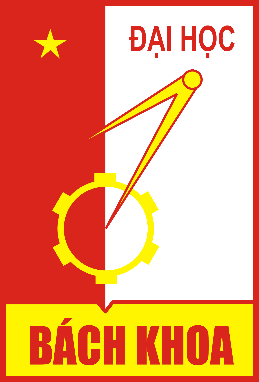
TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

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** 3](#_Toc8809617)

[**1.1. Giới thiệu bài toán** 3](#_Toc8809618)

[**1.2. Mục tiêu bài toán** 3](#_Toc8809619)

[**2. Cơ sở lý thuyết về tìm chủ đề cho tập văn bản với Latent Dirichlet Allocation (LDA)** 3](#_Toc8809620)

[**2.1. Mô hình LDA** 3](#_Toc8809621)

[**2.2. Ước lượng tham số cho mô hình LDA** 4](#_Toc8809622)

[**2.3. Phân loại văn bản sử dụng tần suất chủ đề** 4](#_Toc8809623)

[**3. Phân tích yêu cầu bài toán** 4](#_Toc8809624)

[**4. Quy trình thực hiện** 5](#_Toc8809625)

[**4.1. Thu thập dữ liệu** 5](#_Toc8809626)

[**4.2. Tiền xử lý dữ liệu** 7](#_Toc8809627)

[**4.2.1. Tách từ** 7](#_Toc8809628)

[**4.2.2. Loại bỏ stop words** 7](#_Toc8809629)

[**4.3. Tìm chủ đề cho tập văn bản với LDA** 8](#_Toc8809630)

[**5. Kết quả thực nghiệm và đánh giá** 8](#_Toc8809631)

[**6. Kết luận** 8](#_Toc8809632)

[**TÀI LIỆU THAM KHẢO** 8](#_Toc8809633)

[**PHỤ LỤC** 9](#_Toc8809634)

# **1. Giới thiệu**

## **1.1. Giới thiệu bài toán**

Trong thời đại CNTT hiện nay, chúng ta đều có thể dễ dàng xem được những tin tức thuộc mọi thể loại từ khắp nơi trên thế giới chỉ với vài cú click chuột. Việc xem tin tức trên mạng đã trở thành một phần tất yếu trong cuộc sống của nhiều người, đặc biệt là giới trẻ. Kéo theo đó là sự bùng nổ về số lượng tin tức trên Internet. Với lượng tin tức đồ sộ như vậy, một yêu cầu lớn đặt ra tổ chức, tìm kiếm thông tin có hiệu quả nhất. Việc tìm ra chủ đề của những tin tức đó một cách tự động là một hướng giải quyết cho vấn đề này.

Từ nhu cầu thực tiễn, em tiến hành tìm hiểu các bước thực hiện để tìm ra chủ đề cho tập các 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 và đánh giá kết quả đạt được.

## **1.2. Mục tiêu bài toán**

Có 2 mục tiêu cần thực hiện:

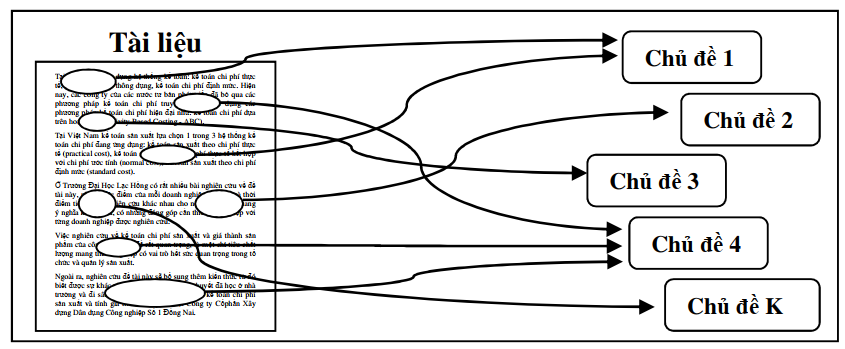
- Nhóm các tin tức thành các chủ đề khác nhau.

- Đặt tên chủ đề cho từng nhóm tin tức ấy.

## **2. Cơ sở lý thuyết về tìm chủ đề cho tập văn bản với Latent Dirichlet Allocation (LDA)**

## **2.1. Mô hình LDA**

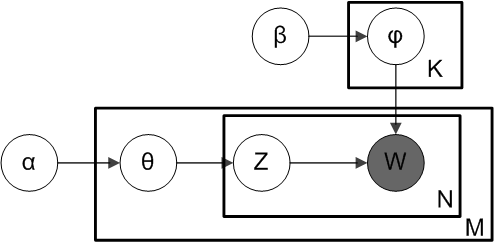
LDA là một mô hình sinh xác suất cho tập dữ liệu rời rạc như text corpora. LDA dựa trên ý tưởng: mỗi tài liệu là sự trộn lẫn của nhiều chủ đề (topic). Về bản chất, LDA là một mô hình Bayesian 3 cấp (three-level hierarchical Bayes model: corpus level, document level, word level) trong đó mỗi phần của mô hình được coi như một mô hình trộn hữu hạn trên cơ sở tập các xác suất chủ đề.



## **2.2. Ước lượng tham số cho mô hình LDA**

Cho một corpus của M tài liệu biểu diễn bởi , trong đó, mỗi tài liệu m trong corpus bao gồm n cặp , từ wi rút từ một tập từ

vựng của các mục từ {t1, …, tv}, V là số lượng các mục từ t trong tập từ vựng. LDA cung cấp một mô hình sinh đầy đủ chỉ ra kết quả tốt hơn các phương pháp trước.



Ước lượng tham số cho mô hình LDA bằng thuật toán Gibbs Sampling, một thuật toán hiệu quả để huấn luyện LDA.

## **2.3. Phân loại văn bản sử dụng tần suất chủ đề**

Dữ liệu cần phân loại cũng phải được qua các bước tiền xử lí như dữ liệu học (tách từ, loại bỏ stop words) để thu được các từ đặc trưng cho văn bản cô đọng nhất mà vẫn thể hiện được đầy đủ ý nghĩa của văn bản. Lần lượt so sánh tần suất xuất hiện của từng chủ đề trên đặc trưng của văn bản vừa thu được. Tần suất của thể loại nào xuất hiện nhiều hơn thì thuộc thể loạiđó.

# **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 10 chủ đề (thể loại) bao gồm: thế giới, thể thao, pháp luật, kinh doanh, công nghệ, giáo dục, xe, sức khỏe, khoa học, 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.

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

|  |  |  |
| --- | --- | --- |
| **ID chủ đề** | **Tên chủ đề** | **Số lượng** |
|  | Thế giới |  |
|  | Thể thao |  |
|  | Pháp luật |  |
|  | Kinh doanh |  |
|  | Công nghệ |  |
|  | Giáo dục |  |
|  | Xe |  |
|  | Sức khỏe |  |
|  | Khoa học |  |
|  | Du lịch |  |
| **Tổng** | |  |

## **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. Vector hóa dữ liệu**

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

# **5. Kết quả thực nghiệm và đánh giá**

# **6. Kết luận**

# **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.

# **PHỤ LỤC**

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

|  |  |  |  |
| --- | --- | --- | --- |
| ID chủ đề | Tên chủ đề | Website | RSS |
|  | Thế giới | vnexpress.net |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |