

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**---------------**

**BẠCH QUỐC HUY**

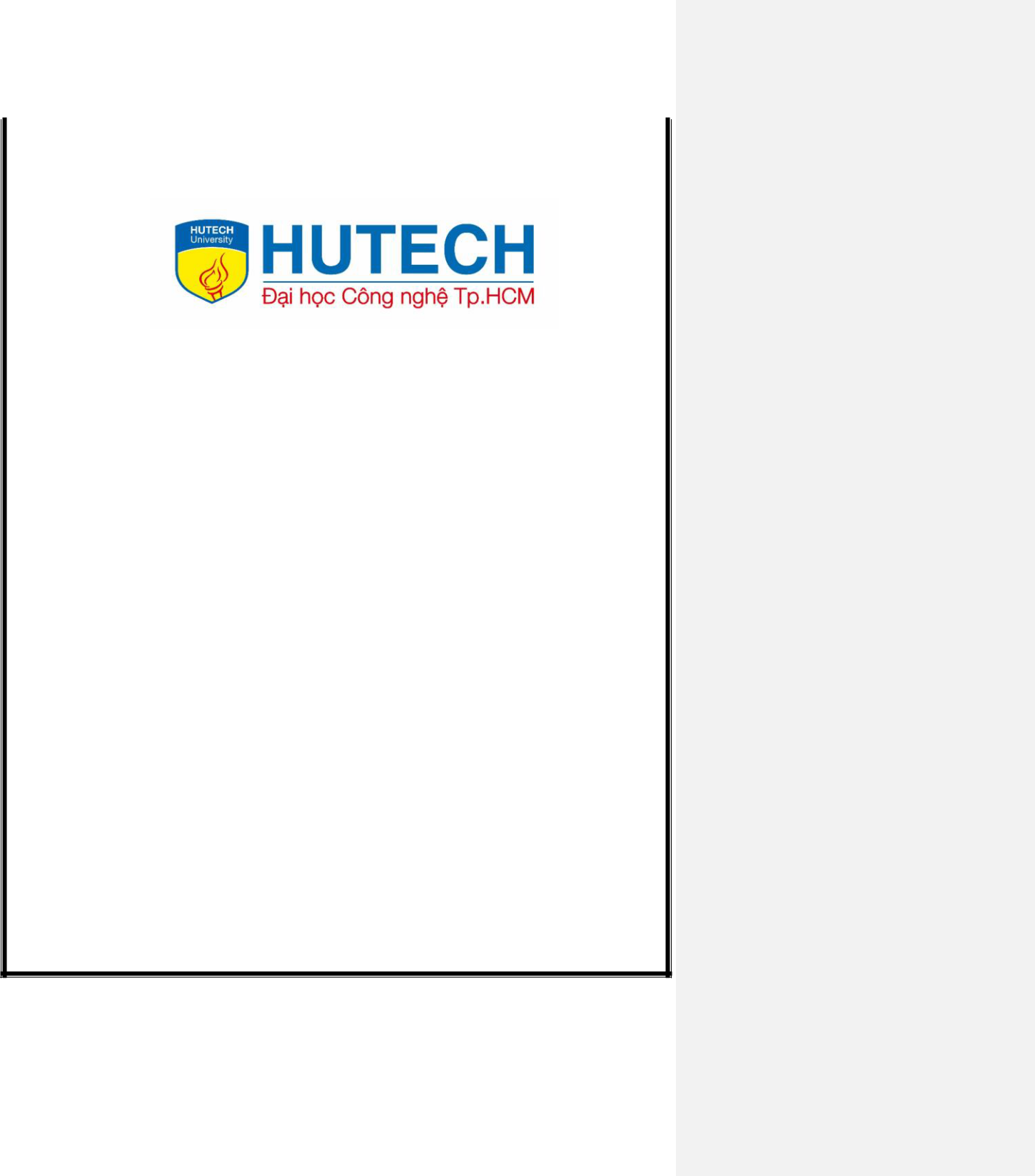
**ÁP DỤNG MÔ HÌNH ĐỒ THỊ VÀO BÀI TOÁN TÓM TẮT VĂN BẢN TIẾNG VIỆT**

**LUẬN VĂN THẠC SỸ**

Chuyên ngành: Công nghệ thông tin

Mã ngành: 60480201

TP. HCM, tháng 8/2018



**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**---------------**

**BẠCH QUỐC HUY**

**ÁP DỤNG MÔ HÌNH ĐỒ THỊ VÀO BÀI TOÁN TÓM TẮT VĂN BẢN TIẾNG VIỆT**

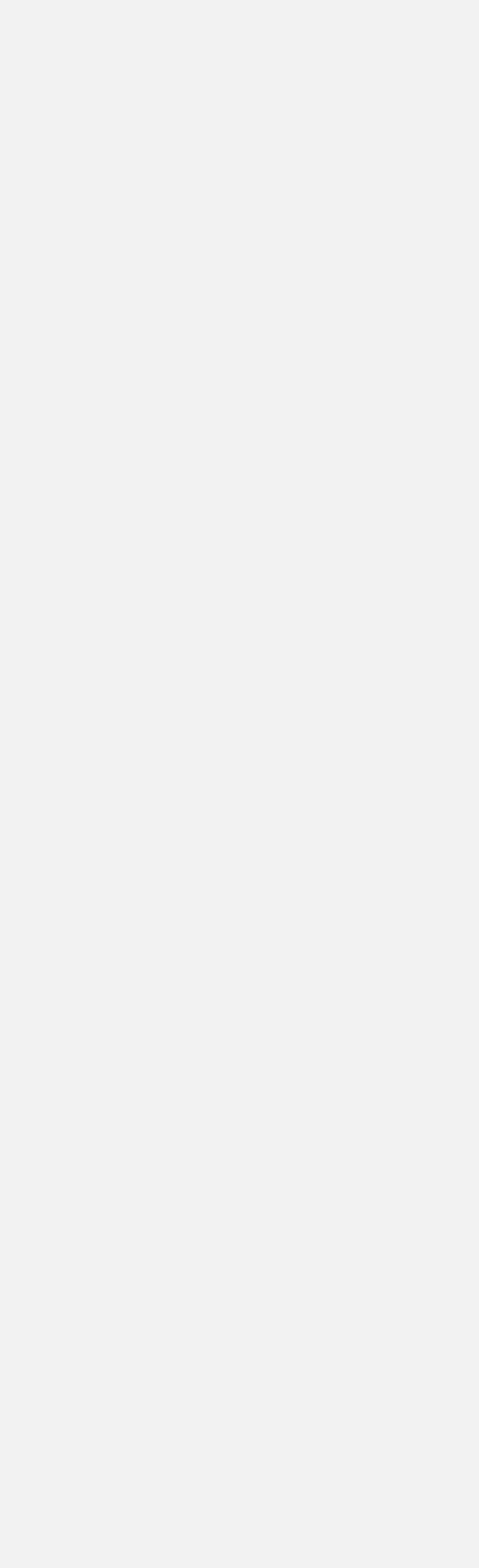
**LUẬN VĂN THẠC SỸ**

Chuyên ngành: Công nghệ thông tin

Mã ngành: 60480201

**CÁN BỘ HƯỚNG DẪN KHOA HỌC: TS. LÊ THỊ NGỌC THƠ**

TP. HCM, tháng 8/2018



CÔNG TRÌNH ĐƯỢC HOÀN THÀNH TẠI

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM**

Cán bộ hướng dẫn khoa học : TS. Lê Thị Ngọc Thơ

*(Ghi rõ họ, tên, học hàm, học vị và chữ ký)*

Luận văn Thạc sĩ được bảo vệ tại Trường Đại học Công nghệ TP. HCM ngày … tháng … năm 2018

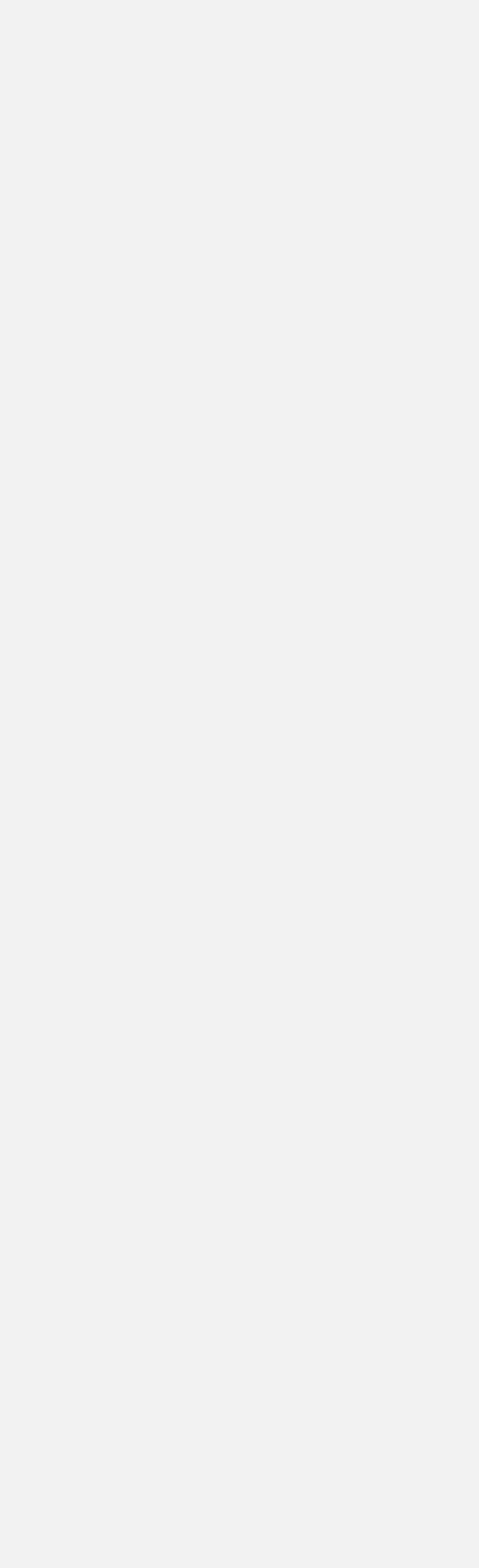
Thành phần Hội đồng đánh giá Luận văn Thạc sĩ gồm:

*(Ghi rõ họ, tên, học hàm, học vị của Hội đồng chấm bảo vệ Luận văn Thạc sĩ)*

|  |  |  |
| --- | --- | --- |
| **TT** | **Họ và tên** | **Chức danh Hội đồng** |
|  |  |  |
| 1 |  | Chủ tịch |
|  |  |  |
| 2 |  | Phản biện 1 |
|  |  |  |
| 3 |  | Phản biện 2 |
|  |  |  |
| 4 |  | Ủy viên |
|  |  |  |
| 5 |  | Ủy viên, Thư ký |
|  |  |  |

Xác nhận của Chủ tịch Hội đồng đánh giá Luận sau khi Luận văn đã được sửa chữa (nếu có).

Chủ tịch Hội đồng đánh giá LV



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TRƯỜNG ĐH CÔNG NGHỆ TP. HCM | | | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM** | | |  |
| **VIỆN ĐÀO TẠO SAU ĐẠI HỌC** | | | **Độc lập – Tự do – Hạnh phúc** | | |  |
|  |  |  |  |  |  |  |
|  |  |  | *TP. HCM, ngày … tháng… năm 2018* | | |  |
|  |  |  |  |

**NHIỆM VỤ LUẬN VĂN THẠC SĨ**

Họ tên học viên: Bạch Quốc Huy

Giới tính: Nam

Ngày, tháng, năm sinh: 23/04/1993

Nơi sinh: Vĩnh Phúc

Chuyên ngành: Công Nghệ Thông Tin

MSHV: 1641860005

**I- Tên đề tài:**

Áp dụng mô hình đồ thị vào bài toán tóm tắt văn bản tiếng Việt.

**II- Nhiệm vụ và nội dung:**

* Nghiên cứu các kỹ thuật loại bỏ các thông tin không quan trọng trong văn bản để giảm kích thước dữ liệu.
* Nghiên cứu phương pháp biểu diễn văn bản thành một đồ thị vô

hướng.

* Nghiên cứu các phương pháp toán học dùng để tính toán độ tương đồng giữa hai câu thông qua đó xác định trọng số của đồ thị.
* Tập trung vào phương pháp xác định độ tương đồng giữa 2 câu thông qua độ đo khoảng cách Euclidean và Cosine.
* Sử dụng thuật toán xếp hạng PageRank để tính toán điểm số cho mỗi đỉnh của đồ thị trích lọc các đỉnh có điểm số cao nhất để xây dựng văn bản tóm tắt.

**III- Ngày giao nhiệm vụ:**

**IV- Ngày hoàn thành nhiệm vụ:**

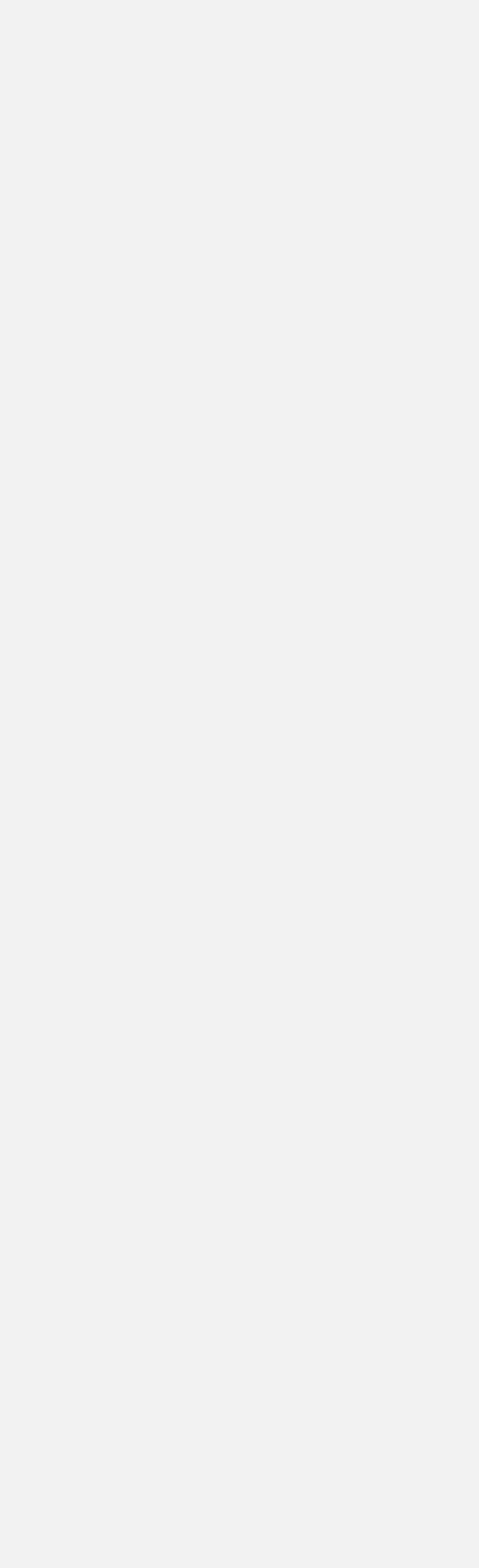
**V- Cán bộ hướng dẫn:** TS. Lê ThịNgọc Thơ

**CÁN BỘ HƯỚNG DẪN**

**KHOA QUẢN LÝ CHUYÊN NGÀNH**

(Họ tên và chữ ký)

(Họ tên và chữ ký)



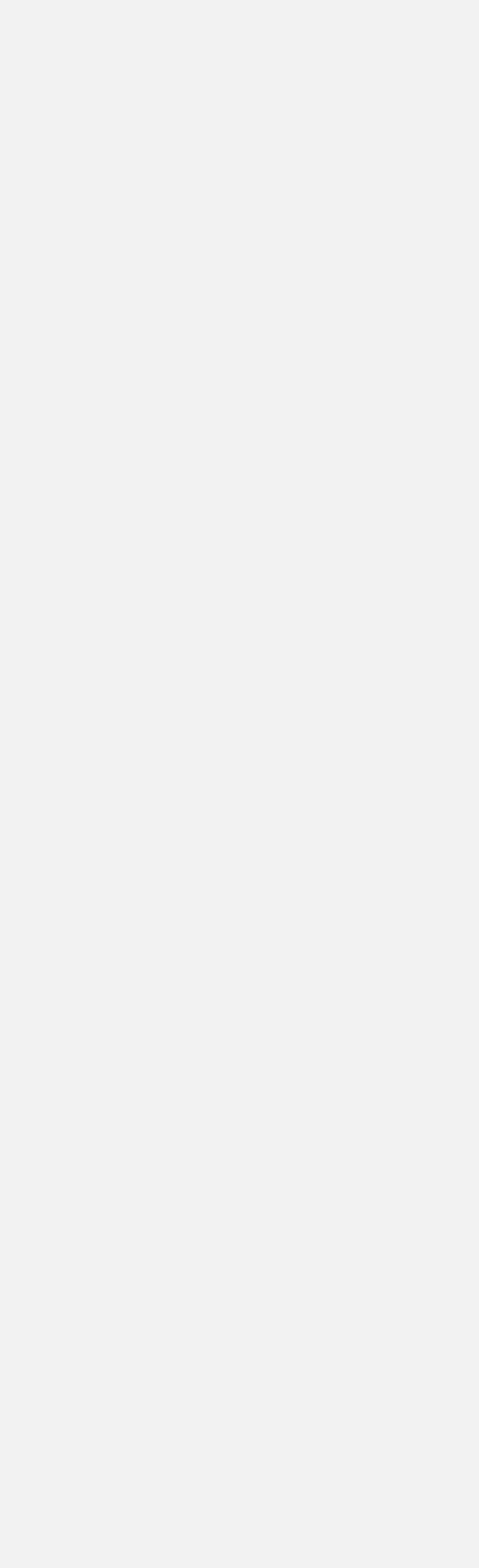
**LỜI CAM ĐOAN**

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi. Các số liệu, kết quả nêu trong Luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tôi xin cam đoan rằng mọi sự giúp đỡ cho việc thực hiện Luận văn này đã được cảm ơn và các thông tin trích dẫn trong Luận văn đã được chỉ rõ nguồn gốc.

**Học viên thực hiện Luận văn**

Bạch Quốc Huy



**LỜI CÁM ƠN**

Luận văn được thực hiện và hoàn thành với sự giúp đỡ quý báu của các thầy cô, đồng nghiệp và bạn bè.

Tôi xin chân thành gửi lời cảm ơn sâu sắc đến TS. Lê Thị Ngọc Thơ, cô đã tận tình hướng dẫn, đồng hành cùng tôi trong suốt quá trình nghiên cứu và thực hiện đề tài này.

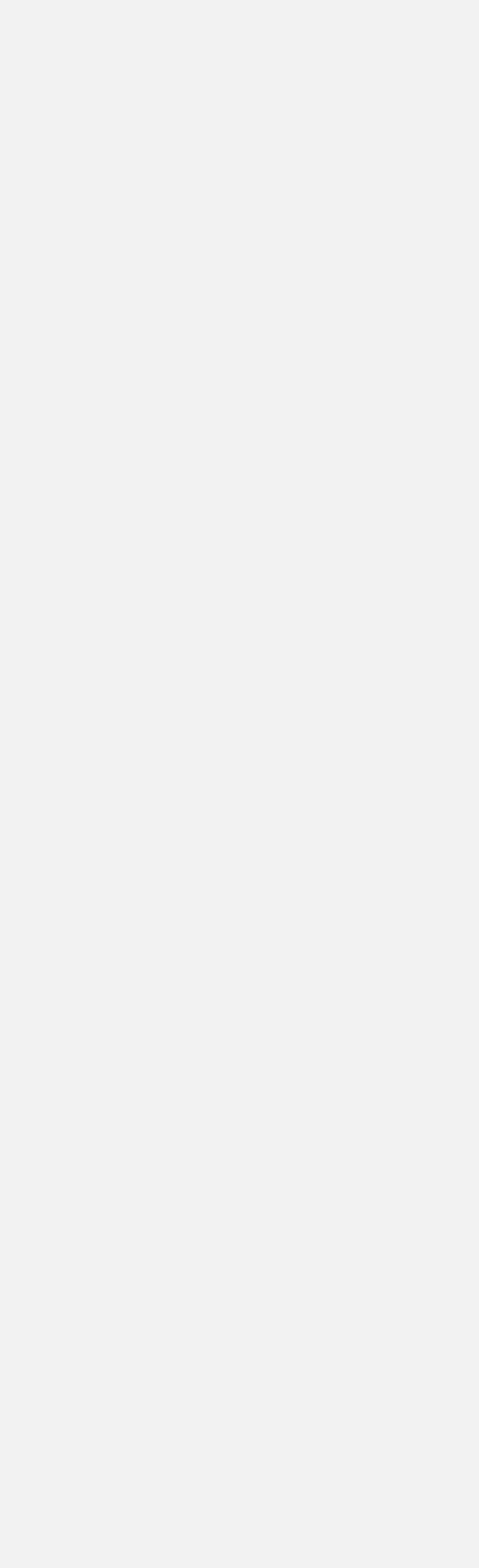
Tôi xin cảm ơn quý thầy cô tham gia giảng dạy lớp 16SCT11, trường Đại học Công nghệ TP. Hồ Chí Minh, đã truyền đạt cho tôi những kiến thức quý báu trong thời gian học tập.

Tôi xin cảm ơn tập thể lãnh đạo, quý thầy cô trường Đại học Công nghệ TP. Hồ Chí Minh đã tạo điều kiện về cơ sở vật chất, trang thiết bị và thời gian để tôi có thể hoàn thành khóa học.

Tôi xin cảm ơn các bạn học trong lớp 16SCT11, trường Đại học Công nghệ TP. Hồ Chí Minh và các bạn của tôi đã cung cấp, hỗ trợ nguồn tài liệu, đóng góp ý kiến trong quá trình nghiên cứu đề tài này.

Một lần nữa tôi xin chân thành cảm ơn rất nhiều sự giúp đỡ nhiệt tình của các thầy cô, đồng nghiệp và bạn bè gần xa trong thời gian qua.

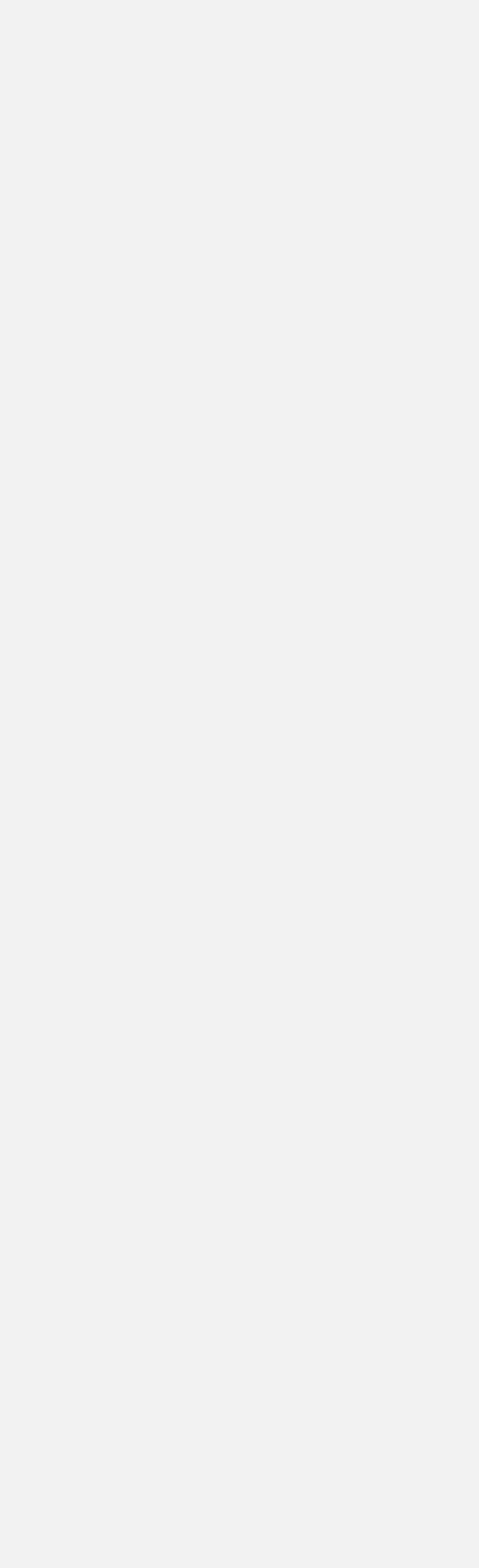
Bạch Quốc Huy



**TÓM TẮT**

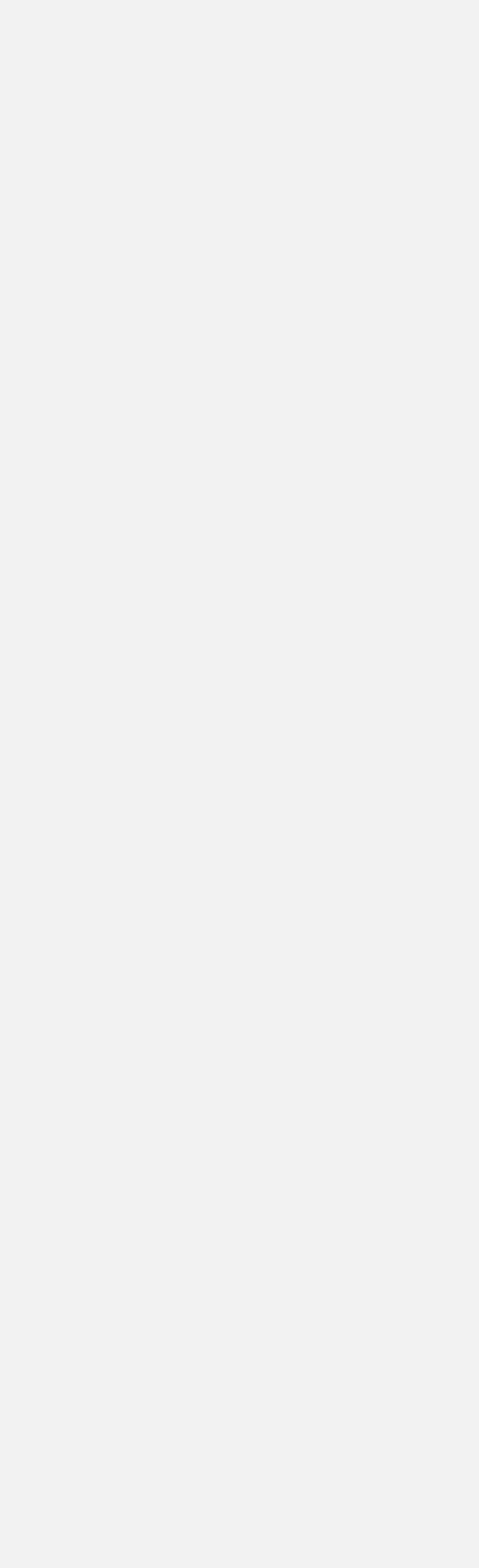
Ngày nay với sự phát triển mạnh mẽ của mạng internet đã đem đến cho chúng ta một kho thông tin khổng lồ để tìm hiểu và khai thác. Tuy nhiên đứng trước rất nhiều lựa chọn như vậy chúng ta sẽ gặp khó khăn để xác định đâu là thông tin, dữ liệu chúng ta cần, có ích và liên quan đến các lĩnh vực mà chúng ta nghiên cứu. Để giải quyết vấn đề này thì cần phải có một công cụ tóm tắt lại phần nào nội dung của các luồng thông tin, dữ liệu nói trên từ đó cho người đọc một cái nhìn khái quát để căn cứ vào đó lựa chọn những thông tin mình cần. Trong bối cảnh này rất nhiều bài toán tóm tắt văn bản đã được nghiên cứu và phát triển tuy nhiên chủ yếu các tác giả hướng đến ngôn ngữ phổ biến đó là tiếng Anh. Các đề tài nghiên cứu tóm tắt văn bản tiếng việt còn rất ít vì vậy kế thừa từ những nghiên cứu trước đây chúng tối tiếp tục nghiên cứu và đề xuất một phương pháp tóm tắt văn bản tiếng việt dựa trên đồ thị. Thông qua nghiên cứu này, chúng tôi hiểu rõ hơn về các phương pháp tóm tắt văn bản, cách xây dựng một đồ thị từ một văn bản, cách tách câu, cách tách từ, cách chọn lọc ra các câu để xây dựng văn bản tóm tắt.

Kết quả thực nghiệm trên tổng thể các văn bản được lấy từ internet ( nguồn được lấy từ trang báo tuoitre.vn ). Mặc dù kết quả thu được chưa đáp ứng được kỳ vọng và còn phụ thuộc vào văn bản của từng lĩnh vực cụ thể, nhưng qua đây chúng tôi hi vọng rằng đây sẽ là một hướng đi và kết quả tham khảo cho những nghiên cứu về sau khi lựa chọn phương pháp tóm tắt văn bản bằng phương pháp đồ thị trong việc giải quyết bài toán tóm tắt văn bản tiếng việt trong tương lai.



**ABSTRACT**

Today, with the strong development of the internet has given us a huge storage of information to explore and exploit. But in the face of so many choices, it is difficult to determine what information we need, which data is useful and relevant to the fields we study. To solve this problem, it is necessary to have a tool to summarize the contents of the information flows, from the above information to give the reader a general view chosing the information we need. In this context, a lot of textual summaries have been researched and developed. However, most of the authors refer to the common language in English.Textual summary research topics in Vietnamese are very few, so inheriting from previous studies we continue to study and propose a method of summarizing Vietnamese text based on graphs. Through this research we gain a better understanding of the methods of summarizing texts, how to construct a graph from a text, how to separate sentences, how to separate words, how to select sentences to create textual summaries . Experimental results on the whole of the text are taken from the internet (source is from the website tuoitre.vn). Although the results have not met the expectations and also depends on the text of specific fields. But here we hope that this will be a direction and a reference results for the following researches when choosing the method of summarizing the text by graph method to solve the problem of summarizing the Vietnamese text in the future.



 **Mục Lục**

 [TỔNG QUAN VỀ TÓM TẮT VĂN BẢN VÀ BÀI TOÁN TÓM TẮT](#page14)

[VĂN BẢN TIẾNG VIỆT](#page14) [14](#page14)

[1.1. Giới thiệu](#page14) [14](#page14)

[1.2. Đặc điểm ngôn ngữ tiếng Việt](#page15) [15](#page15)

[1.2.1. Đặc điểm ngữ âm](#page15) [15](#page15)

[1.2.2. Đặc điểm từ vựng](#page15) [15](#page15)

[1.2.3.Đặc điểm ngữ pháp](#page16) [16](#page16)

[1.3. Phân loại bài toán tóm tắt](#page17) [17](#page17)

[1.3.1. Tóm tắt đơn văn bản](#page20) [20](#page20)

[1.3.2.Tóm tắt theo trích xuất](#page20) [20](#page20)

[1.4. Đánh giá văn bản tóm tắt](#page21) [21](#page21)

 [CÁC KHÁI NIỆM VÀ NGHIÊN CỨU LIÊN QUAN](#page23) [23](#page23)

[2.1. Các khái niệm liên quan](#page23) [23](#page23)

[2.1.2. Độ tương tự giữa hai câu](#page23) [23](#page23)

[2.1.3. Một số phương pháp tính độ tương tự giữa 2 câu.](#page25) [25](#page25)

[2.2. Các nghiên cứu liên quan](#page32) [32](#page32)

 [TÓM TẮT VĂN BẢN SỬ DỤNG TEXTRANK](#page34) [34](#page34)

[3.1. Thuật toán TextRank](#page34) [34](#page34)

[3.1.1 Đồ thị vô hướng](#page35) [35](#page35)

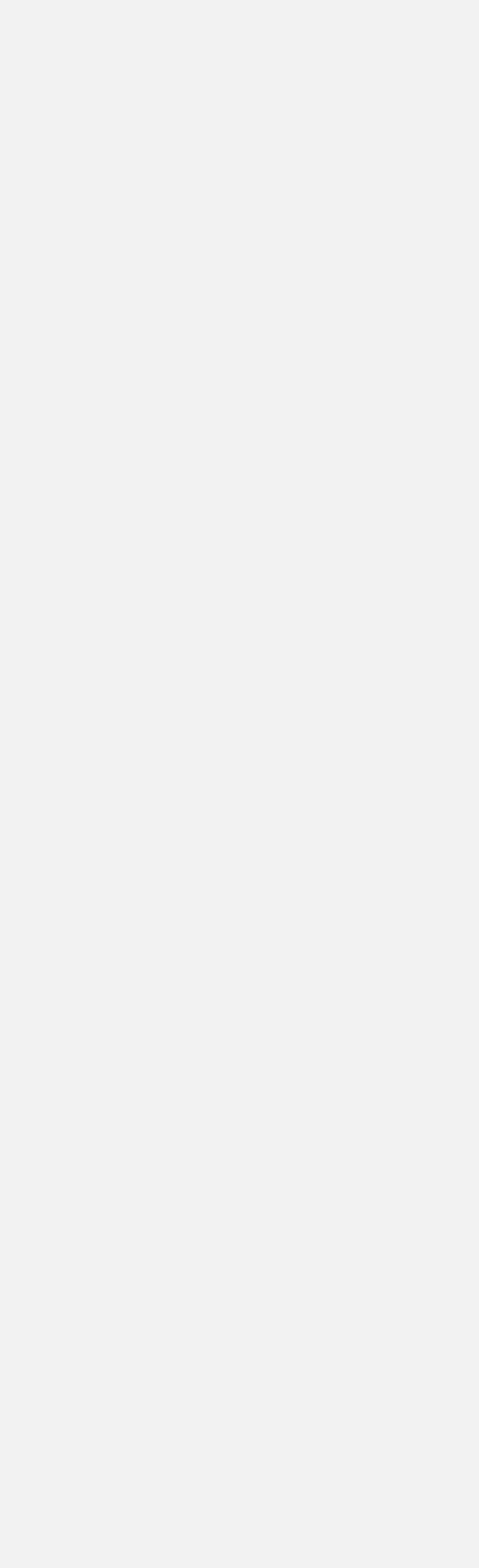
[3.1.2. Đồ thị có trọng số](#page36) [36](#page36)

[3.1.3. Đồ thị hóa văn bản](#page36) [36](#page36)

[3.2. Sử dụng TextRank trích rút câu](#page37) [37](#page37)

 [THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ](#page43) [43](#page43)

[4.1. Dữ liệu thực nghiệm và phương pháp đánh giá](#page43) [43](#page43)



[4.2. Xây dựng hệ thống tóm tắt văn bản Tiếng Việt](#page43) [43](#page43)

[4.3. Thực nghiệm và đánh giá kết quả](#page53) [53](#page53)

[4.4. Khuyến nghị tăng cường chất lượng của văn bản tóm tắt](#page58) [58](#page58)

[4.4.1. Tăng cường độ liên quan giữa các câu](#page58) [58](#page58)

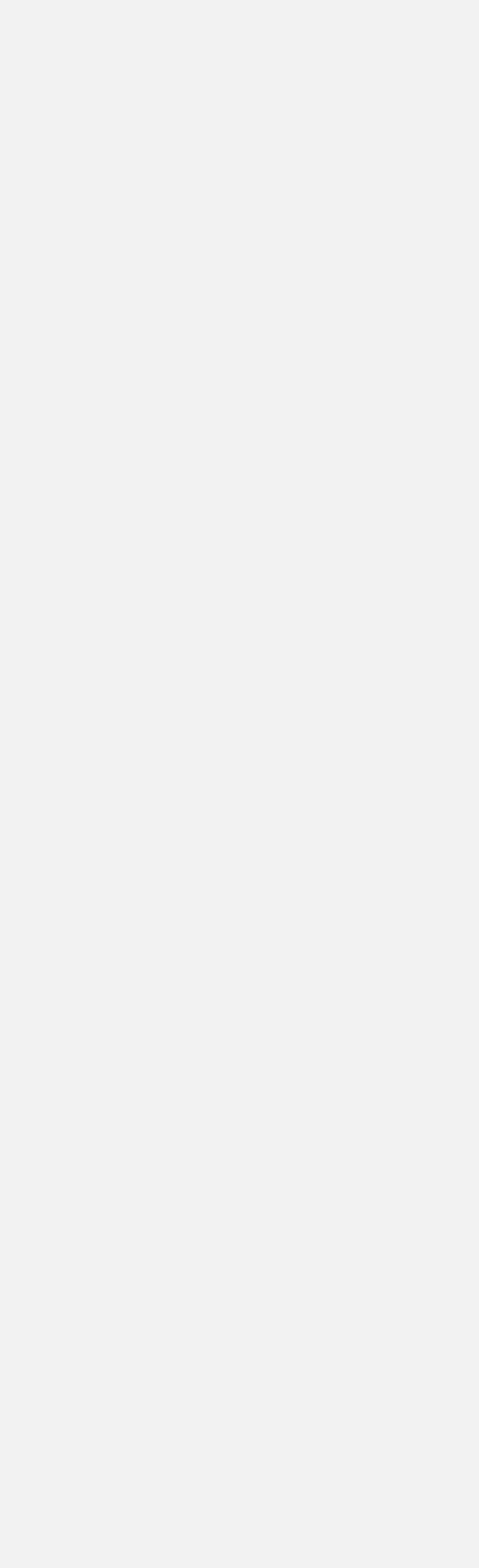
[4.4.2. Tăng cường chất lượng văn bản tóm tắt](#page59) [59](#page59)

 [KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN](#page60) [60](#page60)

[TÀI LIỆU THAM KHẢO](#page62) [62](#page62)

[Phụ lục](#page66) [66](#page66)

|  |  |  |
| --- | --- | --- |
|  | **DANH MỤC CÁC KÝ HIỆU VIẾT TẮT** |  |
|  |  |  |
| **Kí hiệu** | **Giải thích** |  |
|  |  |  |
| *wij* | Trọng số giữa hai đỉnh *Vi* và *Vj* |  |
|  |  |  |
| *S(Vi)* | Trọng số của đỉnh *Vi* trong đồ thị |  |
|  |  |  |
| *In(Vi)* | Số cạnh vào đỉnh *Vi* |  |
|  |  |  |
| *Out(Vj)* | Số cạnh ra từ đỉnh *Vj* |  |
|  |  |  |
| *Similarity* | Độ tương tự giữa câu *Si* và câu *Sj* |  |
| *(Si,Sj)* |  |  |
|  |  |  |
| *Wk* | Từ thứ k trong câu *S* |  |
|  |  |  |
|  | Document Understanding Conferences (Hội nghị chuyên về hiểu văn |  |
| DUC | bản) |  |
|  |  |  |
| BLEU | BiLingual Evaluation Understudy |  |
|  |  |  |
| ROUGE | Recall Oriented Understudy of Gisting Evaluation |  |
|  |  |  |
|  |  |  |

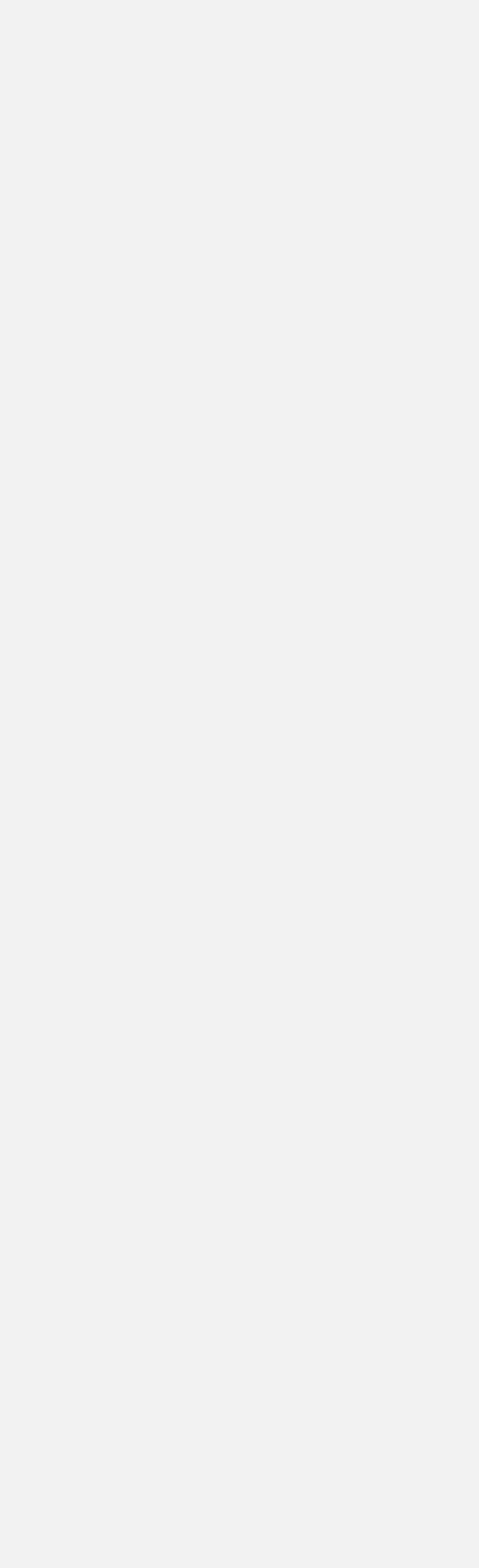


**DANH MỤC CÁC BẢNG**

Bảng 3.1 Kết quả so sánh hệ thống TextRank và các hệ thống tóm tắt khác.

........................................................................................... **Error! Bookmark not defined.**

*Bảng 1.1.* Kết quảthực nghiệm*………………………………………………………..40*



**DANH MỤC CÁC HÌNH**

Hình 2.1. Tần số và xếp hạng N-Gram [15] **Error! Bookmark not defined.**

[Hình 2.2. Hình minh họa Vector Space Model.[28]](#page27) [27](#page27)

[Hình 2.3. Một Framework cho việc học tập của word vectors.](#page28) [28](#page28)

[Hình 2.4. Mô hình distributed bag of words](#page28) [28](#page28)

[Hình 2.5. Mô hình distributed memory.](#page29) [29](#page29)

[Hình 2.6. Công thức tính khoảng cách Euclidean.](#page30) [30](#page30)

[Hình 2.7. Công thức tính khoảng cách cosine.](#page30) [30](#page30)

[Hình 2.8. Khoảng cách Hamming.](#page31) [31](#page31)

[Hình 2.9. Mô tả khoảng cách Mahatan.](#page31) [31](#page31)

[Hình 3.1 Đường cong hội tụ của phương pháp xếp hạng dựa trên đồ thị có hướng-](#page35)

[vô hướng, có trọng số-không có trọng số, 250 đỉnh và 250 cạnh [15].](#page35) [35](#page35)

[Hình 3.2: Đồ thị mô phỏng kết nối giữa các cặp câu trong văn bản.](#page41) [41](#page41)

Hình 1.1. Mô hình tóm tắt văn bản sử dụng phương pháp đồ thị…………….….33



14

 **TỔNG QUAN VỀ TÓM TẮT VĂN BẢN VÀ BÀI TOÁN TÓM TẮT VĂN BẢN TIẾNG VIỆT**

**1.1. Giới thiệu**

Trong những năm thập niên 50 - 60 của thế kỷ XX, các nhà khoa học đã bắt đầu nghiên cứu về tóm tắt văn bản tự động. Năm 1958 , Luhn [18] đã công bố bài báo trình bày phương pháp tóm tắt tự động sử dụng thống kê tần suất và phân bố từ trong văn bản. Đến năm 1969, Edmundson [19] đã công bố nghiên cứu về phương pháp mới trong việc tóm tắt tự động văn bản. Phương pháp này dựa trên tổng hợp của bốn thành phần: vai trò, khoá, tiêu đề và vị trí. Các phương pháp tiếp cận của hai nhà khoa học trên đều thuộc dạng trích rút câu. Các nghiên cứu về tóm tắt văn bản tự động sau một thời gian không có nhiều tiến triển thì đến cuối thế kỷ XX, đầu thế kỷ XXI, với sự bùng nổ mạnh mẽ của CNTT và Internet, lượng thông tin được con người sinh ra và lưu trữ vô cùng lớn. Vấn đề được đặt ra là làm sao để thu nhận thông tin quan trọng nhanh nhất, hiệu quả nhất. Từ đó, bài toán tóm tắt văn bản trở nên cấp thiết và được quan tâm hơn đúng với tầm quan trọng của nó.

Theo Inderjeet Mani [20] , *“Tóm* *tắt văn bản tự động nhằm mục đích trích xuất nội* *dung từ một nguồn thông tin và trình bày các nội dung quan trọng nhất cho người sử dụng theo một khuôn dạng súc tích và gây cảm xúc đối với người sử dụng hoặc một chương trình cần đến".*

Kết quả của quá trình tóm tắt văn bản tự động thường không cho kết quả chất lượng như văn bản tóm tắt bởi con người do bị giới hạn bởi nhiều yếu tố. Chúng ta rất khó khăn để nâng cao chất lượng văn bản tóm tắt tự động mà không bị giới hạn bởi miền ứng dụng. Vì vậy, trong tóm tắt văn bản tự động, các hướng giải quyết thường hướng đến các bài toán cụ thể với một phương pháp cụ thể.



15

**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 hòa về ngữ âm, đến nhạc điệu của câu văn [1].

**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 [1].

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, tan cửa nát nhà… Hiện nay, đây là phương thức chủ yếu để sản sinh ra các đơn vị từ vựng. Theo phương pháp 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.

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 chỉa, chỏng chơ, đỏng đà đỏng đảnh, thơ thẩn, lúng la 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-kĩ thuật, đặc biệt là công nghệ thông tin, thi tiềm năng đó



16

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:

**Nguyên 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 họp 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 họp 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 họp "anh của em" khác với tổ họp "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.



17

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.

**1.3. Phân loại bài toán tóm tắt**

Sự phát triển của các hệ thống xử lý ngôn ngữ tự trên 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.

Hiện tại có nhiều cách phân loại tóm tắt văn bản khác nhau, việc phân loại phụ thuộc vào cơ sở để tóm tắt [3]. Luận văn đề cập đến phân loại tóm tắt dựa trên các cơ sở:

* Định dạng văn bản, nội dung đầu vào
* Định dạng, nội dung đầu ra
* Mục đích tóm tắt

Chi tiết các phân loại dựa trên định dạng, nội dung đầu vào như sau:

* Kiểu văn bản (bài báo, bản tin, thư, báo cáo ...). Với cách phân loại này, tóm tắt văn bản là bài báo sẽ khác với tóm tắt thư, tóm tắt báo cáo khoa học do những đặc trưng văn bản quy định.
  + Định dạng văn bản: dựa vào từng định dạng văn bản khác nhau, tóm tắt cũng chia ra thành các loại khác nhau như: tóm tắt văn bản không theo cấu trúc nhất định và tóm tắt văn bản có cấu trúc. Đối văn bản có cấu trúc, tóm tắt văn



18

bản thường sử dụng một mô hình học dựa vào mẫu cấu trúc đã xây dựng từ trước để tiến hành tóm tắt.

* Số lượng dữ liệu đầu vào: Tóm tắt đơn văn bản khi đầu vào chỉ là một văn bản đơn, trong khi đó đầu vào của tóm tắt đa văn bản là một tập các tài liệu có liên quan đến nhau như: các tin tức có liên quan đến cùng một sự kiện, các trang

web cùng chủ đề hoặc là cụm dữ liệu được trả về từ quá trình phân cụm.

* Miền dữ liệu: tùy theo miền của dữ liệu về cụ thể về một lĩnh vực nào đó, ví dụ như: y tế, giáo dục... hay miền dữ liệu tổng quát, có thể chia tóm tắt ra thành từng loại tương ứng.
* Tóm tắt trên cơ sở mục đích thực chất là làm rõ cách tóm tắt, mục đích tóm tắt là gì, tóm tắt phục vụ đối tượng nào ...
  + Nếu phụ thuộc vào đối tượng đọc tóm tắt thì tóm tắt cho chuyên gia khác cách tóm tắt cho các đối tượng đọc thông thường.
  + Tóm tắt sử dụng trong tìm kiếm thông tin (IR) sẽ khác với tóm tắt phục vụ cho việc sắp xếp.
  + Dựa trên mục đích tóm tắt, còn có thể chia ra thành tóm tắt chỉ thị và tóm tắt thông tin. Tóm tắt chỉ thị chỉ ra loại của thông tin, ví dụ như là loại văn bản chỉ thị “tuyệt mật”. Còn tóm tắt thông tin chỉ ra nội dung của thông

tin.

* Tóm tắt trên cơ sở truy vấn (Query-based) hay tóm tắt chung. Tóm tắt chung có mục đích chính là tìm ra đoạn tóm tắt cho toàn bộ văn bản mà nội dung của đoạn văn bản sẽ bao quát toàn bộ nội dung của văn bản đó. Tóm tắt trên cơ sở truy vấn thì nội dung của văn bản tóm tắt sẽ dựa trên truy vấn của người dùng hay chương trình đưa vào, loại tóm tắt này thường được sử dụng trong quá trình tóm tắt các kết quả trả về từ máy tìm kiếm.

Tóm tắt trên cơ sở đầu ra cũng có nhiều cách phân loại.

* Dựa vào ngôn ngữ: Tóm tắt cũng có thể phân loại dựa vào khả năng tóm tắt các loại ngôn ngữ:
  + Tóm tắt đơn ngôn ngữ (Monolingual): hệ thống có thể tóm tắt chỉ một loại ngôn ngữ nhất định như: tiếng Việt hay tiếng Anh...



19

* + Tóm tắt đa ngôn ngữ (Multilingual): hệ thống có khả năng tóm tắt nhiều loại văn bản của các ngôn ngữ khác nhau, tuy nhiên tương ứng với văn bản đầu vào là ngôn ngữ gì thì văn bản đầu ra cũng là ngôn ngữ tương ứng.
  + Tóm tắt xuyên ngôn ngữ (Crosslingual): hệ thống có khả năng đưa ra các văn bản đầu ra có ngôn ngữ khác với ngôn ngữ của văn bản đầu vào.
* Dựa vào định dạng đầu ra của kết quả tóm tắt: như bảng, đoạn, từ khóa.

Ngoài hai cách phân loại trên, phân loại tóm tắt trên cở sở đầu ra còn có một cách phân loại được sử dụng phổ biến là: tóm tắt theo trích xuất (Extract) và tóm tắt theo tóm lược (Abstract).

* Tóm tắt theo trích xuất: là tóm tắt có kết quả đầu ra là một tóm tắt bao gồm toàn bộ các phần quan trọng được trích ra từ văn bản đầu vào.
* Tóm tắt theo tóm lược: là tóm tắt có kết quả đầu ra là một tóm tắt không giữ nguyên lại các thành phần của văn bản đầu vào mà dựa vào thông tin quan trọng để viết lại một văn bản tóm tắt mới.

Hiện nay, các hệ thống sử dụng tóm tắt theo trích xuất được sử dụng phổ biến và cho kết quả tốt hơn tóm tắt theo tóm lược. Nguyên nhân tạo ra sự khác biệt này là do các vấn đề trong bài toán tóm tắt theo tóm lược như: biểu diễn ngữ nghĩa, suy luận và sinh ra ngôn ngữ tự nhiên được đánh giá là khó và chưa có nhiều kết quả nghiên cứu khả quan hơn so với hướng trích xuất câu của bài toán tóm tắt theo trích xuất. Trong thực tế, theo đánh giá của Dragomir R. Radev [26] chưa có một hệ thống tóm tắt theo tóm lược đạt đến sự hoàn thiện, các hệ thống tóm tắt theo tóm lược hiện nay thường dựa vào thành phần trích xuất có sẵn. Các hệ thống này thường được biết đến với tên gọi tóm tắt theo nén văn bản.

Tóm tắt theo nén văn bản (Text Compaction): là loại tóm tắt sử dụng các phương pháp cắt xén (truncates) hay viết gọn (abbreviates) đối với các thông tin quan trọng sau khi đã được trích xuất.

Mặc dù tính trên cơ sở phân loại có nhiều loại tóm tắt khác nhau nhưng hai loại tóm tắt là tóm tắt đơn văn bản và tóm tắt đa văn bản vẫn được sự quan tâm lớn của các nhà nghiên cứu về tóm tắt tự động.



20

**1.3.1. Tóm tắt đơn văn bản**

Bài toán tóm tắt văn bản đơn cũng giống như các bài toán tóm tắt khác, là một quá trình tóm tắt tự động với đầu vào là một văn bản, đầu ra là một đoạn văn bản ngắn gọn mô tả nội dung chính của văn bản đầu vào. Văn bản đơn có thể là một trang Web, một nội dung đăng trên mạng xã hội, một bài báo, một tài liệu dạng văn bản (ví dụ: .doc,

.txt)... Tóm tắt văn bản đơn là bước làm cơ sở cho việc xử lý tóm tắt đa văn bản và các bài toán tóm tắt phức tạp hơn. Đó là nguyên nhân lý giải cho việc những phương pháp tóm tắt văn bản ra đời đầu tiên đều là các phương pháp tóm tắt đơn văn bản.

Các phương pháp nhằm giải quyết bài toán tóm tắt văn bản đơn cũng tập trung vào hai loại tóm tắt là: tóm tắt theo trích xuất và tóm tắt theo tóm lược.

**1.3.2.Tóm tắt theo trích xuất**

Đa số các phương tóm tắt loại này tập trung vào việc trích xuất ra các câu hay các từ ngữ nổi bật từ các đoạn văn bản và kết hợp chúng lại thành một văn bản tóm tắt. Một số nghiên cứu giai đoạn đầu thường sử dụng các đặc trưng như vị trí của câu trong văn bản, tần số xuất hiện của từ, ngữ hay sử dụng các cụm từ khóa để tính toán trọng số của mỗi câu, qua đó chọn ra các câu có trọng số cao nhất cho văn bản tóm tắt [18,19]. Các kỹ thuật tóm tắt gần đây sử dụng các phương pháp học máy và xử lý ngôn ngữ tự nhiên nhằm phân tích để tìm ra các thành phần quan trọng của văn bản. Sử dụng các phương pháp học máy có thể kể đến phương pháp của Kupiec, Penderson and Chen năm 1995 sử dụng phân lớp Bayes để kết hợp các đặc trưng lại với nhau [24] hay nghiên cứu của Lin và Hovy năm 1997 áp dụng phương pháp học máy nhằm xác định vị trí của các câu quan trọng trong văn bản [25]. Bên cạnh đó việc áp dụng các phương pháp phân tích ngôn ngữ tự nhiên như sử dụng mạng từ Wordnet của Barzilay và Elhadad vào năm 1997 [4].

* **Tóm tắt theo tóm lược**

Các phương pháp tóm tắt không sử dụng trích xuất để tạo ra tóm tắt có thể xem như là một phương pháp tiếp cận tóm tắt theo tóm lược. Các hướng tiếp cận có thể kể đến như dựa vào trích xuất thông tin (information extraction), ontology, hợp nhất và nén thông tin... Một trong những phương pháp tóm tắt theo tóm lược cho kết quả tốt là các



21

phương pháp dựa vào trích xuất thông tin, phương pháp dạng này sử dụng các mẫu đã được định nghĩa trước về một sự kiện hay là cốt truyện và hệ thống sẽ tự động điền các thông tin vào trong mẫu có sẵn rồi sinh ra kết quả tóm tắt. Mặc dù cho ra kết quả tốt tuy nhiên các phương pháp dạng này thường chỉ áp dụng trong một miền nhất định [5].

**1.4. Đánh giá văn bản tóm tắt**

Hiện tại, việc đánh giá kết quả văn bản tóm tắt tự động là việc làm khó khăn. Cách đánh giá tốt nhất là sử dụng ý kiến đánh giá của các chuyên gia ngôn ngữ. Nhưng đây là một phương pháp tốn kém. Vì vậy, ngoài các phương pháp đánh giá thủ công, vấn đề đánh giá tự động kết quả tóm tắt cũng nhận được nhiều sự chú ý. Từ năm 2000, NIST tổ chức hội nghị DUC hàng năm để thực hiện việc đánh giá các hệ thống tóm tắt văn bản. Việc đánh giá tự động nhằm mục đích là tìm ra được một độ đo đánh giá văn bản tóm tắt giống với đánh giá của con người nhất.

Độ phủ (recall) tại các tỷ lệ nén khác nhau là thước đo đánh giá hợp lý, cho nó không chỉ ra được sự khác nhau về hiệu suất. Độ đo này được tính theo công thức:

=

\_ \_ \_

\_ \_\_ \_

**Độ đo ROUGE**

BiLingual Evaluation Understudy (BLEU) [6] là một phương pháp đưa ra để đánh giá các hệ thống dịch tự động. Phương pháp này nhanh, độc lập với ngôn ngữ và sự liên quan với các đánh giá của con người. Recall Oriented Understudy of Gisting Evaluation (ROUGE)

1. được Lin và Hovy đưa ra vào năm 2003 dựa trên khái niệm tương tự BLEU. ROUGE sử dụng n-gram để đánh giá sự tương quan giữa các kết quả của văn bản tóm tắt và tập dữ liệu đánh giá. Phương pháp này cho ra kết quả tốt và được đánh giá cao trong cộng đồng các nhà khoa học trong cùng lĩnh vực.

Công thức đánh giá ROUGE với n-gram được xác định như sau :

∑ ∈ { ă ả ℎ ℎ ế } ∑

∈ Count (gram n)



22

ROUGE − N = ∑ ∈ { ă ả ℎ ℎ ế } ∑ ( ) ∈ S Count match(gram n)

Trong đó:

* *S:* là câu trong văn bản
* n: Độ dài của gram đang xét
  + *Countmatch(gramn):* là sốgram *n* trùng nhau lớn nhất của văn bản cần đánh giávà văn bản tham chiếu
* *Count (gram*n): Sốgram *n* có trong văn bản tham chiếu

Như vậy, độ đo ROUGE-N thuộc dạng độ đo hồi tưởng (recall-related).



23

 **CÁC KHÁI NIỆM VÀ NGHIÊN CỨU LIÊN QUAN**

**2.1. Các khái niệm liên quan**

**2.1.1 .TF-IDF**

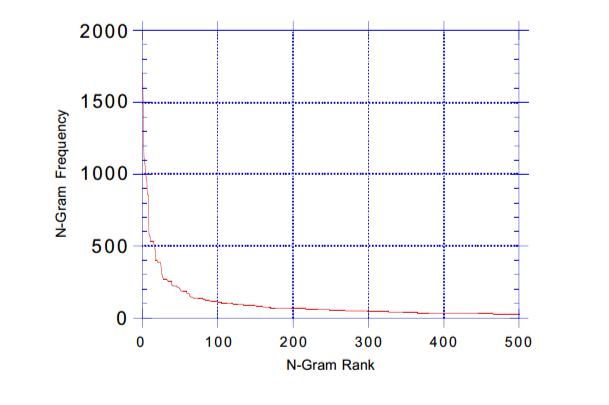
TF-IDF là viết tắt của thuật ngữ tiếng Anh term Term Frequency – Inverse Document Frequency, TF-IDF là trọng số của một từ trong văn bản thu được qua thống kê thể hiện mức độ quan trọng của từ này trong một văn bản, mà bản thân văn bản đang xét nằm trong một tập hợp các văn bản.

Thuật toán này thường được sử dụng vì: trong ngôn ngữ luôn có những từ xảy ra thường xuyên với các từ khác. Và một trong những phát biểu nổi tiếng nhất Zipf’s law phát biểu về vấn đề này như sau:

“*The nth* *most common word in a human language text occurs with a frequency inversely* *proportional to n*.”[21].



24



*Hình 2.1. Tần số và xếp hạng N-Gram [15].*

Có nghĩa là luôn có một tập các từ mà tần số xuất hiện, sử dụng nhiều hơn các từ khác, điều này đúng trong bất kì ngôn ngữ nào. Chính vì vậy ta cần có một phương pháp để làm mịn đường cong tần số trên hay là việc cân bằng mức độ quan trọng giữa các từ.

**Các tính trọng số TF-IDF**

**TF (Term Frequency**): dùng để ước lượng tần xuất xuất hiện của từ trong văn bản. Tuynhiên với mỗi văn bản thì có độ dài khác nhau, vì thế số lần xuất hiện của từ có thể nhiều hơn .Vì vậy số lần xuất hiện của từ sẽ được chia độ dài của văn bản (tổng số từ trong văn bản đó)

|  |  |  |  |
| --- | --- | --- | --- |
| TF(t, d) = | số lần từ t xuất hiện trong văn bản d | |  |
| tổng số từ trong văn bản d |  |  |
|  |  |

**IDF- Inverse Document Frequency**:dùng để ước lượng mức độ quan trọng của từ đónhư thế nào. Khi tính tần số xuất hiện tf thì các từ đều được coi là quan trọng như nhau. Tuy nhiên có một số từ thường được được sử dụng nhiều nhưng không quan trọng để thể hiện ý nghĩa của đoạn văn, ví dụ:

* Từ nối: và, nhưng, tuy nhiên, vì thế, vì vậy…



25

* Giới từ: ở, trong, trên …
* Từ chỉ định: ấy, đó, nhỉ …

Vì vậy ta cần giảm đi mức độ quan trọng của những từ đó bằng cách sử dụng IDF:

IDF(t, D) = Log\_e Tổng số văn bản trong tập mẫu D

Số văn bản có chứa từ t

**2.1.2. Độ tương tự giữa hai câu**

Độ tương đồng ngữ nghĩa giữa các câu đóng một vai trò ngày càng quan trọng trong nghiên cứu Text mining, Web mining và xử lý ngôn ngữ tự nhiên. Nó cũng được sử dụng như là một tiêu chuẩn của trích chọn thông tin để tìm ra những tri thức ẩn trong cơ sở dữ liệu hay trên các kho dữ liệu trực tuyến. Một ứng dụng thực tế là khi tìm kiếm ảnh từ một trang Web, nếu xác định hợp lý sự tương đồng ngữ nghĩa giữa câu truy vấn với các đoạn text ngắn bao quanh ảnh thì hệ thống tìm kiếm sẽ đưa ra kết quả đáp ứng tốt hơn yêu cầu người dùng. Vấn đề tính toán độ tương đồng giữa các câu trong văn bản với nhau hoặc với câu chủ đề của văn bản/nhóm văn bản nhận được sự quan tâm đặc biệt trong các hội nghị khoa học quốc tế, đặc biệt trong các hội nghị thường niên về hiểu văn bản (Document Understanding Workshop - DUC). Việc xây dựng một độ đo chuẩn xác để thể hiện được mối quan hệ tương đồng về ngữ nghĩa giữa các câu sẽ làm cho các ứng dụng trở nên “thông minh” hơn, đặc biệt trên Web [8][9] .

**2.1.3. Một số phương pháp tính độ tương tự giữa 2 câu.**

Để tính toán độ tương tự giữa hai câu theo một số độ đo khoảng cách sẽ đề cập trong mục này đầu tiên chúng ta phải trải biểu diễn các câu hoặc từ thành vector.

Vector hóa câu là làm cho máy tính có thể hiểu được ngôn ngữ tự nhiên nói chung và ngôn ngữ tiếng việt nói riêng. Hiểu một cách đơn giản, vector hóa câu là chuyển đổi dữ liệu từ các kí tự chữ cái thành các vector trọng số. Hiện nay, có khá nhiều cách để ta vector hóa câu nhưng trong nghiên cứu này, chúng tôi chọn phương pháp Doc2Vec. Lý do chúng tôi lựa



26

chọn thuật toán này vì nó có hiệu xuất tương đối tốt trong quá trình vector hóa dữ liệu. Và để có thể hiểu hơn về Doc2Vec hơn, ta phải làm rõ một vài khái niệm sau.

* ***Vector space model***

Nói một cách ngắn gọn, Vector space model (Mô hình không gian vector) là một mô hình đại số (algebraic model) thể hiện thông tin văn bản như một vector, các phần tử của vector này thể hiện mức độ quan trọng của một từ và cả sự xuất hiện hay không xuất hiện (Bag of words) của nó trong một tài liệu.

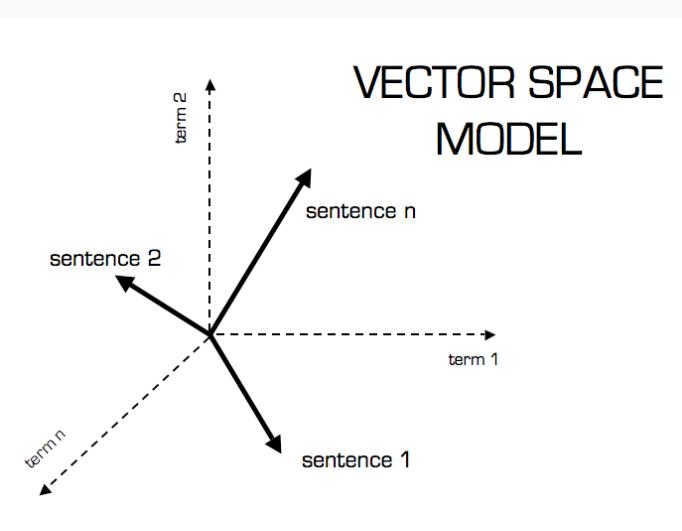
Mô hình này biểu diễn văn bản như những điểm trong không gian Euclid n-chiều, mỗi chiều tương ứng với một từ trong tập hợp các từ. Phần tử thứ i, là di của vector văn bản cho biết số lần mà từ thứ i xuất hiện trong văn bản. Sự tương đồng của hai văn bản được định nghĩa là khoảng cách giữa các điểm, hoặc là góc giữa những vector trong không gian.

Mỗi từ trong không gian vector sẽ có một trọng số, có nhiều phương pháp xếp hạng khác nhau, nhưng TF-IDF (Term Frequency–Inverse Document Frequency) là một phương pháp phổ biến để đánh giá và xếp hạng một từ trong một tài liệu. MySQL fulltext search cũng sử dụng phương pháp này. Về cơ bản thì tf-idf là một kỹ thuật (cụ thể là ranking function) giúp chuyển đổi thông tin dưới dạng văn bản thành một Vector space model thông qua các trọng số. Vector space model và TF-IDF được phát triển bởi Gerard Salton vào đầu thập niên 1960s.

Mặc dù đơn giản, nhưng mô hình không gian vector và những biến thể của nó hiện nay vẫn là cách phổ biến để biểu diễn văn bản trong Data mining và Information Retrieval. Tuy nhiên, một trong những điểm yếu của vector space model số chiều lớn (high-dimensonal), có khoảng cỡ chục triệu chiều trong không gian vector nếu như chúng ta áp dụng nó vào web search engine.



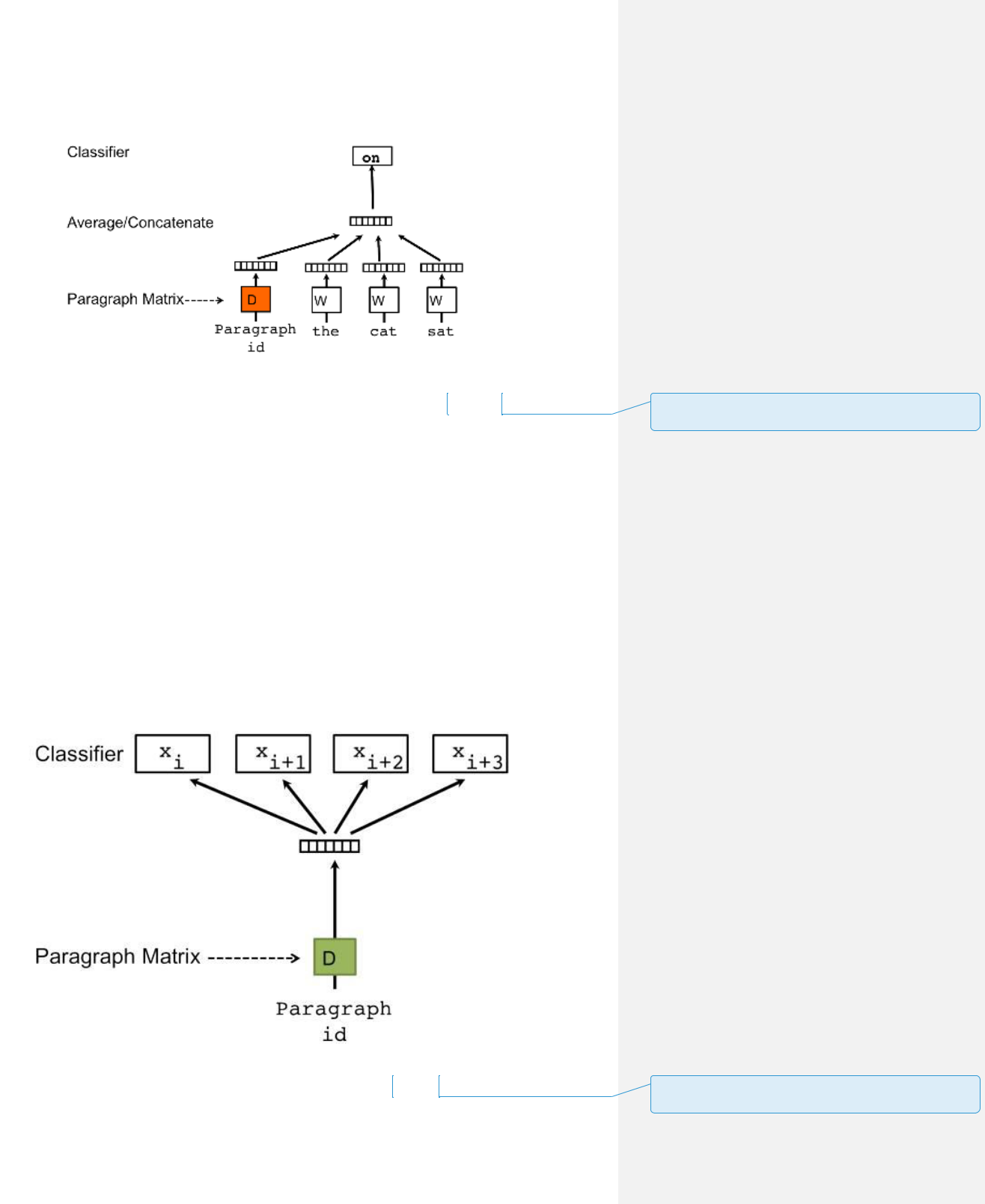
27



*Hình 2.2. Hình minh họa Vector Space Model [28].*

* ***Word Vector***

Sự thành công của Word2vec đã được chứng minh trong rất nhiều công trình NLP. Nhắc lại về Word2vec, nó sử dụng một tập copus, qua một mạng Neural biểu diễn các word thành các vector, các vector giữ lại được tính chất ngữ nghĩa. Tức các từ mang ý nghĩa similar với nhau thì gần nhau trong không gian vector. Trong NLP, đây một trong những phương thức của word embedding. Word2vec hiện nay được sử dụng hết sức rộng rãi. Doc2vec, ngoài từ (word), ta còn có thể biểu diễn các câu (sentences) thậm chí 1 đoạn văn bản (document). Khi đó, bạn có thể dễ dàng vector hóa cả một đoạn văn bản thành một vector có số chiều cố định và nhỏ, từ đó có thể chạy bất cứ thuật toán classification cơ bản nào trên các vector đó.



28

*Hình 2.3. Một Framework cho việc học tập của word vectors.*

* ***Doc2vec***

Doc2vec được giới thiệu bởi Quoc Le và Mikolov [27], cũng như Word2vec, có 2 model chính là: DBOW và DM

DBOW (distributed bag of words): Mô hình này đơn giản là không quan tâm thứ tự các từ, training nhanh hơn, không sử dụng local-context/neighboring. Mô hình chèn thêm 1 "word" là ParagraphID, ParagraphID này đại diện cho văn bản được training. Sau khi training xong có thể hiểu các vector ParagraphID này là vector embedded của các văn bản. Hình ảnh được mô tả trong bài báo:

Hình 2.4. Mô hình distributed bag of words

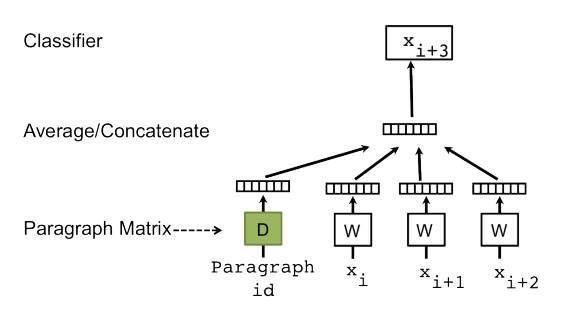
**Commented [N1]:** Hình lây ơ đâu thi ghi tai lieu thamkhảo đó vào dây.

**Commented [N2]:** Thêm tài liệu tham khảo vào đây, tức làchép hình này từ đâu ra.



29

DM (distributed memory): xem một paragraph là một từ, sau đó nối từ này vào tập các từ trong câu. Trong quá trình training, vector của paragraph và vector từ đều được update.



*Hình 2.5. Mô hình distributed memory.*

Sau khi đã vecto hóa dữ liệu chúng tôi sẽ sử dụng các công thức tính khoảng cách để xác định độ tương tự giữa 2 câu. Một số công thức thức tính khoảng cách:

* **Khoảng cách *Euclidean***

***Khoảng cách Euclidean (Euclidean Distance)***:Khoảng cách giữa 2 điểm là chiều dàicủa đường thẳng nối chúng. Trong mặt phẳng, khoảng cách giữa 2 điểm (x1, y1) và (x2, y2) được cho bởi định lý Pythagorean như sau:

= √( 2 − 1)2 + ( 2 − 1)2

Trong không gian ba chiều Euclidean, khoảng cách giữa hai điểm (x1, y1, z1) và (x2, y2,

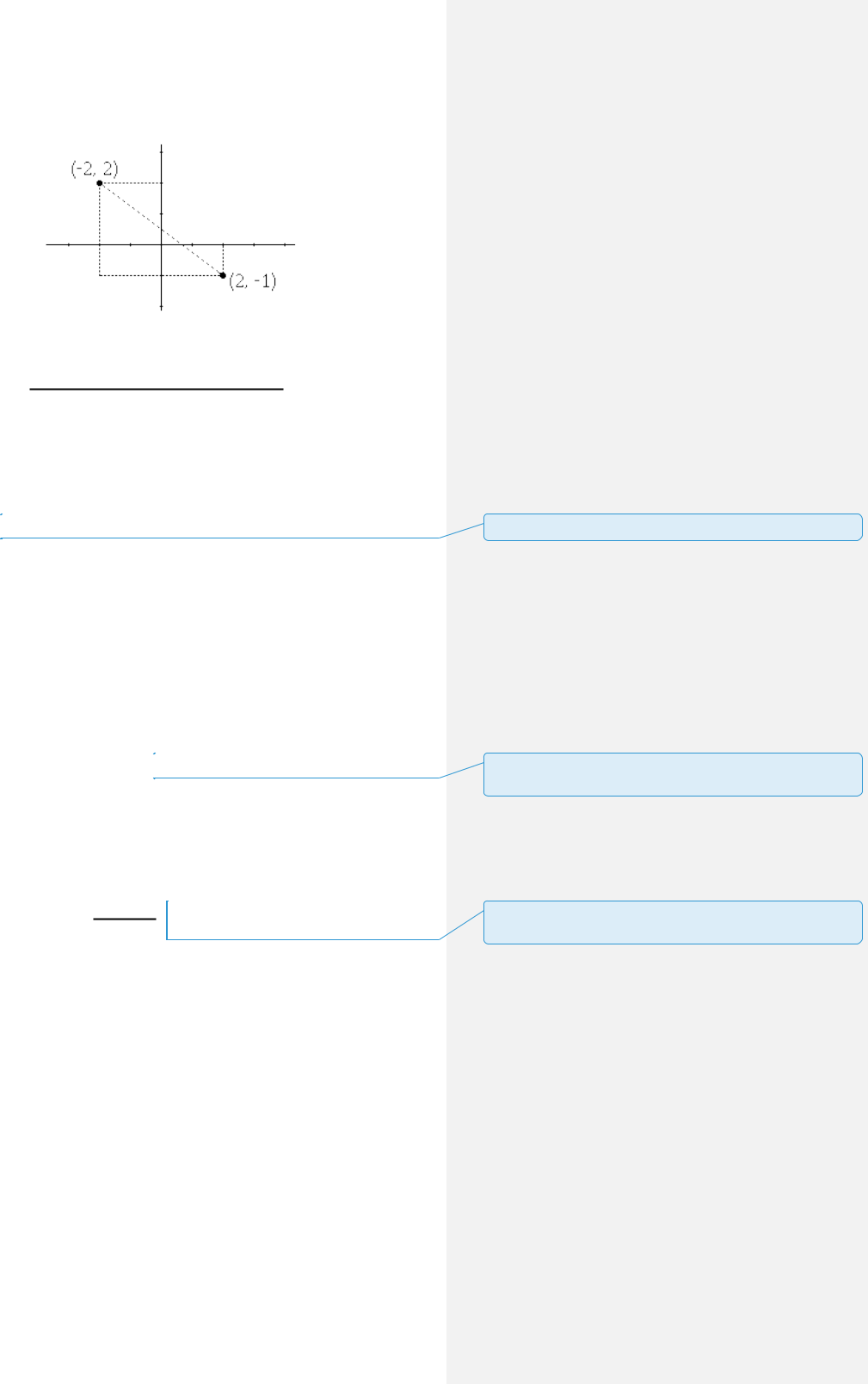
z2) là:

= √( 2 − 1)2 + ( 2 − 1)2 + ( 2 − 1)2

Một cách tổng quát, khoảng cách giữa 2 điểm x, y trong không gian Euclidean R với n chiều được tính như sau:

d=|x-y|=√∑ =1 | − |2

Ví dụ:



30

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dist ((2, [-1), (-2, 2)) = (2 −](http://2.bp.blogspot.com/-YIb1zdScOzQ/UWYr5-MSUZI/AAAAAAAABB0/oX9iHXjfUQE/s1600/EuclideanDistance.gif) (−2))2 | | | | + | ((−1) | − 2)2 |  |
|  | √ | |  |  |  |  |
| =√16 + 9 | | | |  |  |  |  |
|  |  |  | |  |  |  |  |
| =√25 | | | |  |  |  |  |
| = 5 |  |  |  |  |  | **Commented [N3]:** Viết ra cho tửtế, ko chụp hình thếnày |  |

*Hình 2.6. Công thức tính khoảng cách Euclidean.*

* **Khoảng cách Cosine**

**Khoảng cách Cosine (hay Cosine Similarity):** từcông thức tính tích hai vector của

Euclidean như sau:

a.b=||a|| ||b|| cosθ

Chúng ta thấy nếu cho 2 vector A và B với n chiều ta sẽ có *cos* *(θ),* cosine tương tự (Cosine

Similarity) hay khoảng cách theo góc của hai vector sẽ được tính bằng công thức:

.

Similarity=Cosθ=|| || || ||=

*Hình 2.7. Công thức tính khoảng cách cosine.*

* **Khoảng cách Hamming**

***Khoảng cách Hamming (Hamming Distance)***: Là số thay thế nhỏ nhất các thành phần,phần tử thuộc đối tượng, hệ thống, dữ liệu A để có thể trở thành B. Khi đó nó gọi là khoảng cách Hamming giữa A và B. Cũng có thể hiểu khái niệm này như một dạng của phép đo lỗi (error) hay khác biệt (difference) giữa 2 trạng thái, 2 đối tượng khác nhau.

Dễ dàng nhận thấy nếu A và B là hai chuỗi nhị phân thì Hamming Distance của chúng là số giá trị 1 trong kết quả của A XOR B.

**Commented [N4]:** Viết ra cho tửtế, ko chụp hình. Tươngtự với các công thức khác

**Commented [N5]:** Viết ra cho tửtế, ko đc chụp hình dánvào đây



31

Ví dụ:

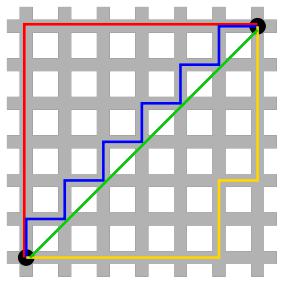
1. = "hello bioz"
2. = "hallo bios"

d (A, B) = 2

*Hình 2.8. Khoảng cách Hamming.*

* **Khoảng cách Manhattan**

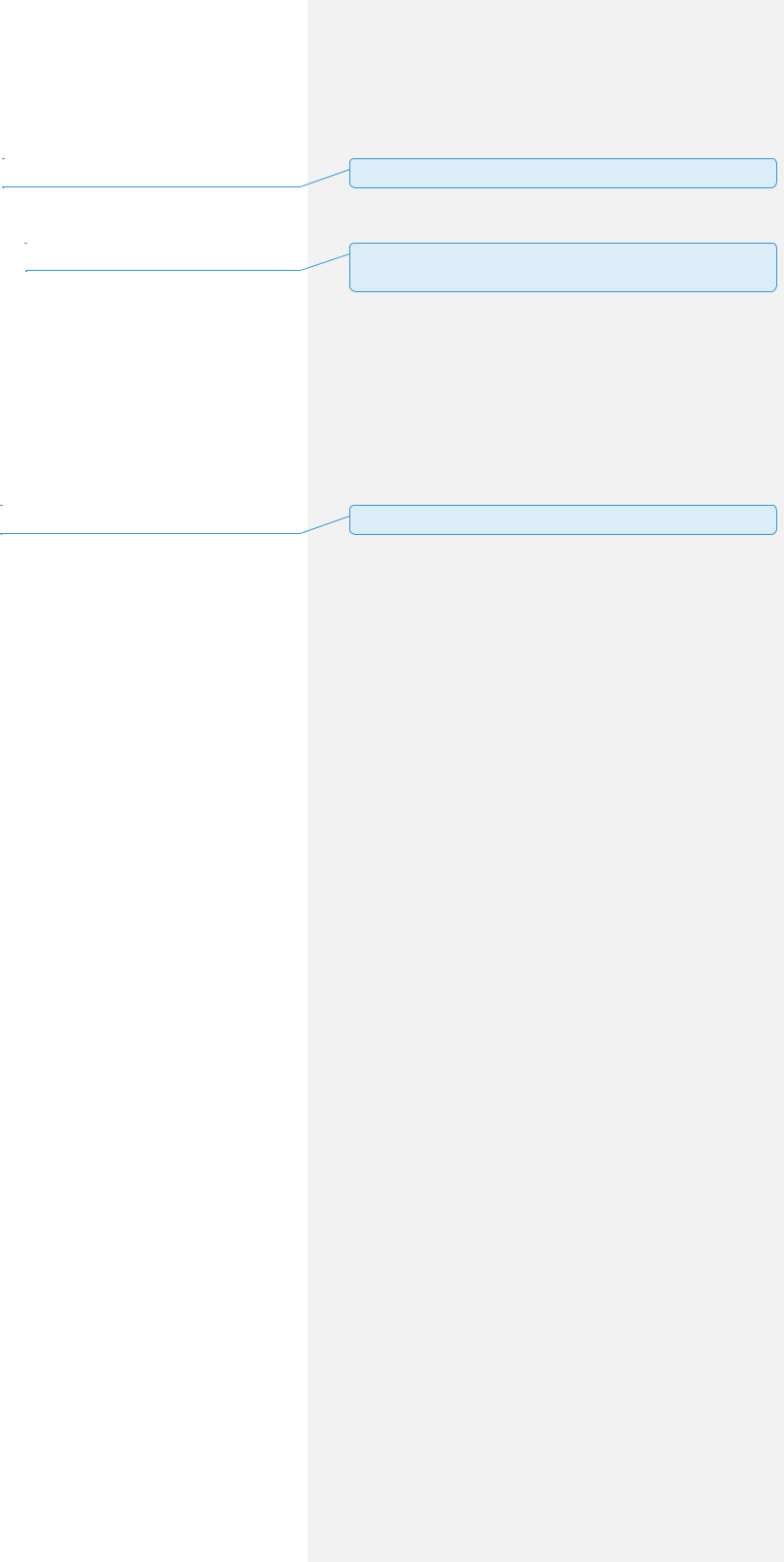
***Khoảng cách Manhattan (Manhattan Distance = Rectilinear Distance = City Block Distance = Taxicab Geometry = Chebyshev Distance)***: Khoảng cách của hai điểm P1 vàP2 trong một mặt phẳng bằng chiều dài của tất cả các đoạn nối p với q dọc theo mặt cắt trục đứng và trục ngang, theo một chiều tới.



*Hình 2.9. Mô tả khoảng cách Mahatan.*

Trong hình trên đường đỏ thể hiện khoảng cách Manhattan, đường xanh nước biển và vàng tương đương với khoảng cách Manhattan, đường xanh lá cây là đường thằng hay khoảng cách Euclidean.

Một cách tổng quát, khoảng cách Manhattan của 2 điểm p và q trong không gian vector n chiều với hệ tọa độ Cartesian xác định được tính như sau:



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 32 |  |  |  |  |  |  |
| d1 (p,q)=||p-q||1=∑ =1 |pi - qi| |  |  |  |  | **Commented [N6]:** Viết ra cho tửtế |  |
|  |  |  |
|  |  |  |  |
| với | | | | |  |  |
|  | | | | |  |  |
| p= (p1,p2,….,pn) and q=(q1, q2,…,qn) | | | | | **Commented [N7]:** Viết ra cho tửtế, ko đc chụp hình dán |  |
|  |  |  |  |  | vào |  |

* **Khoảng cách *Minkowski***

***Khoảng cách Minkowski (Minkowski Distance)***: là một trường hợp tổng quát của cảkhoảng cách Euclidean và khoảng cách Chebyshev, cho bởi công thức:

|  |  |  |  |
| --- | --- | --- | --- |
| d(x,y)=(∑ =1( |xk – yk |r )1/r |  | **Commented [N8]:** Viết ra cho tửtế |  |
|  |  |
|  |  |  |

trong đó **r** gọi là thành tố Minkowski (Minkowski Factor), khi **r = 1** thì **d** tương đương với khoảng cách Euclidean, **r = 2** thì **d** trở thành khoảng cách Chebyshev. Với **x**, **y** là các vector đặc trưng cùng có **n** chiều.

**2.2. Các nghiên cứu liên quan**

Trong những năm gần đây, một số phương pháp khác nhau đã được đề xuất để tự động tóm tắt văn bản tiếng Việt.

Hoang và cộng sự [10] đề xuất một phương pháp tóm tắt văn bản tiếng Việt dựa trên phương pháp trích xuất câu. Ý tưởng của bài báo là dùng một đồ thị vô hướng biểu diễn các câu của văn bản với mỗi đỉnh của đồ thị ứng với một câu. Trọng số của các cạnh được tính dựa vào độ tương tự của hai câu, nó phụ thuộc vào tần số xuất hiện của các từ có trong mỗi câu. Phương pháp này xử lý nhanh tốn ít chi phí tuy nhiên vẫn chưa đảm bảo độ chính xác cao về mặt ngữ nghĩa điều này sẽ ảnh hưởng đến chất lượng nội dung của văn bản tóm tắt. Sau khi xây dựng được đồ thị thì sử dụng PageRank để tiến hành xếp hạng và trích xuất câu quan trọng. Cuối cùng là xây dựng văn bản tóm tắt dựa trên các câu quan trọng được lấy ra.

Huong và cộng sự [11] đề xuất một phương pháp tóm tắt văn bản tiếng Việt dựa trên việc trích xuất cụm từ. Bài báo mô tả một cách tiếp cận để tổng hợp văn bản tiếng Việt tập trung vào cấu trúc bài luận của văn bản. Mặc dù phương pháp đưa ra thu được những kết quả đầy hứa hẹn tuy nhiên hạn chế của đề tài là việc tạo ra một cây thảo luận phục vụ cho việc rút trích mệnh đề là tương đối phức tạp, tập thảo luận cần đầy đủ hơn để nâng cao



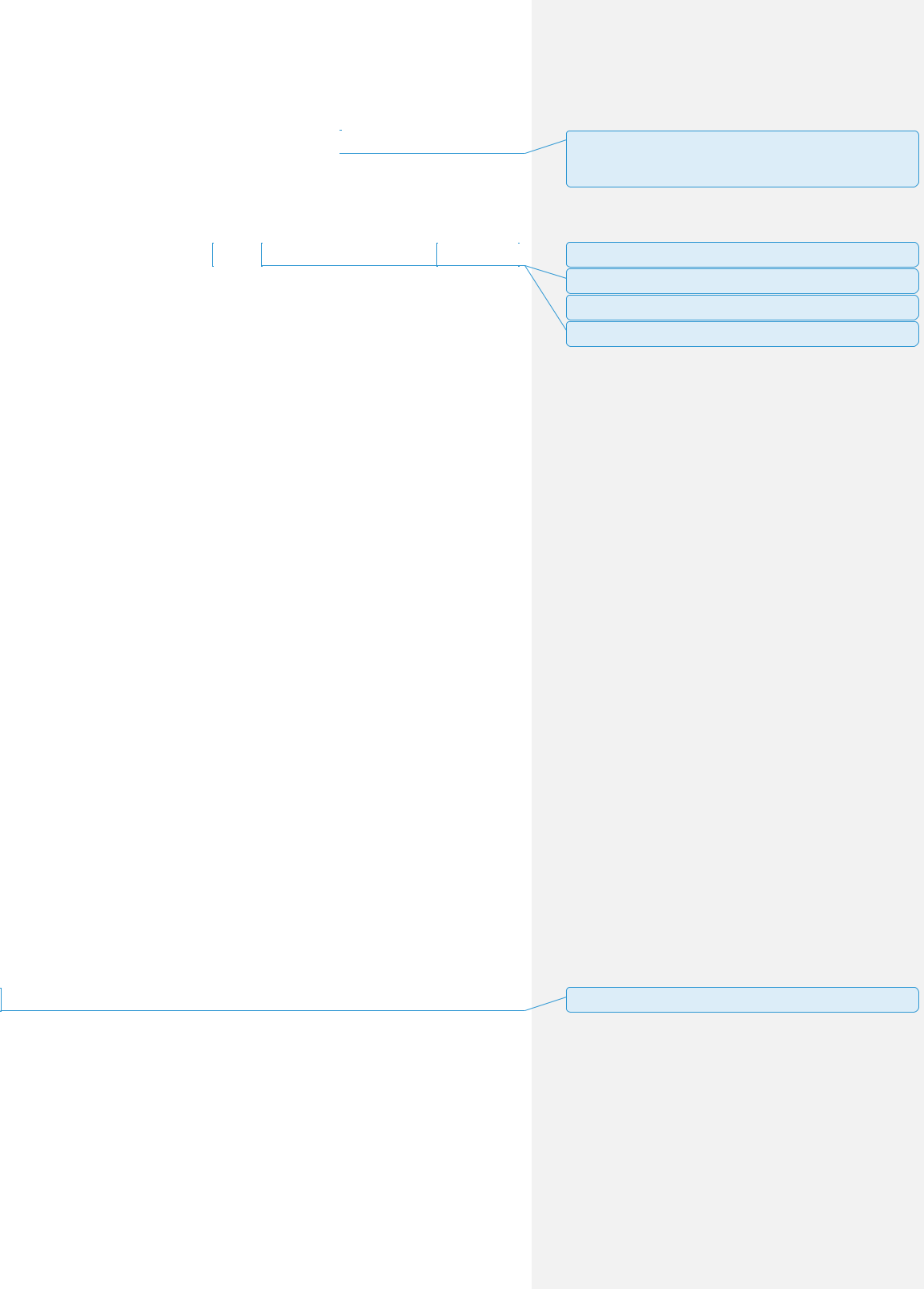
33

chất lượng tóm tắt của văn bản, cần tìm ra thêm nhiều phương pháp để xác định quan hệ giữa các khoảng trong văn bản.

Ha [12] đề xuất một phương pháp tóm tắt đối với văn bản đơn âm tiết dựa trên thuật toán Bayes và sử dụng các từ chủ đề. Phương pháp này là một phương pháp học có giám sát nên đòi hỏi phải có thời gian đển huấn luyện dữ liệu, với mỗi chủ đề khác nhau lại phải huấn luyện một tập dữ liệu riêng biệt khó áp dụng dễ dàng cho nhiều lĩnh vực.

Phung và cộng sự [13] đề xuất một phương pháp sử dụng các từ nhúng (word embeddings) kết hợp với PageRank để xếp hạng các câu và đưa ra tập các câu phù hợp nhất cho việc tóm tắt văn bản tiếng Việt. Phương pháp này cần phải có một tập dữ liệu tóm tắt cho trước để xây dựng nên mối quan hệ giữa câu và từ với mỗi chủ đề thì tập dữ liệu tóm tắt lại thay đổi và cần thời gian thu thập, xây dựng trước khi được sử dụng.

Nguyen và cộng sự [14] đề xuất việc kết hợp ý kiến người trong mục bình luận dưới mỗi bài viết để góp phần tạo ra văn bản tóm tắt chất lượng, là một phương pháp học có giám sát nên nó cũng cần thời gian xây dựng dữ liệu huấn luyện hơn nữa việc xây dựng dữ liệu huấn luyện ở đây cần có sự can thiệp từ con người.



34

 **TÓM TẮT VĂN BẢN SỬ DỤNG TEXTRANK**

Các thuật toán xếp hạng dựa trên đồ thị đã được đưa ra và sử dụng rộng rãi trong những

năm của thế kỷ XX. Trong số đó phổ biến là thuật toán HITS của Kleinberg [22] và PageRank 

của Google do hai nhà đồng sáng lập phát triển (Brin và Page) [23]. Chúng được sử dụng

trong việc phân tích mạng xã hội, cấu trúc liên kết của các trang web, .... Thực tế thì thuật

toán xếp hạng dựa trên đồ thị sẽ xác định đỉnh nào là quan trọng trong đồ thị bằng cách tính toán đệ quy các thông tin trên toàn đồ thị thay vì chỉ sử dụng thông tin trên từng đỉnh. Quá trình này làm cho việc xác định mức độ quan trọng của mỗi đỉnh chính xác hơn.

Từ cách tiếp cận trên, ta có thể áp dụng sang các đồ thị từ vựng và đồ thị ngữ nghĩa trích xuất được từ các tài liệu trong ngôn ngữ tự nhiên. Kết quả của việc sử dụng mô hình xếp hạng dựa trên đồ thị có thể ứng dụng trong nhiều chương trình xử lý ngôn ngữ tự nhiên. Ví dụ như mô hình xếp hạng hướng văn bản được ứng dụng trong các vấn đề như tự động trích xuất từ khoá đến tóm tắt văn bản và xác định từ nhập nhằng ý nghĩa [15].

Trong chương này ta sẽ tìm hiểu mô hình TextRank, các thuật toán và ứng dụng của nó trong việc trích xuất từ khoá và xếp hạng các câu trong một văn bản. Đây là tiền đề cho tóm tắt văn bản tiếng Việt tự động sử dụng phương pháp TextRank.

**3.1. Thuật toán TextRank**

Như trình bày ở trên, thuật toán xếp hạng dựa trên đồ thị là cách đưa ra cách chọn đỉnh quan trọng trong đồ thị dựa trên các thông tin toàn cục của các đỉnh trong đồ thị. Ý tưởng của thuật toán này dựa trên hai yếu tố: bỏ phiếu và đề cử. Mỗi một liên kết đến đỉnh đang xét thì nó được 1 phiếu bầu. Như vậy, càng nhiều phiếu bầu thì đỉnh đó càng quan trọng. Từ cách xác định trên thì trọng số của một đỉnh chính là số phiếu bầu cho đỉnh đó.



Ta có đồ thị *G =* *(V, E)* là đồ thị có hướng. Trong đó:



V: là tập các đỉnh

E: là tập các cạnh của đồ thị, E là tập con của *V × V (E* ⊆ *V × V)*

Với mỗi đỉnh *V****i*** thì ta có:

* In(V) là tập các đỉnh trỏ đến *V****i***
* Out(*V****i****)* là tập các đỉnh mà *V****i*** trỏ đến.

**Commented [N9]:** Đưa sang trang mới, mỗi chương trongluận văn cũng tương tự như chương trong sách vậy, em sửa format lại

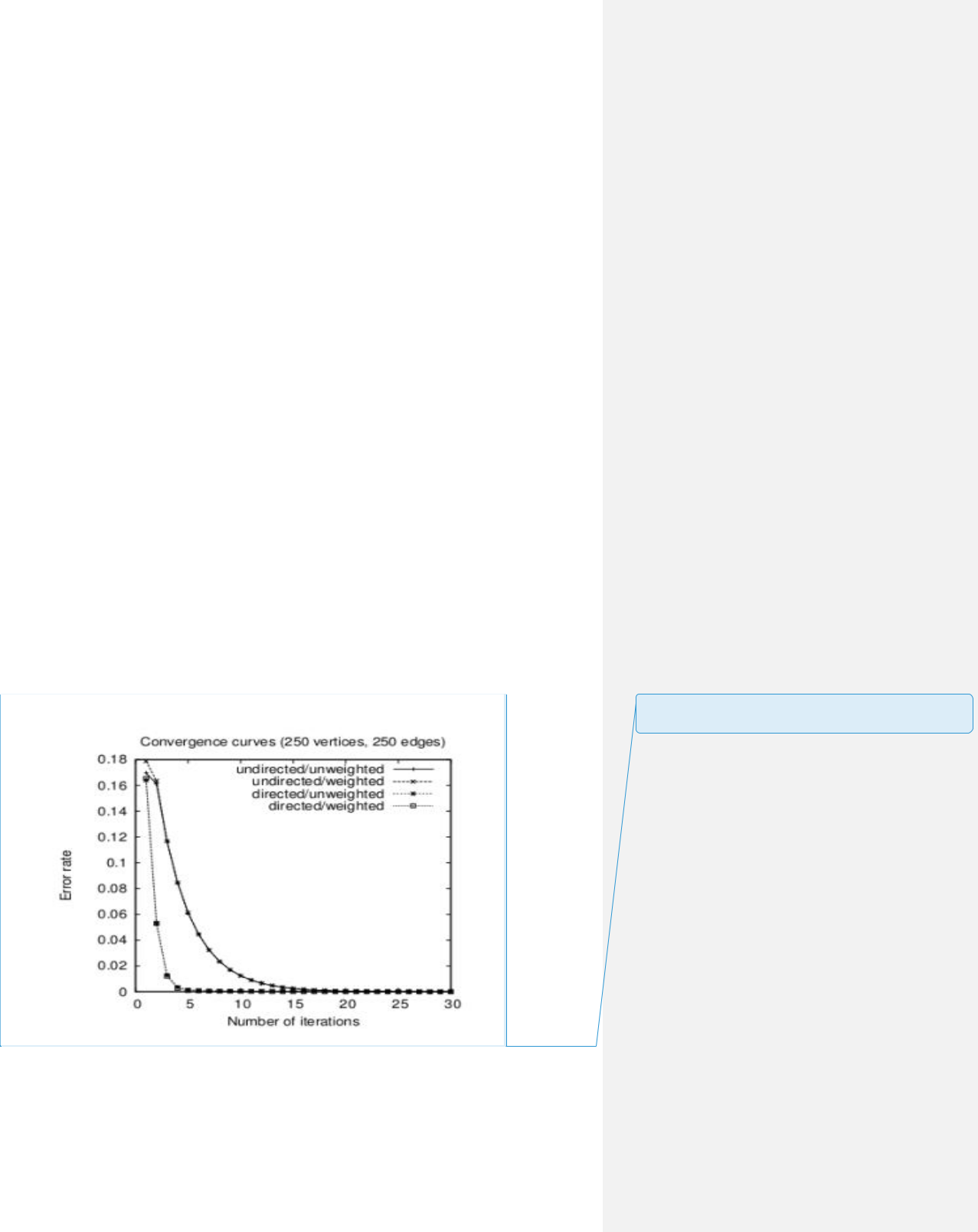
**Commented [N10]:** Thêm tài liệu tham khảo vào đây

**Commented [N11R10]:** OK

**Commented [N12]:** Thêm tài liệu tham khảo vào đây

**Commented [N13R12]:** OK

**Commented [N14]:** y hay V?



35

Trọng số của đỉnh *V****i*** được xác định như sau (Brin and Page, 1998):

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | |  |  |
| S(Vi)=(1-d)+d\* ∑j∈In(Vi) |  | ( ) (1) |  |
| |Out(Vj)| |  |

Trong đó d là nhân tố giảm, có giá trị từ 0 đến 1. Nó là xác xuất mà một đỉnh có liên kết đến một đỉnh bất kỳ trong đồ thị. Đối với các trang web thì d là xác suất người dùng nhấn vào một liên kết bất kỳ và xác suất để người dùng vào một trang web hoàn toàn mới là 1 - d. Theo thực nghiệm với thuật toán PageRank thì d = 0.85. Đây cũng là xác suất được sử dụng trong TextRank.

Ban đầu gán cho tất cả các đỉnh trong đồ thị các giá trị khởi tạo và tính toán lặp lại cho đến khi kết quả hội tụ lại đạt ngưỡng xác định. Sau quá trình tính toán thì trọng số của mỗi đỉnh chính là mức độ quan trọng của đỉnh đó trong toàn đồ thị. Có điều cần lưu ý, đó là giá trị trọng số của mỗi đỉnh sẽ không phụ thuộc vào giá trị khởi tạo ban đầu được gán cho mỗi đỉnh. Ngoài ra thì số lượng các vòng lặp tính toán để ra được trọng số là khác nhau.

**3.1.1 Đồ thị vô hướng**

Việc áp dụng thuật toán TextRank vào đồ thị vô hướng cũng giống như với đồ thị có hướng. Có một điểm cần lưu ý, đó là trong đồ thị vô hướng thì số đỉnh vào bằng số đỉnh ra.

*Hình 3.1 Đường cong hội tụ của phương pháp xếp hạng dựa trên đồ thị có hướng- vô hướng, có trọng số-không có trọng số, 250 đỉnh và 250 cạnh [15].*

**Commented [N15]:** Chép hình nàyở đâuthì phải cite (ghitham chiếu) tới tài liệu đó ở trong caption của hình



36

Trong hình trên thì đường cong hội tụ cho đồ thị được sinh ra ngẫu nhiên với 250 đỉnh và 250 cạnh, với ngưỡng dừng là 10-5 (ngưỡng này được xác định đủ nhỏ để thuật toán dừng tính toán) cho thấy số lần lặp của quá trình tính toán không cao mặc dù số lượng đỉnh và cạnh lớn. Bên cạnh đó thì đường cong độ tụ của đồ thị có hướng và vô hướng gần như trùng nhau. Điều đó cho thấy đồ thị vô hướng hay có hướng đều cho kết quả giống nhau, chỉ khác nhau ở số lần tính toán lặp lại

**3.1.2. Đồ thị có trọng số**

Gần như không có tình huống một trang web có thể có nhiều liên kết đến một trang nào đó trong môi trường web. Vì vậy mà thuật toán Pagerank ban đầu chỉ sử dụng đồ thị không trọng số. Tuy nhiên đối với các văn bản trong ngôn ngữ tự nhiên thì việc một văn bản nào đó có nhiều thành phần tham chiếu đến một văn bản khác là hoàn toàn xảy ra. Do đó, để cải tiến PageRank cho phù hợp với ngôn ngữ tự nhiên, thuật toán TextRank sử dụng đồ thị có trọng số. Trọng số ở đây được định nghĩa là độ dài kết nối giữa hai đỉnh *V****i*** và Vj, ký hiệu *w****ij****.* Từ đó suy ra, công thức (1) cần phải được thay đổi để phù hợp với đồ thị có trọng số trong thuật toán TextRank. Ta được công thức mới như sau:

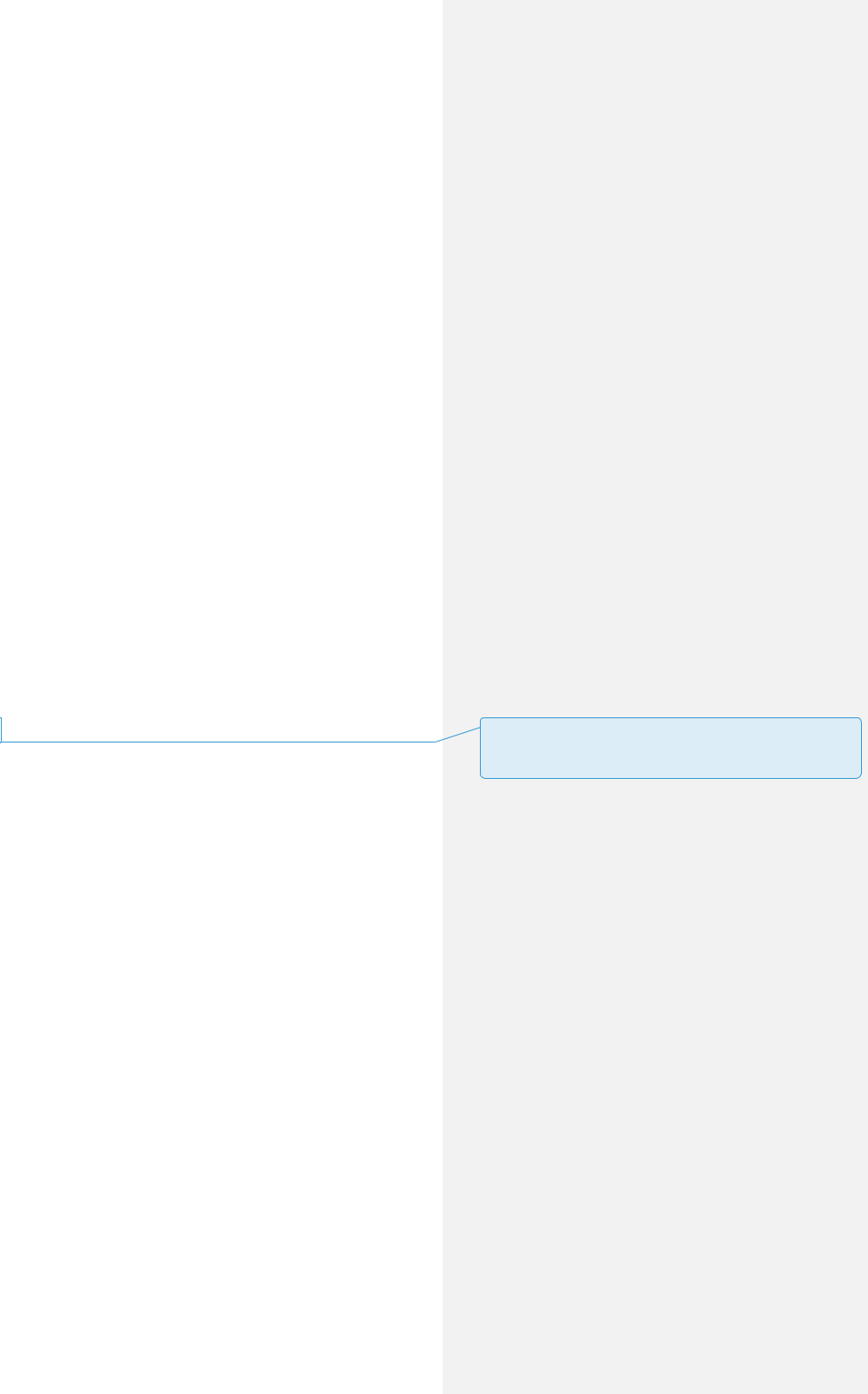
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S(Vi)= (1-d) +d \* ∑ j∈In(Vi) | Wji | ( ) | (2) |  |
|  |  |
| ΣVk∈Out(Vj)Wjk |  |

Như vậy, cũng theo hình 3.1 ở trên thì số lần lặp lại tính toán để có độ tụ đạt ngưỡng 10-

1. của đồ thị có trọng số và đồ thị không trọng số là tương đương nhau.

**3.1.3. Đồ thị hóa văn bản**

Văn bản là một chuỗi các ký tự / từ được sắp xếp với nhau. Vậy nên, để áp dụng được vào thuật toán dùng đồ thị thì cần phải đồ thị hoá văn bản. Việc đồ thị hoá văn bản là xây dựng một đồ thị để đại diện cho văn bản, các liên kết giữa các từ, cụm từ, câu hoặc các quan hệ ngữ nghĩa. Tuỳ thuộc vào các ứng dụng mà kích thước văn bản, các đặc trưng được đưa vào đồ thị là từ, cụm từ hay là cả câu. Cũng giống việc xác định các đỉnh trong đồ thị như trên thì việc xác định các cạnh trong đồ thị là gì cũng phụ thuộc vào miền ứng dụng. Quan hệ được



37

xác định có thể là từ vựng, ngữ nghĩa hoặc ngữ cảnh.

Tùy vào các loại và đặc trưng để đưa vào đồ thị mà có các cách thức làm việc. Nhưng cách thức hoạt động của thuật toán xếp hạng dựa vào đồ thị áp dụng cho ngôn ngữ tự nhiên có các bước như sau:

* Xác định đơn vị văn bản phù hợp nhất với ứng dụng để thêm vào là đỉnh của

đồ thị.

* Xác định quan hệ kết nối giữa các đơn vị văn bản đã xác định ở trên để vẽ các cạnh giữa các đỉnh trong đồ thị. Các cạnh này có thể là vô hướng hoặc có hướng, có trọng số hoặc không trọng số.
* Lặp lại thuật toán xếp hạng cho đến khi độ hội tụ thoả mãn ngưỡng.
* Sắp xếp các đỉnh dựa trên các trọng số đã được tính toán trong bước trên. Như vậy, thuật toán này giúp cho chúng ta làm được hai việc: trích rút từ khoá và trích

rút câu trong văn bản ngôn ngữ tự nhiên. Vấn đề được đề cập ngay sau đây.



**3.2. Sử dụng TextRank trích rút câu**

Đặc điểm của TextRank là sử dụng đồ thị, nên để áp dụng được TextRank thì cần phải đồ thị hoá văn bản. Muốn trích rút được câu thì cần phải xếp hạng được các câu trong văn bản trên toàn đồ thị. Vì thế, mỗi câu sẽ là một đỉnh của đồ thị. Có một điểm cần lưu ý, quan hệ đồng xuất hiện không thể áp dụng trong trường hợp này vì không tồn tại các câu giống nhau 100% trong toàn văn bản. Thay vào đó, một định nghĩa quan hệ khác được đưa ra để xác định kết nối giữa các câu với nhau. Đó là độ tương tự giữa các câu. Ở đây, độ tương tự được xác định bằng độ bao phủ về mặt nội dung giữa các câu với nhau. Mối quan hệ giữa hai câu đó được xem là một “đề cử”: một câu đề cập đến một khái niệm nào đó trong văn bản sẽ “đề cử” cho độc giả một câu khác trong văn bản cũng đề cập đến khái niệm đó. Do đó xuất hiện một liên kết giữa các câu có chung nội dung.

Độ bao phủ của hai câu có thể đo bằng số lượng từ trùng nhau giữa hai câu hoặc có thể chạy chung một hoặc nhiều bộ lọc ngữ nghĩa, cú pháp. Để giảm giá trị của độ tương đồng giữa các câu, tạo thuận lợi trong quá trình tính toán do tồn tại các câu dài thì TextRank sử

**Commented [N16]:** Giãn space dưới heading ra đểnhìnbài đọc thoáng và làm nổi tiêu đề của mục này. Xem lại tương tự cho các mục khác

38

dụng hệ số chuẩn hoá là chia số lượng nội dung bao phủ cho độ dài của từng câu.

Mỗi câu S của từng văn bản sẽ được biểu diễn dưới dạng *Si ={ W1i,W2i,…,WNii* *}* *với W1i* là từ thứ 1 của câu Si và *WNii* là từ thứ n của câu Si.

Độ tương đồng giữa hai câu *Si* *và Sj* theo thuật toán TextRank gốc được tính theo công thức:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | |{Wk|Wk ∈ Si ∧ | Wk ∈ Sj}| | |  |
| Similarity(Si, Sj) = |  |  | (3) |  |
|  |  |  |

log(|Si|)+log(| |)

Trong đó:

* *Similarity(S****i****, S****j****):* là độ tương đồng giữa câu *S****i***và câu *S****j***
* *w****k***: từ thuộc cả câu *S****i***và câu *S****j***
* *log(|S***i***|) +* l o g ( | S**j**| ) :hệ số chuẩn hóa

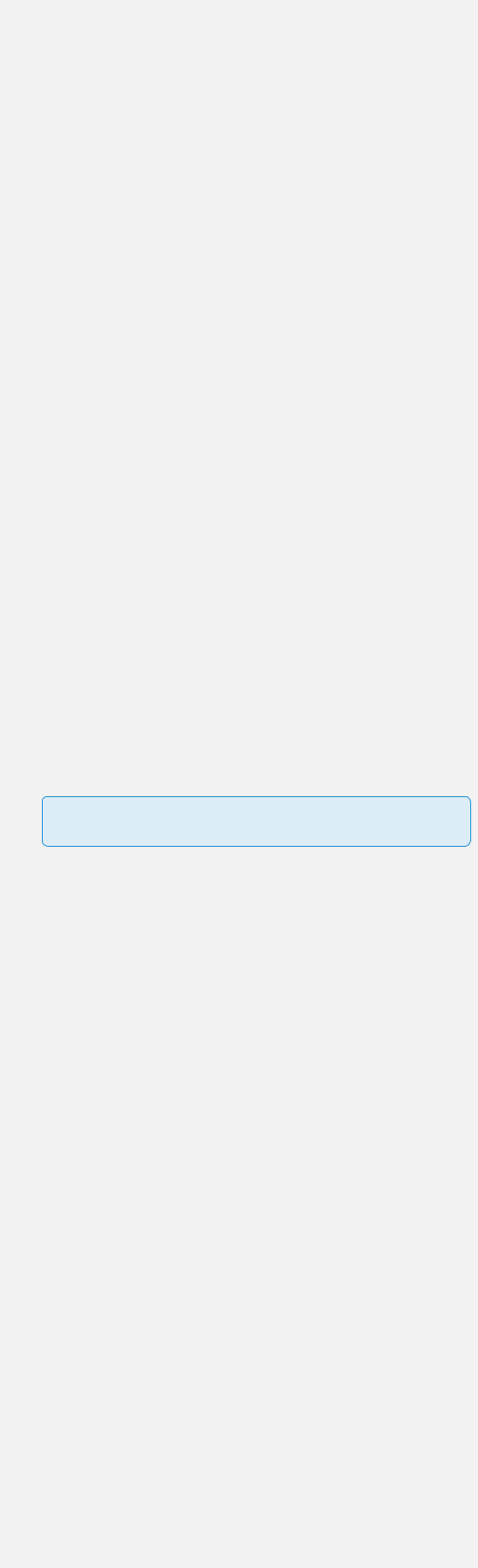


Có nhiều độ đo độ tương tự giữa các câu với nhau như chuỗi nhân, cô-sin, Euclidean  chuỗi con chung dài nhất, ... đều có thể được sử dụng. Trong luận văn ngoài công thức tính

độ tương tự giữa 2 câu của thuật toán TextRank gốc được đề cập trong công thức [3] chúng tôi còn tiến hành khảo sát thêm một số phương pháp tính độ tương tự giữa 2 câu để từ đó đánh giá thêm hiệu quả của hệ thống tóm tắt văn bản tiếng việt dựa trên mô hình đồ thị.

* Phương pháp tính độ tương tự giữa 2 câu dựa vào độ đo cosine: Trước tiên các câu sẽ được vecto hóa. Độ tương tự giữa 2 câu sẽ là độ đo Cosine của 2 vecto đại diện cho 2 câu.
* Phương pháp tính độ tương tự giữa 2 câu dựa vào độ đo Euclidean : Trước tiên các câu sẽ được vecto hóa. Độ tương tự giữa 2 câu sẽ là độ đo Euclidean của 2 vecto đại diện cho 2 câu.
* Phương pháp tính độ tương tự giữa 2 câu dựa vào số từ đơn giống nhau giữa 2 câu: Chúng ta sẽ tiến hành tách các câu thành các từ đơn, số lượng từ giống nhau giữa

**Commented [N17]:** Kiểm tra lại, chỗkhác em ghi làcosine?





39

các câu sẽ là độ tương tự giữa 2 câu.

Ví dụ:

S1: Tôi đi học.

S2: Tôi đi ăn cơm.

Similarity (S1,S2)=2

* Phương pháp tính độ tương tự giữa 2 câu dựa vào số lượng các cụm từ giống nhau giữa 2 câu: Chúng ta sẽ tiến hành xử lý để tách các cụm từ có ý nghĩa giống nhau trong tiếng việt giữa 2 câu, số lượng các cụm từ giống nhau giữa các câu và các từ đơn sẽ là độ tương tự giữa 2 câu.

Ví dụ:

S1: Tuy\_nhiên, trên thực\_tế , đột\_quỵ thiếu máu não là dạng đột\_quỵ thường gặp

hơn.

S2: Cứ 3 bệnh\_nhân đột\_quỵ sẽ có 2 người đột\_quỵ thiếu máu não .

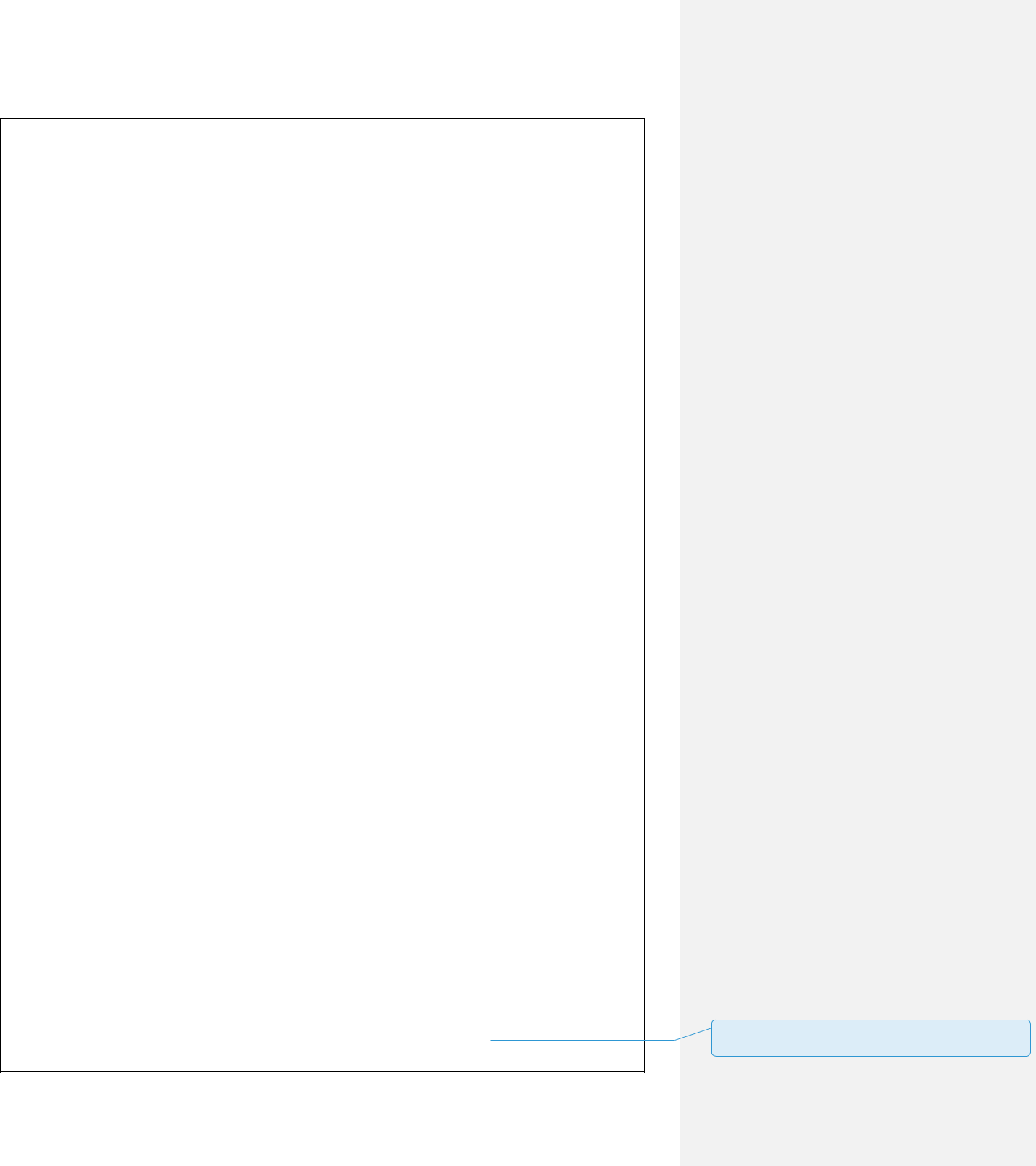
Similarity (S1,S2)=5

Dưới đây là một vài văn bản mẫu được đưa vào thử nghiệm:



1. 100 triệu đồng hỗ trợ SV xuất sắc đến Quảng Ngãi làm việc
2. Hội đồng nhân dân tỉnh Quảng Ngãi vừa ban hành chính sách thu hút, ưu đãi





40

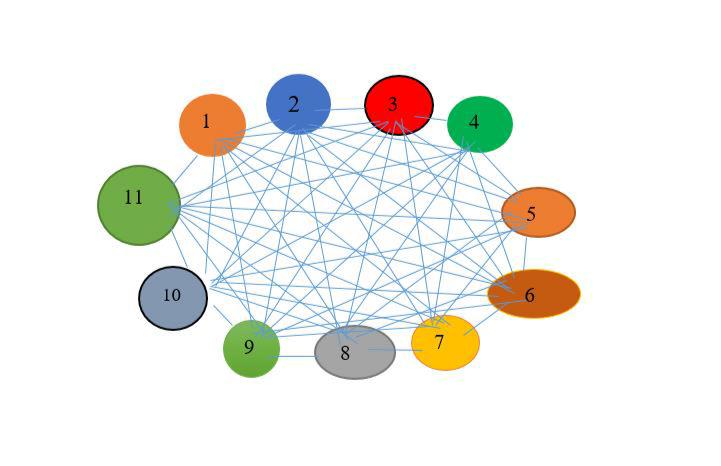
nguồn nhân lực chất lượng cao đến công tác, làm việc tại tỉnh Quảng Ngãi.

1. Theo đó, đối tượng thu hút là những người tốt nghiệp đại học chính quy, trường công lập có uy tín, thạc sĩ trong và ngoài nước đạt loại giỏi, xuất sắc ở các trường có uy tín trong và ngoài nước.
2. Những nhà khoa học, người tài đang công tác trong và ngoài nước đóng góp cho tỉnh bằng những đề tài khoa học hoặc sáng kiến để phát triển kinh tế - xã hội được đánh giá có hiệu quả cao.
3. Các đối tượng thu hút phải cam kết công tác tại các cơ quan, đơn vị thuộc tỉnh từ 10 năm trở lên.
4. Về độ tuổi, Giáo sư, phó giáo sư, tiến sĩ, bác sĩ chuyên khoa II, dược sĩ chuyên khoa II: không quá 50 tuổi đối với nam, 45 tuổi đối với nữ.
5. Thạc sĩ, bác sĩ chuyên khoa I, dược sĩ chuyên khoa I, bác sĩ nội trú và bác sĩ, dược sĩ đại học: không quá 40 tuổi đổi với nam, không quá 35 tuổi đối với nữ.
6. Những người tốt nghiệp đại học không quá 28 tuổi.
7. Được hỗ trợ kinh phí một lần khi nhận nhiệm vụ ở các huyện đồng bằng, thành phố, với mức cụ thể như sau: Giáo sư: 350 triệu đồng/người; phó giáo sư, tiến sĩ: 300 triệu đồng/người; bác sĩ chuyên khoa II, thạc sĩ y khoa, bác sĩ nội trú: 250 triệu đồng/người; bác sĩ chuyên khoa I: 230 triệu đồng/người; bác sĩ đa khoa (tốt nghiệp loại giỏi, xuất sắc): 200 triệu đồng/người; bác sĩ đa khoa (tốt nghiệp loại khá): 150 triệu đồng/người; người tốt nghiệp đại học loại xuất sắc: 100 triệu đồng/người.
8. Nếu nhận nhiệm vụ ở các huyện miền núi trong tỉnh và huyện Lý Sơn thì được hỗ trợ kinh phí một lần với mức 1,5 lần so với mức hỗ trợ tương ứng nói trên.
9. Ngoài ra, đối tượng thu hút là giáo sư, phó giáo sư, tiến sĩ thì được giao đất ở

|  |  |
| --- | --- |
| có thu tiền sử dụng đất theo giá sàn và diện tích giao tối đa 100m2. | **Commented [N18]:** Bảng này, kéo lên từ đầu văn bản cho |
|  | trang giấy đẹp về mặt trình bày |



41



*Hình 3.2: Đồ thị mô phỏng kết nối giữa các cặp câu trong văn bản.*

Kết quả tóm tắt tự động với độ dài 4 câu

1. Theo đó, đối tượng thu hút là những người tốt nghiệp đại học chính quy, trường công lập có uy tín, thạc sĩ trong và ngoài nước đạt loại giỏi, xuất sắc ở các trường có uy

tín trong và ngoài nước.

1. Được hỗ trợ kinh phí một lần khi nhận nhiệm vụ ở các huyện đồng bằng, thành phố, với mức cụ thể như sau: Giáo sư: 350 triệu đồng/người; phó giáo sư, tiến sĩ: 300

triệu đồng/người; bác sĩ chuyên khoa II, thạc sĩ y khoa, bác sĩ nội trú: 250 triệu

đồng/người; bác sĩ chuyên khoa I: 230 triệu đồng/người; bác sĩ đa khoa (tốt nghiệp loại

giỏi, xuất sắc): 200 triệu đồng/người; bác sĩ đa khoa (tốt nghiệp loại khá): 150 triệu

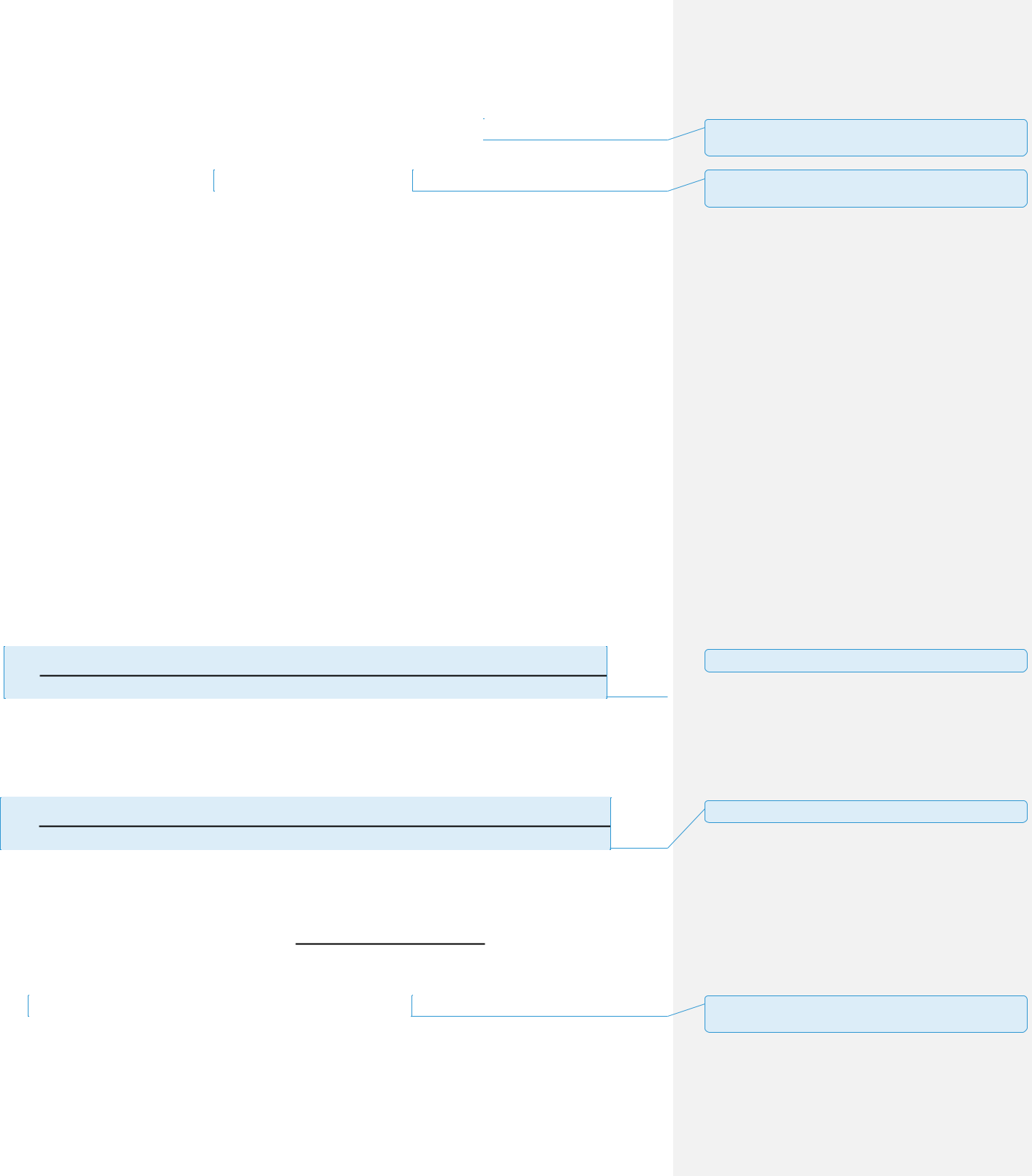
đồng/người; người tốt nghiệp đại học loại xuất sắc: 100 triệu đồng/người.

11. Ngoài ra, đối tượng thu hút là giáo sư, phó giáo sư, tiến sĩ thì được giao đất ở có thu tiền sử dụng đất theo giá sàn và diện tích giao tối đa 100m2.



42

Trong đồ thị ở hình 3.2, trọng số trên các cạnh là chỉ số đo mức độ kết nối giữa các cặp câu với nhau. Đồ thị thể hiện các câu có kết nối với nhau như thế nào. Câu nào có càng nhiều câu khác kết nối với nó thì nó có trọng số càng lớn. Điểm TextRank được tính trên mỗi đỉnh trong đồ thị được tạo ra. Từ đó, các câu có điểm cao nhất sẽ được lựa chọn để tạo thành văn bản tóm tắt. Việc lựa chọn số câu được lấy ra phụ thuộc vào yêu cầu của độ dài văn bản tóm tắt. Trong ví dụ trên, với yêu cầu độ dài văn bản tóm tắt là 3 câu nên các câu số 3, 9, 11 được lựa chọn để tạo thành văn bản tóm tắt. Với 3 câu được trích xuất làm văn bản tóm tắt thì tỉ lệ nén (*CompressionRate*) = 3/11. Ở trong luận văn chỉ sử dụng đến tỉ lệ nén để đánh giá mức độ rút gọn của văn bản tóm tắt so với văn bản gốc.



43

 **THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ**

**4.1. Dữ liệu thực nghiệm và phương pháp đánh giá**

Dữ liệu thực nghiệm chúng tôi sử dụng trong luận văn là các bài báo thuộc tất cả các lĩnh vực được lấy trên trang báo điện tử Tuổi Trẻ Online (http://www.tuoitre.vn). Tập dữ liệu bao gồm 5144 bài báo. Độ dài văn bản tóm tắt được giới hạn là 2 câu. Độ dài này tương đương với độ dài bản tóm tắt được sinh ra lấy từ phần tiêu đề của mỗi bài báo. Dữ liệu được đánh giá bằng phương pháp ROUGE với các tham số:

* Đánh giá toàn bộ văn bản trong mỗi bộ dữ liệu
* Sử đánh giá dựa vào n-gram độ dài 1 từ
* Khoảng tin cậy 95%
* Bao gồm cả từ dừng trong đánh giá
* Kết quả đánh giá cuối cùng là kết quả trung bình của toàn bộ tập dữ liệu
  + Điểm đánh giá được tính theo Recall, Precision, F-score.
* **Recall:** là tỉ lệ các tài liệu có liên quan đến tài liệu truy vấn và trên thực tếđược trích xuất trả về chính bằng ROUGE với n=1.

**Commented [N19]:** Mỗi chương bắt đầuở đầu trang giấy.Tương ứng như là chương trong sách

**Commented [N20]:** Nhớcập nhật Mục lục vì cô thay đổitên đề mục

= ổ ố ừ ố ℎ ữ ó ắ ự độ à ó ắ ủ ườ ố ừ ă ả ó ắ ủ ườ

* **Precision**: là tỉ lệ các tài liệu được trả về thực sự có liên quan đến tài liệutruy vấn.

= ổ ố ừ ố ℎ ữ ó ắ ự độ à ó ắ ủ ườ ố ừ ủ ℎệ ℎố ó ắ ự độ

* **F-score**: một độ đo để cân đối tỉ lệ giữa recall và precision.

 **Commented [N21]:** Sửa thành tiếng Việt luôn đi

**Commented [N22]:** Sao ko dùng tiếng Việt?

=

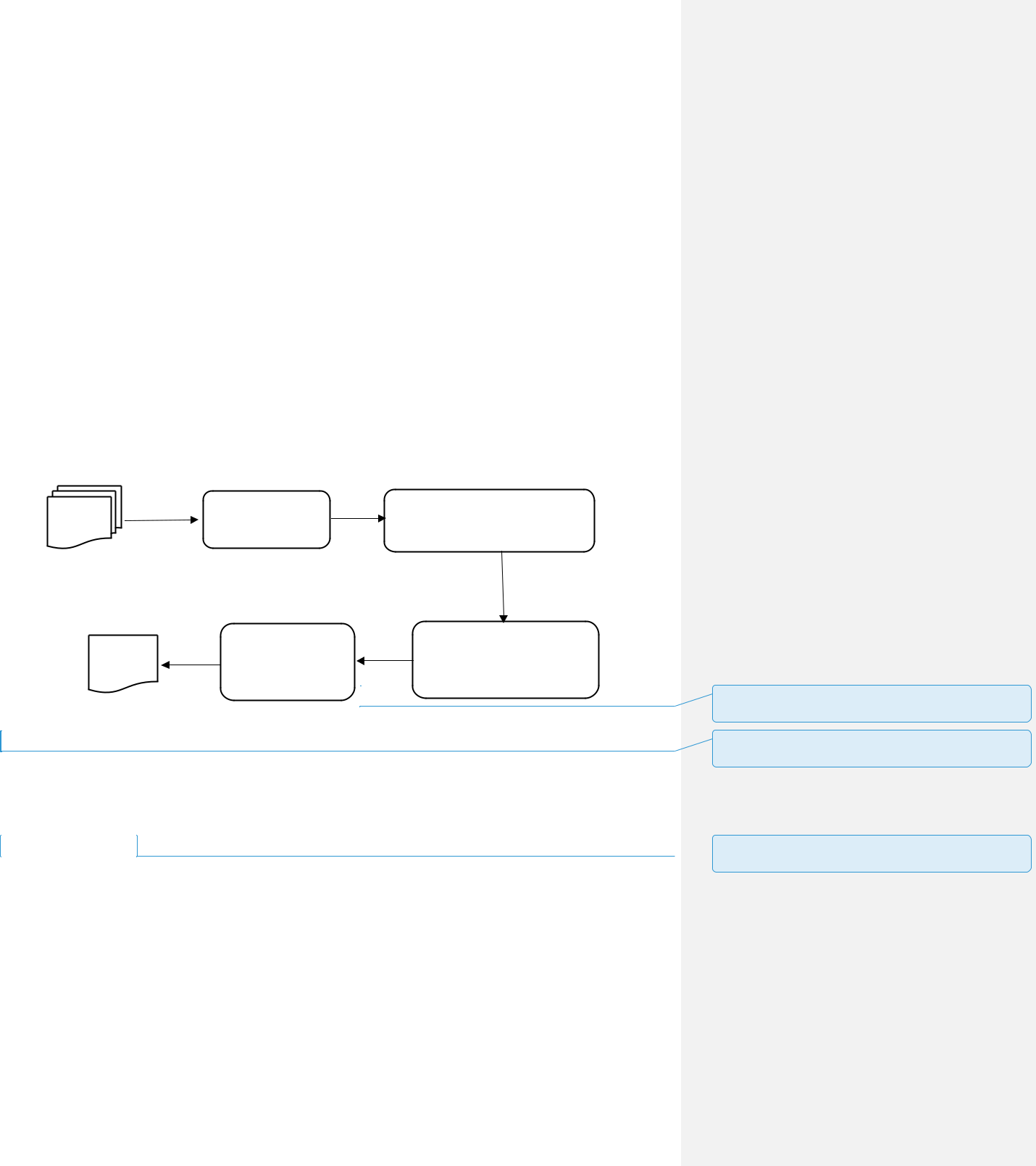
×

( + )/2

**4.2. Xây dựng hệ thống tóm tắt văn bản Tiếng Việt**

Như đã trình bày ở Chương 3, TextRank không bị ảnh hưởng bởi các yếu tố như: cấu

**Commented [N23]:** Phần này nên đặtở Chương 4.Saumục 4.1. Nhớ sửa lại mục từ 4.2 hiện thời lại cho phù hợp



44

trúc văn bản, các thành phần của văn phạm, ngôn ngữ. Chính vì lý do đó mà chúng tôi lựa chọn thuật toán này cho bài toán tóm tắt văn bản Tiếng Việt.

Để đánh giá và chứng minh độ hiệu quả của thuật toán TextRank khi áp dụng vào bài toán tóm tắt văn bản Tiếng việt chúng tôi sẽ tiến hành xây dựng các hệ thống tóm tắt văn bản dựa trên thuật toán TextRank và đánh giá khảo sát chúng để tìm ra hệ thống tối ưu nhất. Cụ thể là trong thuật toán TextRank chúng tôi sẽ lần lượt thay đổi các phương pháp tính độ tương tự giữa 2 câu, với mỗi lần thay đổi đó chúng tôi sẽ ghi nhận lại kết quả và đưa ra đánh giá.

Mô hình tóm tắt văn bản Tiếng Việc dựa trên thuật toán TextRank được xây dựng như hình bên dưới.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Tiền xử lý | Xây dựng đồ thị văn bản |  |
|  | văn bản |  |
|  |  |  |
| Văn bản gốc | |  |  |
|  | Trích rút các | Xếp hạng các đỉnh |  |
|  | câu quan trọng | trong đồ thị |  |
|  |  | **Commented [N24]:** Sửa lại hình vẽ, chữ bé quá đọc ko đc. |  |
|  |  |  |
| Văn bản tóm tắt |  |  |
|  | Hình phía dưới giống y hình này, em kiểm tra lại nội dung |  |
|  |  | **Commented [N25R24]:** Vẫn còn chữbé, em dò lại cho |  |
|  |  |  |
|  |  | hết |  |

*Hình 4.1. Mô hình tóm tắt văn bản sử dụng phương pháp đồ thị.*

Theo hình 4.1 Mô hình tóm tắt văn bản Tiếng Việc dựa trên thuật toán TextRank và Cosine/ Euclidean được xây dựng theo các bước chính như sau. 

Bước 1: Từ văn bản gốc chúng ta sẽ trải qua bước tiền xử lý văn bản. Tại bước này chúng ta tiến hành loại bỏ các hư từ, các từ không có ý nghĩa trong câu rồi tiến hành tách câu, chuyển đổi câu thành các vector.

Bước 2: Sau khi đã tách được các câu chúng ta tiến hành bước xây dựng đồ thị văn bản. Chúng ta sẽ coi mỗi câu của văn bản là một đỉnh của đồ thị. Chúng ta xây dựng các cạnh của đồ thị bằng việc tính độ tương tự giữa các câu của văn bản theo các phương pháp

**Commented [N26]:** Phía trên thì ghi Euclidean,ở đây thìghi Euclidean. Em kiểm tra lại toàn văn cho thống nhất.



45

được đề cập trong mục 3.2 thuộc chương 3.

Bước 3: Sau khi đã xây dựng xong đồ thị chúng ta tiến hành xếp hạng các đỉnh của đồ thị và xếp hạng chúng theo thứ tự từ cao xuống thấp.

Bước 4: Trích xuất các câu theo thứ tự từ cao xuống thấp theo một tỉ lệ định trước.

Bước 5: Xây dựng văn bản tóm tắt.

Thử nghiệm với văn bản mẫu:

300 cơ hội tìm việc với doanh nghiệp nước ngoài

Ông Vincent Floreani - tổng lãnh sự Cộng hòa Pháp tại TP.HCM, giới thiệu về Ngày hội việc làm Pháp - Việt

Trong cuộc họp báo giới thiệu về hoạt động này, ông Vincent Floreani - tổng lãnh sự Cộng hòa Pháp tại TP.HCM, cho biết có khoảng 50 doanh nghiệp Pháp và Việt Nam hoạt động trong nhiều lĩnh vực đa dạng như thương mại, phân phối, kế toán và kiểm toán, truyền thông, giáo dục, tin học… tham gia.

Ngày hội Việc làm Pháp - Việt được Đại sứ quán Pháp tại Việt Nam và Phòng Thương mại và công nghiệp Pháp tại VN (CCIFV) tổ chức lần đầu tiên tại Hà Nội vào tháng 11-2014, tại TP.HCM vào tháng 4-2015 và tại cả hai thành phố vào năm 2016.

Nhận thấy hiệu quả của hoạt động kết nối doanh nghiệp và người tìm việc, cơ quan đại diện ngoại giao của Pháp tiếp tục hỗ trợ cho việc tìm kiếm cơ hội làm việc cho những người trẻ VN tại các công ty nước ngoài hoạt động ở VN.

Chương trình hoạt động ngày 12-5 tại khách sạn Novotel chỉ kéo dài trong 5 giờ nhưng được giới thiệu sẽ phong phú với không chỉ các cuộc phỏng vấn trực tiếp giữa nhà tuyển dụng với ứng viên mà còn có các buổi hội thảo hướng dẫn về phương pháp dự tuyển cho các ứng viên, đồng thời với các chương trình giao lưu cho phép công ty tuyển dụng giới thiệu hoạt động và chiến lược nhân sự.

Giám đốc công ty phần mềm Capgemini (trái) cho biết làm việc trong công ty nước ngoài có yêu cầu ngoại ngữ càng giỏi càng tốt



46

Các hội thảo chuyên đề sẽ được diễn ra song song trong suốt sự kiện nhằm mục đích giúp phát triển khả năng ứng xử và kinh nghiệm cho các ứng cử viên tham gia ứng tuyển

Đây là một hình thức lý tưởng cho việc trao đổi kinh nghiệm và tư vấn về các bước chính trong tiếp cận việc làm.

Ông Guillaume Crouzet - giám đốc CCIFV, khẳng định số doanh nghiệp nước ngoài làm ăn tại VN cũng như số doanh nghiệp VN tham gia ngày hội này cùng số cơ hội việc làm tăng lên trong năm nay cho thấy hiệu quả của việc tổ chức và sự tin cậy của các doanh nghiệp.

Ngày hội mở cửa miễn phí từ 8h sáng 12-5

Tại sao không thử một lần nhỉ?

TTO - Tốt nghiệp đại học ở Việt Nam, nộp đơn xin việc ở các nước trong cộng đồng kinh tế ASEAN (AEC) đang là lựa chọn của không ít cử nhân Việt.

Văn bản đã tách câu và tách từ:

1. 300 cơ\_hội tìm việc với doanh\_nghiệp nước\_ngoài
2. Ông Vincent\_Floreani - tổng\_lãnh\_sự Cộng\_hoà Pháp tại TP. HCM , giới\_thiệu về Ngày hội việc\_làm Pháp - Việt.
3. Trong cuộc họp\_báo giới\_thiệu về hoạt\_động này , ông Vincent\_Floreani -

tổng\_lãnh\_sự Cộng\_hoà Pháp tại TP. HCM , cho\_biết có khoảng 50 doanh\_nghiệp Pháp và Việt\_Nam hoạt\_động trong nhiều lĩnh\_vực đa\_dạng như thương\_mại , phân\_phối , kế\_toán và kiểm\_toán , truyền\_thông , giáo\_dục , tin\_học … tham\_gia .

* 1. Ngày hội Việc\_làm Pháp - Việt được Đại\_sứ\_quán Pháp tại Việt\_Nam và Phòng Thương\_mại và công\_nghiệp Pháp tại VN ( CCIFV ) tổ\_chức lần đầu\_tiên tại Hà\_Nội vào tháng 11-2014 , tại TP. HCM vào tháng 4-2015 và tại cả hai thành\_phố vào năm 2016 .
  2. Nhận thấy hiệu\_quả của hoạt\_động kết\_nối doanh\_nghiệp và người tìm việc
* cơ\_quan đại\_diện ngoại\_giao của Pháp tiếp\_tục hỗ\_trợ cho việc tìm\_kiếm cơ\_hội làm\_việc cho những người trẻ VN tại các công\_ty nước\_ngoài hoạt\_động ở VN .



47

1. Chương\_trình hoạt\_động ngày 12-5 tại khách\_sạn Novotel chỉ kéo\_dài trong 5 giờ nhưng được giới\_thiệu sẽ phong\_phú với không\_chỉ các cuộc phỏng\_vấn trực\_tiếp giữa nhà\_tuyển\_dụng với ứng\_viên mà\_còn có các buổi hội\_thảo hướng\_dẫn về phương\_pháp dự\_tuyển cho các ứng\_viên , đồng\_thời với các chương\_trình giao\_lưu cho\_phép công\_ty tuyển\_dụng giới\_thiệu hoạt\_động và chiến\_lược nhân\_sự .
2. Giám\_đốc công\_ty phần\_mềm Capgemini ( trái ) cho biết làm\_việc trong công\_ty nước\_ngoài có yêu\_cầu ngoại\_ngữ càng giỏi càng tốt .
3. Các hội\_thảo chuyên\_đề sẽ được diễn ra song\_song trong suốt sự\_kiện nhằm mục\_đích giúp phát\_triển khả\_năng ứng\_xử và kinh\_nghiệm cho các ứng\_cử\_viên tham\_gia ứng\_tuyển
4. Đây là một hình\_thức lý\_tưởng cho việc trao\_đổi kinh\_nghiệm và tư\_vấn về các bước chính trong tiếp\_cận việc\_làm .
5. Ông Guillaume\_Crouzet - giám\_đốc CCIFV , khẳng\_định số doanh\_nghiệp nước\_ngoài làm\_ăn tại VN cũng như số doanh\_nghiệp VN tham\_gia ngày hội này cùng số cơ\_hội việc\_làm tăng lên trong năm nay cho thấy hiệu\_quả của việc tổ\_chức và sự tin\_cậy của các doanh\_nghiệp .
6. Ngày hội mở\_cửa miễn\_phí từ 8h sáng 12-5
7. Tại\_sao không thử một lần nhỉ ?
8. TTO - Tốt\_nghiệp đại\_học ở Việt\_Nam , nộp đơn xin việc ở các nước trong cộng\_đồng kinh\_tế ASEAN ( AEC ) đang là lựa\_chọn của không ít cử\_nhân Việt .

Mỗi câu sau khi được vecto hóa có dạng

X={x1, x2,…, xn}

Số lượng chiều của vector sẽ là tùy biến phụ thuộc vào miền dữ liệu mà chúng ta có thể lựa chọn tham số n một cách hợp lý.

Từ văn bản mẫu ta tiến hành trích xuất câu để hợp thành văn bản tóm tắt. Ở đây chúng tôi lấy độ dài văn bản tóm tắt là 4 câu ( tỉ lệ nén =0.36). Ta có kết quả như sau:



48

**Tóm tắt bằng hệ thống sử dụng thuật toán TextRank gốc:**

“'Trong cuộc họp báo giới thiệu về hoạt động này, ông Vincent Floreani - tổng lãnh sự Cộng hòa Pháp tại TP.HCM, cho biết có khoảng 50 doanh nghiệp Pháp và Việt Nam hoạt động trong nhiều lĩnh vực đa dạng như thương mại, phân phối, kế toán và kiểm toán, truyền thông, giáo dục, tin học… tham gia.',

'Nhận thấy hiệu quả của hoạt động kết nối doanh nghiệp và người tìm việc, cơ quan đại diện ngoại giao của Pháp tiếp tục hỗ trợ cho việc tìm kiếm cơ hội làm việc cho những người trẻ VN tại các công ty nước ngoài hoạt động ở VN.',

'Chương trình hoạt động ngày 12-5 tại khách sạn Novotel chỉ kéo dài trong 5 giờ nhưng được giới thiệu sẽ phong phú với không chỉ các cuộc phỏng vấn trực tiếp giữa nhà tuyển dụng với ứng viên mà còn có các buổi hội thảo hướng dẫn về phương pháp dự tuyển cho các ứng viên, đồng thời với các chương trình giao lưu cho phép công ty tuyển dụng giới thiệu hoạt động và chiến lược nhân sự.',

'Ông Guillaume Crouzet - giám đốc CCIFV, khẳng định số doanh nghiệp nước ngoài làm ăn tại VN cũng như số doanh nghiệp VN tham gia ngày hội này cùng số cơ hội việc làm tăng lên trong năm nay cho thấy hiệu quả của việc tổ chức và sự tin cậy của các doanh nghiệp.'

**Tóm tắt bằng hệ thống sử dụng thuật toán TextRank kết hợp với phương pháp tính độ tương tự giữa 2 câu bằng Cosine:**

“'Các hội thảo chuyên đề sẽ được diễn ra song song trong suốt sự kiện nhằm mục đích giúp phát triển khả năng ứng xử và kinh nghiệm cho các ứng cử viên tham gia ứng tuyển',

'Đây là một hình thức lý tưởng cho việc trao đổi kinh nghiệm và tư vấn về các bước chính trong tiếp cận việc làm.',

'Tại sao không thử một lần nhỉ?',

'TTO - Tốt nghiệp đại học ở Việt Nam, nộp đơn xin việc ở các nước trong cộng đồng kinh tế ASEAN (AEC) đang là lựa chọn của không ít cử nhân Việt.'”



49

**Tóm tắt bằng hệ thống sử dụng thuật toán TextRank kết hợp với phương pháp tính độ tương tự giữa 2 câu bằng Euclidean:**

“Ông Vincent Floreani - tổng lãnh sự Cộng hòa Pháp tại TP.HCM, giới thiệu về Ngày hội việc làm Pháp - Việt.

'Chương trình hoạt động ngày 12-5 tại khách sạn Novotel chỉ kéo dài trong 5 giờ nhưng được giới thiệu sẽ phong phú với không chỉ các cuộc phỏng vấn trực tiếp giữa nhà tuyển dụng với ứng viên mà còn có các buổi hội thảo hướng dẫn về phương pháp dự tuyển cho các ứng viên, đồng thời với các chương trình giao lưu cho phép công ty tuyển dụng giới thiệu hoạt động và chiến lược nhân sự.',

'Đây là một hình thức lý tưởng cho việc trao đổi kinh nghiệm và tư vấn về các bước chính trong tiếp cận việc làm.',

'Tại sao không thử một lần nhỉ?”

**Tóm tắt bằng hệ thống sử dụng thuật toán TextRank kết hợp với phương pháp tính độ tương tự giữa 2 câu dựa vào số từ đơn giống nhau giữa 2 câu:**

“'Trong cuộc họp báo giới thiệu về hoạt động này, ông Vincent Floreani - tổng lãnh sự Cộng hòa Pháp tại TP.HCM, cho biết có khoảng 50 doanh nghiệp Pháp và Việt Nam hoạt động trong nhiều lĩnh vực đa dạng như thương mại, phân phối, kế toán và kiểm toán, truyền thông, giáo dục, tin học… tham gia.',

'Nhận thấy hiệu quả của hoạt động kết nối doanh nghiệp và người tìm việc, cơ quan đại diện ngoại giao của Pháp tiếp tục hỗ trợ cho việc tìm kiếm cơ hội làm việc cho những người trẻ VN tại các công ty nước ngoài hoạt

động ở VN.',

'Chương trình hoạt động ngày 12-5 tại khách sạn Novotel chỉ kéo dài trong 5 giờ nhưng được giới thiệu sẽ phong phú với không chỉ các cuộc phỏng vấn trực tiếp giữa nhà tuyển dụng với ứng viên mà còn có các buổi hội thảo hướng dẫn về phương pháp dự tuyển cho



50

các ứng viên, đồng thời với các chương trình giao lưu cho phép công ty tuyển dụng giới thiệu hoạt động và chiến lược nhân sự.',

'Ông Guillaume Crouzet - giám đốc CCIFV, khẳng định số doanh nghiệp nước ngoài làm ăn tại VN cũng như số doanh nghiệp VN tham gia ngày hội này cùng số cơ hội việc làm tăng lên trong năm nay cho thấy hiệu quả của việc tổ chức và sự tin cậy của các doanh nghiệp.'”

**Tóm tắt bằng hệ thống sử dụng thuật toán TextRank kết hợp với** phương pháp tính độtương tự giữa 2 câu dựa vào số lượng các cụm từ giống nhau giữa 2 câu**:**

“'Trong cuộc họp\_báo giới\_thiệu về hoạt\_động này , ông Vincent\_Floreani - tổng\_lãnh\_sự Cộng\_hoà Pháp tại TP. HCM , cho\_biết có khoảng 50 doanh\_nghiệp Pháp và Việt\_Nam hoạt\_động trong nhiều lĩnh\_vực đa\_dạng như thương\_mại , phân\_phối , kế\_toán và kiểm\_toán , truyền\_thông , giáo\_dục , tin\_học … tham\_gia .

'Ngày hội Việc\_làm Pháp - Việt được Đại\_sứ\_quán Pháp tại Việt\_Nam và Phòng Thương\_mại và công\_nghiệp Pháp tại VN ( CCIFV ) tổ\_chức lần đầu\_tiên tại Hà\_Nội vào tháng 11-2014 , tại TP. HCM vào tháng 4-2015 và tại cả hai thành\_phố vào năm 2016 .\r',

'Nhận thấy hiệu\_quả của hoạt\_động kết\_nối doanh\_nghiệp và người tìm việc , cơ\_quan đại\_diện ngoại\_giao của Pháp tiếp\_tục hỗ\_trợ cho việc tìm\_kiếm cơ\_hội làm\_việc cho những người trẻ VN tại các công\_ty nước\_ngoài hoạt\_động ở VN .\r',

'Ông Guillaume\_Crouzet - giám\_đốc CCIFV , khẳng\_định số doanh\_nghiệp nước\_ngoài làm\_ăn tại VN cũng như số doanh\_nghiệp VN tham\_gia ngày hội này cùng số cơ\_hội việc\_làm tăng lên trong năm nay cho thấy hiệu\_quả của việc tổ\_chức và sự tin\_cậy của các doanh\_nghiệp .”



51

52

**Nhận xét:**

Khi đưa văn bản trên vào 5 hệ thống tóm tắt mà chúng ta xây dựng dựa trên thuật toán TextRank thì chúng ta thu được kết quả là có 3 hệ thống tóm tắt đưa ra văn bản tóm tắt với cùng một nội dung đó là TextRank gốc sử dụng công thức (3) trong chương 3, TextRank sử dụng phương pháp tính độ tương tự giữa 2 câu dựa vào số lượng các cụm từ giống nhau giữa 2 câu, TextRank sử dụng phương pháp tính độ tương tự giữa 2 câu dựa vào số lượng các từ giống nhau giữa 2 câu và hai hệ thống sử dụng độ đo Cosine và Euclidean cho kết quả khác nhau.



Qua đọc nội dung thì thấy được chất lượng bản tóm tắt sử dụng TextRank gốc



có chất lượng tốt hơn Cosine và Euclidean. Văn bản đọc lên thấy trôi chảy về mặt nội  dung, ý nghĩa. Tuy nhiên một số câu có mức độ thông tin phù hợp hơn lại không được



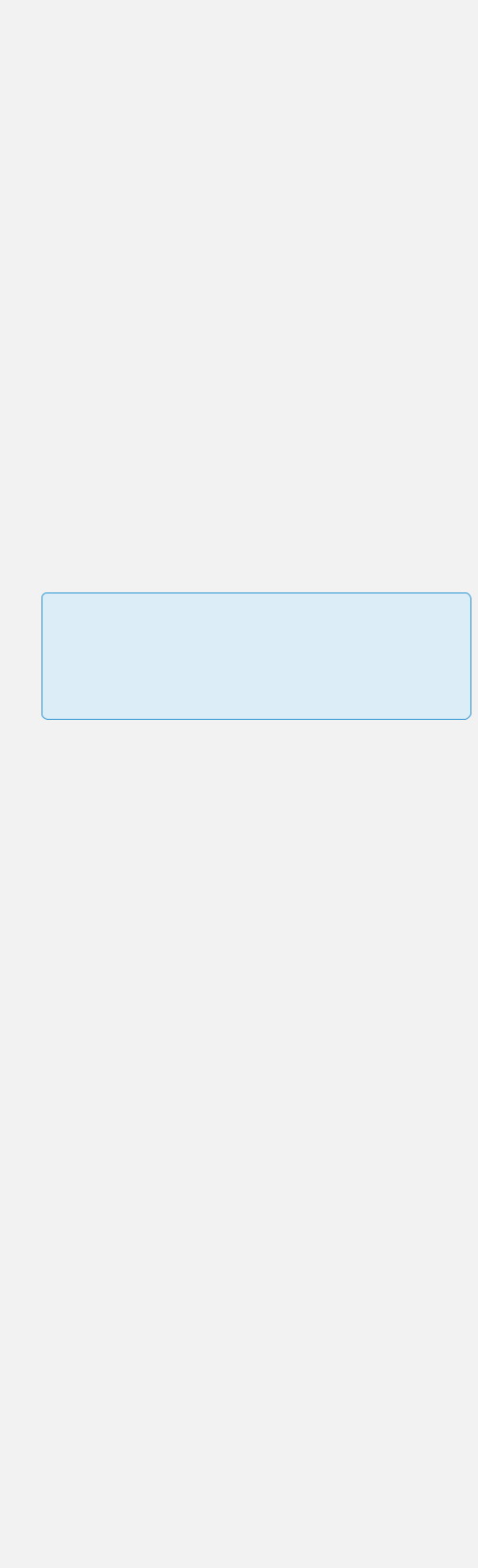
lựa chọn do độ quan trọng không cao. Do độ dài văn bản tóm tắt bị giới hạn ở mức 4 câu nên các câu có hàm lượng thông tin cao chưa được góp mặt.

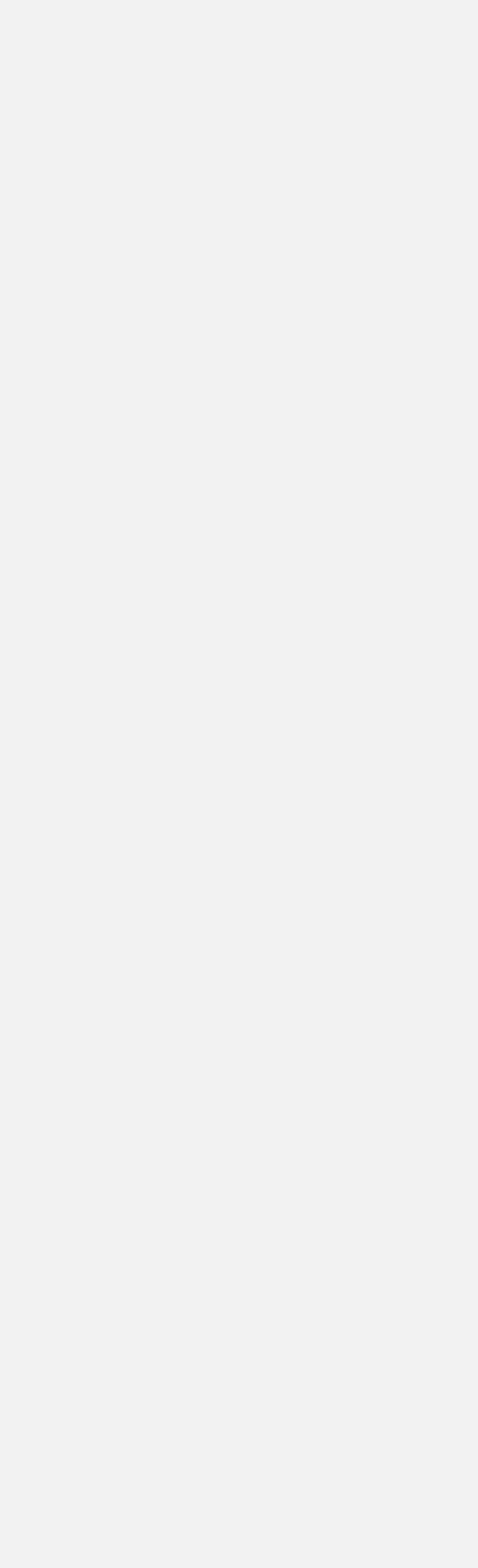
Các câu dài thường được ưu tiên lựa chọn để trích xuất do khi tính toán độ tương đồng thì khả năng các câu này có độ tương đồng cao hơn so với các câu khác. Đây cũng là một nhược điểm trong thuật toán TextRank. Điều này làm giảm đi một phần chất lượng của văn bản tóm tắt.

Bên cạnh đó, do là trích rút câu, các câu không được chỉnh sửa nên trong câu sẽ tồn tại từ nối, quan hệ, từ không mang nhiều ý nghĩa trong câu. Các từ này lại chiếm số lượng không nhỏ trong văn bản. Điều này cũng làm cho độ đo tương tự giảm một phần chính xác.

Mặc dù có một vài khuyết điểm trên, thuật toán TextRank vẫn cho kết quả tóm tắt ở mức độ tốt về mặt hình thức và nội dung. Người đọc hoàn toàn hiểu được nội dung của văn bản gốc trình bày vấn đề gì khi đọc văn bản tóm tắt tự động.

**Commented [N27]:** Phương pháp TextRank, CosinevàEuclidean là gì? Em cần nói rõ, là dùng TextRank gốc, tức là xác định cạnh/mối liên hệ giữa 2 câu bằng cách đếm số từ chung. Còn Cosine là cách dùng TextRank trong đó mối liên hệ giữa 2 câu đc xác định bằng cách dùng Cosine để tính độ tương tự giữa 2 câu. Tương tự với Euclidean ….





53

**4.3. Thực nghiệm và đánh giá kết quả**

Các dữ liệu thực nghiệm được thực hiện dựa trên thuật toán TextRank kết hợp cùng các phương pháp khác nhau để tính độ tương tự giữa 2 câu trong một văn bản.

* TextRank sử dụng công thức thứ (3) được đề cập trong mục 3.2 chương 3 để tính toán độ tương tự giữa 2 câu chúng ta gọi là hệ thống S1.
* TextRank sử dụng độ đo Cosine để tính độ tương tự giữa 2 câu chúng ta gọi là hệ thống S2.
* TextRank sử dụng độ đo Euclidean để tính độ tương tự giữa 2 câu chúng ta gọi là hệ thống S3.
* TextRank sử dụng phương pháp tính độ tương tự giữa 2 câu dựa vào số lượng các từ giống nhau giữa 2 câu chúng ta gọi là hệ thống S4.
* TextRank sử dụng phương pháp tính độ tương tự giữa 2 câu dựa vào số lượng các cụm từ giống nhau giữa 2 câu chúng ta gọi là hệ thống S5.

|  |  |  |  |
| --- | --- | --- | --- |
| **Hệ thống** | **Recall** | **Precision** | **F-score** |
|  |  |  |  |
| S1 | 0.63 | 0.26 | 0.36 |
|  |  |  |  |
| S2 | 0.3 | 0.22 | 0.25 |
|  |  |  |  |
| S3 | 0.31 | 0.21 | 0.25 |
|  |  |  |  |
| S4 | 0.64 | 0.26 | 0.37 |
|  |  |  |  |
| S5 | 0.35 | 0.16 | 0.22 |
|  |  |  |  |

Bảng 4.1. Kết quả thực nghiệm.

54

**Nhận xét:**

Độ tương tự cosine và Euclidean cho kết quả không được cao như kỳ vọng khi sử dụng Rouge để đánh giá. Điều này cũng dễ hiểu bởi vì các đoạn văn bản tóm tắt do con người tạo ra dùng để so sánh trong tập dữ liệu là các đoạn tóm tắt theo dạng ngữ nghĩa không phải là dạng trích rút câu như phương pháp chúng tôi đang nghiên cứu cụ thể như một đoạn tóm tắt mẫu sau:

Ngày hội việc làm Pháp - Việt 2018 tại TP.HCM khai mạc sáng 12-5 tại khách sạn Novotel Saigon Center (Q.3) với khoảng 50 doanh nghiệp Pháp và VN giới thiệu 300 cơ hội việc làm.

Theo kết quả đánh giá ở trên, TextRank gốc và hệ thống TextRank sử dụng phương pháp tính độ tương tự giữa 2 câu dựa vào tổng số từ giống nhau giữa 2 câu đạt được kết quả cao nhất do đặc điểm của chúng là sẽ ưu tiên đánh giá xếp hạng các câu có nội dung dài trong văn bản bởi vậy số lượng các từ giống nhau giữa văn bản tóm tắt của hệ thống và của con người sẽ nhiều hơn kéo theo tỉ lệ sẽ tốt hơn nhiều. Ngoài ra chúng tôi sử dụng doc2vec để biểu diễn câu thành vecto mà phương pháp này chỉ đạt hiểu quả cao khi không gian dữ liệu thực nghiệm là lớn nên với tập dữ liệu là 5144 văn bản có lẽ chưa tối ưu được thuật toán này dẫn tới các giá trị các chiều của vecto rất nhỏ từ đó ảnh hưởng trực tiếp đến độ tương đồng giữa hai câu ( nhỏ hơn



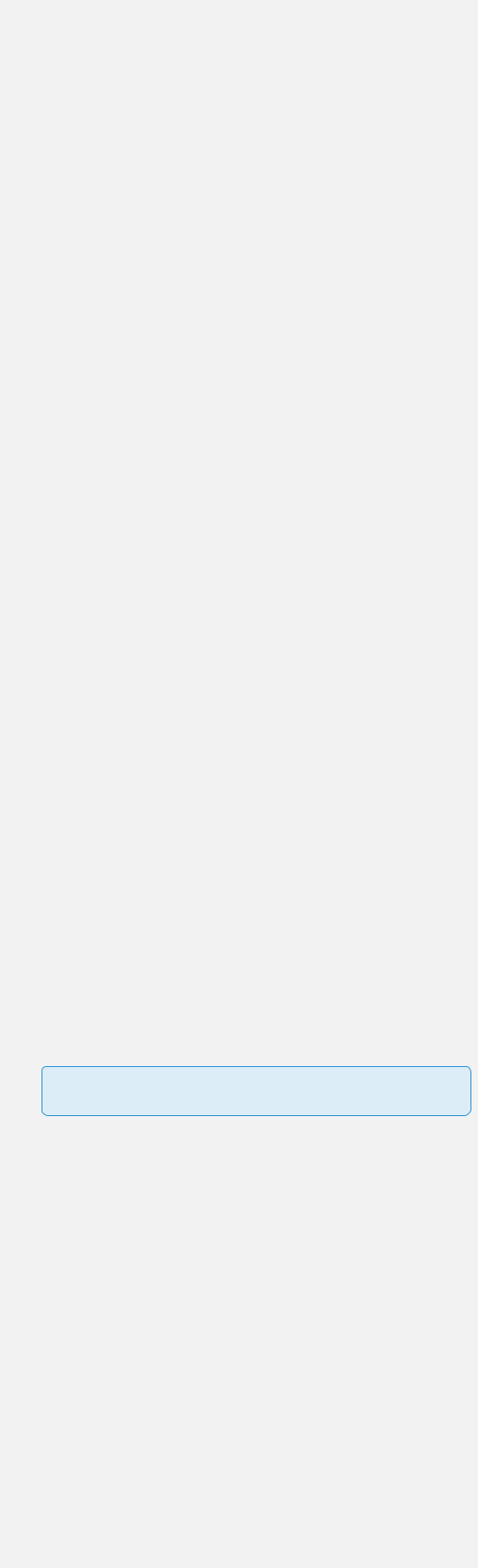
so với TextRank với Recall thu được là 0.3). Tuy nhiên các kết quả tóm tắt được sinh  ra khi chúng tôi sử dụng độ đo Cosine và Euclidean cũng rất khả quan trên một vài khía cạnh như là:



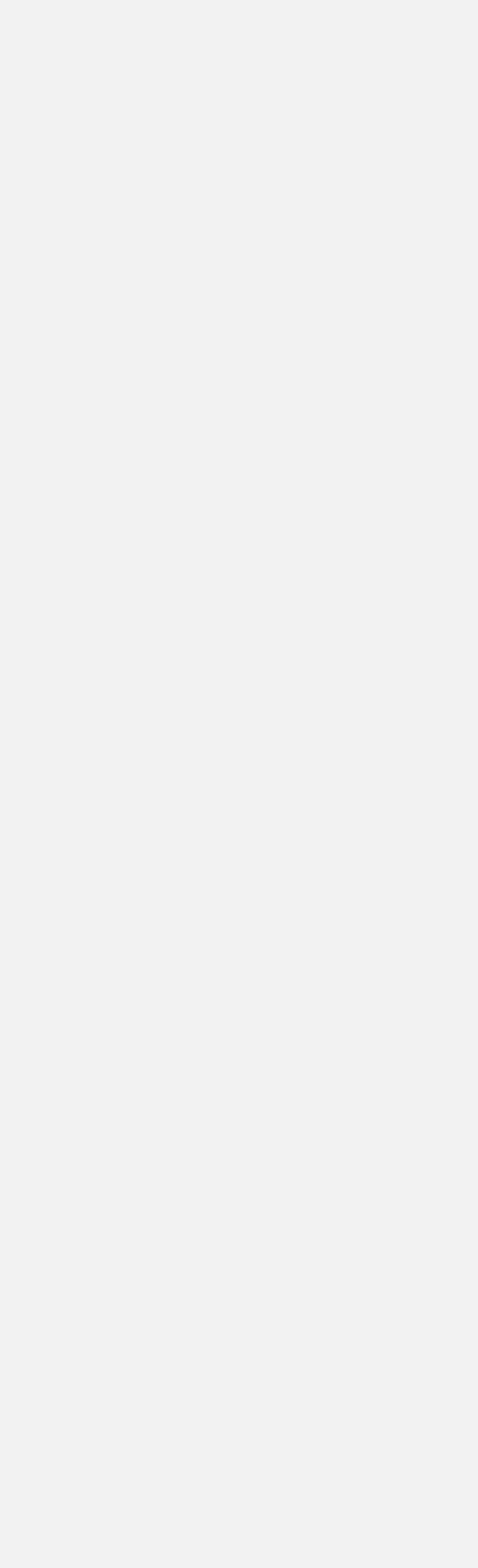
* Độ dài của tóm tắt ngắn hơn TextRank
* Vẫn cung cấp đầy đủ ý nghĩa thông tin của đoạn văn cho người

đọc.

**Commented [N28]:** Viết luôn kết quảkhi dùng doc2vec



1. chiều vào đây.



55

* Tốc độ tóm tắt nhanh áp dụng được cho nhiều miền ứng dụng Một số văn bản tóm tắt sinh ra bởi nghiên cứu:

**Văn bản ban đầu:**

“Yêu khoa học từ những nhọc nhằn

Đó là chiếc máy tuốt hạt bắp để nông dân đỡ nhọc nhằn, là máy xới cày liếp của cô con gái dành cho người cha, hay thiết bị tách lọc nước biển thành nước ngọt cho người dân miền biển...

Đó là những câu chuyện tiêu biểu trong 180 dự án do 306 học sinh của 31 tỉnh thành từ Đà Nẵng trở vào thực hiện tham gia Hội thi khoa học kỹ thuật cấp quốc gia dành cho học sinh trung học khu vực phía Nam diễn ra ở Đồng Tháp trong hai ngày 16 và 17-3.

Cho bớt nhọc nhằn

Chiếc máy tuốt hạt bắp của hai cô cậu học trò đến từ Trường THCS Đào Duy Từ (Bình Định) thỉnh thoảng vang tiếng lèng xèng theo nhịp giải thích của hai người sáng chế nhỏ tuổi mỗi lần có khách ghé thăm.

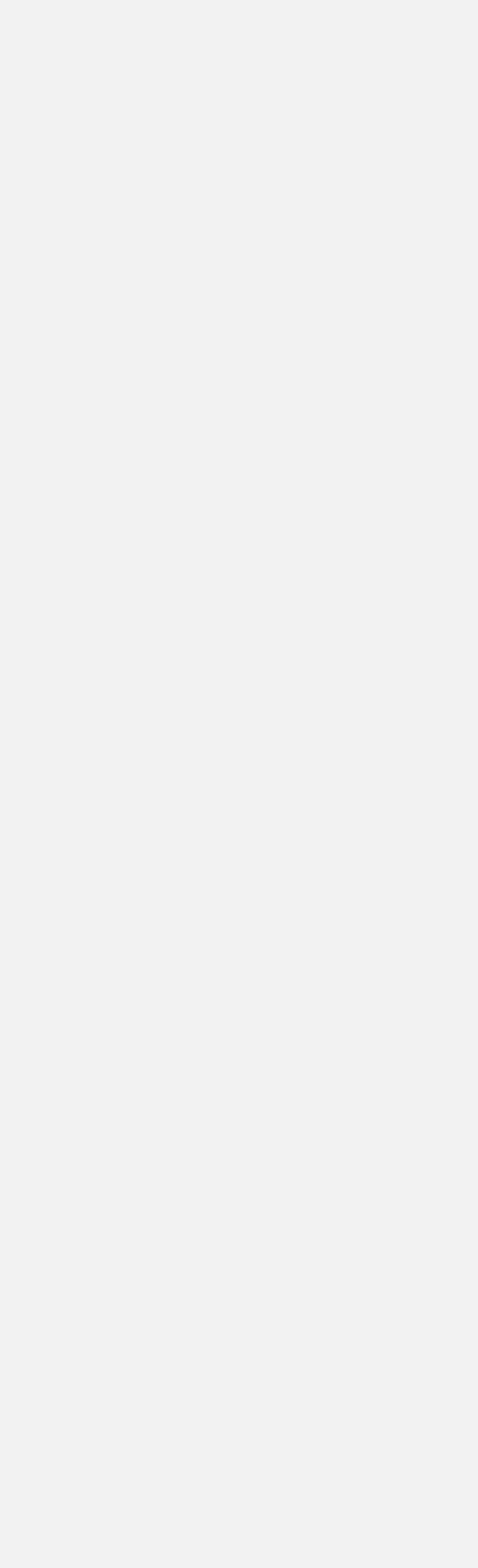
“Quê em trồng nhiều bắp lắm, bà con đa số tuốt bắp bằng tay

Có chỗ cũng có máy tuốt nhưng hạt nát vụn nhiều, hiệu quả rất thấp lại không thể dùng hạt để làm giống, em đánh bạo chế tạo máy tuốt này trước dùng cho nương bắp nhà em, sau là cho nông dân quê em đỡ cực” - Huỳnh Đức Phú, học sinh thực hiện dự án, chia sẻ.

Máy cùng lúc tuốt được nhiều trái, hạn chế hạt bắp văng tung tóe, cùi bắp không bị cắt nát, có thể tận dụng để làm chất đốt

Say sưa nghe Phú thuyết trình nguyên lý hoạt động của máy, thầy Nguyễn Đình Phùng, hiệu trưởng Trường THPT Chu Văn An (An Giang), hết lời khen ngợi.

“Dự án rất gần gũi và thiết thực với cuộc sống



56

Nếu máy này đưa vào thực tiễn tôi nghĩ có nhiều dự án xã hội sẽ ủng hộ” - thầy Phùng chia sẻ.

Võ Thị Thảo Lil, Trường THPT Giồng Thị Đam (Đồng Tháp), chia sẻ: “Là con gái không thể giúp cha cày cuốc như con trai nên em mày mò thiết kế máy này, vừa xới đất được lại tích hợp cày đất thành từng liếp cho cha đỡ cực”.

Cha Lil có hai công đất làm hoa màu nhưng thị trường chưa có máy móc phục vụ chuyên biệt cho công việc trồng màu

Phần lớn nông dân phải cày cuốc bằng tay vừa tốn công vừa tốn thời gian

Để đào rãnh lên liếp trồng hoa màu trung bình mười ngày cha Lil mới làm được 1.000m2.

Nhọc nhằn là vậy nhưng Lil lại không thể phụ giúp cha

Nghĩ vậy nên em quyết định thực hiện đề tài vốn chẳng dễ dàng với một đứa con gái.

“Cũng thất bại nhiều lần, lúc thành công thấy cha cười tươi rói khen con gái giỏi, biết thương cha cực khổ, em thấy rất mãn nguyện” - Lil kể

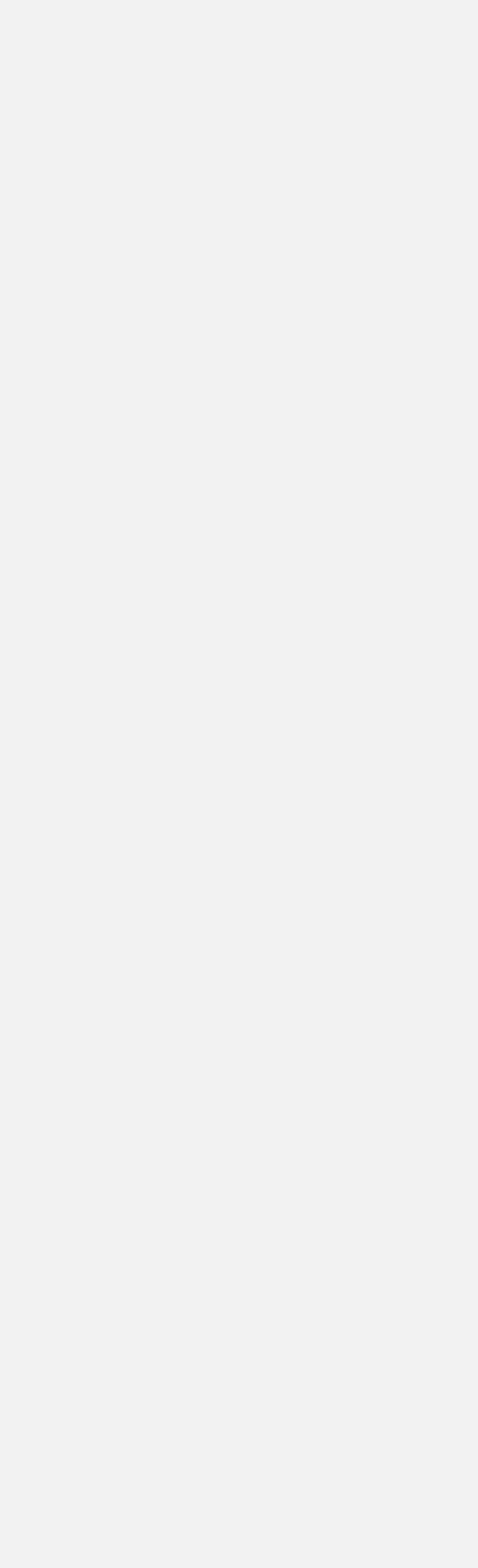
Chiếc máy có thể vừa xới đất vừa cày lên liếp với công suất gấp 40 lần lao động bằng chân tay.

Bước ra khỏi trang sách

Đưa đôi bàn tay có nhiều vết cắt, xước, em Nguyễn Ngọc Bích, Trường THPT Đầm Dơi (Cà Mau), cho biết đó là vết tích của nhiều lần cắt kiếng, cắt thanh thép để hoàn thành thiết bị tách lọc nước biển thành nước ngọt bằng năng lượng mặt trời.

“Tụi em tự cắt hết nên nhìn không đẹp lắm

Từ lúc bắt tay vào nghiên cứu tụi em thấy thích thú lắm



57

Kiến thức không còn bó buộc trong sách vở, phải vận dụng những điều đã học, đã thấy ngoài thực tế rồi tư duy tìm cách làm của riêng mình” - Bích hào hứng kể.

Dự án này Bích nảy ra từ những khó khăn của cuộc sống xung quanh

Vào mùa khô hằng năm, người dân miền biển phải chắt chiu từng giọt nước ngọt

“Em luôn tự đặt câu hỏi có thiết bị nào lọc nước biển thành nước ngọt mà không tốn nguyên liệu chất đốt không.

Rồi em chia sẻ điều này với bạn Trương Đức Tiến, không ngờ bạn cũng trăn trở như em

Hai đứa em quyết định phải nghiên cứu” - Bích chia sẻ tiếp

Không chỉ dừng ở việc lọc nước biển, Tiến và Bích còn tích hợp vào thiết bị bộ phận để lọc nước phèn, nước nhiễm bẩn thành nước sạch

Dự án này đoạt giải nhì hội thi.

Nếu như Bích, Tiến tìm lời giải biến nước mặn thành nước ngọt thì Nguyễn Trí Khang, học sinh Trường Thực hành sư phạm Trà Vinh, lại tìm cách đưa nước lên ruộng, tưới cây bằng sức người.

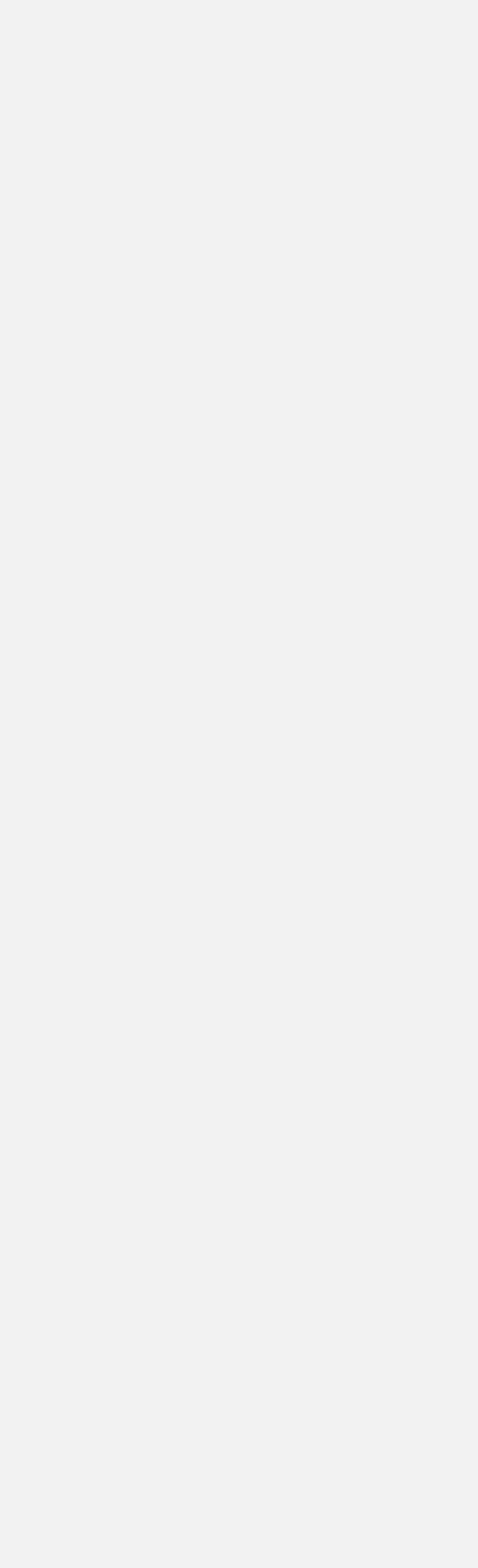
“Một là người dân không đủ tiền mua máy bơm, hai là máy bơm cồng kềnh không vận chuyển ra ruộng xa được

Máy bơm nước thân thiện của em không tốn nhiên liệu nhưng hiệu suất có thể đạt được 50% so với máy bơm thông thường

Giá thành của máy cũng chỉ mấy trăm ngàn đồng thôi” - Khang cho biết.

Cô Trang Thị Thu Trang, giáo viên hướng dẫn của Khang, chia sẻ ban đầu khi Khang trình bày ý tưởng cô có phần e ngại, nhưng cô đã dần bị thuyết phục hoàn toàn với cách làm táo bạo của học trò.

“Tự mày mò, ra sản phẩm chưa như ý muốn em lại tháo ra làm lại



58

Từng chút một cho đến khi hoàn thành

Từ đó tôi thấy em có một tình yêu đối với khoa học, càng yêu khoa học càng thôi thúc em phải học tốt hơn không chỉ trong sách vở mà từ thực tế” - cô Trang nói.

Nhiều thiết bị hạn chế tai nạn giao thông

Hội thi lần này còn có nhiều sáng kiến lý thú góp phần giảm thiểu tai nạn giao thông

Học sinh Trường THPT chuyên Lê Quý Đôn (Ninh Thuận) thiết kế thiết bị lắp đặt trên xe sẽ nhận cảm biến hoạt động của tàu lửa và tự động vô hiệu hóa cáp dây ga để hạn chế việc vượt các đường cắt ngang.

Hai bạn Lê Nguyễn Đức Tân và Nguyễn Lý Minh Nguyệt, Trường THCS Chánh Nghĩa (Bình Dương), mong muốn thông qua dự án mũ bảo hiểm thông minh, người tham gia giao thông hễ bước lên xe là phải đội mũ bảo hiểm.

Vì nếu không đội xe sẽ không khởi động được

Bộ thiết bị hai bạn sáng chế giá thành chỉ khoảng 82.000 đồng

Toàn bộ chi phí thực hiện dự án do hai bạn tiết kiệm tiền quà sáng.”

**Văn bản tóm tắt sinh ra bởi nghiên cứu.**

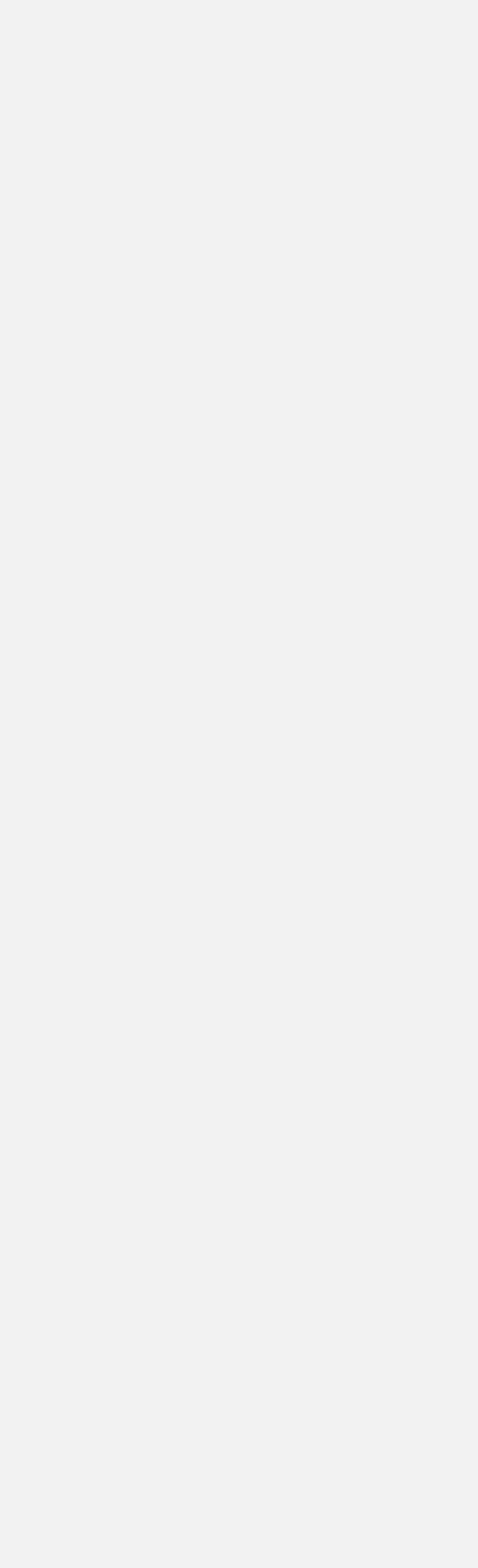
“Dự án này Bích nảy ra từ những khó khăn của cuộc sống xung quanh.

Một là người dân không đủ tiền mua máy bơm, hai là máy bơm cồng kềnh không vận chuyển ra ruộng xa được.”

**4.4. Khuyến nghị tăng cường chất lượng của văn bản tóm tắt**

**4.4.1. Tăng cường độ liên quan giữa các câu**

Hiện tại, trong luận văn, chúng tôi có khảo sát 5 phương pháp để tính độ tương



59

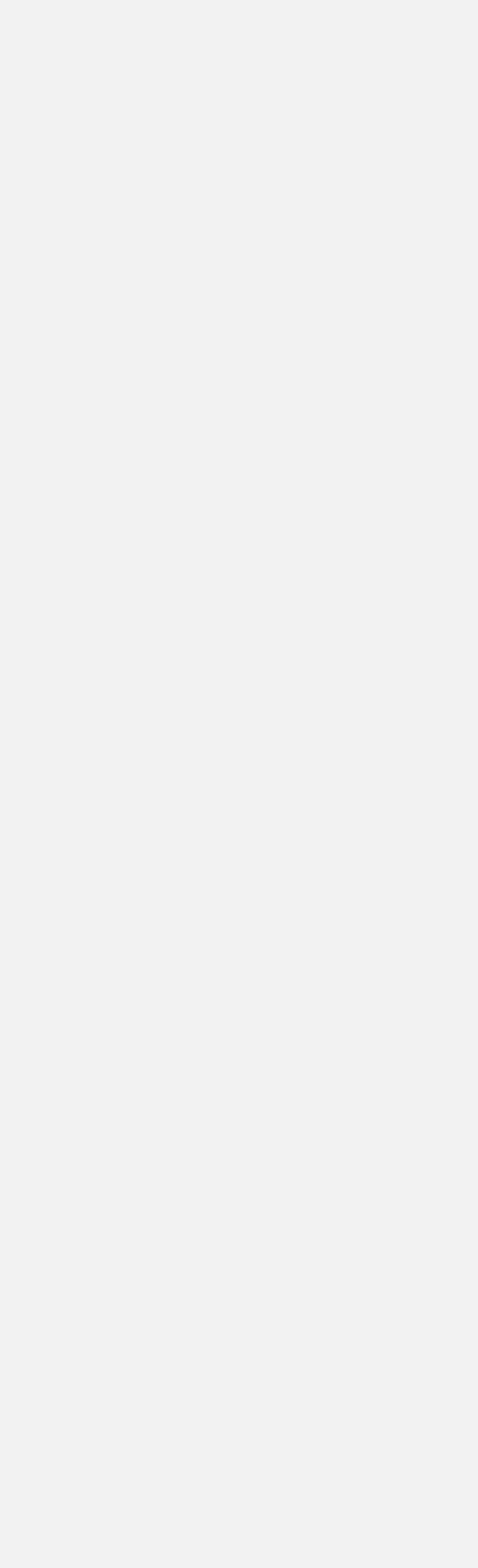
đồng giữa các câu trong văn bản: đếm số từ chung theo công thức TextRank gốc, đếm số từ chung giữa 2 câu, đếm số cụm từ chung giữa 2 câu, Cosine và Euclidean. Tuy nhiên cả 5 phương pháp trên đều không dựa vào ngữ nghĩa hay văn phạm, ngữ pháp của văn bản. Đặc biệt là chưa sử dụng các đặc trưng của Tiếng Việt. Điều này làm cho thông tin đưa vào tính độ tương đồng không đầy đủ. Vì vậy chúng tôi khuyến nghị sử dụng các phương pháp có sử dụng các đặc trưng ngôn ngữ để tăng cường độ liên quan giữa các câu trong văn bản. Một số phương pháp được đề xuất: sử dụng kho dữ liệu Wordnet Tiếng Việt, sử dụng mạng ngữ nghĩa Wikipedia.

* Sử dụng kho dữ liệu WordNet Tiếng Việt.
* Sử dụng mạng ngữ nghĩa Wikipedia. Đối với sử dụng phương pháp này thì cần phải tính độ tương đồng giữa các khái niệm trong mạng ngữ nghĩa

Wikipedia

**4.4.2. Tăng cường chất lượng văn bản tóm tắt**

Sau khi tăng cường được độ tương đồng giữa các câu trong văn bản thì chất lượng văn bản tóm tắt sẽ được nâng cao. Nguyên nhân là do các câu quan trọng trong câu được xếp hạng cao hơn. Tuy nhiên, do phương pháp TextRank là trích rút câu nên khi ghép các câu lại với nhau sẽ không được tự nhiên về mặt ngôn ngữ. Vì vậy, cần phải loại bỏ được yếu tố này để đảm bảo được văn bản tóm tắt tự động giống người hơn. Chúng tôi đề xuất kết hợp phương pháp tóm tắt trích rút câu sử dụng TextRank với phương pháp tóm lược câu sử dụng cấu trúc cú pháp. Phương pháp tóm lược câu này sẽ giúp rút gọn câu, đưa câu trở về dạng đơn giản, ngắn gọn. Đồng thời nó cũng tạo cho các câu trong văn bản tóm tắt sử dụng trích rút câu không còn cảm giác gượng gạo bởi sự xuất hiện của các từ thừa do cách sử dụng các cấu trúc cú pháp phức tạp.



60

 **KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN**

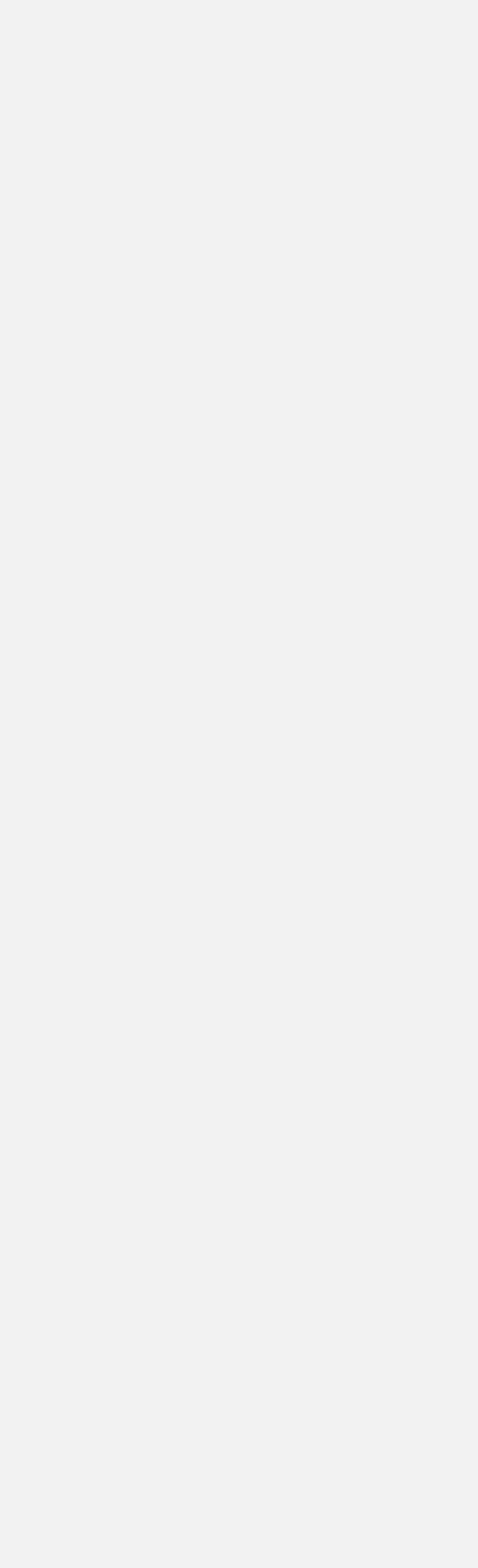
**Những vấn đề đã giải quyết được trong luận văn**

Luận văn đã nghiên cứu khảo sát, giải quyết vấn đề tóm tắt văn bản tiếng Việt sử dụng phương pháp đồ thị cụ thể là TextRank. Bài toán này có tính ứng dụng thực tế cao và không cần những kiến thức chuyên sâu về ngôn ngữ học. Phương pháp trong luận văn sử dụng chủ yếu dựa vào các thông tin trong chính văn bản được tóm tắt. Dựa vào việc tìm ra các mối quan hệ của thông tin chứa trong các câu trong văn bản mà thuật toán tính toán được mức độ quan trong của từng câu trong văn bản. Từ đó, hệ thống sẽ đưa ra được một văn bản tóm tắt tự động có mức độ chính xác cao, chứa thông tin đầy đủ so với bản gốc. Bên cạnh đó, luận văn cũng có trình bày hoàn chỉnh mô hình của một hệ thống tóm tắt văn bản tự động. Từ đó có thể ứng dụng nó vào nhiều lĩnh vực trong thực tế: các công cụ tóm tắt online, tóm tắt tài liệu, biểu mẫu trong một công ty, tự động tóm tắt nội dung các cuốn sách để tham khảo trước khi đọc…

**Hướng phát triển**

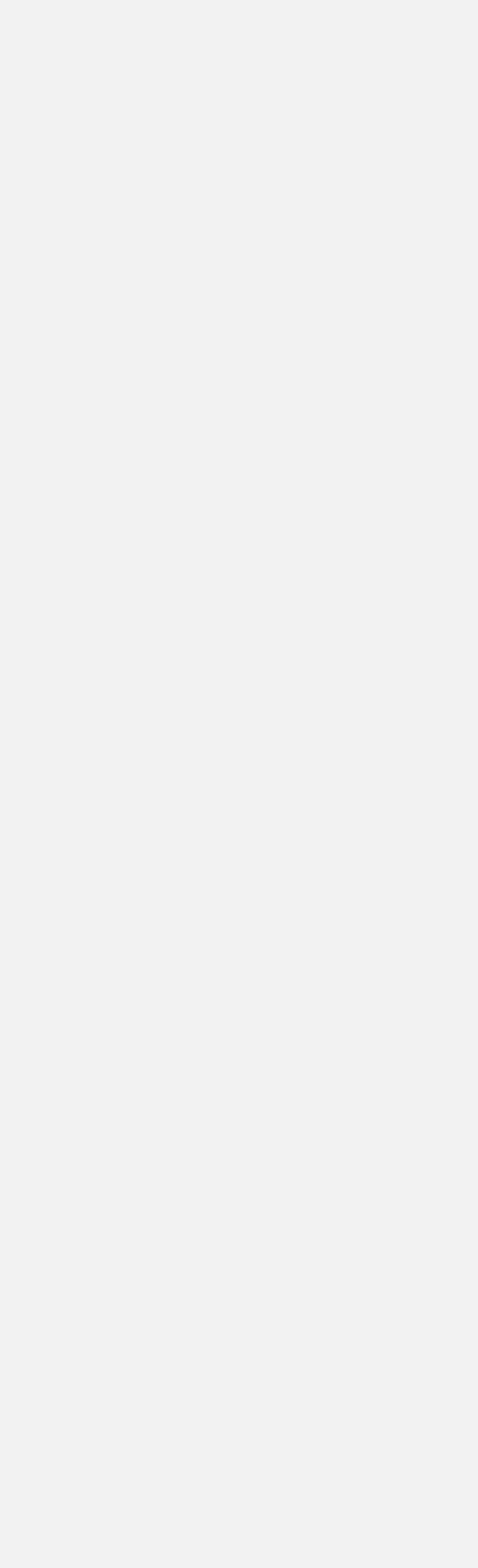
- Mở rộng không gian tập dữ liệu để cải thiện kết quả của độ đo cosine, Euclidean

* Nghiên cứu áp dụng các phương pháp khác vào việc nâng cao độ tương đồng giữa các câu, từ đó tìm ra được các câu quan trọng trong văn bản.
* Nghiên cứu áp dụng các phương pháp giúp nâng cao chất lượng văn bản tóm tắt bằng việc rút gọn các câu trong văn bản tóm tắt. Giúp cho văn bản tóm tắt giống người hơn.
* Triển khai xây dựng hệ thống tóm tắt đa văn bản sử dụng phương pháp TextRank kết hợp với các phương pháp nâng cao chất lượng. Từ đó cung cấp một sản phẩm tự động tổng hợp tin tức theo chủ đề từ các bài báo trên Internet thành một bài viết ngắn gọn, súc tích, đầy đủ thông tin. Công cụ này giúp cho người đọc không cần phải vất vả tìm kiếm thông tin hữu ích trong một khối



61

lượng lớn thông tin có trên Internet.



62

**TÀI LIỆU THAM KHẢO**

1. Hoàng Tất Thắng.2013. 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 họcHuế, 2013.
2. La Đức Dũng.2012.” *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. Trần Mai Vũ *. 2009. “Tóm tắt đa văn bản dựa vào trích xuất câu”.* *Luận văn thạc sĩ, Trường Đại học Công nghệ, Đại học quốc gia Hà Nội 2009.*
4. Regina Barzilay and Michael Elhadad*.1999. “Using Lexical Chains for* *Text Summarization”. In Advances in Automatic Text Summarization (Inderjeet Mani*

*and* Mark T. Maybury, editors): 111-121, The MIT Press, 1999.

1. Kathleen R. McKeown and Dragomir R. Radev .1995. “*Generating*

*summaries of multiple news articles”*. ACM Conference on Research andDevelopment in Information Retrieval (SIGIR’95): 74-82, Seattle, Washington, July 1995.

1. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu *.2002.*

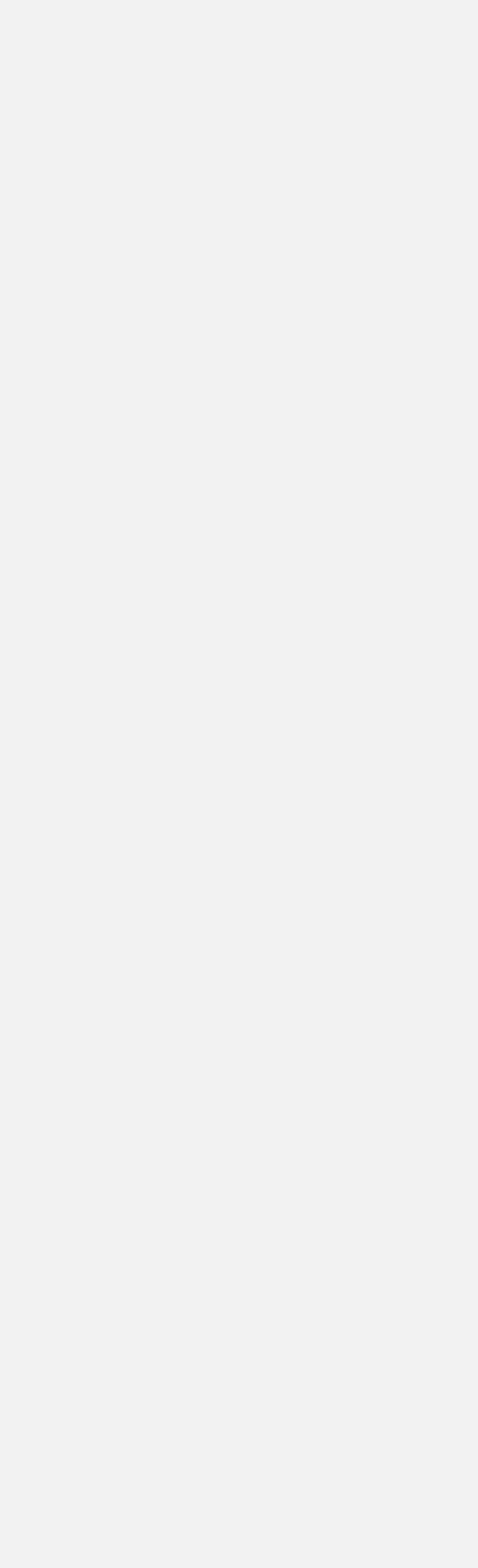
*“Bleu: a method for automatic evaluation of machine translation”.* Proceedings ofthe 40th Annual Meeting of the Association for Computational Linguistics (ACL): 311-318, 2002.

1. Chin-Yew Lin and Eduard Hovy. 2003*. “Automatic evaluation of*

*summaries using n-gram co-occurrence statistics”.* In Human Technology Coference2003.

1. Pierre Senellart.2007. “*Understanding the Hidden Web”.* *PhD thesis in* *Computer science, Université Paris-Sud,* Orsay, France, December 2007.
2. Krishna Sapkota, Laxman Thapa, Shailesh Bdr. Pandey .2006.

“*Efficient* *Information Retrieval Using Measures of Semantic Similarity”*.



63

Conference on Software, Knowledge, Information Management and Applications*,* *Chiang Mai, Thailand, December 2006, 9498.*

1. T. A. N. Hoang, H. K. Nguyen, Q. V. Tran. 2010. “*An Efficient* *Vietnamese Text Summarization Approach Based on Graph Model*”. Kỷyếu hội nghị

International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), tr. 1-6.

1. H. T. Le, R. C. Sam, P. T. Nguyen. 2010. “*Extracting Phrases in* *Vietnamese Document for Summary Generation*”.Kỷyếu hội nghịInternationalConference on Asian Language Processing, tr. 207-210.
2. Ha Nguyen Thi Thu. 2014. “*An Optimization Text Summarization* *Method Based on Naïve Bayes and Topic Word for Single Syllable Language*”.

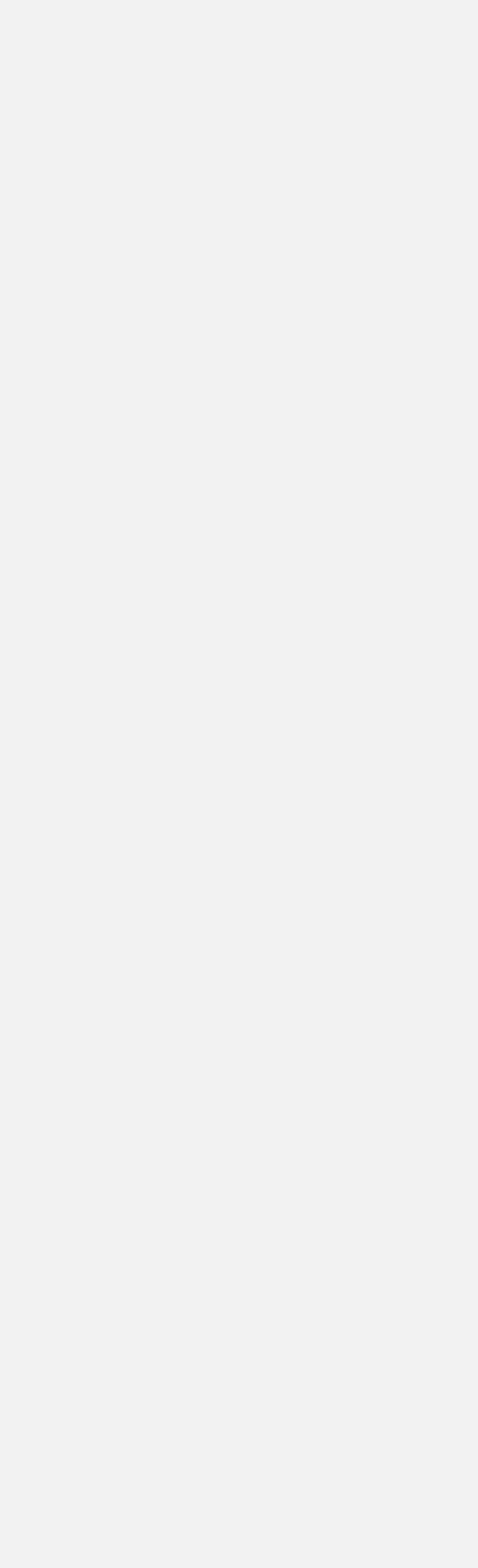
Applied Mathematical Sciences, Vol. 8, 2014, no. 3, tr. 99 - 115.

1. V. Phung, L. D. Vine. 2015. “*A Study on the Use of Word Embeddings* *and PageRank for Vietnamese Text Summarization*”. Kỷyếu hội nghịADCS '15Proceedings of the 20th Australasian Document Computing Symposium Article No. 7, tr. 1-8.
2. M. T. Nguyen, V. D. Lai, P. K. Do, D. V. Tran, M. L. Nguyen. 2016.

“*VSoLSCSum: Building a Vietnamese Sentence-Comment Dataset for Social* *Context Summarization*”.Kỷyếu hội thảo The 12th Workshop on Asian LanguageResources (COLING) 2016, tr. 38-48.

1. R. Mihalcea, P.Tarau. 2004. “*TextRank: Bringing order into texts*”. Kỷ yếu hội nghị EMNLP 2004, tr. 404–411.
2. F. Wei, W. Li, Q. Lu, Y. He. 2010. “*A document-sensitive graph model* *for multi-document summarization*”. Tạp chí Knowledge and Information Systems,2010, Volume 22, Number 2, tr. 245-259.
3. *H. Zha. 2002. “ Generic summarization and keyphrase extraction using*

*mutual reinforcement principle and sentence clustering*”, Kỷ yếu hội nghị The 25thannual international ACM SIGIR conference on Research and development in

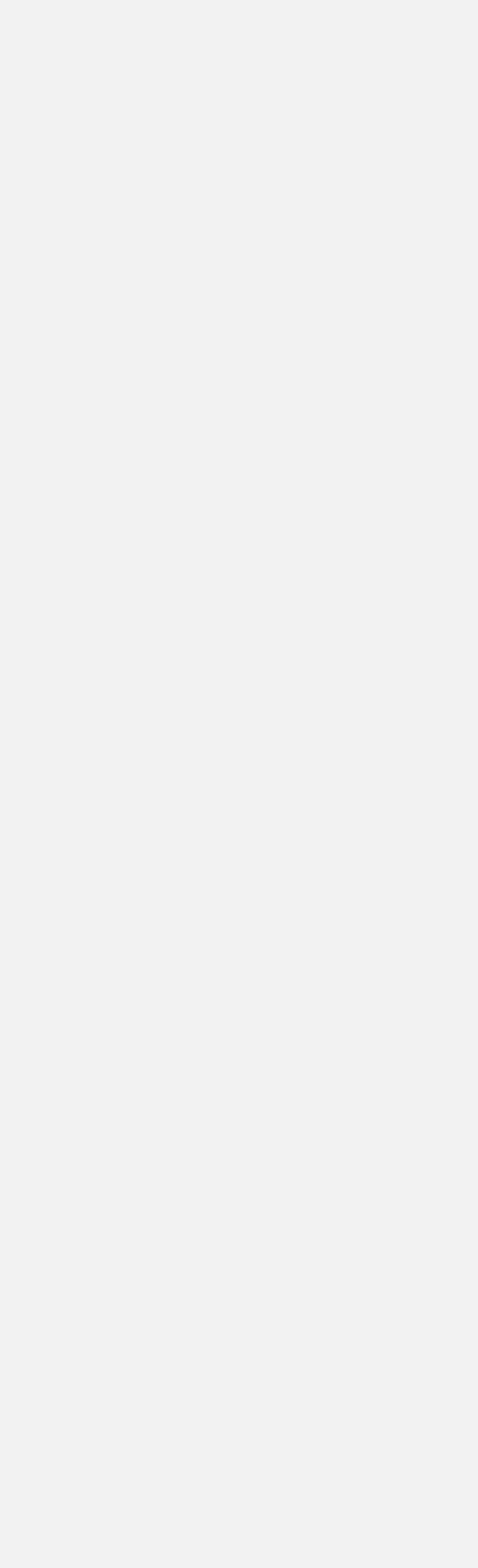


64

information retrieval, tr. 113-120.

1. *H. P. Luhn. 1958. “H. P. Luhn and Automatic Indexing’.* TheInternational Conference on Scientific Information (ICSI), Washington, DC, in November 1958*.*
2. H.P. Edmundson. 1969. “*New Methods in Automatic Abstracting*”.
3. Inderjeet Mani and Mark T. Maybury (eds) (1999).” *Advances in* *Automatic Text Summarization”*, MIT Press, 1999, ISBN 0-262-13359-8.
4. K.G.Zipf, “*Human Behaviour and the Principle of Least Effort*”.
5. Kleinberg, Jon (December 1999). [*"Hubs, Authorities, and*](http://www.cs.brown.edu/memex/ACM_HypertextTestbed/papers/10.html) [*Communities*".](http://www.cs.brown.edu/memex/ACM_HypertextTestbed/papers/10.html) *Cornell* University. Retrieved 2008-11-09.
6. [Brin,](https://en.wikipedia.org/wiki/Sergey_Brin) [S.Page, L.](https://en.wikipedia.org/wiki/Larry_Page) (1998). [*"The anatomy of a large-scale hypertextual*](http://infolab.stanford.edu/pub/papers/google.pdf) [*Web search engine*"](http://infolab.stanford.edu/pub/papers/google.pdf) *.* Computer Networks and ISDN Systems.
7. Jan O. Pendersen, Kupiec Julian and Francine Chen (1995). “*A* *trainable document summarizer”,* Research and Development in InformationRetrieval: 68 - 73, 1995.
8. Chin-Yew Lin and Eduard Hovy (1997). “*Identifying topics by* *position*”, Fifth Conference on Applied Natural Language Processing: 283-290, 1997.
9. Kathleen R. McKeown and Dragomir R. Radev (1995). “*Generating* *summaries of multiple news articles*”, ACM Conference on Research andDevelopment in.
10. Quoc Le and Tomas Mikolov, “*Distributed Representations of* *Sentences and Documents*.” Proc. ICML’14, vol. 32, tr. 1188-1196, 2014.
11. .https://butchiso.com/2013/10/tim-hieu-ve-mo-hinh-khong-gian-

vector.html



65



**Phụ lục**

**Danh sách các thư viện sử dụng trong hệ thống**

|  |  |  |
| --- | --- | --- |
| **STT** | **Tên thư viện** | **Ngôn ngữ** |
|  |  |  |
| 1 | Nltk | Nodejs |
|  |  |  |
| 2 | compute-cosine-similarity | Nodejs |
|  |  |  |
| 3 | dean Euclidean-distance | Nodejs |
|  |  |  |
| 4 | vntokenizer | Nodejs |
|  |  |  |

 **Commented [N29]:** Sang trang mới

