giao-duc/du-hoc.rss |
| https://thanhnien.vn/rss/giao-duc/tuyen-sinh.rss |
| https://thanhnien.vn/rss/giao-duc/nguoi-thay.rss |
| https://thanhnien.vn/rss/giao-duc/chon-nghe.rss |
| nld.com.vn | https://nld.com.vn/giao-duc-khoa-hoc.rss |
| Sức khỏe | vnexpress.net | https://vnexpress.net/rss/suc-khoe.rss |
| vietnamnet.vn | https://vietnamnet.vn/rss/suc-khoe.rss |
| tuoitre.vn | https://tuoitre.vn/rss/suc-khoe.rss |
| 24h.com.vn | https://www.24h.com.vn/upload/rss/suckhoedoisong.rss |
| thanhnien.vn | https://thanhnien.vn/rss/suc-khoe/lam-dep.rss |
| https://thanhnien.vn/rss/doi-song/gioi-tinh.rss |
| https://thanhnien.vn/rss/suc-khoe/khoe-dep-moi-ngay.rss |
| https://thanhnien.vn/rss/suc-khoe/yeu-da-day.rss |
| nld.com.vn | https://nld.com.vn/suc-khoe.rss |
| Du lịch | vnexpress.net | https://vnexpress.net/rss/du-lich.rss |
| tuoitre.vn | https://tuoitre.vn/rss/du-lich.rss |
| 24h.com.vn | https://www.24h.com.vn/upload/rss/dulich24h.rss |