

**ĐẠI HỌC THÁI NGUYÊN  
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN  
THÔNG**

=====

**NGÔ THANH HẢO**

**TÌM HIỂU PHƯƠNG PHÁP PHÂN LOẠI NAÏVE BAYES  
VÀ NGHIÊN CỨU XÂY DỰNG ỨNG DỤNG TÓM TẮT  
VĂN BẢN TIẾNG VIỆT**

**LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH**

**THÁI NGUYÊN - 2015**

## LỜI CẢM ƠN

Lời đầu tiên tôi xin gửi lời cảm ơn chân thành và lòng biết ơn sâu sắc TS Nguyễn Thị Thu Hà, người đã chỉ bảo và hướng dẫn tận tình cho tôi và đóng góp ý kiến quý báu trong suốt quá trình học tập, nghiên cứu và thực hiện luận văn này.

Tôi xin trân trọng cảm ơn Ban giám hiệu Trường Đại học Công Nghệ Thông Tin và Truyền Thông Đại học Thái Nguyên, khoa CNTT đã giúp đỡ và tạo các điều kiện cho chúng tôi được học tập và làm khóa luận một cách thuận lợi.

Và cuối cùng tôi xin gửi lời cảm ơn đến gia đình, người thân và bạn bè – những người luôn bên tôi và là chỗ dựa giúp cho tôi vượt qua những khó khăn nhất. Họ luôn động viên tôi khuyến khích và giúp đỡ tôi trong cuộc sống và công việc cho tôi quyết tâm hoàn thành luận văn này.

Tuy nhiên do thời gian có hạn, mặc dù đã nỗ lực cố gắng hết mình nhưng chắc rằng luận văn khó tránh khỏi những thiếu sót. Rất mong được sự chỉ bảo, góp ý tận tình của Quý thầy cô và các bạn.

***Tôi xin chân thành cảm ơn!***

## **LỜI CAM ĐOAN**

Tôi xin cam đoan luận văn là kết quả nghiên cứu của tôi, không sao chép của ai. Nội dung luận văn có tham khảo và sử dụng các tài liệu liên quan, các thông tin trong tài liệu được đăng tải trên các tạp chí và các trang website theo danh mục tài liệu của luận văn.

**Tác giả luận văn**

**Ngô Thanh Hảo**

## MỤC LỤC

<b>LỜI CẢM ƠN .....</b>	<b>I</b>
<b>LỜI CAM ĐOAN .....</b>	<b>III</b>
<b>MỤC LỤC .....</b>	<b>IV</b>
<b>DANH MỤC HÌNH VẼ .....</b>	<b>VI</b>
<b>DANH MỤC BẢNG BIỂU .....</b>	<b>VI</b>
<b>DANH MỤC TỪ VIẾT TẮT.....</b>	<b>VIII</b>
<b>LỜI MỞ ĐẦU .....</b>	<b>1</b>
<b>CHƯƠNG 1 : TỔNG QUAN VỀ TÓM TẮT VÀ TÓM TẮT VĂN BẢN TIẾNG VIỆT .....</b>	<b>3</b>
<b>1.1 Giới thiệu.....</b>	<b>3</b>
1.1.1 Tổng quan bài toán tóm tắt văn bản .....	3
1.1.2 Tỷ lệ trong tóm tắt văn bản .....	6
<b>1.2 Đặc điểm ngôn ngữ tiếng Việt.....</b>	<b>7</b>
1.2.1 Đặc điểm ngữ âm .....	7
1.2.2 Đặc điểm từ vựng.....	8
1.2.3 Đặc điểm ngữ pháp .....	9
1.2.4 Xử lý ngôn ngữ tiếng Việt trên máy tính .....	10
<b>1.3 Một số phương pháp tóm tắt văn bản .....</b>	<b>12</b>
<b>1.4 Đánh giá tóm tắt văn bản .....</b>	<b>14</b>
1.4.1 Đánh giá theo cách thủ công .....	14
1.4.2 Phương pháp đánh giá BLEU .....	14
1.4.3 Phương pháp đánh giá ROUGE.....	15
1.4.4 Độ đo precision và độ đo recall .....	16
<b>CHƯƠNG 2 : PHƯƠNG PHÁP TÓM TẮT VĂN BẢN TIẾNG VIỆT DỰA TRÊN NAIVE BAYES .....</b>	<b>18</b>
<b>2.1 Một số phương pháp tóm tắt văn bản điển hình .....</b>	<b>18</b>
2.1.1 Phương pháp tóm tắt văn bản bằng cây quyết định .....	18
2.1.2 Phương pháp tóm tắt văn bản bằng mạng nơ ron .....	19
2.1.3 Phương pháp phân tích ngôn ngữ tự nhiên mức sâu.....	19
2.1.4 Phương pháp tóm tắt ngắn .....	22

2.1.5 Phương pháp dựa trên mô hình markov ẩn .....	23
2.1.6 Phương pháp tóm tắt dựa trên rút gọn câu .....	24
2.1.7 Phương pháp tóm tắt văn bản bằng naïve bayes:.....	24
<b>2.2 Phương pháp tóm tắt văn bản sử dụng lý thuyết phân loại Naïve Bayes</b> .....	<b>25</b>
2.2.1 Phân loại Naïve Bayes .....	25
2.2.2 Lựa chọn các đặc trưng cho trích chọn .....	31
<b>2.3 Huấn luyện và tính trọng số các câu trong tập huấn luyện.....</b>	<b>39</b>
<b>2.4 Lựa chọn các câu tạo tóm tắt.....</b>	<b>41</b>
<b>CHƯƠNG 3. XÂY DỰNG VÀ CÀI ĐẶT HỆ THỐNG TÓM TẮT VĂN BẢN TIẾNG VIỆT DỰA TRÊN LÝ THUYẾT NAÏVE BAYES .....</b>	<b>44</b>
<b>3.1 Mô hình hệ thống tóm tắt văn bản tiếng Việt dựa trên lý thuyết Naïve Bayes .....</b>	<b>44</b>
<b>3.2 Phân tích thiết kế hệ thống tóm tắt văn bản tiếng Việt dựa trên Naïve Bayes .....</b>	<b>50</b>
<b>3.3 Một số giao diện của hệ thống tóm tắt văn bản tiếng Việt dựa trên Naïve Bayes .....</b>	<b>52</b>
3.3.1 Giao diện trang chủ hệ thống tóm tắt văn bản tiếng Việt .....	52
3.3.2 Giao diện trang quản trị hệ thống tóm tắt văn bản tiếng Việt.....	53
<b>3.4 Kết quả thực nghiệm phương pháp tóm tắt văn bản tiếng Việt dựa trên Naïve Bayes.....</b>	<b>59</b>
3.4.1 Xây dựng tập dữ liệu phục vụ huấn luyện .....	59
3.4.2 Xây dựng bộ từ điển danh từ.....	60
3.4.3 Tiền xử lý và chuẩn hóa dữ liệu.....	60
3.4.4 Đánh giá kết quả của hệ thống tóm tắt văn bản dựa trên Naïve Bayes .	61
<b>KẾT LUẬN .....</b>	<b>62</b>
<b>TÀI LIỆU THAM KHẢO .....</b>	<b>63</b>
<b>TIẾNG VIỆT .....</b>	<b>63</b>
<b>PHỤ LỤC .....</b>	<b>64</b>

## DANH MỤC HÌNH VẼ

Hình 1.1 Hệ Thống Tóm Tắt Văn Bản Text Compactor .....	4
Hình 2.1. Cây Cấu Trúc Tu Từ .....	22
Hình 2.2. Mô Hình Markov Ẩn Sử Dụng Trong Trích Rút Câu. ....	23
Hình 2.3. Ma Trận Ví Dụ. ....	33
Hình 2.4. Mô Hình Giảm Chiều Véc Tơ.....	33
Hình 2.5. Văn Bản Ví Dụ.....	35
Hình 2.6 Quan Hệ Giữa Số Văn Bản Và Số Thuật Ngữ.....	36
Hình 2.7 Tách Từ Dựa Trên Hệ Thống Phân Tích Câu Visp. ....	36
Hình 2.8. Thuật Toán Tính Trọng Số Của Câu.....	40
Hình 2.9 Thuật Toán Trích Rút Câu .....	42
Hình 3.1. Mô Hình Tóm Tắt Văn Bản Thông Thường.....	45
Hình 3.2. Mô Hình Tóm Tắt Văn Bản Trong Luận Văn Đề Xuất.....	47
Hình 3.3 Cơ sở dữ liệu của hệ thống.....	50
Hình 3.4 Sơ Đồ Usecase Tổng Quát. ....	51
Hình 3.5. Usecase Trường Hợp Huấn Luyện.....	52
Hình 3.6. Giao Diện Trang Chủ Của Hệ Thống .....	53
Hình 3.7 Giao Diện Chính Của Trang Quản Trị.....	54
Hình 3.8 Lấy Tin Tự Động. ....	54
Hình 3.9 Giao Diện Hiển Thị Dữ Liệu Lấy Về. ....	55
Hình 3.10 Giao Diện Huấn Luyện Văn Bản. ....	56
Hình 3.11 Giao Diện Quản Lý Từ. ....	56
Hình 3.12 Hiển Thị Tin Tức Sau Khi Cập Nhật. ....	57
Hình 3.13 Giao Diện Tóm Tắt Tin Tức. ....	58
Hình 3.14 Giao Diện Tóm Tắt Văn Bản.....	58

## DANH MỤC BẢNG BIỂU

Bảng 1.1. Hiện Trạng Các Kho Ngữ Liệu Tiếng Việt. ....	12
Bảng 2.1 : Ví dụ về bảng huấn luyện.....	28
Bảng 3.1. Bảng Kết Quả Thực Nghiệm .....	61

## DANH MỤC TỪ VIẾT TẮT

Kí hiệu	Diễn giải
tf	Tần suất từ ( <i>Term frequency</i> )
Idf	tần suất nghịch đảo văn bản ( <i>inverse document frequency</i> )
TREC	Hội thảo tra cứu văn bản ( <i>Text REtrieval Conference</i> )
DUC	Hội thảo hiểu văn bản ( <i>Document Understanding Conference</i> )
BLEU	Phương pháp đánh giá dịch máy tự động ( <i>Bilingual Evaluation Under Study</i> )
NIST	Viện công nghệ tiêu chuẩn quốc gia ( <i>National Institute of Standards and Technology</i> )
Rouge	Phương pháp đánh giá kết quả tóm tắt ROUGE ( <i>Recall – Oriented Understudy for Gisting Evaluation</i> )



## LỜI MỞ ĐẦU

Ngày nay thông tin đã và đang đóng vai trò cực kỳ quan trọng trong xã hội. Sự phát triển mạnh mẽ của Internet mang lại cho con người những thông tin quan trọng và bổ ích, với lượng lớn thông tin này mang lại cho con người những tiện ích tra cứu thông tin. Các hệ thống tìm kiếm, tra cứu được nghiên cứu, đề xuất và xây dựng thỏa mãn phần nào yêu cầu của người dùng đặt ra trong hiện tại. Tuy nhiên, nó khiến chúng ta khó khăn trong việc tìm kiếm và tổng hợp thông tin.

Các nhà nghiên cứu đã đề xuất các giải pháp để xây dựng các hệ thống, công cụ khai phá dữ liệu như: phân loại dữ liệu, phân cụm dữ liệu, nén dữ liệu, tra cứu thông tin, tóm tắt văn bản... Một trong những công cụ quan trọng đó là tóm tắt văn bản.

Đối với dữ liệu dạng văn bản, tóm tắt văn bản là tóm tắt các thông tin chính từ trong văn bản gốc để nhận được một văn bản ở dạng ngắn gọn và chất lọc các thông tin quan trọng từ trong văn bản gốc.

Tóm tắt văn bản nhận được nhiều sự quan tâm nghiên cứu của các nhà khoa học nhóm nghiên cứu và các công ty trên thế giới. Bài toán tóm tắt văn bản tiếng Việt cũng không ngoại lệ vì không thể khai thác thông tin tiếng Việt hiệu quả nếu không có phương pháp tóm tắt văn bản tiếng Việt.

Trong khuôn khổ đề tài luận văn, tôi sử dụng cách tiếp cận rút gọn câu dựa trên Naive Bayes để:

- Nâng cao chất lượng của hệ thống tóm tắt văn bản tiếng Việt tự động bằng cách học giám sát. Trên thực tế để giải quyết bài toán này đã có rất nhiều phương pháp được đưa ra như sử dụng thuật toán Naive Bayes, phương

pháp cây quyết định(Decision tree), Phương pháp tóm tắt văn bản bằng mạng nơron nhân tạo(Artificial Neural Network), phương pháp tóm tắt ngắn, Phương pháp phân tích ngôn ngữ tự nhiên mức sâu, phương pháp học không giám sát, phương pháp máy học. Mỗi phương pháp đều cho kết quả khá tốt, tuy nhiên phương pháp tóm tắt văn bản tiếng Việt bằng thuật toán Naïve Bayes có chất lượng của tóm tắt văn bản là cao hơn.

- Giảm độ phức tạp tính toán về mặt thời gian.
- Xây dựng hệ thống tự động tổng hợp tin tức trực tuyến và tóm tắt.
- Xây dựng tập dữ liệu huấn luyện gồm 200 văn bản tiếng Việt.

**Luận văn được chia thành 3 chương với các nội dung sau:**

*Chương 1: Tổng quan về tóm tắt và tóm tắt văn bản tiếng Việt*

*Chương 2: Phương pháp tóm tắt văn bản tiếng việt dựa trên Naive Bayes*

*Chương 3: Xây dựng ứng dụng tóm tắt văn bản tiếng Việt dựa trên Naive Bayes.*

## **Chương 1 : TỔNG QUAN VỀ TÓM TẮT VÀ TÓM TẮT VĂN BẢN TIẾNG VIỆT**

Trong chương này, luận văn trình bày các khái niệm tổng quan về tóm tắt văn bản và tóm tắt văn bản tiếng Việt, các kỹ thuật tóm tắt văn bản dựa trên máy học như: Naïve Bayes, Cây quyết định, phương pháp can thiệp mức ngôn ngữ tự nhiên,...thông qua đó, luận văn cũng trình bày một số phương pháp đánh giá tóm tắt cơ bản hiện nay.

### ***1.1 Giới thiệu***

#### **1.1.1 Tổng quan bài toán tóm tắt văn bản**

##### **1.1.1.1 Khái niệm**

Mạng Internet cùng với bước tiến mạnh mẽ của công nghệ lưu trữ làm cho lượng thông tin lưu trữ ngày càng lớn. Lượng thông tin khổng lồ đó đã mang lại lợi ích không nhỏ cho con người nhưng đồng thời nó cũng khiến chúng ta khó khăn trong việc tìm kiếm và tổng hợp thông tin. Giải pháp cho vấn đề chính là Tóm tắt văn bản tự động. Việc áp dụng tóm tắt văn bản giúp người dùng tiết kiệm thời gian đọc tăng hiệu quả tìm kiếm.

**Định nghĩa 1.1** [Tóm tắt văn bản (Text summarization)]: Tóm tắt văn bản là quá trình rút ra những thông tin quan trọng từ một văn bản để tạo thành một văn bản ngắn gọn hơn theo nhiệm vụ cụ thể và yêu cầu của người sử dụng [9].

Text Compactor: Free Onli x

tor.com

Follow these simple steps to create a summary of your text.

**Step 1**  
Type or paste your text into the box.

Báo điện tử Hoa Nam còn dẫn thông tin của Tập đoàn Dầu khí ngoài khơi quốc gia TQ (CNOOC) cho biết, giàn khoan bán chìm - chính là giàn khoan Hải Dương 981 mà TQ từng đơn phương hạ đặt trái phép ở vùng đặc quyền kinh tế và thềm lục địa VN - gần đây đã tìm thấy một mỏ khí nước sâu tại khu vực cách nam đảo Hải Nam 150km thuộc bắc Biển Đông.

Báo nhắc lại thông tin từ Tân hoa xã rằng, Xie Yuhong, quản lý CNOOC cho biết, một khối lượng lớn khí tự nhiên đã được tìm thấy hồi tháng trước tại giếng Lingshui 17-2-1. Quan chức CNOOC nói giếng này không nằm ở khu vực tranh chấp.

Đây là mỏ khí nước sâu lớn đầu tiên mà giàn khoan Hải Dương 981 của CNOOC phát hiện. Hiện phía TQ đang xác định chính xác trữ lượng mỏ khí Lingshui, nhưng Xie nói rằng, một mỏ khí lớn đồng nghĩa với việc có ít nhất 30 tỉ mét khối.

Tân hoa xã còn dẫn lời Chủ tịch CNOOC Wang Yilin cho biết, việc phát hiện mỏ khí đã mở cánh cửa khai thác tài nguyên dầu khí nước sâu ở Biển Đông. Giàn khoan trị giá 6 tỉ nhân dân tệ đã hoạt động ở Biển Đông kể từ tháng 5/2012.

**Step 2**  
Drag the slider, or enter a number in the box, to set the percentage of text to keep in the summary.

53 %

**Step 3**  
Read your summarized text. If you would like a different summary, repeat Step 2. When you are happy with the summary, copy and paste the text into a word processor, or [text to speech program](#), or [language translation tool](#)

TQ có thể đưa giàn khoan trở lại khi quan hệ TQ-VN không căng thẳng".

Báo điện tử Hoa Nam còn dẫn thông tin của Tập đoàn Dầu khí ngoài khơi quốc gia TQ (CNOOC) cho biết, giàn khoan bán chìm - chính là giàn khoan Hải Dương 981 mà TQ từng đơn phương hạ đặt trái phép ở vùng đặc quyền kinh tế và thềm lục địa VN - gần đây đã tìm thấy một mỏ khí nước sâu tại khu vực cách nam đảo Hải Nam 150km thuộc bắc Biển Đông.

Báo nhắc lại thông tin từ Tân hoa xã rằng, Xie Yuhong, quản lý CNOOC cho biết, một khối lượng lớn khí tự nhiên đã được tìm thấy hồi tháng trước tại giếng Lingshui 17-2-1. Hiện phía TQ đang xác định chính xác trữ lượng mỏ khí Lingshui, nhưng Xie nói rằng, một mỏ khí lớn đồng nghĩa với việc có ít nhất 30 tỉ mét khối.

Tân hoa xã còn dẫn lời Chủ tịch CNOOC Wang Yilin cho biết, việc phát hiện mỏ khí đã mở cánh cửa khai thác tài nguyên dầu khí nước sâu ở Biển Đông.

Hình 1.1 Hệ thống tóm tắt văn bản Text Compactor

Những nghiên cứu sớm nhất về tóm tắt văn bản được đề xuất bởi Luhn vào năm 1958, tại Viện nghiên cứu của IBM, trong phương pháp của mình,

Luhn đã coi tần suất là đặc trưng chính trong một văn bản và cũng là độ đo quan trọng có ý nghĩa. Ý tưởng này đã mở đầu cho các công trình liên quan sau này. Luhn đã biên dịch từ một danh sách các từ chứa nội dung (content words) được sắp xếp theo tần xuất giảm dần và đánh chỉ số độ đo quan trọng của chúng. Ở mức một câu, nhân tố quan trọng (significance factor) được dựa trên độ đo quan trọng của các từ có mặt trong câu đó và khoảng cách giữa chúng với các từ có độ đo quan trọng thấp. Tất cả các câu được sắp xếp theo thứ tự của nhân tố quan trọng và các câu có vị trí cao nhất sẽ được lựa chọn trong hệ thống tóm tắt tự động [9].

Một nghiên cứu liên quan khác của Baxendale cũng được đề xuất vào năm 1958 tại viện nghiên cứu IBM và công bố trong cùng một tạp chí, cung cấp một góc nhìn khác khi tập trung vào tìm kiếm các thành phần ngữ nghĩa ngầm của các văn bản: Vị trí câu. Theo mục đích này, tác giả đã thu thập 200 đoạn để tìm ra tới 85% trong các đoạn đó, các câu chủ đề nằm ở vị trí đầu đoạn và 7% nằm ở vị trí cuối đoạn. Do đó, đơn giản nhất sẽ chọn câu đứng ở đầu đoạn hoặc cuối đoạn để tạo ra tóm tắt. Đặc trưng về vị trí câu cũng là một trong những đặc trưng tổ hợp trong các hệ thống tóm tắt dựa trên máy học sau này [9].

Nghiên cứu cơ bản của Edmundson năm 1969, mô tả một hệ thống sinh ra văn bản tóm tắt dựa trên cách tiếp cận trích rút câu. Đầu tiên tác giả phát triển một giao thức để tạo trích rút thủ công ứng dụng cho một tập gồm 400 văn bản kỹ thuật. Tiếp theo, các đặc trưng tần suất từ và vị trí quan trọng được sử dụng lại từ các nghiên cứu trước và bổ sung thêm hai đặc trưng nữa. Trọng số câu được tính toán dựa trên các đặc trưng này. Khi đánh giá, độ chính xác của phương pháp tương đương với 44% so với trích rút thủ công [9].

### 1.1.1.2 Phân loại tóm tắt

Tuỳ theo yêu cầu và mục đích sử dụng, tóm tắt văn bản được phân thành các kiểu khác nhau:

- Tóm tắt trình bày (indicative summary),
- Tóm tắt thông tin (informative summary),
- Tóm tắt hướng truy vấn (queries –oriented summary),
- Tóm tắt khái lược (generic summary),
- Tóm tắt dựa trên trích rút câu (extraction summary)
- Tóm tắt dựa trên trừu tượng (abstraction summary).

Trong các kiểu tóm tắt văn bản này, tóm tắt trình bày quan tâm tới diễn giải văn bản mà bỏ qua ngữ cảnh, tóm tắt thông tin đưa ra tóm tắt nội dung ở dạng ngắn nhất. Tóm tắt hướng truy vấn chỉ đưa ra nội dung mà người đọc quan tâm. Tóm tắt khái lược đưa ra tổng quan văn bản, tóm tắt dựa trên trích rút trích chọn ra những phần quan trọng trong văn bản như câu, mệnh đề, thuật ngữ,... Tóm tắt dựa trên trừu tượng tạo ra một văn bản tóm tắt đảm bảo về mặt cú pháp, ngữ nghĩa, câu được xử lý một cách tinh vi. [6].

### 1.1.2 Tỷ lệ trong tóm tắt văn bản

Thông thường, khi tóm tắt văn bản người ta đề cập tới hai yêu cầu chính sau:

- Văn bản tóm tắt phải ngắn hơn văn bản gốc.
- Văn bản tóm tắt phải giữ được thông tin quan trọng của văn bản gốc.

Do đó, trong quá trình tóm tắt văn bản người ta thường quan tâm tới hai tỷ lệ tóm tắt: tỷ lệ nén và tỷ lệ thông tin. Tỷ lệ nén (compression ratio) biểu thị chiều dài của văn bản tóm tắt được rút ngắn so với văn bản gốc. Tỷ lệ thông tin (retention ratio) biểu thị lượng thông tin giữ lại được từ văn bản gốc [11]. Dưới đây là định nghĩa về hai tỷ lệ tóm tắt này.

**Định nghĩa 1.2** [Tỉ lệ nén (compression ratio)]: Tỉ lệ nén là sự mô tả độ nén về mặt chiều dài của văn bản tóm tắt so với văn bản gốc [11].

Tỉ lệ nén  $r_l$  được xác định theo công thức (1-1) dưới đây.

$$r_l = \frac{L_s}{L_o}, \quad (1-1)$$

trong đó:  $r_l$  là tỉ lệ nén,  $L_s$  là chiều dài của văn bản tóm tắt và  $L_o$  là chiều dài của văn bản gốc.

**Định nghĩa 1.3** [Tỉ lệ thông tin (retention ratio)]: Tỉ lệ thông tin là sự mô tả lượng thông tin được lấy ra so với văn bản gốc [11].

Tỉ lệ thông tin được xác định theo công thức (1-2) ở dưới.

$$r_c = \frac{C_s}{C_o}, \quad (1-2)$$

trong đó:  $r_c$  là tỉ lệ thông tin,  $C_s$  là số các từ mang thông tin của văn bản tóm tắt và  $C_o$  là số các từ mang thông tin của văn bản gốc.

## **1.2 Đặc điểm ngôn ngữ tiếng Việt**

### **1.2.1 Đặc điểm ngữ âm**

Trong tiếng Việt có một loại đơn vị đặc biệt gọi là "tiếng". Về mặt ngữ âm, mỗi tiếng là một âm tiết. Hệ thống âm vị tiếng Việt phong phú và có tính cân đối, tạo ra tiềm năng của ngữ âm tiếng Việt trong việc thể hiện các đơn vị có nghĩa. Nhiều từ tượng hình, tượng thanh có giá trị gợi tả đặc sắc. Khi tạo câu, tạo lời, người Việt rất chú ý đến sự hài hoà về ngữ âm, đến nhạc điệu của câu văn [5].

### 1.2.2 Đặc điểm từ vựng

Mỗi tiếng, nói chung là một yếu tố có nghĩa. Tiếng là đơn vị cơ sở của hệ thống các đơn vị có nghĩa của tiếng Việt. Từ tiếng, người ta tạo ra các đơn vị từ vựng khác để định danh sự vật, hiện tượng..., chủ yếu nhờ phương thức ghép và phương thức láy [5].

Việc tạo ra các đơn vị từ vựng ở phương thức ghép luôn chịu sự chi phối của quy luật kết hợp ngữ nghĩa, chẳng hạn: đất nước, máy bay, nhà lầu xe hơi, nhà tan cửa nát... Hiện nay, đây là phương thức chủ yếu để sản sinh ra các đơn vị từ vựng. Theo phương thức này, tiếng Việt triệt để sử dụng các yếu tố cấu tạo từ thuần Việt hay vay mượn từ các ngôn ngữ khác để tạo ra các từ, ngữ mới, chẳng hạn: tiếp thị, karaoke, thư điện tử (e-mail), thư thoại (voice mail), phiên bản (version), xa lộ thông tin, siêu liên kết văn bản, truy cập ngẫu nhiên v.v.

Việc tạo ra các đơn vị từ vựng ở phương thức láy thì quy luật phối hợp ngữ âm chi phối chủ yếu việc tạo ra các đơn vị từ vựng, chẳng hạn: chôm chĩa, chông chơ, đồng đa đồng đánh, thơ thần, lúng lá lúng liếng, v.v.

Vốn từ vựng tối thiểu của tiếng Việt phần lớn là các từ đơn tiết (một âm tiết, một tiếng). Sự linh hoạt trong sử dụng, việc tạo ra các từ ngữ mới một cách dễ dàng đã tạo điều kiện thuận lợi cho sự phát triển vốn từ, vừa phong phú về số lượng, vừa đa dạng trong hoạt động. Cùng một sự vật, hiện tượng, một hoạt động hay một đặc trưng, có thể có nhiều từ ngữ khác nhau biểu thị. Tiềm năng của vốn từ ngữ tiếng Việt được phát huy cao độ trong các phong cách chức năng ngôn ngữ, đặc biệt là trong phong cách ngôn ngữ nghệ thuật. Hiện nay, do sự phát triển vượt bậc của khoa học-kỹ thuật, đặc biệt là công nghệ thông tin, thì tiềm năng đó còn được phát huy mạnh mẽ hơn. Ngoài ra,



có những từ vẫn mang âm tiếng Hán do đó phải giải nghĩa theo tiếng Hán, chẳng hạn:

### **Nguyễn Tiêu**

*“Kim dạ nguyên tiêu nguyệt chính viên,*

*Xuân giang xuân thủy tiếp xuân thiên.*

*Yên ba thâm xứ đàm quân sự*

*Dạ bán quy lai nguyệt mãn thuyền”.*

*Hồ Chí Minh – 1948.*

### **1.2.3 Đặc điểm ngữ pháp**

Từ của tiếng Việt không biến đổi hình thái. Đặc điểm này sẽ chi phối các đặc điểm ngữ pháp khác. Khi từ kết hợp từ thành các kết cấu như ngữ, câu, tiếng Việt rất coi trọng phương thức trật tự từ và hư từ [2].

Việc sắp xếp các từ theo một trật tự nhất định là cách chủ yếu để biểu thị các quan hệ cú pháp. Trong tiếng Việt khi nói "Anh ta lại đến" là khác với "Lại đến anh ta". Khi các từ cùng loại kết hợp với nhau theo quan hệ chính phụ thì từ đứng trước giữ vai trò chính, từ đứng sau giữ vai trò phụ. Nhờ trật tự kết hợp của từ mà "củ cải" khác với "cải củ", "tình cảm" khác với "cảm tình". Trật tự chủ ngữ đứng trước, vị ngữ đứng sau là trật tự phổ biến của kết cấu câu tiếng Việt.

Phương thức hư từ cũng là phương thức ngữ pháp chủ yếu của tiếng Việt. Nhờ hư từ mà tổ hợp "anh của em" khác với tổ hợp "anh và em", "anh vì em". Hư từ cùng với trật tự từ cho phép tiếng Việt tạo ra nhiều câu cùng có nội dung thông báo cơ bản như nhau nhưng khác nhau về sắc thái biểu cảm. Ví dụ, so sánh các câu sau đây:

- Ông ấy không hút thuốc.
- Thuốc, ông ấy không hút.
- Thuốc, ông ấy cũng không hút.

Ngoài trật tự từ và hư từ, tiếng Việt còn sử dụng phương thức ngữ điệu. Ngữ điệu giữ vai trò trong việc biểu hiện quan hệ cú pháp của các yếu tố trong câu, nhờ đó nhằm đưa ra nội dung muốn thông báo. Trên văn bản, ngữ điệu thường được biểu hiện bằng dấu câu. Chúng ta thử so sánh hai câu sau để thấy sự khác nhau trong nội dung thông báo:

- Đêm hôm qua, cầu gãy.
- Đêm hôm, qua cầu gãy.

#### **1.2.4 Xử lý ngôn ngữ tiếng Việt trên máy tính**

Sự phát triển của các hệ thống xử lý ngôn ngữ tự nhiên trên thế giới, đặc biệt là đối với ngôn ngữ tiếng Anh cho thấy sự cần thiết của xử lý ngôn ngữ tiếng Việt. Hiện nay, do sự phức tạp, khó khăn của xử lý văn bản tiếng Việt và các nghiên cứu về tiếng Việt hiện nay vẫn còn mới mẻ, các kết quả về nghiên cứu tiếng Việt vẫn mang tính chất tìm hiểu, chưa hệ thống và định hướng rõ ràng. Một số nghiên cứu là những đề tài cử nhân, thạc sĩ tại một số trường Đại học. Hầu hết các đề tài mới xây dựng được mô hình, thử và kiểm tra trên những tập ngữ liệu nhỏ do các cá nhân và tập thể tự xây dựng, không có các tài nguyên và công cụ cần thiết cho xử lý tiếng Việt.

Bắt đầu từ năm 2006 nhánh đề tài "Xử lý văn bản" là một phần của đề tài KC01.01/06-10 "Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lý tiếng nói và văn bản tiếng Việt" đã được triển khai. Cho đến nay, nhánh đề tài này đã thu được một số kết quả bao gồm kho ngữ liệu và công cụ phục vụ cho xử lý văn bản như sau:

Nhóm các sản phẩm về tài nguyên:

- Từ điển điện tử gồm 35,000 mục từ cho người sử dụng máy tính.
- Kho tài nguyên gồm 10,000 câu có chú giải (Viet treebank).
- Kho ngữ liệu gồm 100,000 cặp câu Anh - Việt.

Nhóm các công cụ cho cộng đồng về xử lý ngôn ngữ tự nhiên:

- Hệ phân tách từ Việt.
- Hệ phân loại từ Việt.
- Hệ phân cụm từ Việt.
- Hệ phân tích cú pháp tiếng Việt.

Do tính phức tạp và không phổ biến của tiếng Việt, mà những nghiên cứu về tóm tắt văn bản tiếng Việt so với tiếng Anh vẫn còn nhiều hạn chế. Hiện nay, hầu hết các nghiên cứu về tóm tắt tiếng Việt tập trung chủ yếu vào trích rút câu và rút gọn câu. Chúng ta vẫn gặp nhiều khó khăn, ngoài việc các công cụ phục vụ tách từ loại hiệu quả chưa cao và chưa có kho ngữ liệu chuẩn phục vụ cho tóm tắt, hiệu năng của các phương pháp cũng cần được cải tiến.

Trong lĩnh vực xử lý ngôn ngữ tự nhiên tiếng Việt, tùy từng mục đích khác nhau mà cần phải có kho ngữ liệu tương ứng, chẳng hạn, với mục đích rút gọn câu, người ta phải xây dựng kho ngữ liệu tiếng Việt phục vụ việc rút gọn câu. Bên cạnh đó, phải lựa chọn nguồn tài liệu phù hợp với lĩnh vực xác định trước hoặc bao phủ nhiều lĩnh vực khác nhau. Các tài liệu có thể được nhập thủ công vào máy tính hoặc được quét (scan) và nhận dạng để chuyển thành tập tin văn bản. Hoặc có thể sử dụng các nguồn tài nguyên trên Internet để xây dựng nguồn dữ liệu kết hợp với sự đánh giá của con người để đánh giá lại các dữ liệu được khai thác từ Internet [3].

Để tóm tắt văn bản tiếng Việt, cần thiết phải có các kho ngữ liệu tiếng Việt và các công cụ phục vụ cho tóm tắt văn bản tiếng Việt. Dưới đây là bảng danh mục và hiện trạng các kho ngữ liệu và các công cụ xử lý tiếng Việt cần thiết.

STT	Kho ngữ liệu / công cụ	Hiện trạng	
		<i>Có</i>	<i>Chưa</i>
1	Công cụ tách từ	X	
2	Công cụ gán nhãn từ loại	X	
3	Kho ngữ liệu phân loại văn bản		X
4	Kho ngữ liệu tóm tắt văn bản		X
5	Viet WordNet Tool		X
6	Công cụ đánh giá tự động		X

*Bảng 1.1. Hiện trạng các kho ngữ liệu tiếng Việt.*

### **1.3 Một số phương pháp tóm tắt văn bản**

#### **1.3.1 Hiện trạng nghiên cứu**

Vấn đề tóm tắt văn bản tự động nhận được nhiều sự quan tâm của các nhà công nghệ thông tin trên thế giới. Có thể thấy rõ nhất là qua công cụ AutoSummarize trong phần mềm Microsoft Word của tập đoàn Microsoft. Có thể nói sơ qua cơ chế làm việc của công cụ này là nó sẽ tính điểm cho các câu chứa từ được lặp lại nhiều lần. Những câu được nhiều điểm nhất sẽ được gợi ý đưa ra cho người dùng. Tuy nhiên đối với các văn bản tiếng Việt thì công cụ này cho kết quả không có tính chính xác cao.

Ngoài ra cũng có các bài báo đề cập đến các công trình nghiên cứu liên quan đến vấn đề xử lý ngôn ngữ tự nhiên trong việc rút trích tự động ý chính

trong văn bản. Những nghiên cứu sớm nhất về tóm tắt văn bản được đề xuất bởi Luhn vào năm 1958, tại Viện nghiên cứu của IBM, trong phương pháp của mình, Luhn đã coi tần suất là đặc trưng chính trong một văn bản và cũng là độ đo quan trọng có ý nghĩa. Ý tưởng này đã mở đầu cho các công trình liên quan sau này. Luhn đã biên dịch từ một danh sách các từ chứa nội dung (content words) được sắp xếp theo tần xuất giảm dần và đánh chỉ số độ đo quan trọng của chúng. Ở mức một câu, nhân tố quan trọng được dựa trên độ đo quan trọng của các từ có mặt trong câu đó và khoảng cách giữa chúng với các từ có độ đo quan trọng thấp. Tất cả các câu được sắp xếp theo thứ tự của nhân tố quan trọng và các câu có vị trí cao nhất sẽ được lựa chọn trong hệ thống tóm tắt tự động [9].

Một nghiên cứu liên quan khác của Baxendale cũng được đề xuất vào năm 1958 tại viện nghiên cứu IBM và công bố trong cùng một tạp chí, cung cấp một góc nhìn khác khi tập trung vào tìm kiếm các thành phần ngữ nghĩa ngầm của các văn bản: Vị trí câu. Theo mục đích này, tác giả đã thu tập 200 đoạn để tìm ra tới 85% trong các đoạn đó, các câu chủ đề nằm ở vị trí đầu đoạn và 7% nằm ở vị trí cuối đoạn. Do đó, đơn giản nhất sẽ chọn câu đứng ở đầu đoạn hoặc cuối đoạn để tạo ra tóm tắt. Đặc trưng về vị trí câu cũng là một trong những đặc trưng tổ hợp trong các hệ thống tóm tắt dựa trên máy học sau này [9].

Nghiên cứu cơ bản của Edmundson năm 1969, mô tả một hệ thống sinh ra văn bản tóm tắt dựa trên cách tiếp cận trích rút câu. Đầu tiên tác giả phát triển một giao thức để tạo trích rút thủ công ứng dụng cho một tập gồm 400 văn bản kỹ thuật. Tiếp theo, các đặc trưng tần suất từ và vị trí quan trọng được sử dụng lại từ các nghiên cứu trước và bổ sung thêm hai đặc trưng nữa

là ngăn chặn từ và vai trò từ, trọng số câu được tính toán dựa trên các đặc trưng này. Khi đánh giá, độ chính xác của phương pháp tương đương với 44% so với trích rút thủ công [9].

Các đề tài đều có ưu điểm nhất định nhưng hầu hết các đề tài đều tập trung xử lý ngôn ngữ tiếng nước ngoài, đa số là các văn bản tiếng Anh. Để áp dụng cho các tài liệu tiếng Việt thì không có được độ chính xác mong muốn do đặc điểm ngôn ngữ tiếng Việt phức tạp và có rất nhiều điểm khác biệt so với ngôn ngữ khác. Một số phần mềm tóm tắt văn bản được đưa lên Internet để sử dụng miễn phí như phần mềm Text Compactor[16].

Hiện nay, các nghiên cứu về tóm tắt văn bản tiếng Việt chưa nhiều. Đã có một số các nghiên cứu được công bố song vẫn còn nhiều hạn chế. Một số công trình nghiên cứu tập trung chính vào vấn đề trích rút các câu trong văn bản gốc và tổng hợp lại thành văn bản tóm tắt của nhóm tác giả Lê Thanh Hà, Huỳnh Thắng và Lương Chi Mai, năm 2005 [13]. Tác giả Nguyễn Thị Thu Hà với công trình tóm tắt văn bản tiếng Việt dựa trên học giám sát bằng mạng nơ ron và một số công trình liên quan khác [1].

## ***1.4 Đánh giá tóm tắt văn bản***

### **1.4.1 Đánh giá theo cách thủ công**

Hội thảo DUC (Document Understanding Conference) đã đưa ra đánh giá về các hệ thống tóm tắt trên tập dữ liệu dùng chung kể từ năm 2001. Nhiều chuyên gia phát triển những phương pháp đánh giá khác nhau. Đánh giá của hội thảo DUC dựa trên chuyên gia con người. Do đó, chỉ dùng chú thích của một người tạo các mô hình với tập dữ liệu kiểm tra khác nhau.

### **1.4.2 Phương pháp đánh giá BLEU**

Độ đo BLEU (Bilingual Evaluation Under Study) do Papineni và cộng sự đề xuất năm 2001. Trong độ đo này họ sử dụng trọng số xuất hiện n-gram. BLEU gắn với NIST (National Institute of Standards and Technology). Một phương pháp liên quan đến đánh giá tóm tắt tự động và được gọi là độ đo NIST. NIST là phương pháp dựa trên BLEU.

Ý tưởng chính của BLEU là đánh giá độ tương tự giữa một văn bản ứng cử (candidate) và tập các bản tham khảo dưới dạng trung bình có trọng số của các n-gram trong văn bản cho bởi hệ thống và trong tập các văn bản tham khảo được cho bởi con người theo công thức (1-1) như sau:

$$p_n = \frac{\sum_{C \in \text{Candidates}} \sum_{n\text{-gram} \in C} \text{Count}_{clip}(n\text{-gram})}{\sum_{C \in \text{Candidates}} \sum_{n\text{-gram} \in C} \text{Count}(n\text{-gram})} \quad (1-1)$$

Trong đó  $\text{Count}_{clip}(n\text{-gram})$  là số n-gram xuất hiện lớn nhất trong văn bản cho bởi hệ thống và văn bản tham khảo và  $\text{Count}(n\text{-gram})$  là số n-gram trong văn bản cho bởi hệ thống. Khi sử dụng phương pháp đánh giá BLEU để đánh giá chất lượng tóm tắt, ta coi văn bản tóm tắt là văn bản ứng viên, văn bản gốc là văn bản nguồn. Trong một số trường hợp người ta sử dụng phương pháp BLEU trong đánh giá chất lượng tóm tắt thủ công.

### 1.4.3 Phương pháp đánh giá ROUGE

Các phương pháp đánh giá tóm tắt truyền thống thường gắn với đánh giá thủ công do chuyên gia con người thực hiện thông qua một số độ đo khác nhau, chẳng hạn: mức độ súc tích, mức độ liền mạch, ngữ pháp, mức độ dễ đọc và nội dung. Tuy nhiên, phương pháp đánh giá kết quả tóm tắt thủ công được báo cáo tại hội thảo DUC 2003 đòi hỏi hơn 3000 giờ. Chi phí này quá cao. Vì thế, đánh giá tóm tắt tự động là một yêu cầu cấp thiết. Lin và Hovy đề

xuất một phương pháp đánh giá mới gọi là ROUGE (Recall-Oriented Understudy for Gisting Evaluation). Hiện nay phương pháp đo này được sử dụng như một phương pháp chuẩn đánh giá kết quả tóm tắt tự động cho văn bản tiếng Anh.

Một cách hình thức, ROUGE-N là một độ đo đối với các n-gram trong văn bản tóm tắt ứng viên và trong tập các văn bản tóm tắt tham khảo, được tính theo công thức (1-2) ở dưới đây.

$$ROUGE - N = \frac{\sum_{S \in \text{ReferenceSummaries}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \text{ReferenceSummaries}} \sum_{gram_n \in S} Count(gram_n)} \quad (1-2)$$

Trong công thức (1-2), n biểu thị cho chiều dài của n-gram,  $gram_n$  và  $Count_{match}(gram_n)$  là số chuỗi n-gram lớn nhất xuất hiện trong văn bản tóm tắt ứng viên và tập các văn bản tóm tắt tham khảo.

#### 1.4.4 Độ đo precision và độ đo recall

Đối với phương pháp tóm tắt văn bản dựa trên trích rút câu, các câu được trích chọn kết nối với nhau, tạo nên văn bản tóm tắt, không cần hiệu chỉnh thêm. Trong trường hợp này, người ta sử dụng độ đo triệu hồi và chính xác để đánh giá chất lượng bản tóm tắt.

Độ đo triệu hồi là tỉ số giữa số lượng các câu đồng thời được trích rút bởi con người và hệ thống trên số các câu chỉ được lựa chọn bởi con người.

$$Recall = \frac{SCHO}{SCH} \quad (1-3)$$

trong đó:

*SCHO*: số lượng những câu được cả hệ thống và con người trích rút.

*SCH*: số lượng những câu được con người trích rút.



Độ đo chính xác là tỉ số giữa số lượng các câu được cả hệ thống và con người trích rút trên số các câu được hệ thống trích rút.

$$Precision = \frac{SCHO}{SCS} \quad (1-4)$$

trong đó:

*SCHO*: số lượng những câu được cả hệ thống và con người trích rút.

*SCS*: số lượng những câu được hệ thống trích rút.

Trong chương này luận văn đã đưa ra tổng quan về tóm tắt văn bản tiếng Việt, hiện trạng nghiên cứu tóm tắt văn bản ở trong nước cũng như ngoài nước, hiện trạng tóm tắt văn bản tiếng Việt hiện nay cũng đã và đang được quan tâm nghiên cứu và phát triển bởi các nhóm xử lý ngôn ngữ tự nhiên tiếng Việt trong nước (JAIST). Luận văn cũng đã đưa ra đặc điểm của ngôn ngữ tiếng Việt, một số phương pháp tóm tắt văn bản đánh giá tóm tắt văn bản. Ở chương 2 của luận văn sẽ đi sâu vào phương pháp tóm tắt văn bản dựa trên Naïve Bayes.

## **Chương 2 : PHƯƠNG PHÁP TÓM TẮT VĂN BẢN TIẾNG VIỆT DỰA TRÊN NAIVE BAYES**

Trong chương này, luận văn trình bày một số phương pháp tóm tắt văn bản điển hình và đi sâu vào phương pháp tóm tắt văn bản Naïve Bayes, trên cơ sở đó tìm hiểu phương pháp rút gọn đặc trưng trong xử lý tiếng Việt và đưa giải pháp tích hợp với bộ công cụ gán nhãn từ loại VnTagger để xây dựng hệ thống thử nghiệm.

### ***2.1 Một số phương pháp tóm tắt văn bản điển hình***

#### **2.1.1 Phương pháp tóm tắt văn bản bằng cây quyết định**

Lin và Hovy (1997) đã nghiên cứu một đặc trưng rất quan trọng, vị trí của câu. Độ quan trọng của câu bằng chính vị trí của nó trong văn bản, tác giả đã gọi là “position method”, nảy sinh từ ý tưởng rằng các văn bản sinh ra một cấu trúc diễn ngôn, và một câu gần chủ đề hơn khuynh hướng tập trung xuất hiện trong vị trí có thể định được (ví dụ tiêu đề, tóm tắt...). Do đó, cấu trúc diễn ngôn quan trọng thay đổi theo lĩnh vực, đặc trưng vị trí câu không thể được định nghĩa đơn giản như (Baxendale, 1958). Nghiên cứu này đã có một đóng góp quan trọng bằng kỹ thuật xác định vị trí tối ưu và cách đánh giá thế nào cho hiệu quả. Một kho dữ liệu tin tức lớn được sử dụng, kho được sưu tập bởi Zif-Davis từ chương trình TIPSTER, nó bao gồm văn bản về máy tính (computer) và liên quan tới phần cứng, thêm vào là tập các từ khóa chủ đề và abstract nhỏ khoảng 6 câu. Có hai cách đánh giá được sử dụng là precision và recall.

Trong nghiên cứu tiếp theo của Lin (1999) đã bác bỏ giả thiết rằng các đặc trưng là độc lập lẫn nhau và đã đưa ra mô hình trích rút câu sử dụng cây

quyết định thay thế cho phân loại Naïve – Bayes. Lin đã khảo sát rất nhiều đặc trưng và hiệu ứng của chúng trong trích rút câu. Dữ liệu được sử dụng trong công việc này được sử dụng tập dữ liệu văn bản chuẩn, đã được phân loại theo các chủ đề khác nhau, cung cấp bởi hệ thống đánh giá TIPSTER-SUMMAC. Các thực nghiệm mô tả là hệ thống SUMMARIST được phát triển tại Trường đại học Southern California.

### **2.1.2 Phương pháp tóm tắt văn bản bằng mạng nơ ron**

Svore và các cộng sự (2007) đưa ra một thuật toán dựa trên mạng neural và sử dụng tập dữ liệu đưa ra để giải quyết vấn đề tóm tắt trích rút, tốt hơn tiêu chuẩn thống kê các đặc trưng quan trọng.

Các tác giả đã sử dụng tập dữ liệu bao gồm 1365 tài liệu thu thập được từ CNN.com, mỗi tài liệu bao gồm tiêu đề, dấu thời gian, các đoạn quan trọng do con người tạo ra và văn bản. Con người tạo ra đoạn quan trọng không đúng theo nguyên văn trích rút từ trong bài báo. Svore đã huấn luyện một mô hình từ các nhãn và các đặc trưng cho mỗi câu trong bài báo, có thể suy luận ra sắp xếp của các câu trong văn bản kiểm tra. Sắp xếp được hoàn thành sử dụng RankNet (Burgess et al., 2005), một cặp dựa trên thuật toán mạng neural thiết kế để sắp xếp một tập đầu vào sử dụng phương pháp giảm gradient trong huấn luyện. Với tập huấn luyện họ sử dụng ROUGE-1 (Lin, 2004) để tính độ tương tự của các câu trong văn bản và đoạn được viết bởi con người. Những độ tương tự này được sử dụng như một nhãn mềm trong suốt quá trình huấn luyện, khác với những đề cập khác các câu là các nhãn cứng.

### **2.1.3 Phương pháp phân tích ngôn ngữ tự nhiên mức sâu**

Đây là kỹ thuật phân tích bao gồm phân tích ngôn ngữ tự nhiên. Phần lớn những kỹ thuật này cố gắng tạo ra một mô hình văn bản súc tích liên mạch.

Barzilay và Elhadad (1997) đã mô tả một công việc sử dụng việc xem xét phân tích ngôn ngữ để nâng cao hiệu năng tóm tắt. Trong đó chuỗi từ vựng (lexical chains) được sử dụng rất nhiều: nó là một chuỗi các từ liên quan trong văn bản, các từ kề nhau hoặc các câu hoặc chiều dài khoảng cách (toàn bộ văn bản). Phương pháp này được thực hiện với các bước sau: tách văn bản, nhận dạng chuỗi từ vựng và sử dụng các chuỗi từ vựng để nhận dạng các câu thích hợp để trích rút. Họ cố gắng sử dụng kết hợp cả phương pháp phân tích thống kê và cả cấu trúc ngữ nghĩa của văn bản.

Các tác giả mô tả khái niệm súc tích trong văn bản có nghĩa móc nối các thành phần khác nhau của văn bản. Ví dụ trong câu

John bought a Jag. He loves the car.

Ở đây, từ car xem xét tới từ Jag trong câu trước và ví dụ minh họa súc tích từ vựng. Hiện tượng súc tích xảy ra không chỉ ở mức từ nhưng cũng không chỉ ở mức các chuỗi từ, kết quả trong các chuỗi từ vựng, các tác giả đã sử dụng một nguồn biểu diễn tóm tắt. Các từ liên quan và chuỗi các từ liên quan ngữ nghĩa được nhận dạng trong văn bản, và một vài chuỗi được trích rút để biểu diễn văn bản. Để tìm ra các chuỗi từ vựng, các tác giả sử dụng Wordnet (Miller, 1995) ứng dụng 3 bước sau đây:

1. Chọn tập các từ ứng cử.
2. Đối với mỗi từ ứng cử, tìm ra chuỗi tương ứng dựa vào một tiêu chuẩn liên quan giữa các thành viên của các chuỗi.
3. Nếu tìm thấy, chèn từ trong chuỗi và cập nhật nó.

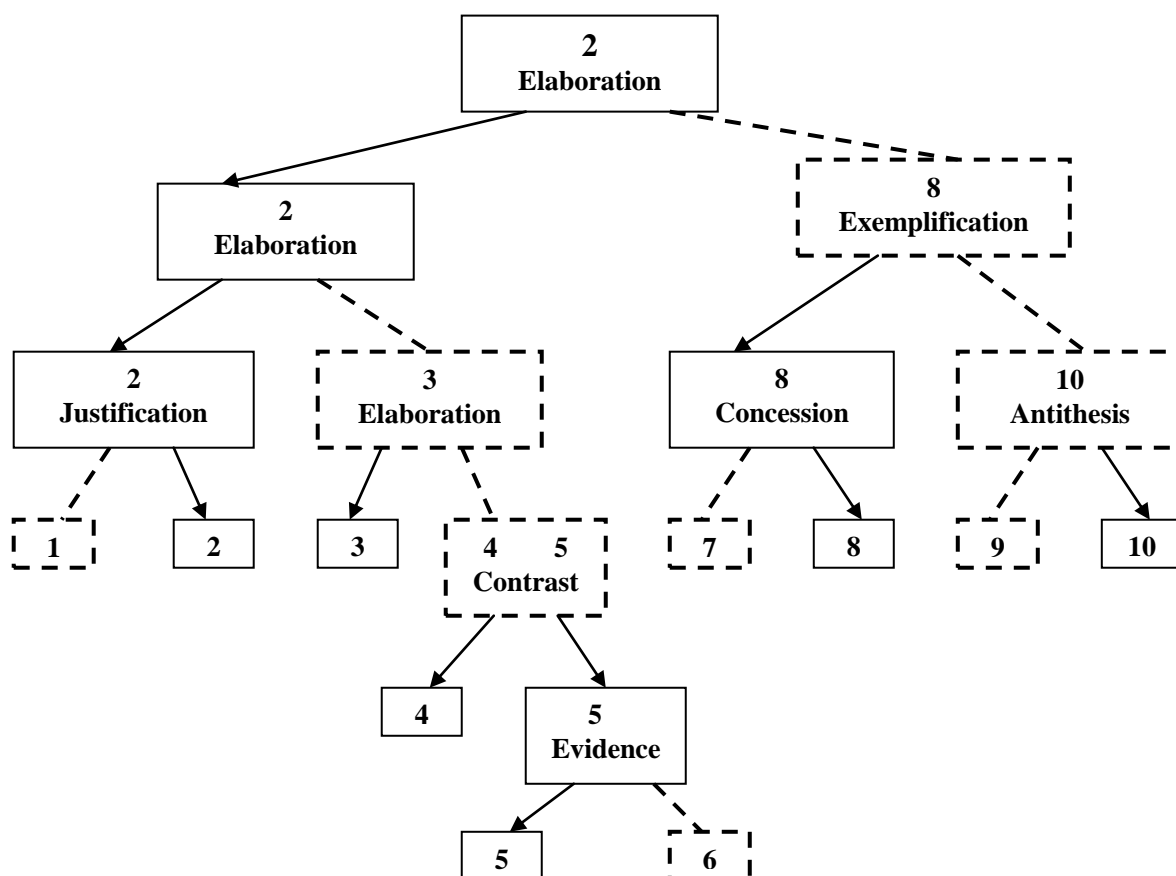
Sự tương thích được đo dựa vào Wordnet. Các danh từ đơn và danh từ ghép được sử dụng như một điểm bắt đầu tới tập ứng cử. Trong bước cuối cùng, các chuỗi từ vựng tốt sẽ được sử dụng để tạo ra các tóm tắt. Các chuỗi

từ vựng được tính trọng số bằng chiều dài. Sau đó, tác giả chọn ra các câu quan trọng.

Trong bài báo khác, Ono và các cộng sự (1994) tiến tới một mô hình tính toán đoạn diễn thuyết cho bài văn tiếng Nhật, trong đó họ thực nghiệm một cách cẩn thận các thủ tục trích rút cấu trúc tu từ trong diễn thuyết, một cây nhị phân biểu diễn quan hệ giữa các câu (cây cấu trúc tu từ được sử dụng trong Marcu, 1998). Cấu trúc này đã trích rút sử dụng chuỗi các bước xử lý ngôn ngữ tự nhiên: phân tích câu, trích rút quan hệ tu từ, tách, sinh ra các ứng cử viên và ưu tiên lời phê bình. Đánh giá đã dựa trên độ quan trọng tương đối của các quan hệ tu từ. Trong bước tiếp theo, các nút của cây cấu trúc tu từ được tĩa để rút gọn câu, giữ lại những thành phần quan trọng. Thực hiện tương tự cho các đoạn cuối cùng được tóm tắt. Đánh giá đã thực hiện trên các câu tinh và 30 bài báo biên dịch của bản tin tiếng Nhật đã được sử dụng như tập dữ liệu.

Marcu (1998) đã mô tả một tiếp cận tóm tắt không giống các phương pháp cũ, không giả thiết giả thiết rằng các câu trong một tài liệu tạo thành một chuỗi. Bài báo này sử dụng diễn thuyết dựa trên khám phá các đặc trưng truyền thống đã được sử dụng trong tóm tắt bài luận. Diễn thuyết được sử dụng trong bài báo này là Thuyết cấu trúc tu từ

Marcu (1998) mô tả chi tiết thủ tục phân tích tu từ thành cây tu từ. Hình 1.1 minh họa một ví dụ cây diễn thuyết trong văn bản.



Hình 2.1. Cây cấu trúc tu từ

Các số trong các nút cho thấy số câu trong văn bản ví dụ. Văn bản phía dưới của số trong các nút được lựa chọn là các quan hệ tu từ. Các nút có dấu chấm là thứ yếu và các nút thường là trung tâm.

#### 2.1.4 Phương pháp tóm tắt ngắn

Wibrock và Mittal (1999) khẳng định rằng tóm tắt trích rút không thực sự tốt trong đó, các trích rút không đủ súc tích khi văn bản tóm tắt là ngắn. Chúng biểu diễn một hệ thống tóm tắt như dạng sinh ra các tiêu đề. Kho dữ liệu sử dụng trong nghiên cứu này là các bài báo tin tức từ Reuters và Associate Press, sẵn có tại LDC. Hệ thống học theo mô hình thống kê các quan hệ giữa các khối văn bản nguồn và khối tiêu đề. Cố gắng để mô hình cả hai loại và khả năng

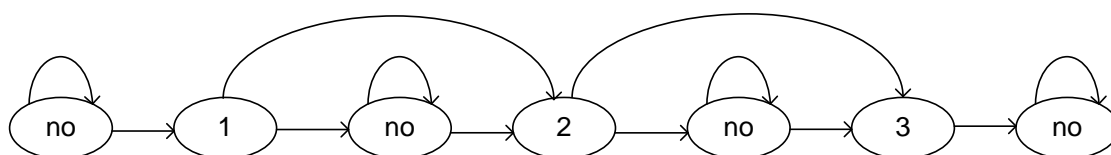
xuất hiện của các tokens trong các tài liệu đích. Cả hai mô hình, một cho trích chọn nội dung và một mô hình khác cho thực hiện bề mặt.

Mô hình trích chọn nội dung là mô hình học từ văn bản và tóm tắt (Brown, 1993). Mô hình này là mô hình đơn giản nhất thông qua việc ánh xạ giữa một từ trong văn bản và một vài từ khả năng xuất hiện trong văn bản tóm tắt. Để đơn giản mô hình này, tác giả đã giả thiết xác suất xuất hiện của một từ trong văn bản tóm tắt phụ thuộc vào cấu trúc của nó.

Mô hình thực hiện bề mặt là mô hình bigram. Viterbi tìm kiếm được sử dụng hiệu quả để tối ưu tóm tắt. Giả thiết Markov ảnh hưởng bằng cách sử dụng backtracking tại mọi trạng thái để tạo đường dẫn liên tục tốt nhất. Để đánh giá hệ thống, tác giả so sánh đầu ra của nó với tiêu đề thực tế trong tập các văn bản đầu vào.

### 2.1.5 Phương pháp dựa trên mô hình Markov ẩn (HMM - Hidden Markov Model)

Khác với các kỹ thuật được đề cập ở trên, dựa trên ý tưởng về các đặc trưng không liên tục Conroy và O'leary đưa ra một phương pháp tóm tắt văn bản dựa trên mô hình Markov ẩn. Các tác giả sử dụng mô hình chuỗi để tính toán phụ thuộc cục bộ giữa các câu. Ba đặc trưng được sử dụng là: Vị trí câu trong văn bản, Số thuật ngữ trong câu và Xác suất của thuật ngữ trong câu đã có trong thuật ngữ văn bản.



Hình 2.2. Mô hình Markov ẩn sử dụng trong trích rút câu.

Trong mô hình này, tác giả sử dụng tập dữ liệu huấn luyện TREC và xác định giá trị lớn nhất đối với mỗi xác suất dịch chuyển. Để đánh giá kết quả tóm tắt, các tác giả so sánh với tóm tắt trích rút bởi con người.

### **2.1.6 Phương pháp tóm tắt dựa trên rút gọn câu**

Trong nghiên cứu của Knight và Marcu, họ đã xây dựng một kho dữ liệu tiêu chuẩn và đề xuất phương pháp đánh giá cho rút gọn câu. Họ sử dụng kho dữ liệu của Ziff – Davis với hơn 4000 tài liệu kỹ thuật và trích rút được 1,067 cặp câu gốc- rút gọn. Nhiệm vụ được xác định là cho một câu dài  $l$ , nén theo phiên bản  $c$  và giữ lại nghĩa của câu, ngữ pháp tốt. Họ cũng đề xuất hai kỹ thuật học khác nhau để sinh ra câu rút gọn, một phương pháp sử dụng kênh nhiễu (noisy channel), phương pháp còn lại sử dụng cây quyết định.

### **2.1.7 Phương pháp tóm tắt văn bản bằng Naïve Bayes:**

Kupiec (1995) đã mô tả một phương pháp bắt nguồn từ Edmundson (1969) đó là học từ dữ liệu. Sử dụng hàm phân loại mỗi câu về các lớp khác nhau. Giả sử  $s$  là một câu,  $S$  là tập các câu tạo nên văn bản tóm tắt, và  $F_1 \dots F_k$  là các đặc trưng. Những đặc trưng dựa trên phương pháp Edmundson (1969) và được bổ sung thêm một số các đặc trưng khác: chiều dài câu và sự xuất hiện của từ viết hoa. Mỗi câu sau khi tính toán sẽ có một giá trị nhất định, và được sắp xếp theo thứ tự giảm dần, chỉ có  $n$  câu đứng đầu được trích rút. Để đánh giá hệ thống Kupiec đã sử dụng một kho dữ liệu văn bản bao gồm các tài liệu kỹ thuật cùng với các văn bản tóm tắt đã được tóm tắt bởi con người [9].

Aoen và các cộng sự (1999) cũng sử dụng phương pháp phân loại của Naïve- Bayes, nhưng thêm vào đó một số đặc trưng. Họ xây dựng một hệ thống gọi là DimSum được dựa trên các đặc trưng: như tần suất từ (tf) và tần



suất nghịch đảo văn bản (idf) để thu được các từ quan trọng. idf được tính từ trong tập dữ liệu lớn các văn bản trọng tâm cùng chủ đề. Họ cũng thực hiện một số phân tích bề mặt như tồn tại độ tương tự nhau giữa các câu trong văn bản, duy trì súc tích. Các thông kê tên viết tắt trong văn bản tựa như U.S thành United States hoặc IBM là International Business Machines. Từ đồng nghĩa và hình thái từ cũng được sử dụng trong khi xem xét thuật ngữ từ vựng, nhận dạng sử dụng Wordnet ( Miler, 1995 ). Kho dữ liệu sử dụng trong thực nghiệm được lấy từ các trang tin, và đánh giá dựa vào TREC.

## ***2.2 Phương pháp tóm tắt văn bản sử dụng lý thuyết phân loại Naïve Bayes***

### **2.2.1 Phân loại Naïve Bayes**

Phân loại Naïve Bayes (Naïve Bayes Classifier) là một thuật ngữ trong xử lý số liệu thống kê Bayesian với một phân lớp xác suất dựa trên các ứng dụng định lý Bayes. Naïve Bayes là phương pháp phân loại dựa vào xác suất được sử dụng rộng rãi trong lĩnh vực máy học, có thể được đào tạo hiệu quả trong một thiết lập học có giám sát, phương pháp phân loại này được sử dụng lần đầu tiên trong lĩnh vực phân loại bởi Maron vào năm 1961 sau đó trở nên phổ biến dùng trong nhiều lĩnh vực như trong các công cụ tìm kiếm.

Naïve Bayes sử dụng xác suất có điều kiện giữa từ và chủ đề để dự đoán xác suất chủ đề của một văn bản cần phân loại. Điểm quan trọng của phương pháp này chính là ở chỗ giả định rằng sự xuất hiện của tất cả các từ trong văn bản là độc lập với nhau. Như thế Naïve Bayes không tận dụng được sự phụ thuộc của nhiều từ vào một chủ đề cụ thể làm cho việc tính toán của Naïve Bayes hiệu quả và nhanh chóng hơn các phương pháp khác với độ phức tạp theo số mũ vì nó không sử dụng việc kết hợp các từ để đưa ra phán đoán. Mặc dù phương pháp phân loại Naïve Bayes khá đơn giản nhưng nó có khả năng phân loại tốt hơn nhiều các phương pháp phân hoạch khác. Với mỗi loại

văn bản thuật toán Naïve Bayes tính cho mỗi lớp văn bản một xác suất mà tài liệu cần phân hoạch có thể thuộc loại đó, tài liệu đó sẽ được gán cho lớp văn bản nào có xác suất cao nhất.

Thuật toán Naïve Bayes được xem là thuật toán đơn giản so với các phương pháp khác. Bộ phân lớp Bayes có thể dự báo các xác suất là thành viên của lớp, chúng giả định các thuộc tính là độc lập nhau (độc lập điều kiện lớp). Thuật toán Naïve Bayes được dựa trên định lý Bayes, định lý được phát biểu như sau:

$$P(Y|X) = \frac{P(X|Y) * P(Y)}{P(X)} \quad (2-1)$$

Trong đó

Y đại diện một giả thuyết mà sự kiện liên quan X đã xảy ra

$P(X)$ : Xác suất X xảy ra

$P(Y)$ : Xác suất Y xảy ra

$P(X|Y)$ : Xác suất X xảy ra khi Y xảy ra (xác suất có điều kiện, khả năng X khi Y đúng)

$P(Y|X)$ : Xác suất hậu nghiệm của Y nếu biết X

Áp dụng trong bài toán phân loại, các dữ liệu cần có

D: Tập dữ liệu huấn luyện đã được vecto hóa dưới dạng

$$\vec{x} = (x_1, x_2, \dots, x_n)$$

$C_i$ : tập các tài liệu của D thuộc lớp  $C_i$  với  $i = \{1, 2, \dots\}$

Các thuộc tính  $x_1, x_2, \dots, x_n$  độc lập xác suất đôi một với nhau

Theo định lý Bayes:

$$P(C_i|X) = \frac{P(X|C_i).P(C_i)}{P(X)} \quad (2-2)$$

Theo tính chất độc lập điều kiện:

$$P(X|C_i) = \prod_{k=1}^n p(x_k|C_i) = P(x_1|C_i) \cdot P(x_2|C_i) \dots P(x_n|C_i) \quad (2-3)$$

Khi đó luật phân lớp cho các tài liệu mới  $X^{\text{new}} = \{x_1, x_2, \dots, x_n\}$  là

$$\max(P(C_i) \prod_{k=1}^n P(x_k|C_i)) \quad (2-4)$$

Trong đó

$P(C_i)$ : được tính dựa trên tần suất xuất hiện tài liệu trong tập huấn luyện

$P(x_k|C_i)$ : được tính từ những tập thuộc tính đã được tính trong quá trình huấn luyện

Trên cơ sở của định lý Bayes ta đi vào thuật toán Naïve Bayes. Các bước tiến hành thuật toán:

#### Bước 1

Huấn luyện Naïve Bayes (dựa vào tập dữ liệu)

+ Tính xác suất  $P(C_i)$

+ Tính xác suất  $P(x_k|C_i)$

#### Bước 2

$X^{\text{new}}$  được gán vào lớp có giá trị lớn nhất theo công thức

$$\max(P(C_i) \prod_{k=1}^n P(x_k|C_i))$$

Xét một ví dụ kinh điển là ví dụ dự đoán xem quyết định của người chơi có đi chơi tennis hay không với các điều thời tiết đã được dự báo trước. ta có bảng dữ liệu huấn luyện:

Day	Outlook	Temp.	Humidity	Wind	Play tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

*Bảng 2.1 : Ví dụ về bảng huấn luyện*

Bước 1:

Tính xác suất  $P(C_i)$

- với  $C_1 = \text{"yes"}$

$$P(C_1) = P(\text{"yes"}) = 9/14$$

- với  $C_2 = \text{"no"}$

$$P(C_1) = P(\text{"no"}) = 5/14$$

Tính xác suất  $P(x_k | C_i)$

- Với thuộc tính Outlook: có các giá trị sunny, overcast, rain

$$P(\text{sunny} | \text{yes}) = 2/9$$

$$P(\text{sunny} | \text{no}) = 3/5$$

$$P(\text{overcast} | \text{yes}) = 4/9$$

$$P(\text{overcast} | \text{no}) = 0/5$$

$$P(\text{rain} | \text{yes}) = 3/9$$

$$P(\text{rain} | \text{no}) = 2/5$$

- Với thuộc tính Temp: có các giá trị hot, cool, mild

$$P(\text{hot} | \text{yes}) = 2/9$$

$$P(\text{hot} | \text{no}) = 2/5$$

$$P(\text{cool} | \text{yes}) = 3/9$$

$$P(\text{cool} | \text{no}) = 1/5$$

$$P(\text{mild} | \text{yes}) = 4/9$$

$$P(\text{mild} | \text{no}) = 1/5$$

- Với thuộc tính Humidity: có các giá trị normal, high

$$P(\text{normal} | \text{yes}) = 6/9$$

$$P(\text{normal} | \text{no}) = 1/5$$

$$P(\text{high} | \text{yes}) = 3/9$$

$$P(\text{high} | \text{no}) = 4/5$$

- Với thuộc tính Wind: có các giá trị weak, strong

$$P(\text{wesk} \mid \text{yes})=6/9$$

$$P(\text{weak} \mid \text{no})=2/5$$

$$P(\text{strong} \mid \text{yes})=3/9$$

$$P(\text{strong} \mid \text{no})=3/5$$

Bước 2: Phân lớp  $x^{\text{new}}=\{\text{sunny, cool, high, strong}\}$

Tính xác suất

$$P(\text{yes}). P(x^{\text{new}} \mid \text{yes})=0.005$$

$$P(\text{no}). P(x^{\text{new}} \mid \text{no})=0.021$$

Từ kết quả này ta có  $x^{\text{new}}$  thuộc lớp no

Luận văn sử dụng phân loại Naïve Bayes để tính xác suất của câu s với k đặc trưng khác nhau  $F_1, F_2, \dots, F_k$  để phân loại câu s có được lựa chọn hay không được lựa chọn.

$$P(s \in S \mid F_1, F_2, \dots, F_k) = P(F_1, F_2, \dots, F_k \mid s \in S) \times P(s \in S) / P(F_1, F_2, \dots, F_k) \quad (2-4)$$

Giả thiết rằng các đặc trưng là độc lập với nhau, công thức trên được chuyển đổi thành

$$P(s \in S \mid F_1, F_2, \dots, F_k) = \prod P(F_j \mid s \in S) P(s \in S) / \prod P(F_j) \quad (2-5)$$

Làm tròn công thức trên theo luật logarit:

$$P(s \in S \mid F_1, F_2, \dots, F_k) = \log(P(s)) + \sum \log P(F_j \mid s) \quad (2-6)$$

Trong đó:

$P(s) = C(s)/C(w)$  trong đó  $C(s)$  là số các câu trong tập huấn luyện và  $C(s)$  là trong lớp C,  $C(w)$  là tổng các câu trong tập huấn luyện.

$P(F_j \mid s) = C(F_j, s)/C(s)$ . Trong đó  $C(F_j, s)$  là số lần xuất hiện của đặc trưng  $F_j$  trong câu của lớp C.

Luận văn sử dụng phân loại Naive Bayes để phân loại thành hai lớp riêng biệt (lớp được trích rút và lớp không được trích rút). Từ đó, tính toán xác suất theo

mỗi trường hợp  $P(s \in S | F_j)$  và  $P(s \notin S | F_j)$ . Câu sẽ được lựa chọn nếu như  $P(s \in S | F_j) > P(s \notin S | F_j)$ .

## 2.2.2 Lựa chọn các đặc trưng cho trích chọn

### 2.2.2.1 Khái niệm giảm chiều đặc trưng

Biểu diễn văn bản là phương pháp thể hiện nội dung hoặc đặc trưng riêng của văn bản đó bằng mô hình khác thay thế cho biểu diễn dạng text thông thường. Khi biểu diễn văn bản bằng mô hình véc tơ không gian, người ta thường sử dụng các véc tơ biểu diễn đặc trưng của thuật ngữ (term) hay từ (word), giá trị của mỗi đặc trưng này gọi là trọng số thuật ngữ (term weight), để mô tả tần suất của thuật ngữ xuất hiện trong văn bản.

#### **Định nghĩa 2.1 [Trọng số của thuật ngữ (term weight)]**

Trọng số của thuật ngữ là cách thể hiện độ quan trọng của thuật ngữ đó trong văn bản hoặc trong một tập văn bản.

#### **Định nghĩa 2.2 [Độ quan trọng của từ]**

Độ quan trọng của từ biểu thị sự ảnh hưởng của từ này đối với văn bản chứa nó. Độ quan trọng của từ tỉ lệ thuận với tần suất xuất hiện của từ này trong một hoặc một tập văn bản.

**Ví dụ 2.1:** Giả sử có một đoạn văn bản liên quan tới thể thao. Ta có thể tìm trên trang web bốn thuật ngữ liên quan: bóng đá, quần vợt, sân vận động, Chelsea. Tần suất của chúng lần lượt là: 8, 6, 7, 2. Ta có thể dùng một véc tơ đặc trưng của văn bản để biểu diễn sự xuất hiện của bốn từ này như sau:

$$\vec{d}_j = (8, 6, 7, 2)$$

Một cách tổng quát của ví dụ trên, có thể biểu diễn véc tơ cho một văn bản  $d_j$  như sau:

$$\overrightarrow{d_j} = (w_{1,j}, w_{2,j}, w_{3,j}, \dots, w_{n,j})$$

Trên đây là ví dụ về biểu diễn một văn bản dựa trên đặc trưng tần suất thuật ngữ. Trên thực tế, có nhiều phương pháp biểu diễn văn bản khác nhau như: phương pháp Boolean, mô hình xác suất, mô hình không gian véc tơ, LSI, ...

Xem xét một số ứng dụng ví dụ như trong một hệ thống xử lý dữ liệu (tín hiệu tiếng nói, ảnh hoặc nhận dạng mẫu nói chung) tập các đặc trưng nếu coi là tập hợp các véc tơ giá trị thực. Giả thiết rằng, hệ thống chỉ hiệu quả nếu số chiều của mỗi véc tơ riêng lẻ không quá lớn. Vấn đề của giảm chiều xuất hiện khi dữ liệu có số chiều lớn hơn khả năng xử lý của hệ thống [3]. Xét một ví dụ điển hình sau:

Một hệ thống nhận dạng phân loại khuôn mặt dựa trên ảnh đa cấp xám kích cỡ  $m \times n$ , tương ứng với  $m \times n$  chiều véc tơ giá trị thực. Trong thực nghiệm, một ảnh có thể có  $m=n=256$  hoặc 65536 chiều. Nếu sử dụng mạng một perceptron đa lớp để thực hiện hệ thống phân loại, trọng số sẽ quá nhiều [3].

Giả sử một ma trận dữ liệu  $A^n$  bao gồm  $n$  hàng (điểm dữ liệu) và trong  $R^D$ ,  $D$  là các chiều (các đặc trưng hoặc các thuộc tính). Ma trận  $A$  được biểu diễn như hình dưới đây.



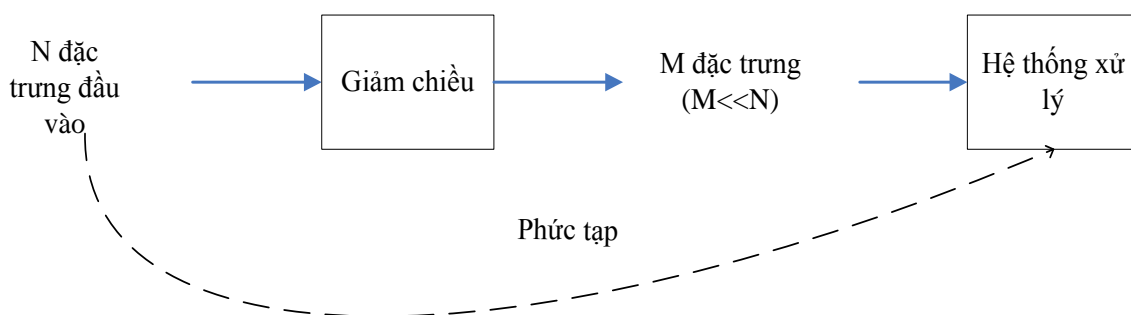
$$A =$$

	$t_1$	$t_2$	$t_3$	$t_4$	$\dots$	$t_n$
$u_1$	*	*	*	*	$\dots$	*
$u_1$	*	*	*	*	$\dots$	*
$u_3$	*	*	*	*	$\dots$	*
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$u_4$	*	*	*	*	$\dots$	*

Hình 2.3. Ma trận ví dụ.

Chi phí tính toán là  $O(n^2D)$ . Trong trường hợp nếu  $n = 0.6$  triệu,  $D = 70$  triệu.  $n^2D = 2.5 \times 10^{19}$ . Như vậy, quá lớn cho quá trình xử lý.

Do đó, vấn đề giảm chiều là vấn đề tương đối cần thiết trong các bài toán làm việc với dữ liệu có nhiều đặc trưng ví dụ như ảnh, tiếng nói, văn bản,.... Vấn đề giảm chiều véc tơ được đơn giản hóa như hình 2.4



Hình 2.4. Mô hình giảm chiều véc tơ

Mô hình giảm chiều trên thể hiện một cách tổng quát nhất sự phụ thuộc của  $N$  đặc trưng đầu vào đối với một hệ thống xử lý. Nếu chúng ta không xử lý giảm chiều mà giữ nguyên  $N$  đặc trưng đầu vào để đưa vào hệ thống xử lý,

độ phức tạp sẽ cao. Nếu ta giảm số chiều của véc tơ đặc trưng xuống còn  $M$  chiều. Với  $M$  nhỏ hơn rất nhiều so với  $N$  ban đầu, hệ thống xử lý sẽ dễ dàng hơn, hạn chế được độ phức tạp và mang lại độ chính xác cao hơn, nếu ta biết giảm chiều một cách hợp lý.

#### **2.2.2.2 Phương pháp giảm chiều biểu diễn đặc trưng sử dụng trong luận văn**

Giảm chiều đặc trưng là vấn đề quan trọng trong xử lý các dữ liệu đầu vào của một hệ thống. Giả sử một văn bản gồm  $n$  từ khác nhau, nếu coi mỗi từ là một đặc trưng của văn bản thì văn bản đó sẽ bao gồm  $n$  đặc trưng. Xét văn bản theo ví dụ 2.2 dưới đây.

Ví dụ 2.2: Cho văn bản theo hình sau:

Thủ lĩnh phiến quân Ukraine bàn giao hai hộp đen của chiếc máy bay MH17 cho các chuyên gia Malaysia, trong khi thi thể của các nạn nhân đang được chuyển về Hà Lan.

"Chúng đây, những chiếc hộp đen", ông Aleksander Borodai sáng sớm nay tuyên bố trong căn phòng đặc kín phóng viên tại trụ sở của nước Cộng hòa Nhân dân Donetsk tự xưng. Một phiến quân có vũ trang liền đưa các hộp đen ra và đặt lên bàn.

Ông Borodai và các chuyên gia Malaysia sau đó ký vào một văn bản được xem là thỏa thuận nhằm kết thúc các thủ tục sau những cuộc đàm phán kéo dài giữa hai bên.

"Tôi có thể thấy rằng các hộp đen vẫn còn nguyên vẹn, dù hơi trầy xước chút ít. Chúng ở trong điều kiện tốt", đại tá Mohamed Sakri, thuộc Hội đồng An ninh Quốc gia Malaysia cho biết và cảm ơn Borodai vì đã trao trả các

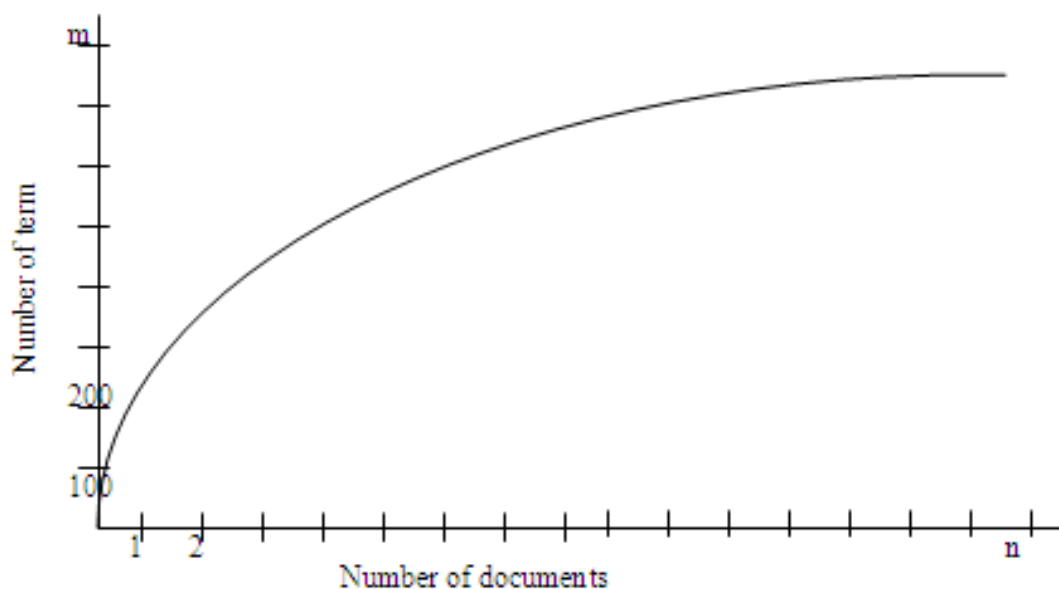
thiết bị này.

Thủ lĩnh phe ly khai cũng cho biết thêm rằng, một chuyến tàu chở thi thể của 282 nạn nhân trên chiếc máy bay của Malaysia Airlines đã có mặt ở Donetsk và đang trên đường đến Kharkiv, cách đó khoảng 300 km về phía tây bắc. Các chuyên gia Malaysia và Hà Lan sẽ đi cùng đoàn tàu. Hiện còn thi thể của 16 người đang được tìm kiếm.

*Hình 2.5. Văn bản ví dụ*

Văn bản trên gồm 9 câu và 157 từ tiếng Việt. Thông thường, các văn bản dài hơn 9 câu, do vậy rất mất thời gian khi xử lý và tính toán với số lượng từ lớn. Các phương pháp giảm chiều đặc trưng cho văn bản tiếng Việt trước đây thường sử dụng kỹ thuật loại bỏ đi các từ dừng, từ không cần thiết trong văn bản, do số lượng các từ dừng không quá nhiều, nên số đặc trưng được giảm cũng không nhiều.

Hình dưới đây mô tả sự biến thiên giữa số đặc trưng trong văn bản tương ứng với số lượng văn bản. Độ biến thiên này được coi là tỉ lệ thuận giữa số lượng văn bản và số đặc trưng. Nếu ta xét ở phạm vi nhỏ và chi tiết hơn với đối tượng là một văn bản thì số véc tơ đặc trưng của văn bản tỉ lệ thuận với chiều dài của văn bản. Vậy, khi chiều dài văn bản tăng lên, thì số véc tơ đặc trưng cũng tăng lên.



Hình 2.6 Quan hệ giữa số văn bản và số thuật ngữ

Ví dụ 2.3: Để chi tiết hơn, giả sử coi văn bản là một câu như sau:

*“Thủ lĩnh phiến quân Ukraine bàn giao hai hộp đen của chiếc máy bay MH17 cho các chuyên gia Malaysia, trong khi thi thể của các nạn nhân đang được chuyển về Hà Lan.”*

Để tách văn bản trên thành các từ, luận văn sử dụng công cụ tách từ trên hệ thống VLSP. Dựa trên hệ thống VLSP, văn bản trên được tách thành 25 từ bao gồm: “Thủ lĩnh”, “phiến quân”, “ukraine”, “bàn giao”, “hai”, “hộp đen”, “chiếc”, “máy bay”, “MH17”, “cho”, “các”, “chuyên gia”, “Malaysia”, “trong”, “khi”, “thi thể”, “của”, “các”, “nạn nhân”, “đang”, “được”, “chuyển”, “về”, “Hà Lan”.

T	h	ủ	l	ĩ	n	h	ph	i	ễ	n	_	q	u	â	n	U	k	r	a	i	n	e	b	à	n	_	g	i	a	o	H	a	i	h	ộ	p	_	đ	e	n	c	ù	a	c	h	i	ế	c	m	á	y	_	b	a	y	M	H	1	7	C	h	o	c	á	c	ch	u	y	ê	n	_	g	i	a	M	a	l	a	y	s	i	a	T	r	o	n	g	K	h	i	t	h	i	th	ể	c	ù	a	c	á	c	n	à	n	_	n	h	â	n	Đ	a	n	g	đ	ư	ợ	c	ch	u	y	ê	n	V	ề	H	à	_	L	a	n
---	---	---	---	---	---	---	----	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	----	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	----	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	----	---	---	---	---	---	---	---	---	---	---	---	---

Hình 2.7 Tách từ dựa trên hệ thống phân tích câu VLSP.

Văn bản trong ví dụ 2.3 cần xử lý 25 từ khác nhau. Văn bản trong ví dụ 2.3 có thể được xử lý ngắn gọn hơn bằng cách chỉ sử dụng các danh từ trong văn bản trong xử lý.

**Định nghĩa 2.1:** Danh từ (Nouns) Các danh từ là từ biểu thị một người, địa điểm, sự kiện, động vật hay ý tưởng. Trong ngôn ngữ học, một danh từ là một phần của nhóm từ vựng thường xuất hiện trong văn bản, là chủ từ của các mệnh đề, đối tượng của động từ hay của một giới từ (theo Wikipedia).

Ví dụ: Trời (N) mưa (V).

**Định nghĩa 2.2:** (Ý nghĩa của danh từ) Trong một câu, một mệnh đề hay một văn bản, danh từ mang ý nghĩa về mặt nội dung, thông tin, tiêu đề hay chủ đề của câu, mệnh đề hay văn bản chứa nó.

Từ định nghĩa trên, muốn tìm hiểu thông tin của một văn bản muốn mô tả về sự kiện gì ta có thể dựa trên tập các danh từ được trích rút ra từ văn bản đó, là có thể hiểu được một phần nội dung mà văn bản muốn trình bày.

Với ý tưởng này, ta sử dụng phương pháp giảm chiều đặc trưng cho văn bản tiếng Việt bằng cách chỉ trích rút ra các danh từ trong văn bản để mô tả ý nghĩa của văn bản và xử lý trên tập danh từ đó. Trở lại ví dụ 2.3, văn bản trên có thể được xử lý để tách thành tập các danh từ, sẽ làm giảm đi độ phức tạp tính toán về mặt thời gian đồng thời có thể tăng độ chính xác hơn.

Ví dụ 2.4: Tách danh từ trong văn bản ở ví dụ 2.3

Văn bản tại ví dụ 2.3 gồm 11 danh từ bao gồm: “*Thủ lĩnh*”, “*phiến quân*”, “*ukraine*”, “*hộp đen*”, “*máy bay*”, “*MH17*”, “*chuyên gia*”, “*Malaysia*”, “*thi thể*”, “*nạn nhân*”, “*Hà Lan*”.

Với phương pháp tách từ như trên. Số lượng đặc trưng cần xử lý giảm tới hơn 50%.

### 2.2.2.3 Các đặc trưng trích chọn

Trong phương pháp này, nhận diện các câu sử dụng trích rút dựa trên việc kết hợp ba đặc trưng: độ quan trọng thông tin, lượng thông tin và vị trí của câu:

*Độ quan trọng thông tin (information significant):*

Đối với văn bản tiếng Anh, các danh từ của một câu thường là những từ mang thông tin của câu đó và các từ còn lại trong câu thường phản ánh về mặt ngữ nghĩa của câu. Trong luận văn sử dụng ý tưởng này để giải quyết bài toán gán nhãn từ và lựa chọn các từ quan trọng trong câu. Được biểu diễn bởi độ quan trọng của các từ trong câu và nó được tính bằng số các câu xuất hiện từ đó trên tổng các câu được tách ra từ tập văn bản huấn luyện và tính trọng số của từ. Thông thường, để tính trọng số của từ người ta thường dùng kỹ thuật tfidf. Kỹ thuật tính trọng số này đòi hỏi kho ngữ liệu dùng cho huấn luyện phải lớn. Tuy nhiên, hiện nay kho ngữ liệu tiếng Việt dùng cho tóm tắt văn bản tiếng Việt còn thiếu cho nên chúng tôi sử dụng kỹ thuật tính trọng số của từ cải tiến, được mô tả theo công thức sau:

$$I(w_i) = \frac{N_S(w_i)}{\sum_{w_j \in d} N_S(w_j)} + \frac{N_D(w_i)}{N_D} \quad (2-7)$$

trong đó:

- $I(w_i)$  là trọng số của từ  $w_i$ .
- $N_S(w_i)$  là số lần xuất hiện của từ  $w_i$  trong câu  $S$ .
- $N_D(w_i)$  là số văn bản xuất hiện  $w_i$  trong tập văn bản huấn luyện  $D$ .
- $N_D$  là tổng số các văn bản trong tập huấn luyện  $D$ .

*Lượng thông tin trong câu:* Theo định nghĩa 2.2, các danh từ trong câu chứa nội dung, thông tin của câu. Do đó, trong phương pháp này sử dụng lượng thông tin của câu được tính bằng tổng các danh từ có trong câu. Trong cách đề xuất khái niệm “lượng thông tin” trong câu thay thế cho phương pháp “chiều dài câu” bởi một số các nghiên cứu của Edmundson, Baxendale cho rằng, các câu dài thường chứa nhiều thông tin quan trọng. Tuy nhiên, một số các nghiên cứu gần đây cho thấy, một số các câu ngắn hơn nhưng mang nội dung, thông tin nhiều hơn. Do vậy, trong luận văn sử dụng khái niệm “lượng thông tin” để đảm bảo các câu được trích rút chứa nhiều thông tin hơn.

*Vị trí của câu:* Các câu trong văn bản ở mỗi vị trí khác nhau có ý nghĩa khác nhau. Thông thường, các câu có vị trí đầu các đoạn văn hoặc trên tiêu đề thường có ý nghĩa phản ánh nhiều nội dung quan trọng trong đoạn văn hoặc văn bản, do vậy, đặc trưng vị trí câu được tính theo công thức  $Pos(s_i) = 1 / i$ .  $i = 1$  khi  $s_i$  nằm ở đầu các đoạn.

### **2.3 Huấn luyện và tính trọng số các câu trong tập huấn luyện**

Trong pha huấn luyện, các văn bản được tách thành các câu và được gán nhãn thủ công bởi con người cho những câu được trích rút và những câu không được trích rút.

Pha huấn luyện được tiến hành theo các bước sau:

Bước 1: Xây dựng tập dữ liệu dùng cho huấn luyện bao gồm  $n$  văn bản khác nhau được ký hiệu  $D = \{d_1, d_2, \dots, d_n\}$

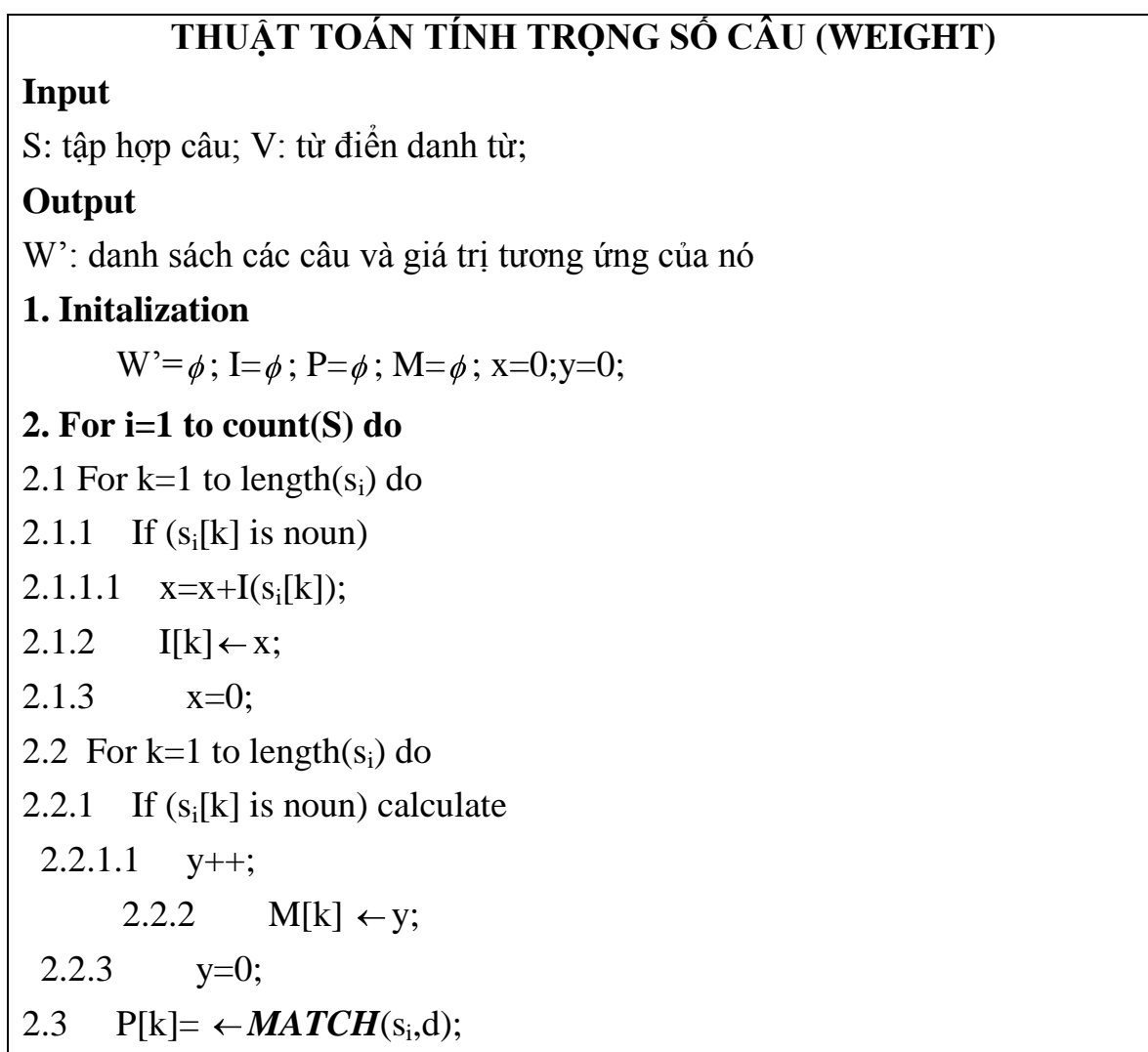
Bước 2: Sử dụng công cụ tách từ để xây dựng tập danh từ  $V$ .

Bước 3: Lựa chọn thủ công các câu có độ quan trọng trong mỗi văn bản  $d_s$  để tạo ra các tóm tắt.

Bước 4: Chia thành hai lớp khác nhau: Các câu được lựa chọn sẽ gán nhãn (+) và không được lựa chọn sẽ gán nhãn (-). Lưu giữ vị trí các câu kể cả các câu được gán nhãn (+) và (-) trong mỗi văn bản  $d_s$ .

Bước 5: Tính giá trị độ quan trọng thông tin của mỗi danh từ trong cả các câu gán nhãn (+) và gán nhãn (-). Lưu giữ các giá trị này trong bảng cơ sở dữ liệu của hệ thống.

Hình dưới đây là thuật toán tính trọng số của câu, với đầu vào là tập các câu và tập các danh từ, đầu ra là các câu và giá trị tương ứng của nó.



Hình 2.8. Thuật toán tính trọng số của câu



Mô tả thuật toán : Trong thuật toán trên, có sử dụng một số hàm ***count()*** là hàm đếm số từ trong câu. Hàm ***length()*** là hàm trả về giá trị chiều dài của câu, hàm ***match()*** là hàm đối sánh để lấy giá trị từ trong cơ sở dữ liệu.

#### 2.4 Lựa chọn các câu tạo tóm tắt

Trong pha tóm tắt, các câu quan trọng được lựa chọn từ văn bản gốc để tạo ra văn bản tóm tắt. Quá trình này gồm bốn bước như sau:

Bước 1: T là văn bản gốc

Bước 2: Tách T thành tập các câu  $S = \{s_1, s_2, \dots, s_n\}$

Bước 3: Đối với mỗi câu  $s_i$  trong T, tính toán xác suất của mỗi đặc trưng  $F_j$ , rồi sử dụng phân loại Naive Bayes để tính toán xác suất của  $s_i$  đối với cả hai lớp (+) và (-).

Bước 4: Trong trường hợp xác suất của  $s_i$  đối với lớp (+) lớn hơn xác suất của  $s_i$  đối với lớp (-),  $s_i$  sẽ được lựa chọn để tạo ra tóm tắt.

Hình dưới đây mô tả thuật toán trích rút câu tạo tóm tắt.

#### THUẬT TOÁN TRÍCH RÚT CÂU

##### Input:

- C: văn bản gốc;
- K: Mục từ chủ đề và giá trị của nó;
- n: số câu đã được gán nhãn (+);
- n': số câu đã được gán nhãn (-);

##### Output:

- T: văn bản tóm tắt

##### Initialization

$T = \phi$ ;

```

S=Split( C); // chia câu từ C
F= $\phi$ ; F'= $\phi$ ; m=0;

1.   For each sentence  $s_i$  in C do
1.1   For j=1 to length  $s_i$  do
1.1.1   If  $w(j) \in V$  then
1.1.1.1   match_hush( $w(j)$ , K)           // Đối sánh với bảng K
1.1.1.2    $F(k) \leftarrow n(j)$            // Tần số w (k) xuất hiện trong câu có
nhãn (+)
1.1.1.3    $F'(k) \leftarrow n'(j)$        // Tần số w (j) xuất hiện trong câu có
nhãn (-)
1.1.1.4    $m=m+1$ ;                     //tính từ chủ đề
1.1.1.5    $W(s_i) = W(s_i) + \log(\frac{F(k)}{n})$ ; // độ quan trọng trong câu nhãn (+)
1.1.1.6    $W'(s_i) = W'(s_i) + \log(\frac{F'(k)}{n'})$ ; // độ quan trọng trong câu nhãn (-)
1.2    $W(s_i) = W(s_i) + \log(\frac{n}{n+n'}) + \log(\frac{m}{n}) + \log(\frac{Pos(s_i)}{n})$ ; // Xác suất  $s_i$  với lớp nhãn (+)
1.3    $W'(s_i) = W'(s_i) + \log(\frac{n}{n+n'}) + \log(\frac{m}{n'}) + \log(\frac{Pos(s_i)}{n'})$ ; //Xác suất của  $s_i$  lớp nhãn (-)
1.4   If  $W(s_i) > W'(s_i)$  then
1.4.1    $T = T \cup W'(s_i)$ ;
1.5    $m=0$ ;

```

Hình 2.9 Thuật toán trích rút câu

Trong thuật toán tại hình 2.8, có sử dụng một số hàm như hàm *Split*() tách các câu từ trong văn bản gốc. Hàm *Pos*() lấy vị trí của câu trong văn bản, hàm *match\_hush*() là hàm đối sánh để lấy giá trị từ trong cơ sở dữ liệu đã được lưu trữ dưới dạng bảng băm làm giảm thời gian xử lý dữ liệu của máy tính.

Trong chương này luận văn đã trình bày chi tiết phương pháp tóm tắt văn bản tiếng Việt dựa trên lý thuyết Naïve Bayes và giảm chiều đặc trưng bằng cách trích chọn các danh từ để tính toán và xử lý phân loại thành tập câu được lựa chọn. Trong phần huấn luyện có sự tham gia của con người ở quá trình gán nhãn dữ liệu huấn luyện, do đó phương pháp này đảm bảo hiệu quả hơn, chất lượng hơn so với các phương pháp học không giám sát đã được đề xuất trước đó. Đồng thời đưa ra thuật toán tính trọng số của câu và thuật toán trích rút câu trong quá sử dụng trong quá trình huấn luyện và quá trình tóm tắt văn bản.

### **Chương 3. XÂY DỰNG VÀ CÀI ĐẶT HỆ THỐNG TÓM TẮT VĂN BẢN TIẾNG VIỆT DỰA TRÊN LÝ THUYẾT NAÏVE BAYES**

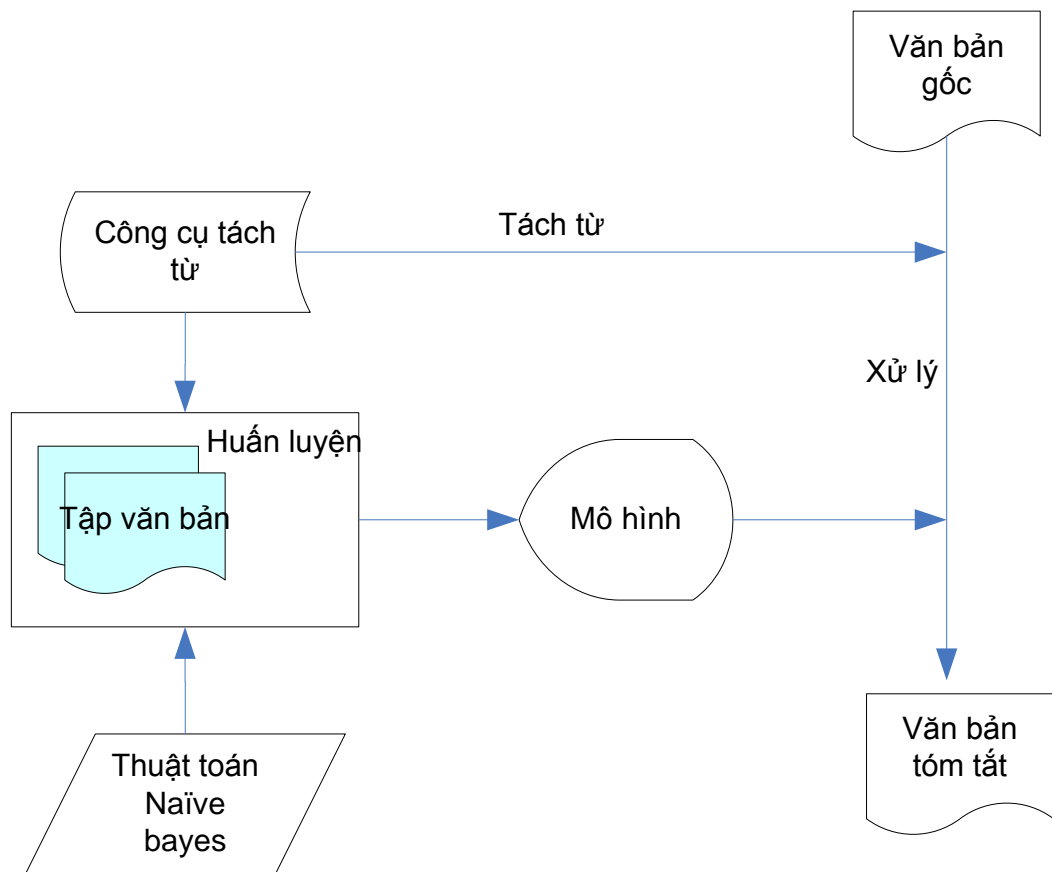
Trong chương này, luận văn trình bày phương pháp phân tích và xây dựng hệ thống thử nghiệm tóm tắt văn bản tiếng Việt dựa trên lý thuyết Naïve Bayes. Hệ thống được xây dựng trực tuyến, tự động lấy dữ liệu phục vụ quá trình học và tóm tắt văn bản tiếng Việt tự động hiển thị trên trang của người dùng.

#### ***3.1 Mô hình hệ thống tóm tắt văn bản tiếng Việt dựa trên lý thuyết Naïve Bayes***

Một cách tổng quan nhất khi làm việc với văn bản tiếng Việt, các nhà nghiên cứu thường sử dụng công cụ tách từ để khai thác các đặc trưng trong văn bản tiếng Việt, do vậy trong các mô hình tóm tắt văn bản tiếng Việt thông thường, các công cụ tách từ thường được sử dụng ở cả hai pha: pha huấn luyện và pha tóm tắt.

Trong pha huấn luyện: Công cụ tách từ tách các từ trong tập văn bản huấn luyện và tính tần suất xuất hiện các từ đó trong văn bản.

Trong pha tóm tắt: Văn bản gốc phải sử dụng công cụ tách từ để tách từ, tính tần suất xuất hiện các từ trong câu để đối sánh qua mô hình và lựa chọn các câu quan trọng để trích rút.



*Hình 3.1. Mô hình tóm tắt văn bản thông thường*

Do đặc thù của tiếng Việt, công cụ tách từ thường được sử dụng trong cả hai pha của hệ thống tóm tắt. Trong quá trình tóm tắt, khi sử dụng công cụ tách từ một văn bản đầu vào phải thông qua ba bước:

Tiền xử lý văn bản, loại bỏ các từ dừng, từ nhiễu như là một phương pháp giảm chiều đặc trưng.

Sử dụng công cụ tách từ.

Áp dụng mô hình học Naïve Bayes và sinh ra văn bản tóm tắt.

Đối với các mô hình tóm tắt dựa trên ngôn ngữ đa âm tiết như tiếng Anh, tiếng Pháp và một số ngôn ngữ khác trên thế giới, bước thứ hai trên

thường không sử dụng, do đó hệ thống tóm tắt văn bản bằng ngôn ngữ đa âm tiết thường có tốc độ nhanh hơn.

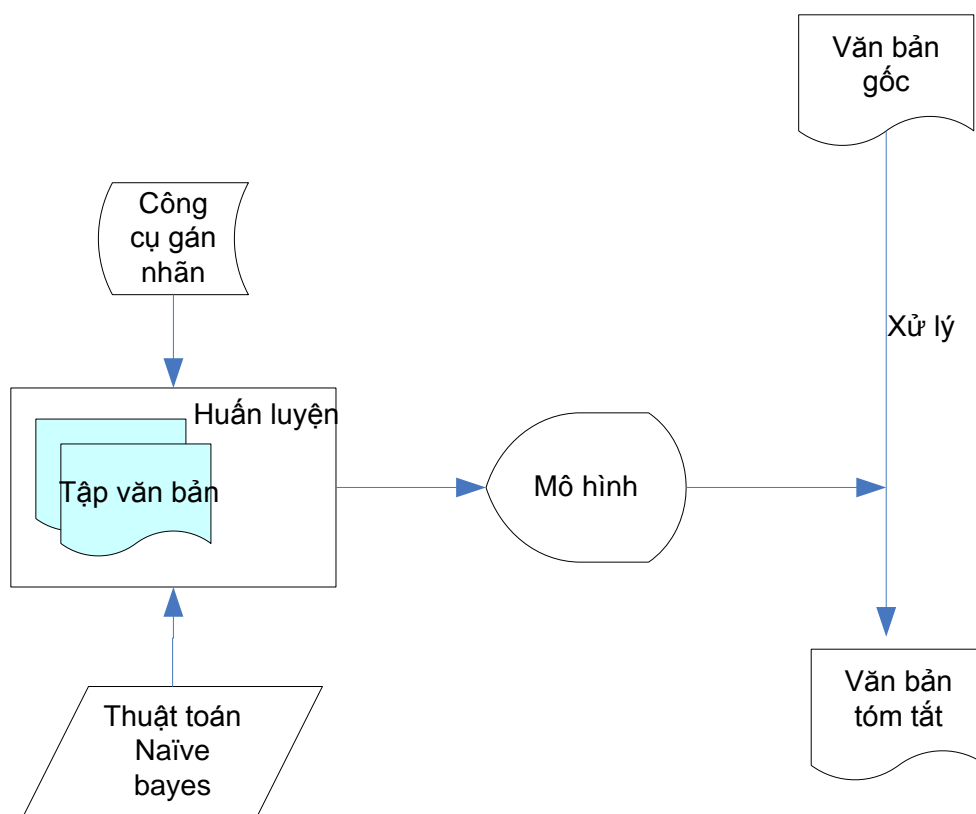
Để giảm bớt một bước xử lý đối với ngôn ngữ đơn âm tiết, cụ thể là ngôn ngữ tiếng Việt, luận văn tìm hiểu phương pháp đã được đề xuất tại [10] để xây dựng hệ thống. Hình dưới đây mô tả phương pháp tóm tắt văn bản tiếng Việt bằng phương pháp Naïve Bayes sử dụng công cụ gán nhãn từ loại tiếng Việt thay thế cho công cụ tách từ tiếng Việt. Sử dụng phương pháp này có thuận lợi sau:

Công cụ gán nhãn từ loại nhận diện ra danh từ trong tập huấn luyện và chỉ sử dụng tập danh từ để xử lý từ là một cách giảm chiều đặc trưng sẽ được mô tả ở mục 2.2.1

Các danh từ được tách ra trong quá trình huấn luyện được lưu trữ tại cơ sở dữ liệu của hệ thống.

Trong quá trình tóm tắt, văn bản gốc được đối sánh với tập danh từ đã được lưu trữ tại cơ sở dữ liệu để xử lý.

Với cách tiếp cận này, quá trình tóm tắt sẽ loại bỏ được quy trình tách từ và xử lý từ, do vậy tốc độ tính toán của hệ thống sẽ nhanh hơn.



Hình 3.2. Mô hình tóm tắt văn bản trong luận văn đề xuất

### 3.1.1 Lựa chọn ngôn ngữ lập trình và yêu cầu của hệ thống

#### 3.1.1.1 Lựa chọn ngôn ngữ lập trình

Hiện nay, hầu hết các hệ điều hành trên máy tính xách tay, máy tính để bàn đều sử dụng hệ điều hành Microsoft Windows. Do đó, để tránh trường hợp xung đột với hệ thống, nên luận văn lựa chọn ngôn ngữ lập trình C# 2012 và thiết kế cơ sở dữ liệu của hệ thống bằng hệ quản trị cơ sở dữ liệu SQL Server.

Microsoft Visual Studio 2012 là ngôn ngữ hoàn thiện và hoạt động theo hướng đối tượng, đây cũng là ngôn ngữ lập trình thông dụng trên Windows, hỗ trợ quản lý cơ sở dữ liệu, lập trình internet. Chương trình có nhiều tính năng

mới, các điều khiển mới cho phép ta viết chương trình ứng dụng kết hợp các giao diện, ngoài ra sử dụng chương trình Microsoft Visual Studio 2012 sẽ tiết kiệm được thời gian và công sức so với các chương trình khác. Bên cạnh đó Microsoft Visual Studio 2012 còn hỗ trợ tính năng kết nối môi trường dữ liệu SQL, việc liên kết có thể thực hiện bằng nhiều cách.

Hệ quản trị cơ sở dữ liệu SQL Server 2008 là phần mềm tương tác với người sử dụng chạy trên môi trường Windows, nó tăng thêm sức mạnh trong công tác tổ chức và tìm kiếm thông tin, các công tác kiểm tra dữ liệu, giá trị mặc định, khuôn nhập dữ liệu của chương trình hoàn toàn đáp ứng yêu cầu. Quản lý được khối lượng dữ liệu lớn và tần suất truy cập cao, đáp ứng dịch vụ trực tuyến và đảm bảo yêu cầu về an toàn dữ liệu. Chính vì lẽ đó mà luận văn chọn sử dụng ngôn ngữ lập trình Microsoft Visual Studio 2012 và hệ quản trị cơ sở dữ liệu SQL Server 2008.

#### **3.1.1.2 Yêu cầu của hệ thống**

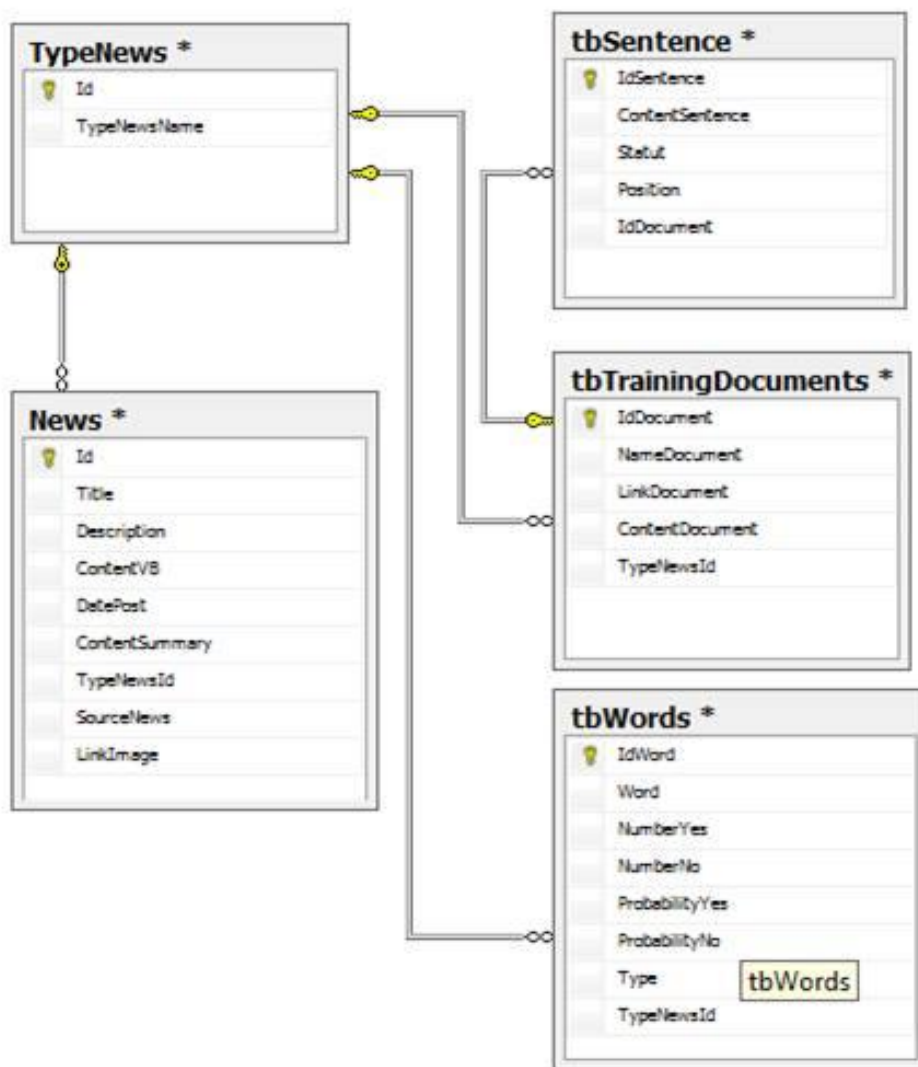
Hệ thống gồm hai pha riêng biệt, pha huấn luyện và pha tóm tắt. Trong pha huấn luyện, các văn bản được tải về từ nguồn dữ liệu internet, tự động loại các thẻ html, các hình ảnh và lưu trữ trong hệ thống dưới dạng đường dẫn. Chương trình cho phép quản lý và lưu trữ các văn bản huấn luyện đồng thời tính xác suất chọn và xác suất không chọn cho các từ quan trọng và lưu trữ để sử dụng trong quá trình tóm tắt. Các văn bản tóm tắt có thể được lưu lại và quản lý. Ngoài việc tóm tắt văn bản được tải từ internet thì chương trình cho phép được tóm tắt các văn bản sẵn có.

#### **3.1.1.3 Cơ sở dữ liệu của hệ thống**

Cơ sở dữ liệu của hệ thống mô tả các thông tin lưu trữ của hệ thống trong cơ sở dữ liệu. Bao gồm thông tin:



- Bảng “*tbSentence*” lưu các câu được tách trong văn bản huấn luyện.
- Bảng “*TypeNews*” lưu các thể loại tin tức( công nghệ thông tin, thể thao, xã hội...).
- Bảng “*Tranning documents*” lưu các văn bản huấn luyện.
- Bảng “*tbWord*” lưu các từ quan trọng.



Hình 3.3 Cơ sở dữ liệu của hệ thống.

#### 3.1.1.4 Các chức năng chính của hệ thống

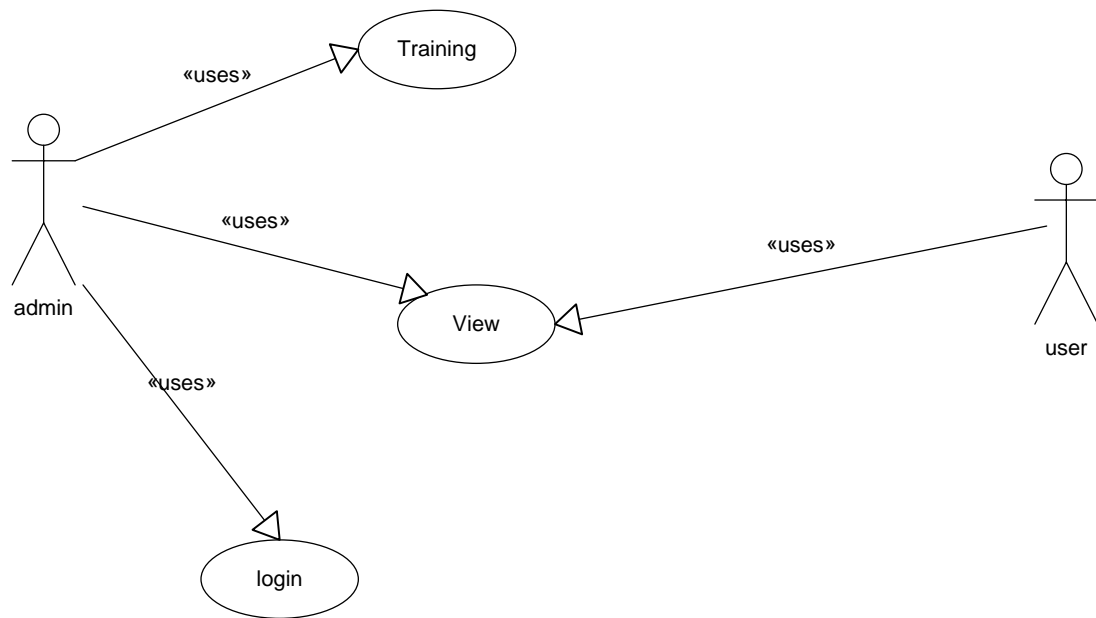
- Chức năng thêm văn bản huấn luyện.
- Chức năng thêm từ mới.
- Chức năng xóa văn bản huấn luyện.
- Chức năng cập nhật lại văn bản huấn luyện.
- Module tách từ: tích hợp từ công cụ Vntagger.
- Module xử lý tách câu trong văn bản.

#### 3.1.1.5 Tập từ điển danh từ

Hệ thống sử dụng công cụ Vntagger được tải về từ trang web vlsp do nhóm tác giả của đề tài KC01 thực hiện. Công cụ này được tích hợp vào hệ thống để tự động lọc ra các danh từ trong tập văn bản huấn luyện rồi lưu trữ vào cơ sở dữ liệu của hệ thống.

### 3.2 Phân tích thiết kế hệ thống tóm tắt văn bản tiếng Việt dựa trên Naïve Bayes

Dựa trên các chức năng được mô tả, hệ thống được phân tích thành sơ đồ tổng quát như sau

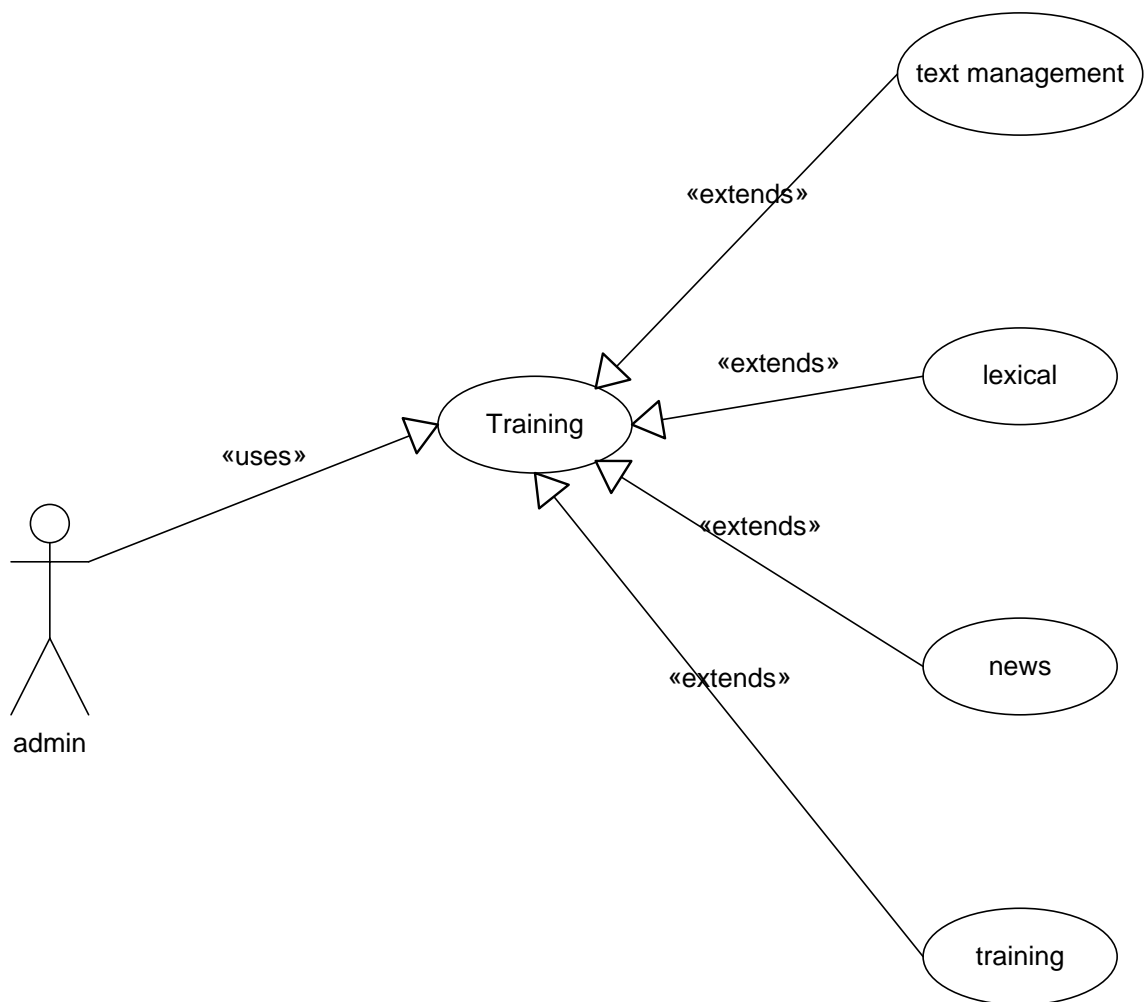


*Hình 3.4 Sơ đồ usecase tổng quát.*

Trong sơ đồ trên, hệ thống gồm hai cấp người dùng chủ yếu: người quản trị và người sử dụng hệ thống (người đọc tin tức).

- Chức năng quản trị:

- Login: chức năng đăng nhập vào hệ thống
- View: Chức năng xem tin tức
- Training: Chức năng huấn luyện, bao gồm các tính năng quản lý văn bản, quản lý từ điển, học từ tập văn bản,...



Hình 3.5. Usecase trường hợp huấn luyện.

### 3.3 Một số giao diện của hệ thống tóm tắt văn bản tiếng Việt dựa trên Naïve Bayes

#### 3.3.1 Giao diện trang chủ hệ thống tóm tắt văn bản tiếng Việt

Hệ thống được chạy trên địa chỉ máy chủ ảo localhost. Tuy nhiên có thể tự động cập nhật dữ liệu từ các trang Internet nếu được kết nối mạng. Hình 3.3 dưới đây mô tả giao diện trang chủ của hệ thống



*Hình 3.6. Giao diện trang chủ của hệ thống*

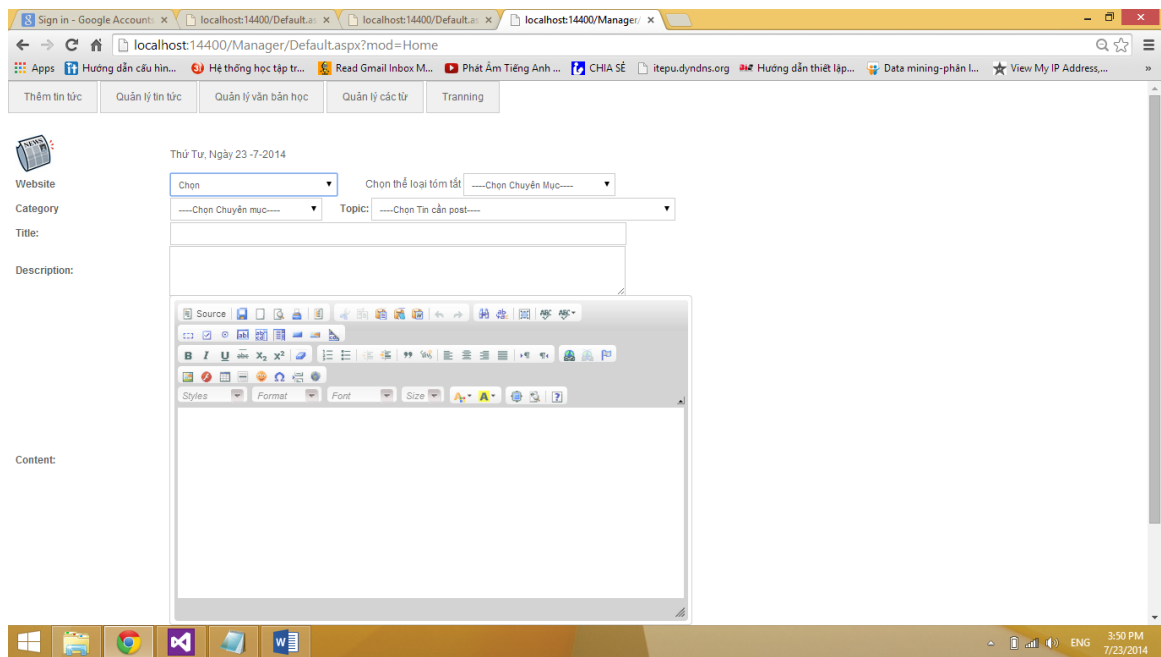
Trong giao diện trang chủ, nguồn tin tức được tổng hợp về từ một số website như: <http://24h.com.vn>, <http://vnexpress.net>, <http://dantri.com>, <http://thanhnien.com.vn>, ... Nguồn tin được lấy về từ trang web nào đều được hiển thị nguồn tại phần tiêu đề cuối của tin tức đó.

### 3.3.2 Giao diện trang quản trị hệ thống tóm tắt văn bản tiếng Việt

Phần quản trị của hệ thống gồm có các chức năng sau:

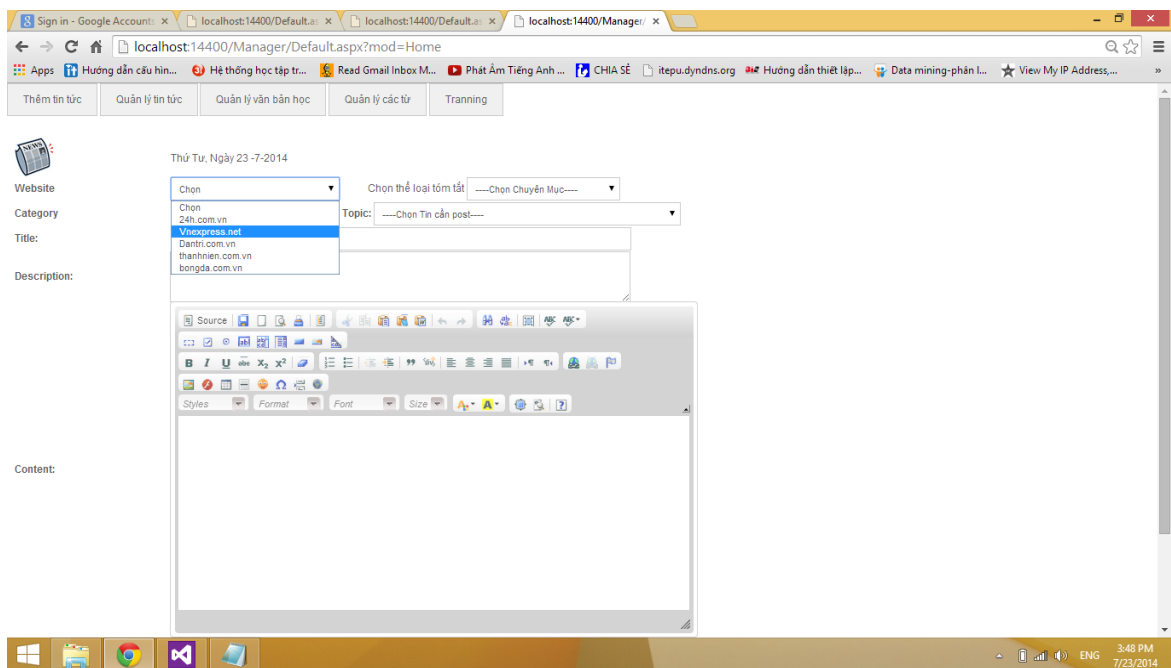
- Thêm tin tức tại trang chủ của hệ thống
- Quản lý mục tin tức
- Quản lý văn bản huấn luyện
- Quản lý từ điển
- Huấn luyện
- Tóm tắt văn bản

Hình 3.4 dưới đây là giao diện trang quản trị



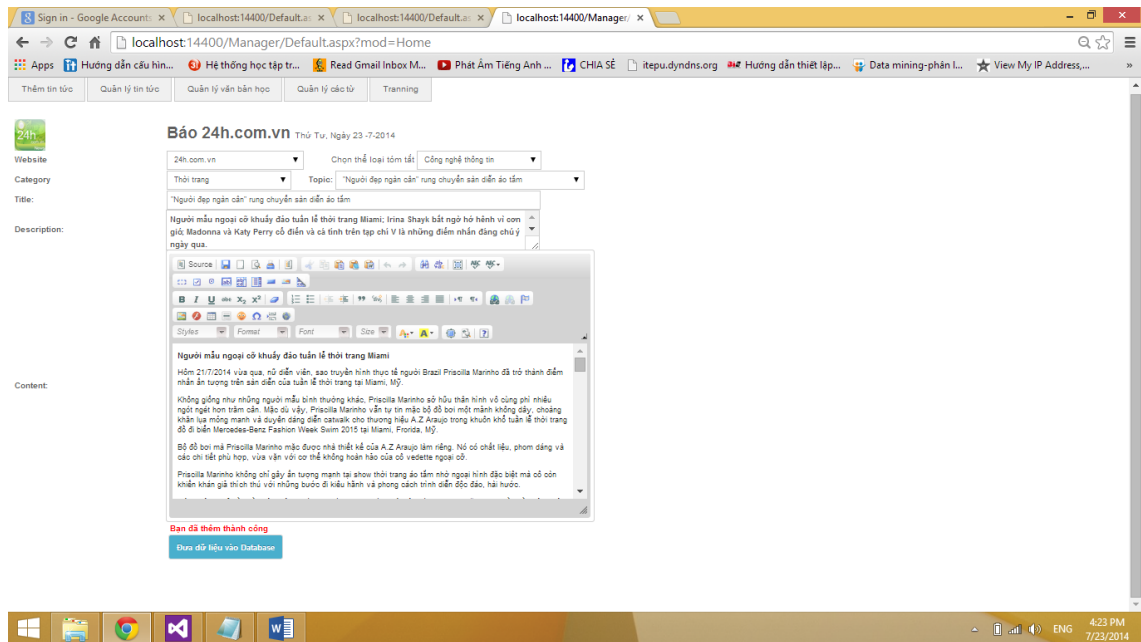
Hình 3.7 Giao diện chính của trang quản trị.

Đối với phần tính năng thêm tin tức, các tin tức được lấy tự động từ một số các trang web đã được hệ thống mặc định thông qua RSS.



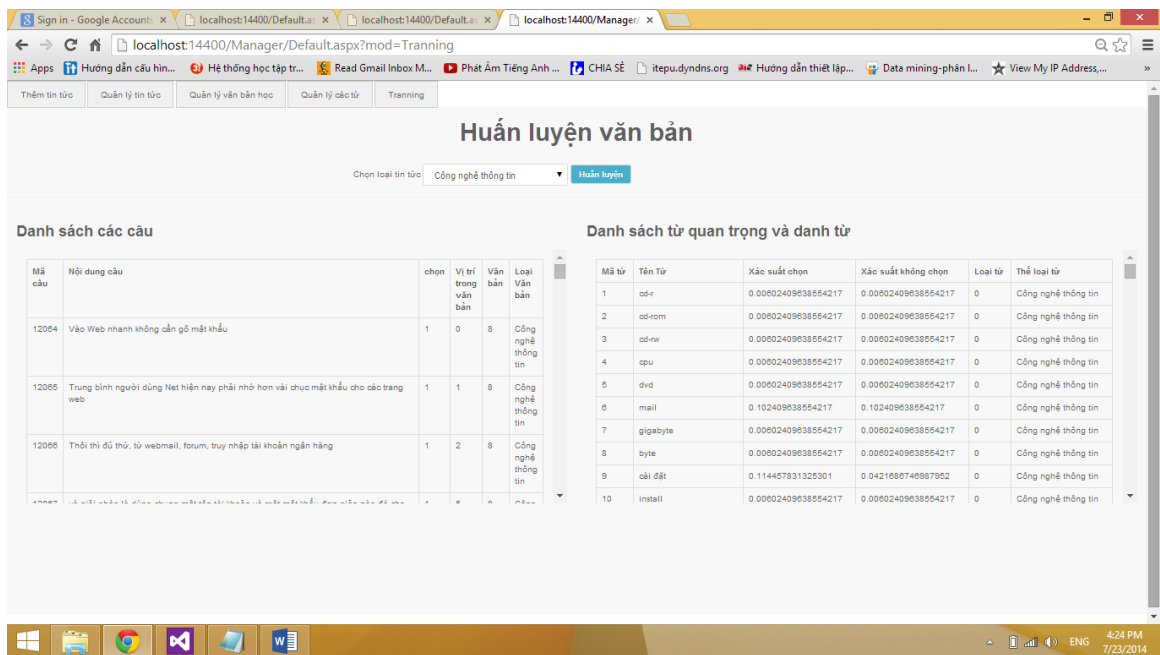
Hình 3.8 Lấy tin tự động.

Hình 3.6 trên đây là giao diện lựa chọn tin tức lấy về. Tin tức được lựa chọn nguồn tin, thể loại tin. Sau khi lựa chọn, nội dung chính của tin được thể hiện trên giao diện lựa chọn tin của hệ thống. Tiếp theo, các tin tức có thể được lưu tại cơ sở dữ liệu của hệ thống sau khi thực hiện thao tác “Đưa dữ liệu vào database”.



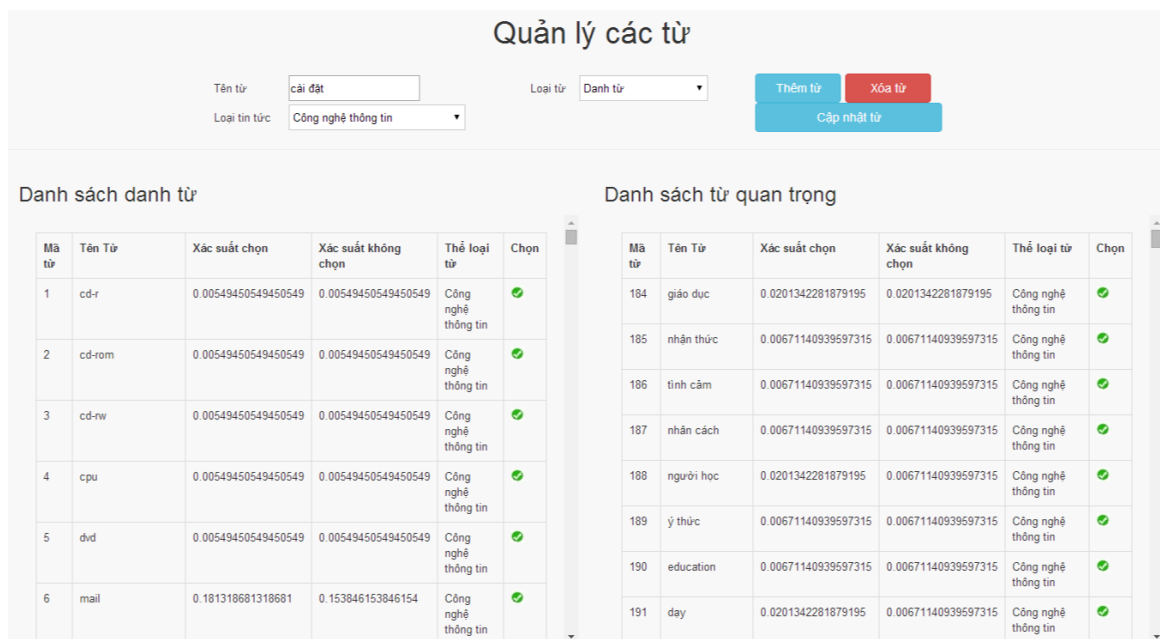
Hình 3.9 Giao diện hiển thị dữ liệu lấy về.

Chức năng huấn luyện văn bản được thể hiện trong hình 3.7. Các văn bản được tách thành các câu và tính xác suất dựa trên lý thuyết Naïve Bayes.



Hình 3.10 Giao diện huấn luyện văn bản.

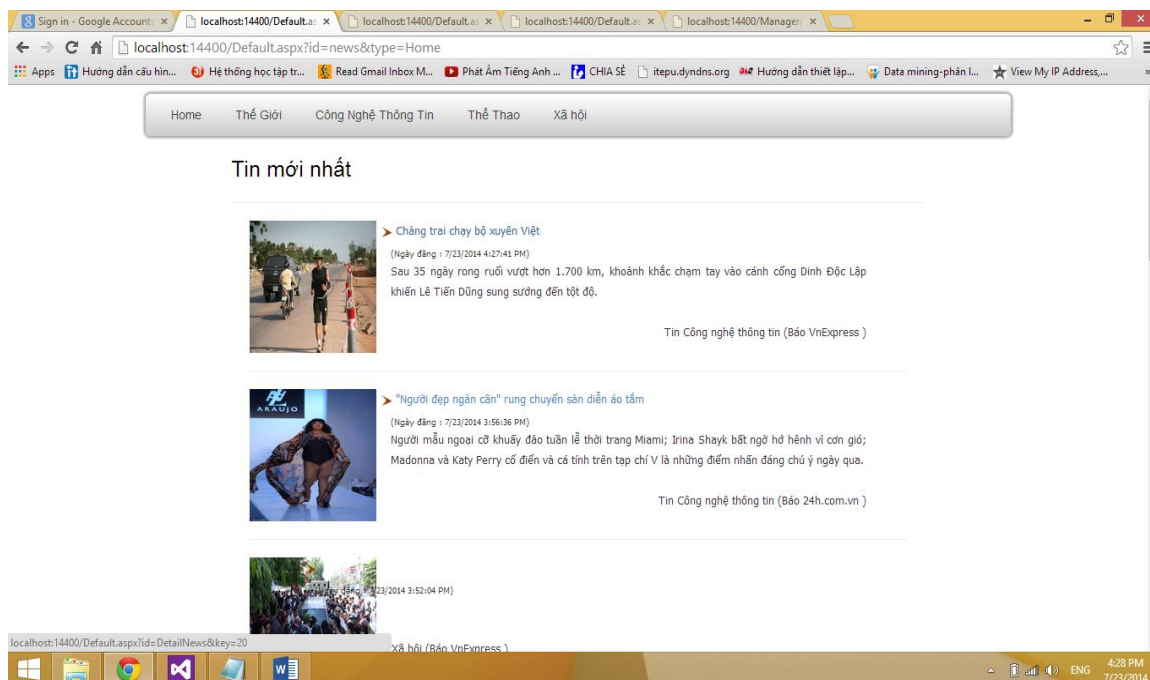
Danh sách các từ quan trọng cũng được chương trình quản lý và tính xác suất chọn và xác suất không chọn theo từng thể loại lĩnh vực khác nhau được thể hiện trong hình 3.9 sau



Hình 3.11 Giao diện quản lý từ.

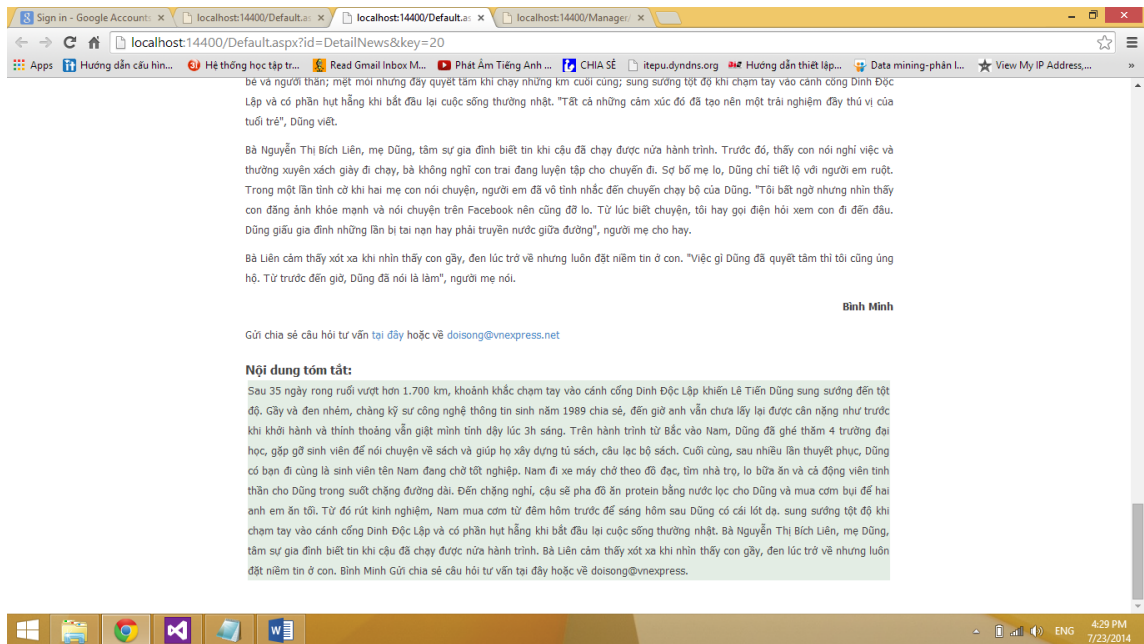


Tin tức sau khi cập nhật được hiển thị ra ngoài trang chủ của hệ thống, người dùng được quyền truy cập hệ thống để xem các thông tin dưới dạng tổng hợp từ một số nguồn dữ liệu khác nhau.



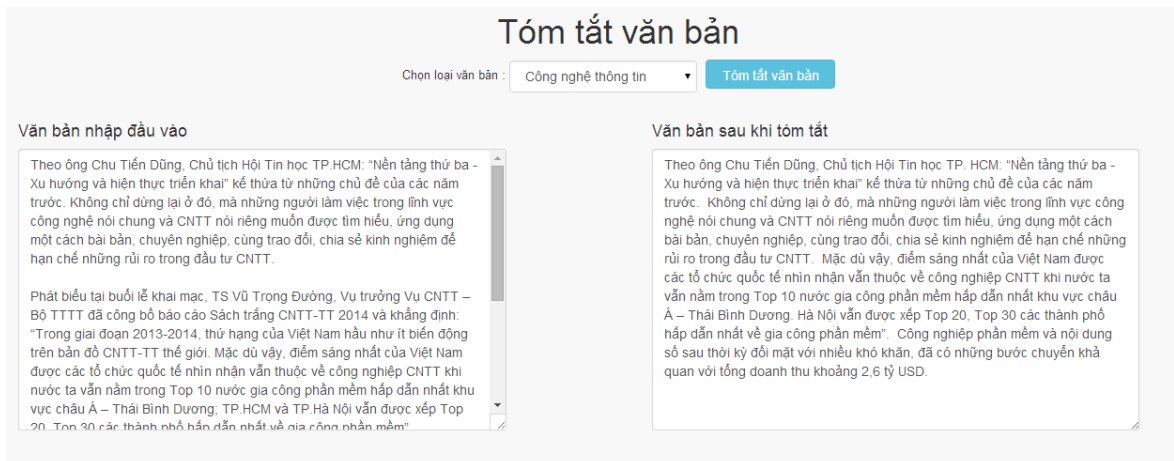
*Hình 3.12 Hiển thị tin tức sau khi cập nhật.*

Sau khi người dùng lựa chọn mục tin tức cần đọc. Nội dung tin được hiển thị dưới dạng full text (bản đầy đủ) và bản short text (văn bản tóm tắt) như hình 3.9



Hình 3.13 Giao diện tóm tắt tin tức.

Ngoài việc tóm tắt văn bản trên các trang web khi được nối mạng thì hệ thống cho phép tóm tắt những văn bản có sẵn, dưới đây là giao diện thể hiện chức năng này.



Hình 3.14 Giao diện tóm tắt văn bản

### 3.4 Kết quả thực nghiệm phương pháp tóm tắt văn bản tiếng Việt dựa trên *Naïve Bayes*

#### 3.4.1 Xây dựng tập dữ liệu phục vụ huấn luyện

Các nghiên cứu trước đây thường làm việc với tập dữ liệu đã qua tiền xử lý, do đó, thời gian chuẩn bị dữ liệu thường được làm bằng cách thủ công, mất thời gian và chi phí lớn, hơn nữa khó khăn khi bổ sung học tăng cường cho những hệ thống đòi hỏi phải cập nhật tri thức thường xuyên. Trong luận văn này, tập dữ liệu được sử dụng bằng cách tải về (download) tự động trên hệ thống và được lưu trữ vào cơ sở dữ liệu của hệ thống dưới dạng đường dẫn lưu văn bản. Các văn bản khi tải về được tự động loại bỏ các thẻ html và chỉ lấy phần nội dung của văn bản.

Các văn bản trên được lưu thành tập các văn bản phục vụ cho quá trình huấn luyện.

Gọi  $D = \{d_1, d_2, \dots, d_n\}$  là tập các văn bản huấn luyện. Tập các văn bản trên được tách thành các câu. Đối với mỗi văn bản  $d_s$  thuộc tập  $D$  tách thành các câu

$$S_{ds} = \{s_{1ds}, s_{2ds}, \dots, s_{kds}\}$$

Với mỗi câu  $s_{ids}$  được tính toán trọng số dựa trên 3 đặc trưng:

Độ quan trọng thông tin

Lượng thông tin trong câu

Vị trí câu trong văn bản.

Tiếp theo các câu được gán nhãn thủ công bằng cách dựa trên con người trích chọn ra các câu họ cho rằng có ý nghĩa trong văn bản và lưu vào tập (+). Các câu không được lựa chọn lưu vào tập (-).

Dữ liệu từ tập D gồm n văn bản sau quá trình chuẩn bị dữ liệu được gán nhãn thành hai tập con gồm các câu có nhãn (+) và các câu có nhãn (-).

### 3.4.2 Xây dựng bộ từ điển danh từ

Để tăng tốc cho hệ thống và quá trình xây dựng tập từ điển gồm các danh từ, luận văn đã sử dụng công cụ Vntagger được tải về từ trang web vlsp [15] và nhúng vào mã nguồn của chương trình thành bộ công cụ tích hợp của hệ thống.

### 3.4.3 Tiền xử lý và chuẩn hóa dữ liệu

Tập văn bản đầu vào là văn bản dạng thô, để đơn giản cho việc xử lý dữ liệu, với mỗi văn bản đầu vào, ta sẽ thực hiện qua bước tiền xử lý ký tự để đưa văn bản về dạng tiêu chuẩn. Ở đây tiêu chuẩn là tiêu mà trong đó không có 2 dấu cách nào liền nhau, có dấu câu khi kết thúc tiêu, trước dấu câu không có dấu cách. Để có được tiêu chuẩn, chuẩn bị cho việc tách từ, ta thực hiện qua các bước sau:

Chuyển hết các ký tự chữ hoa thành chữ thường.

Dùng các dấu câu (bao gồm dấu “.” “,” “:”...) để tách văn bản thành một tập hợp các câu. Ta có thể tách như vậy vì 2 âm tiết cách nhau bởi một dấu câu sẽ không bao giờ thuộc về cùng một từ

- Tiến hành chuẩn hoá với mỗi câu:
- Khi có >1 dấu cách đứng kề nhau, loại bớt đi, chỉ để lại một dấu cách loại bỏ những dấu cách ở đầu và cuối câu.

### 3.4.4 Đánh giá kết quả của hệ thống tóm tắt văn bản dựa trên Naïve Bayes

Luận văn sử dụng phương pháp đánh giá truyền thống là độ đo Precision để đánh giá chất lượng của tóm tắt, độ chính xác của hệ thống so với con người.

Để đánh giá với từng mức của tóm tắt, trong khi một số các hệ thống khác hoặc phương pháp khác như textcompactor [16], VTSONline [14], Le Thanh Ha [13] thường sử dụng tóm tắt theo tỉ lệ được định nghĩa như sau:

Tỉ lệ  $r = \text{chiều dài văn bản tóm tắt} / \text{chiều dài văn bản gốc} \%$ .

Kết quả được thể hiện như bảng sau

Phương pháp	Tỉ lệ			
	80%	60%	40%	20%
Luận văn	0.88	0.86	0.82	0.6
HLT	0.82	0.75	0.69	0.54
Baseline	0.81	0.8	0.84	0.63
Textcompactor	0.85	0.82	0.65	0.57
VTSONline	0.72	0.68	0.51	0.48

*Bảng 3.1. Bảng kết quả thực nghiệm*

Dựa vào bảng kết quả thực nghiệm trên thấy rằng, phương pháp luận văn sử dụng được cài đặt hiệu quả trên hệ thống thực có hiệu quả và gần với kết quả đánh giá của con người.

## KẾT LUẬN

Các phương pháp khai phá dữ liệu hiện nay ngày càng gần với yêu cầu của người dùng là mong muốn cho thông tin hữu ích nhất trong vô vàn lượng thông tin trên Internet. Trong đó, dữ liệu dạng văn bản chiếm tới trên 80% kho dữ liệu lớn đã có. Để khai phá hiệu quả thông tin này cần tới nhiều công cụ khác nhau để khai phá, trong đó có công cụ tóm tắt văn bản.

Trong luận văn này đã trình bày một phương pháp tóm tắt văn bản tiếng Việt dựa trên lý thuyết Naïve Bayes để phân lớp các câu có độ quan trọng so với tập dữ liệu đã được huấn luyện bởi người dùng cho chất lượng tóm tắt tốt hơn các phương pháp đã được đề xuất dựa trên cách tiếp cận học không giám sát.

Luận văn cũng đã xây dựng và cài đặt hệ thống chạy trên môi trường web, góp phần đưa những nghiên cứu gần hơn với thực tế và áp dụng trong thực tế với kết quả thử nghiệm chấp nhận được. Văn bản tóm tắt dễ đọc dễ hiểu và gần với kết quả tóm tắt của con người.

Dù đã hết sức cố gắng để hoàn thành luận văn và xây dựng hệ thống tóm tắt văn bản tiếng Việt tự động, tuy nhiên, do thời gian nghiên cứu có hạn nên không thể tránh khỏi những sai sót. Kính mong các thầy cô, đồng nghiệp, bạn bè đóng góp để luận văn hoàn thiện hơn.

***Trân trọng cảm ơn!***

## TÀI LIỆU THAM KHẢO

### Tiếng Việt

- [1]. Phạm Công Cảnh, *Phương pháp rút gọn câu tiếng Việt dựa trên mạng Bayesian*, luận văn thạc sĩ, Học viện kỹ thuật quân sự, 2014.
- [2]. La Đức Dũng, *Khai phá dữ liệu văn bản bằng công cụ tập thô*, luận văn thạc sĩ, Đại học công nghệ thông tin và truyền thông Thái Nguyên, 2012.
- [3]. Lê Mạnh Hùng, *Tra cứu văn bản tiếng Việt dựa trên mô hình phân cụm phân cấp*, luận văn thạc sĩ, học viện bưu chính viễn thông, 2013.
- [4]. Lương Chi Mai (2009), *Nghiên cứu phát triển một số sản phẩm thiết yếu về xử lý tiếng nói và văn bản tiếng Việt*, Chương trình KH&CN cấp nhà nước KC01/06-10, Đề tài KC01/06-10.
- [5]. Hoàng Tất Thắng, Nguyễn Thị Bạch Nhạn, Nguyễn Quốc Dũng Lê Thị Hoài Nam, Trần Thị Quỳnh Nga, *Tài liệu hướng dẫn ôn tập và thi tốt nghiệp môn Tiếng Việt và phương pháp giảng dạy tiếng Việt ở Tiểu học*, trường đại học Huế, 2013
- [6]. Nguyễn Thị Ngọc Tú, *Tóm tắt văn bản tiếng Việt dựa trên mạng nơ ron*, luận văn thạc sĩ, đại học công nghệ thông tin – đại học quốc gia thành phố Hồ Chí Minh, 2014.

### Tiếng Anh

- [7]. Chin-Yew Lin, Eduard Hovy (2003/5/27), *Automatic evaluation of summaries using n-gram co-occurrence statistics*, In Proceedings of the Human Technology Conference, Association for Computational Linguistics Volume 1, 71-78.
- [8]. Clarke, J., & Lapata, M. (2008), *Global inference for sentence compression: An integer linear programming approach*, Journal of Artificial Intelligence Research, 31, 399-429.

- [9]. Dipanjan Das and Andre F.T. Martins (2007), *A Survey on Automatic Text Summarization*, Language Technologies Institute, Carnegie Mellon University.
- [10]. Ha N.T.T, *An optimization text summarization method based on Naïve Bayes and topic word for Single syllable Language*, Applied Mathematical Sciences, Vol 8, No 3, pp 99-115, 2014.
- [11]. Hovy, E. and Lin, C. , *Automated text summarization and the summarist system*, TIPSTER '98 Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998, pp.197–214, 1998.
- [12]. Knight, K., & Marcu, D. (2002), *Summarization beyond sentence extraction: a probabilistic approach to sentence compression*, Artificial Intelligence, 139 (1), 91-107.
- [13]. Thanh, Le Ha; Quyet, Thang Huynh; Chi, Mai Luong, *A Primary Study on Summarization of Documents in Vietnamese*, Proceedings of the First World Congress of the International Federation for Systems Research : The New Roles of Systems Sciences For a Knowledge-based Society 2005-11.

## Website

- [14]. <http://labs.baomoi.com/demoTS.aspx>
- [15]. <http://vlsp.vietlp.org:8080/demo/>
- [16]. <http://www.textcompactor.com/>
- [17]. <http://www.tools4noobs.com/summarize>

## PHỤ LỤC

### 1. Phần học văn bản

```
public void Trainings(string typenewsId)
{
```



```

DataTable topword = dt.GetDataTable("Select * From tbWords where
Type='0' and TypeNewsId='"+int.Parse(tynewsId)+"'");
    if (topword != null)
    {
        foreach (DataRow row in topword.Rows)
        {
            row[2] = 0;
            row[3] = 0;
            row[4] = 0;
            row[5] = 0;
            row[6] = 0;
        }
    }
    string s1;
    int tong = 0;
    int dem = 0;
    for (int i = 0; i < 2; i++)
    {
        string sentence = st.GetSentence(i +
"",tynewsId);
        if (topword != null)
        {
            foreach (DataRow row in topword.Rows)
            {
                s1 = Convert.ToString(row[1]);
                Regex thegex = new Regex(s1.ToLower());

                MatchCollection theMatches =
thegex.Matches(sentence);

```

```

int str = 0;
foreach (Match theMatch in theMatches)
{
    str++;
}

if (str > 0)
{
    dem++;
    tong = Convert.ToInt32(row[i + 2]) +
Convert.ToInt32(str.ToString());
    row[i + 2] =
Convert.ToInt32(str.ToString());
}
}
}
double xs = 0.0;
for (int i = 0; i < 2; i++)
{
    foreach (DataRow row1 in topword.Rows)
    {
        if (tong > 0)
        {
            xs = (Convert.ToDouble(row1[i + 2]) + 1) /
(Convert.ToDouble(dem) + Convert.ToDouble(tong));
        }
        else
            xs = 0.0;
    }
}

```

```

        row1[i + 4] = xs.ToString();
    }
}
tw.Update(topword, typenewsId);
DataTable noun = dt.GetDataTable("Select * From
tbWords where Type='1' and
TypeNewsId='"+int.Parse(typenewsId)+"'");
if (noun != null)
{
    foreach (DataRow row in noun.Rows)
    {
        row[2] = 0;
        row[3] = 0;
        row[4] = 0;
        row[5] = 0;
        row[6] = 1;
    }
}

string s2;
int tong2 = 0;
int dem2 = 0;
for (int i = 0; i < 2; i++)
{
    string sentence = st.GetSentence(i +
"", typenewsId);
    if (noun != null)
    {
        foreach (DataRow row in noun.Rows)

```

```

        {
            s2 = Convert.ToString(row[1]);
            Regex thegex = new Regex(s2.ToLower());

            MatchCollection theMatches =
thegex.Matches(sentence);

            int str = 0;
            foreach (Match theMatch in theMatches)
            {
                str++;
            }
            if (str > 0)
            {
                dem2++;
                tong2 = Convert.ToInt32(row[i + 2]) +
Convert.ToInt32(str.ToString());
                row[i + 2] =
Convert.ToInt32(str.ToString());
            }
        }
    }
    double xs2 = 0.0;
    for (int i = 0; i < 2; i++)
    {
        foreach (DataRow row1 in noun.Rows)
        {
            if (tong2 > 0)

```

```

        {
            xs2 = (Convert.ToDouble(row1[i + 2]) + 1)
/ (Convert.ToDouble(dem2) + Convert.ToDouble(tong2));
        }
        else
            xs2 = 0.0;
            row1[i + 4] = xs2.ToString();
        }
    }
    tw.Update(noun, typenewsId);
}

```

## 2. Phần Tóm tắt

```

public string SummaryText(string text, string typenewsId)
{
    DataTable topword = dt.GetDataTable("Select * From
tbWords where Type='0' and
TypeNewsId='" + int.Parse(typenewsId) + "'");
    DataTable noun = dt.GetDataTable("Select * From
tbWords where Type='1' and TypeNewsId='" + int.Parse(typenewsId) +
"'");
    string output = "";
    string[] input = text.Split('.', '?', '!', ';');
    for (int i = 0; i < input.Length; i++)
    {
        string s1 = "";
        string s2 = "";
        double ProbabilityYes = 0;
        double ProbabilityNo = 0;
    }
}

```

```

        if (topword != null)
        {
            foreach (DataRow row in topword.Rows)
            {
                s1 =
Convert.ToString(row[1]);
                Regex thegex = new Regex(s1);
                MatchCollection theMatches =
thegex.Matches(input[i]);
                int s = 0;
                foreach (Match theMatch in theMatches)
                {
                    ProbabilityYes +=
Convert.ToDouble(row[4].ToString());
                    ProbabilityNo +=
Convert.ToDouble(row[5].ToString());
                }
            }
        }
        if (noun != null)
        {
            foreach (DataRow row in noun.Rows)
            {
                s2 = Convert.ToString(row[1]);
                Regex thegex = new Regex(s2);
                MatchCollection theMatches =
thegex.Matches(input[i]);
                int s = 0;

```

```

        foreach (Match theMatch in theMatches)
        {
            ProbabilityYes +=
Convert.ToDouble(row[4].ToString());
            ProbabilityNo +=
Convert.ToDouble(row[5].ToString());
        }
    }
    if (ProbabilityYes > ProbabilityNo)
    {
        output += input[i] + ". ";
    }
}
return output;
}

```