ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN MẠNH CƯỜNG

TÓM TẮT VĂN BẢN TIẾNG VIỆT TỰ ĐỘNG DỰA TRÊN MÔ HÌNH ĐỒ THỊ

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Hà Nội, 06/2019

ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

NGUYỄN MẠNH CƯỜNG

TÓM TẮT VĂN BẢN TIẾNG VIỆT TỰ ĐỘNG DỰA TRÊN MÔ HÌNH ĐỒ THỊ

Ngành: Khoa học máy tính

Chuyên ngành: Khoa học máy tính

Mã Số: 8480101.01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC: **PGS.TS NGUYỄN PHƯƠNG THÁI**

Hà nội – 06/2019

LỜI CẢM ƠN

Luận văn này được tôi thực hiện dưới sự hướng dẫn của PGS.TS Nguyễn Phương Thái.

Tôi xin bày tỏ lòng biết ơn tới thầy Nguyễn Phương Thái, thầy đã tận tình hướng dẫn, để tôi có thể hoàn thiện luận văn này.

Tôi xin cảm ơn các đồng nghiệp của tôi, đã tạo mọi điều kiện thuận lợi giúp tôi có thể thu xếp thời gian vừa công tác, vừa học tập.

Tôi xin gửi lời cảm ơn đến bố mẹ, những người luôn đồng hành, ủng hộ tôi trong suốt quá trình học tập và nghiên cứu.

Xin chân thành cảm ơn!

Tác giả

Nguyễn Mạnh Cường

LỜI CAM ĐOAN

Tôi - Nguyễn Mạnh Cường - cam đoan luận văn này là công trình nghiên cứu của bản thân tôi dưới sự hướng dẫn của PGS.TS. Nguyễn Phương Thái.

Các kết quả nêu trong luận văn là trung thực, và không sao chép toàn văn của bất kỳ công trình nào khác.

Tôi xin hoàn toàn chịu trách nhiệm và chịu mọi hình thức kỷ luật theo quy định cho lời cam đoan này.

Hà Nội, ngày 10 tháng 06 năm 2019

MỤC LỤC

DANH MỤC KÝ HIỆU, VIẾT TẮT

Kí hiệu	Giải thích
DUC	Document Understanding Conferences
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
TF.IDF	Term frequency–inverse document frequency

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG

MỞ ĐẦU

Theo số liệu báo cáo [18] của Global Digital từ We Are Social và Hootsuite, trong tháng 1 năm 2019 có 4,39 tỷ người dùng internet trên toàn thế giới, tăng 366 triệu người dùng so với cùng kỳ năm 2018, điều đó cho thấy sự phát triển nhanh chóng của mạng internet. Sự phát triển này kéo theo sự tăng trưởng mạnh về số lượng các blog, trang web và các tài liệu văn bản. Từ đó gia tăng nhu cầu tìm kiếm, xử lý và tổng hợp thông tin của con người. Để cải thiện khả năng tìm kiếm cũng như tăng hiệu quả cho các công việc xử lý thông tin, tóm tắt văn bản tự động là một giải pháp hàng đầu.

Tóm tắt văn bản là quá trình tạo ra một văn bản ngắn hơn từ một hoặc nhiều văn bản gốc đáp ứng một số yêu cầu nào đó của người dùng, mà vẫn đảm bảo nội dung và ý nghĩa của văn bản gốc. Bài toán tóm tắt văn bản đóng vai trò quan trọng trong khoa học khai phá dữ liệu. Là một bài toán thực tiễn, có khả năng thương mại, áp dụng cho các hệ thống tìm kiếm thông minh, hệ gợi ý, tổng hợp thông tin. Thay vì một tài liệu đầy đủ, chỉ có một văn bản tóm tắt ngắn gọn cần được xử lý. Chẳng hạn, bằng cách cung cấp các đoạn mô tả ngắn gọn nội dung truy vấn, công cụ tìm kiếm có thể giúp người dùng xác định các tài liệu ưa thích trong thời gian ngắn.

Trên thế giới, các nghiên cứu đầu tiên về tóm tắt văn bản được công bố vào những năm 50 của thế kỉ trước. Cho tới nay, tóm tắt văn bản vẫn không ngừng được nghiên cứu, phát triển, và đã đạt được thành tựu đáng kể trong việc tóm tắt các văn bản tiếng Anh, tiếng Trung…

Tại Việt Nam, tóm tắt văn bản cũng rất được quan tâm, cụ thể cho bài toán tóm tắt văn bản tiếng Việt. Tuy nhiên, do sự phức tạp về cấu trúc, ngữ pháp của tiếng Việt, do thiếu tài nguyên về những kho ngữ liệu, tập mẫu nên những nghiên cứu về tóm tắt văn bản tiếng Việt vẫn còn hạn chế cả về mặt số lượng lẫn chất lượng. Vì thế tôi lựa chọn đề tài luận văn "Tóm tắt văn bản tiếng Việt tự động dựa trên mô hình đồ thị" bởi tính cấp thiết và tính ứng dụng cao của nó.

Luận văn bao gồm 4 chương:

Chương 1. Tổng quan về tóm tắt văn bản

Trình bày về các khái niệm cơ bản của tóm tắt văn bản, phân loại bài toán tóm tắt văn bản, các ứng dụng của tóm tắt văn bản và các phương pháp đánh giá một hệ thống tóm tắt văn bản.

Chương 2: Các phương pháp tóm tắt văn bản

Trình bày về các phương pháp tóm tắt văn bản, các hướng tiếp cận cho việc giải quyết bài toán tóm tắt văn bản, một số đặc điểm của tiếng Việt, hiện trạng các nghiên cứu về tóm tắt văn bản tiếng Việt.

Chương 3: Xây dựng mô hình tóm tắt văn bản tiếng Việt dựa theo phương pháp đồ thị.

Trình bày chi tiết về mô hình tóm tắt trích rút đơn văn bản tiếng Việt dựa trên mô hình đồ thị trên cơ sở áp dụng thuật toán iSpreadRank. Phần này đi sâu về thiết kế mô hình tóm tắt và các giai đoạn xử lý, bên cạnh đó luận văn cũng trình bày chi tiết thuật toán trong từng giai đoạn.

Chương 4: Đánh giá kết quả đạt được.

Tiến hành thực nghiệm và đánh giá kết quả thực nghiệm.

CHƯƠNG 1. TỔNG QUAN VỀ TÓM TẮT VĂN BẢN

Trong chương này, luận văn trình bày tổng quan về tóm tắt văn bản, bao gồm các khái niệm cơ bản, phân loại tóm tắt văn bản và các phương pháp đánh giá độ chính xác của tóm tắt văn bản.

1.1. Khái niệm tóm tắt văn bản.

Có rất nhiều định nghĩa khác nhau về tóm tắt văn bản. Tuỳ thuộc vào mục đích yêu cầu của bài toán hay góc nhìn nhận của đối tượng sử dụng mà chúng ta có các định nghĩa khác nhau:

- Tóm tắt văn bản là quá trình trích rút những thông tin quan trọng nhất từ một hoặc nhiều nguồn để tạo ra phiên bản cô đọng, ngắn gọn phục vụ cho một hoặc nhiều người dùng cụ thể, hay một hoặc nhiều nhiệm vụ cụ thể [1] .
- Tóm tắt văn bản là cô đọng văn bản nguồn thành một phiên bản ngắn hơn bảo tồn nội dung thông tin và ý nghĩa tổng thể của nó [16].
- Tóm tắt văn bản tự động là nhiệm vụ tạo ra một bản tóm tắt ngắn gọn và trôi chảy trong khi bảo tồn nội dung thông tin chính và ý nghĩa tổng thể [11].

Ví dụ:

Văn bản gốc:

Thành lập Tiểu ban An toàn và an ninh hạt nhân trực thuộc Ban Chỉ đạo Nhà nước Dự án điện hạt nhân Ninh Thuận Trưởng Ban Chỉ đạo Nhà nước Dự án điện hạt nhân Ninh Thuận đã ký Quyết định số 106/QĐ-BCĐĐHNNT ngày 29/5/2013 về việc thành lập Tiểu ban An toàn và an ninh hạt nhân.
Tiểu ban An toàn và an ninh hạt nhân có nhiệm vụ giúp Ban Chỉ đạo Nhà nước Dự án điện hạt nhân Ninh Thuận (Ban Chỉ đạo Nhà nước) chỉ đạo, đôn đốc, kiểm tra việc: xây dựng, tiến độ ban hành và thực hiện các văn bản quy phạm pháp luật, quy chuẩn và tiêu chuẩn quốc gia, các văn bản hướng dẫn về an toàn, an ninh cho dự án điện hạt nhân, tham gia và thực hiện các điều ước quốc tế về an toàn hạt nhân; xây dựng và thực hiện các chương trình về đảm bảo an toàn bức xạ hạt nhân, bảo đảm an ninh và bảo vệ nhà máy điện hạt nhân, xây dựng trung tâm ứng phó quốc gia; thực hiện quan trắc cảnh báo phóng xạ môi trường và đánh giá tác động môi trường của Dự án điện hạt nhân Ninh Thuận; thẩm

định, thanh tra và giám sát an toàn và an ninh hạt nhân.
Tiểu ban cũng có trách nhiệm tham mưu, tư vấn cho Ban Chỉ đạo Nhà nước về các vấn đề liên quan đến công tác bảo đảm an toàn, an ninh hạt nhân; xây dựng và kiểm tra việc thực hiện chính sách, chương trình về bảo đảm an toàn bức xạ hạt nhân, bảo đảm an ninh và ứng phó sự cố cho Dự án điện hạt nhân Ninh Thuận.
Trưởng Tiểu ban là Ủy viên Ban Chỉ đạo Nhà nước, Thứ trưởng Bộ Khoa học và Công nghệ; Phó Trưởng Tiểu ban thường trực là Cục trưởng Cục An toàn bức xạ và hạt nhân. Các ủy viên của Tiểu ban là đại diện các Bộ, cơ quan, địa phương liên quan đến nhiệm vụ của Tiểu ban.
Bộ máy giúp việc của Tiểu ban có Tổ giúp việc (hoặc bộ phận thường trực) thuộc Cục An toàn bức xạ và hạt nhân.

Văn bản tóm tắt:

Trưởng Ban Chỉ đạo Nhà nước Dự án điện hạt nhân Ninh Thuận đã ký Quyết định về việc thành lập Tiểu ban An toàn và an ninh hạt nhân.
Tiểu ban có nhiệm vụ giúp Ban Chỉ đạo Nhà nước Dự án điện hạt nhân Ninh Thuận chỉ đạo, đôn đốc, kiểm tra đồng thời tham mưu, tư vấn về các vấn đề liên quan đến công tác bảo đảm an toàn, an ninh hạt nhân, xây dựng và kiểm tra việc thực hiện chính sách, chương trình về bảo đảm an toàn bức xạ hạt nhân, bảo đảm an ninh và ứng phó sự cố cho Dự án điện hạt nhân Ninh Thuận.
Trưởng Tiểu ban là Thứ trưởng Bộ Khoa học và Công nghệ; Phó Trưởng Tiểu ban thường trực là Cục trưởng Cục An toàn bức xạ và hạt nhân. Các ủy viên của Tiểu ban là đại diện các Bộ, cơ quan, địa phương liên quan đến nhiệm vụ của Tiểu ban.

1.2. Phân loại bài toán tóm tắt văn bản

Có thể phân chia bài toán tóm tắt văn bản thành nhiều loại. Mỗi loại được sử dụng cho các mục đích khác nhau, các yêu cầu khác nhau, bởi vậy cũng có các phương pháp, kỹ thuật tương ứng với mỗi loại. Không có một hệ thống tóm tắt văn bản nào có thể đáp ứng được hết tất cả các yêu cầu của con người.

Theo kết quả (out put)

Tóm tắt trích rút (Extract): Là một bản tóm tắt bao gồm các đơn vị quan trọng trong văn bản như câu, đoạn văn được trích rút y nguyên từ văn bản gốc

[16].

Tóm tắt tóm lược (Abtract): Tương tự như cách con người tóm tắt, văn bản mới được tạo ra bằng cách viết lại văn bản gốc. Nói cách khác, chúng ta diễn giải và biểu diễn văn bản tóm tắt bằng các kỹ thuật ngôn ngữ tự nhiên tiên tiến để tạo ra một văn bản mới truyền tải thông tin quan trọng nhất từ văn bản gốc [11].

Theo mục đích tóm tắt

Tóm tắt thông tin (Information): Tóm tắt bao gồm tất cả thông tin nổi bật của văn bản gốc ở nhiều mức độ chi tiết khác nhau.

Tóm tắt đánh giá: Tóm tắt nhằm mục đích đánh giá vấn đề chính của văn bản gốc theo quan điểm của người đánh giá.

Theo nội dung

Tóm tắt chung (Generalized): Tóm tắt nhằm mục đích đưa ra các nội dung quan trọng phản ánh toàn bộ nội dung của văn bản gốc. Hay nói cách khác mục đích của loại tóm tắt này là sao cho văn bản tóm tắt chứa đựng những nội dung mà tác giả muốn người đọc biết và hiểu.

Tóm tắt truy vấn (Qurery-based): Tóm tắt nhằm mục đích đưa ra các kết quả dựa vào câu truy vấn của người dùng. Tóm tắt này thường được sử dụng trong quá trình tìm kiếm thông tin.

Theo miền dữ liệu

Tóm tắt trên một miền dữ liệu (Domain): Tóm tắt nhắm vào một miền nội dung cụ thể nào đó, như tin tức thể thao, tin tức giáo dục, bản tin tài chính...

Tóm tắt trên một thể loại (Genre): Đối tượng cần tóm tắt là một loại văn bản cụ thể, ví dụ như văn bản báo chí, email, website..

Tóm tắt độc lập (Independent): Tóm tắt có thể áp dụng cho nhiều loại văn bản và trên nhiều miền dữ liệu.

Theo số lượng

Tóm tắt đơn văn bản: Văn bản tóm tắt được tạo ra từ một văn riêng lẻ.

Tóm tắt đa văn bản: Văn bản tóm tắt được tạo ra từ nhiều văn bản cùng liên quan tới một chủ đề.

Theo ngôn ngữ

Tóm tắt đơn ngôn ngữ: Văn bản nguồn chỉ được trình bày bởi duy nhất một ngôn ngữ, văn bản tóm tắt được sinh ra mang ngôn ngữ của văn bản đó.

Tóm tắt đa ngôn ngữ: Hệ thống tóm tắt có thể áp dụng tóm tắt cho nhiều văn bản ở nhiều ngôn ngữ khác nhau. Mỗi văn bản gốc chỉ chứa duy nhất một loại ngôn ngữ.

Tóm tắt xuyên ngôn ngữ: Trong mỗi văn bản gốc chứa nhiều ngôn ngữ khác nhau. Hệ thống cần có khả năng nhận dạng cụ thể từng loại ngôn ngữ và cho ra văn bản tóm tắt phù hợp. Đây là loại tóm tắt văn bản khó nhất trong ba loại phân chia theo ngôn ngữ.

1.3. Ứng dụng của tóm tắt văn bản

Tóm tắt văn bản có rất nhiều ứng dụng trong thực tế. Có thể nêu ra một số ứng dụng điển hình như sau:

Tóm tắt phục vụ máy tìm kiếm (Search engine)

Về khía cạnh công nghệ: Với kho dữ liệu lớn, nếu trước khi tìm kiếm không có bước tóm tắt và trích lọc thì đồng nghĩa với việc vông cụ tìm kiếm phải duyệt qua nội dung của tất cả các tài liệu hay bản ghi để tìm thông tin liên quan đến từ khoá, việc này gây tốn thời gian và và lãng phí tài nguyên. Trong trường hợp này tóm tắt văn bản đóng vai trò như một giải pháp tối ưu giúp nâng cao hiệu quả cho các máy tìm kiếm, thay vì phải duyệt tất cả nội dung từ đầu đến cuối, máy tìm kiếm chỉ cần duyệt nội dung tóm tắt của của các văn bản đó.

Về khía cạnh trải nghiệm của người dùng: Khi hiển thị kết quả tìm kiếm thay vì hiển thị toàn bộ nội dung, máy tìm kiếm hiển thị một phần nội dung (được in đậm) có thể coi đó như một bản tóm tắt ngắn, cho phép người dùng một bản xem trước, giúp người dùng có thể nhanh chóng chọn được tài liệu thích hợp.

Hiện nay, một số trang web hay công cụ tìm kiếm nổi tiếng như google, Cốc cốc đều đã ứng dụng rất tốt tóm tắt văn bản vào hệ thống của họ.

Tóm tắt tin tức (Multimedia New Summaries)

Giá trị của thông tin trong thương mại rất quan trọng, ví dụ từ việc tổng hợp một lượng tin tức đủ lớn, chúng ta có thể có các bản thống kê phục vụ các nhu cầu khác nhau như thống kê về xu hướng mua hàng, thống kê về các sự kiện được quan tâm trong một khoảng thời gian nào đó. Trên thực tế đã có nhiều công ty, tổ chức coi tin tức như một loại hàng hoá bằng cách cung cấp cho khách hàng

những thông tin được xuất bản trong ngày có nội dung liên quan đến một lĩnh vực được "đặt hàng" trước.

Tóm tắt tài liệu

Đối tượng của tóm tắt tài liệu bao gồm sách, báo, tài liệu khoa học. Thông thường mỗi tài liệu như sách, tài liệu khoa học đều có một phần tóm tắt ngay tại những trang đầu. Phần tóm tắt này cung cấp cho người đọc cái nhìn tổng quan về nội dung sách, tài liệu đó.

Giản lược nội dung cho các thiết bị cầm tay

Đặc điểm của các thiết bị cầm tay như điện thoại, máy tính bảng… là thường nhỏ gọn, hạn chế về diện tích hiển thị. Do vậy việc truyền tải nội dung dạng văn bản đặc biệt văn bản dài có những hạn chế nhất định, một bản tóm tắt ngắn gọn là cần thiết trong trường hợp này.

1.4. Các phương pháp đánh giá tóm tắt văn bản

1.4.1. Đánh giá thủ công

Các chuyên gia trực tiếp đánh giá văn bản tóm tắt dựa vào chất lượng đoạn văn, trên cơ sở những tham số về ngữ pháp, không dư thừa và sự gắn kết. Họ sẽ xem xét lỗi ngữ pháp trong văn bản như sai từ, lỗi dấu câu, bản tóm tắt tạo ra không được chứa thông tin dư thừa, thể hiện rõ ràng sự liên kết giữa các câu, và sự liên kết với chủ đề của văn bản gốc. Tuy nhiên, phương pháp này có một số hạn chế như việc đánh giá do con người thực hiện thường không ổn định và đặc biệt tiêu tốn rất nhiều thời gian và tiền bạc.

1.4.2. Đánh giá đồng chọn

Phương pháp này chỉ có thể đánh giá độ chính xác cho văn bản tóm tắt theo hướng trích rút, các câu được kết nối với nhau tạo nên văn bản tóm tắt và không cần hiệu chỉnh gì thêm. Phương pháp này đánh giá độ chính xác giữa văn bản tóm tắt với văn bản gốc dựa trên ba đặc trưng là: Độ đo chính xác (Precision), độ đo triệu hồi (Recall) và độ đo F-measure.

Độ đo chính xác (precision): Được tính dựa trên tổng số câu trùng nhau của văn bản tóm tắt lý tưởng và văn bản tóm tắt của hệ thống, chia cho tổng số câu văn bản tóm tắt của hệ thống.

$$Precision = \frac{|SH \cap SM|}{|SM|} \quad (1.1)$$

Trong đó:

$|SM|$: Là số lượng câu của văn bản tóm tắt do hệ thống trích rút.

$|SH|$: Là số lượng câu của bản tóm tắt lý tưởng do con người trích rút.

$|SH{:}\cap SM|$ Là số lượng câu trùng nhau giữa hai văn bản do hệ thống và con người trích rút.

Độ đo triệu hồi (Recall): Được tính dựa trên tổng số câu trùng nhau của văn bản tóm tắt lý tưởng và văn bản tóm tắt của hệ thống, chia cho tổng số câu của văn bản tóm tắt lý tưởng do con người thực hiện.

$$Recall = \frac{|SH \cap SM|}{|SH|} \quad (1.2)$$

Độ đo f-score: Là độ đo kết hợp giữa độ đo chính xác và độ đo triệu hồi. Người ta gọi f-score là một hàm điều hoà của độ đo chính xác và độ đo triệu hồi. Các giá trị f-score nhận đượ trong đoạn [0,1], hiển nhiên giá trị tốt nhất là 1.

$$f - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (1.3)$$

Trong tóm tắt văn bản, người ta cũng thường dùng các trọng số khác nhau cho precision và recall trong khi tính f-score. Giá trị trọng số β là một số không âm. $\beta > 1$ nghĩa là precision quan trọng hơn, $\beta < 1$ nghĩa là recall quan trọng hơn.

$$f - score = \frac{(\beta^2 + 1) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \quad (1.4)$$

1.4.3. Đánh giá dựa trên nội dung

Phương pháp đánh giá LCS (Longest Common Subsequence): LCS tìm ra độ dài của chuỗi con chung dài nhất giữa hai văn bản X và Y, độ dài của chuỗi con chung dài nhất càng lớn thì hai văn bản X, Y càng giống nhau.

$$LCS(X,Y) = \frac{length(X) + length(Y) - edit(X,Y)}{2} \quad (1.5)$$

Trong đó:

$length(X)$: Là độ dài chuỗi X.

$length(Y)$: Là độ dài chuỗi Y.

$edit(X,Y)$: Là số lần tối thiểu của việc xoá hoặc chèn thêm để biến X

thành Y.

Phương pháp ROUGE [22]: Trong điều kiện hạn hẹp về thời gian và chi phí, việc đánh giá chất lượng văn bản tóm tắt theo cách thủ công do con người thực hiện là một phương án không khả thi, chưa kể rằng phương pháp đánh giá này thường không ổn định, phụ thuộc vào kiến thức của người đánh giá. ROUGE tính toán dựa trên việc thống kê các n-gram đồng xuất hiện giữa văn văn tóm tắt do hệ thống thực hiện và văn bản tóm tắt lý tưởng. Hiện nay, phương pháp này được coi như một phương pháp đáng tin cậy để đánh giá độ chính xác của một hệ thống tóm tắt văn bản tự động. ROUGE-N được tính theo công thức:

$$ROUGE - N = \frac{\sum_{S \in SH} \sum_{g_n \in S} C_m(g_n)}{\sum_{S \in SH} \sum_{g_n \in S} C(g_n)} \quad (1.6)$$

Trong đó:

SH: Là tập tất cả văn bản tóm tắt lý tưởng.

$C_m(g_n)$: Là số lượng n-gram đồng xuất hiện lớn nhất giữa văn bản tóm tắt hệ thống và tập văn bản tóm tắt lý tưởng.

$C(g_n)$: Là số lượng n-gram trong văn bản tóm tắt lý tưởng.

Phương pháp đánh giá BLEU (Bilingual Evaluation Understudy)[23]: Đây là một phương pháp nổi tiếng để đánh giá độ chính xác của hệ thống dịch máy. Tuy vậy, chúng ta cũng có thể áp dụng nó để đánh giá độ chính xác của một hệ thống tóm tắt văn bản tự động. Hướng tiếp cận tương tự ROUGE, BLEU đánh giá độ tương đồng giữa văn bản tóm tắt hệ thống và tập các bản tóm tắt lý tưởng dựa vào sự đồng xuất hiện của các n-gram trong bản tóm tắt hệ thống và trong tập các bản tóm tắt lý tưởng.

$$BLEU - N = \frac{\sum_{n-gram \in C} Count_{clip}(n - gram)}{\sum_{n-gram \in C} Count(n - gram)}$$

Trong đó:

C: Là văn bản tóm tắt hệ thống.

$Count_{clip}(n - gram)$: Là số lượng lớn nhất của n-gram đồng xuất hiện giữa văn bản tóm tắt hệ thống và các văn bản tóm tắt lý tưởng.

$Count(n - gram)$: Là số lượng của n-gram trong văn bản tóm tắt hệ thống.

CHƯƠNG 2. CÁC PHƯƠNG PHÁP TÓM TẮT VĂN BẢN

Trong chương này, luận văn trình bày về các phương pháp tóm tắt văn bản, các hướng tiếp cận giải quyết bài toán tóm tắt văn bản, hiện trạng nghiên

cứu tóm tắt văn bản tiếng Việt.

2.1. Tóm tắt trích rút.

Kỹ thuật tóm tắt trích rút bằng cách chọn một tập hợp con các câu trong văn bản gốc. Những bản tóm tắt này chứa những câu quan trọng nhất của văn bản gốc. Đầu vào có thể là một tài liệu duy nhất hoặc nhiều tài liệu.

Theo [11] cho đến nay, tóm tắt trích rút vẫn cho kết quả tốt, hiệu quả ổn định hơn so với tóm tắt trừu tượng. Điều này do thực tế là các phương pháp tóm tắt trừu tượng phải đối mặt với các vấn đề như biểu diễn ngữ nghĩa, suy luận và tạo ngôn ngữ tự nhiên, mức độ khó hơn rất nhiều các phương pháp dựa trên dữ liệu như trích rút câu. Thực tế ngày nay, không có hệ thống tóm tắt nào hoàn toàn trừu tượng (viết lại hoàn toàn) [11], một số 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.

Để hiểu rõ hơn về cách thức hoạt động của các hệ thống tóm tắt loại trích rút, tôi mô tả ba nhiệm vụ khá độc lập mà tất cả các hệ thống tóm tắt trích rút cần thực hiện:

- Biến đổi văn bản hay nói cách khác là dùng các thuật toán về thống kê, đồ thị hoá, học máy… để biểu diễn văn bản.
- Tính trọng số về tính quan trọng của câu.
- Chọn một tập con trong văn bản gốc để trở thành văn bản tóm tắt.

a. Đồ thị hoá

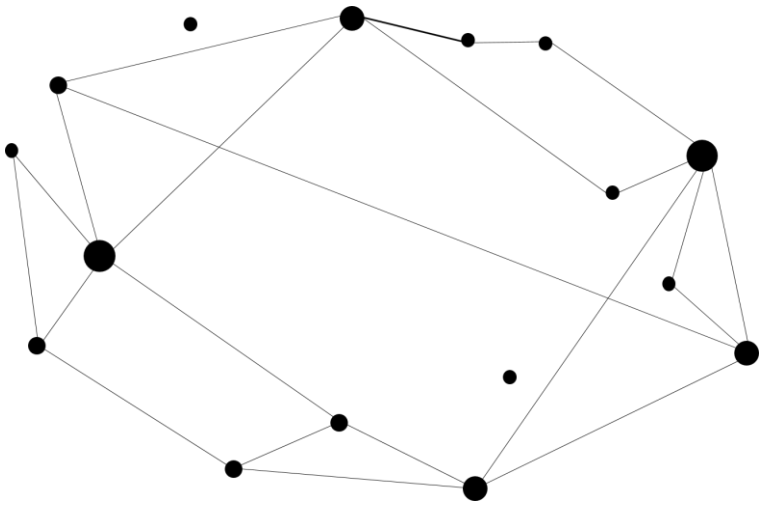
Đồ thị hoá văn bản hay biểu diễn văn bản dưới dạng đồ thị thuộc bước tiền xử lý mà tất cả các hệ thống tóm tắt theo mô hình đồ thị phải thực hiện. Trong đó, mỗi đồ thị biểu diễn một văn bản hoặc biểu diễn nhiều văn bản.

Với bài toán tóm tắt văn bản, ý tưởng của phương pháp đồ thị hoá là biểu diễn hay mô hình hoá văn bản dưới dạng một đồ thị. Đỉnh của đồ thị có thể đại diện cho một câu, một từ hoặc kết hợp câu và từ. Các cạnh của đồ thị thể hiện mối quan hệ về mặt ngữ nghĩa giữa các câu, trọng số của cạnh được xác định bởi giá trị sự tương đồng giữa hai câu. Kỹ thuật phổ biến hay dùng để xác định độ tương đồng giữa hai câu là tính độ đo cosine kết hợp với TF.IDF.

Một đồ thị cho chúng ta biết hai thông tin:

- Đồ thị con (sub-graphs) thể hiện sự phân vùng về chủ đề, tài liệu.

- Các câu quan trọng trong văn bản, câu quan trọng thường là câu có nhiều kết nối với các câu khác.



Hình 1.Đồ thị biểu diễn các câu trong văn bản

Đối với tóm tắt dành riêng cho truy vấn có thể câu chỉ cần chọn trong các đồ thị con, trong khi tóm tắt chung (generic summaries) câu cần chọn có thể lấy từ các đồ thị con.

Một số nghiên cứu điển hình gần đây như:

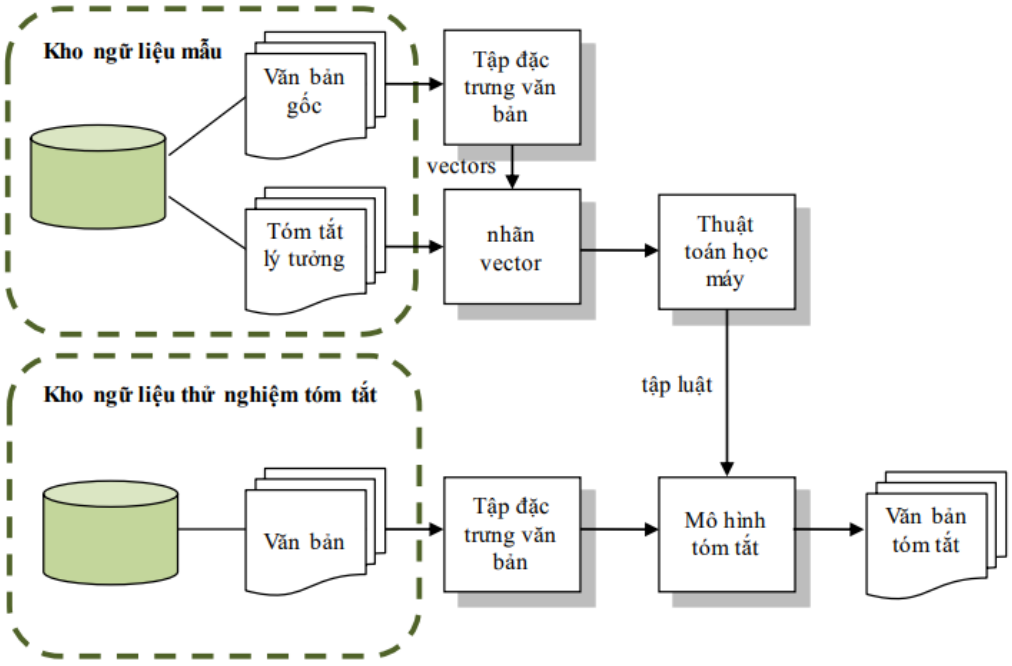
Nghiên cứu [27] của Kang Yang sử dụng thuật toán TextRank để trích chọn câu cho văn bản tóm tắt.

Nghiên cứu [15] của nhóm tác giả Rafael Ferreira đã đưa ra một mô hình đồ thị mới cho các ứng dụng xử lý văn bản, nhóm tác giả dựa vào bốn đặc điểm (4 chiều) (tương tự, giống nhau về ngữ nghĩa,đồng tham chiếu, thông tin diễn ngôn) để tạo ra đồ thị.

Nghiên cứu [17] của nhóm tác giả Xu Han đã sử dụng hệ thống FrameNet để xác định độ tương quan giữa các câu, sau cùng nhóm tác giả áp dụng thuật toán PageRank để xếp hạng và trích chọn câu cho văn bản tóm tắt.

b. Học máy

Với các tiến bộ của học máy, học máy cũng là một trong những phương pháp hiệu quả để xử lý bài toán tóm tắt văn bản dựa vào trích xuất câu. Các thuật toán tóm tắt dựa trên học máy sử dụng kỹ thuật như Naïve-Bayes, mô hình Markov ẩn HMM, K-mean…



Hình 2.Framework chung cho hệ thống tóm tắt văn bản bằng phương pháp học máy

(Nguồn ảnh: [1])

Một trong số những hạn chế với hầu hết các phương pháp tóm tắt văn bản hiện có là việc coi các câu là độc lập với nhau [26], vì vậy các chủ đề được nhúng trong các tài liệu bị coi nhẹ. Để cải thiện hạn chế đó, người ta có thể sử dụng mô hình Naïve-Bayes, bởi ý tưởng chính của mô hình Naïve-Bayes là tập trung vào việc xác định các câu, chuỗi từ liên quan đến chủ để của văn bản. Daume et al. [13] đề xuất BayeSum, một mô hình tóm tắt Bayes cho tóm tắt tập trung vào truy vấn. Wang và cộng sự [26] đã giới thiệu một mô hình tóm tắt dựa trên chủ đề áp dụng Bayes. Hệ thống của họ đạt được hiệu suất hiệu quả và vượt trội so với nhiều phương pháp tóm tắt khác.

Với K-mean, nghiên cứu [25] của nhóm tác giả Xinghao Song, đề xuất phương pháp vector hoá đồ thị bằng Node2Vec, mỗi vector đại diện cho một câu trong văn bản, sau đó dùng thuật toán K-mean để xác định các câu trọng tâm (câu trọng tâm tương ứng với trọng tâm K của các cụm).

Mô hình Markov ẩn (HMM), một nghiên cứu cho kết quả khá tốt khi sử dụng HMM là [12] của nhóm tác giả John M Conroy. Ý tưởng chính của nhóm nghiên cứu là xác định khả năng chọn các câu tiếp theo sẽ được chọn trong văn bản tóm tắt dựa trên việc đã xuất hiện của các câu trong văn bản tóm tắt trước đó.

2.2. Tóm tắt tóm lược

Các phương pháp tóm tắt tóm lược cố gắng để hiểu đầy đủ các văn bản cần tóm tắt, ngay cả các văn bản chủ đề không rõ ràng. Sau đó, tạo ra các câu mới cho bản tóm tắt theo tỉ lệ của người dùng yêu cầu [1]. Một cách ngắn gọn, yêu cầu của tóm tắt tóm lược là sao cho hệ thống tóm tắt càng giống với cách con người tóm tắt càng tốt.

Ví dụ văn bản gốc:

Trong báo cáo dự toán ngân sách 2013 trình bày chiều 22.10, Chính phủ cho biết chưa thể cân đối đủ nguồn để bố trí 60.000 tỉ đồng tăng lương tối thiểu lên 1,3 triệu đồng từ tháng 5 năm sau. Theo tính toán của Chính phủ, nếu thực hiện tăng lương lên 1,3 triệu đồng và nâng phụ cấp công vụ từ 25% lên 30% từ 1.5.2013, ngân sách nhà nước cần bố trí khoảng 60.000 tỉ đồng.
Chủ nhiệm Ủy ban các Vấn đề xã hội của Quốc hội Trương Thị Mai cho rằng: "Bộ Lao động - Thương binh và Xã hội đã nói là sẽ tăng lương cho khu vực doanh nghiệp, còn với khu vực nhà nước, Chính phủ tính lại rồi mới báo cáo Quốc hội cho ý kiến.
Nếu tăng theo lộ trình quy định thì năm 2013 cần tới 60 ngàn tỉ đồng để chi cho việc tăng lương. Với tình hình thu ngân sách nhà nước hiện nay thì đây là bài toán khó.
Tuy vậy, về mặt chủ quan thì cũng cần cân nhắc, tính toán, sắp xếp lại các khoản chi cho hợp lý để có thể tăng lương cho người lao động".

Văn bản tóm tắt:

Ngân sách nhà nước cần khoảng 60.000 tỉ đồng đề có thể tăng lương cơ bản lên 1.3 triệu. Bà Trương Thị Mai cho rằng: "Bộ Lao động – Thương binh và Xã hội sẽ tăng lương cho khu vực doanh nghiệp, còn doanh nghiệp nhà nước sẽ tính lại". Đây là một bài toán khó, vì vậy Chính phủ cần cân nhắc, tính toán, sắp xếp hợp lý các khoản chi để có thể tăng lương cho người lao động.

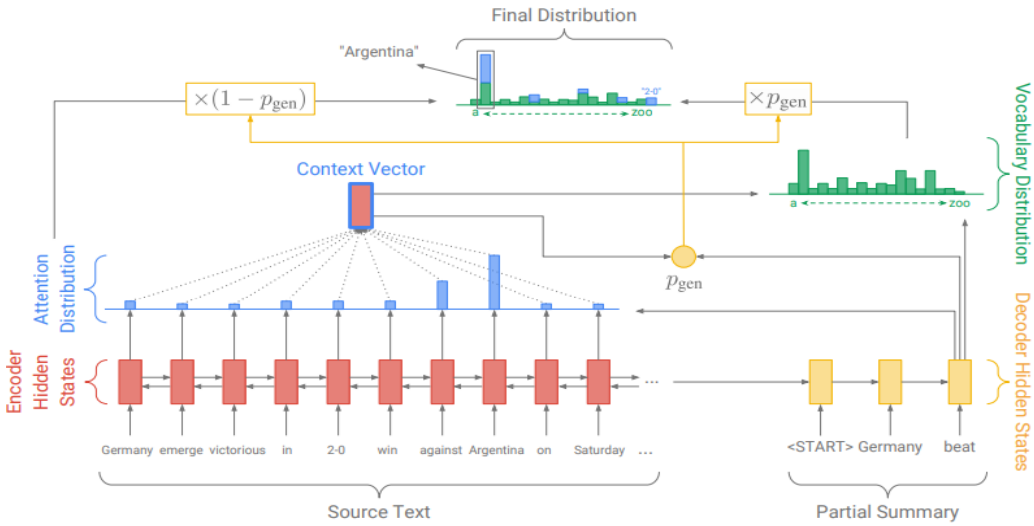
Các kỹ thuật liên quan đến tóm tắt tóm lược bao gồm phân tích cú pháp, phân tích ngữ nghĩa, và sinh ngôn ngữ tự nhiên. Hiện nay, có hai hướng tiếp cận chính cho bài toán tóm tắt tóm lược là tiếp cận dựa trên cấu trúc, và tiếp cận dựa trên ngữ nghĩa.

a. Phương pháp tiếp cận dựa trên cấu trúc: Một ý tưởng điển hình của phương pháp này là cố gắng xây dựng một hệ thống sinh ra văn bản tóm tắt

bằng cách tự động hoàn thiện nội dung vào các mẫu cho trước, các mẫu được xây dựng có cấu trúc với các vị trí được sử dụng để xác định các thông tin quan trọng cần trích rút. Mỗi một chủ đề, một vấn đề cần có một mẫu riêng. Ngoài ra phương pháp này còn có các kỹ thuật, ý tưởng khác như dựa trên cây văn bản, dựa trên Ontology, dựa trên tập luật.

b. Phương pháp tiếp cận dựa trên ngữ nghĩa: Hệ thống sinh ra văn bản tóm tắt dựa trên những phân tích về ngữ nghĩa của văn bản đầu vào, trong đó đặc biệt quan tâm tới việc xác định các cụm danh từ và cụm động từ để làm cơ sở cho các kỹ thuật sinh ngôn ngữ . Một số kỹ thuật áp dụng cho phương pháp này như kỹ thuật dựa trên mô hình ngữ nghĩa đa phương thức, dựa trên thông tin, dựa trên đồ thị ngữ nghĩa.

c. Phương pháp tiếp cận dựa trên học sâu (deep learning): Sequence-to-Sequence là một kỹ thuật điển hình áp dụng cho các mô hình tóm tắt văn bản đi theo hướng này. Mặc dù Sequence-to-Sequence đã được áp dụng thành công cho nhiều bài toán trong xử lý ngôn ngữ tự nhiên, chẳng hạn như dịch máy, nhưng với bài toán tóm tắt văn bản vẫn còn nhiều hạn chế. Thực tế là mô hình này có thể đạt được điểm ROUGE cao trên các bản tóm tắt với đầu vào nhỏ, nhưng thường không có khả năng tóm tắt khi đầu vào lớn.



Hình 3.Một mô hình tóm tắt văn bản sử dụng kỹ thuật Sequence-to-Sequence with Attention

(Nguồn ảnh: [24])

Hình 3 là một mô hình tóm tắt văn bản sử dụng kỹ thuật Sequence-to-Sequence trong nghiên cứu [24], nhóm tác giả xây dựng mô hình này với 3 thành phần chính:

- Bộ mã hóa – LSTM (Long Short Term Memory) là trường hợp đặc biệt của RNN (Recurrent Neural Networks, có khả năng học với sự phụ thuộc lâu dài của các nơ-ron trích xuất thông tin từ văn bản gốc. Điều này được thể hiện bằng màu đỏ trong mô hình. LSTM đọc một từ tại một thời điểm và nó cập nhật trạng thái ẩn dựa trên từ hiện tại và các từ đã đọc trước đó.
- Bộ giải mã - Lớp LSTM Uni-directional tạo ra một từ tóm tắt tại một thời điểm. Bộ giải mã LSTM bắt đầu hoạt động khi nhận được tín hiệu rằng văn bản nguồn đã được đọc toàn bộ. Nó sử dụng thông tin từ bộ mã hóa cũng như những gì đã được viết trước đó để xác định phân phối xác suất cho từ tiếp theo. Bộ giải mã được hiển thị màu vàng, và phân phối xác suất màu xanh lá cây.
- Cơ chế Attention: Đầu vào của bộ giải mã là trạng thái ẩn cuối cùng từ bộ mã hóa có thể là vector 256 hoặc 512 chiều, thông thường vector nhỏ này khó có thể chứa tất cả thông tin. Thông qua cơ chế attention, bộ giải mã có thể truy cập các trạng thái ẩn trung gian của bộ mã hóa và sử dụng tất cả thông tin đó để quyết định từ nào tiếp theo. Attention được thể hiện bằng màu xanh da trời trong mô hình.

2.3. Một số nghiên cứu tóm tắt văn bản tiếng Việt hiện nay

2.3.1. Đặc điểm của tiếng Việt

2.3.1.1 Đặc điểm về từ

Một từ trong tiếng Anh chỉ gồm một tiếng, khác với tiếng Anh, tiếng Việt bao gồm hai loại từ là từ đơn và từ ghép. Từ đơn chỉ gồm một tiếng tạo thành. Từ ghép được tạo ra bằng cách ghép hai hoặc nhiều tiếng có quan hệ với nhau về ngữ nghĩa. Như vậy, một từ trong tiếng Việt có thể có thể được cấu thành bởi lớn hơn một tiếng. Ví dụ: Từ "giảng viên" là một từ ghép gồm hai tiếng "giảng" và "viên". Trong hầu hết các hệ thống tóm tắt văn bản, tách từ là một công việc quan trọng cần thực hiện tại bước tiền xử lý.

Nghĩa của một từ là nội dung sự vật, sự việc, tính chất, mối quan hệ…mà từ biểu thị, có hai cách giải thích nghĩa của một từ:1) trình bày khái niệm; 2) đưa ra từ đồng nghĩa hoặc trái nghĩa [8]. Từ đồng nghĩa là những từ có nghĩa tương tự nhau, có thể thay thế cho nhau trong một số hoàn cảnh nhất định. Từ trái nghĩa là những từ có nghĩa trái ngược nhau.

Tiếng Việt có nhiều loại từ, nhưng trong luận văn tôi xin đưa ra khái niệm của ba loại từ chính là danh từ, động từ và tính từ.

- **Danh từ :** Là những từ chỉ đối tượng, khái niệm. Danh từ có thể bao gồm từ chỉ lượng ở phía trước, và các từ như "này, ấy, đó" ở

phía sau. Trong câu danh từ đóng vai trò là chủ ngữ. Khi làm vị ngữ danh từ cần có từ "là" đứng trước. Danh từ gồm hai loại là danh từ chỉ số lượng và danh từ chỉ sự vật. Danh từ chỉ sự vật lại bao gồm hai loại là danh từ chỉ tên riêng và danh từ chung.

- **Động từ:** Là những từ diễn tả trạng thái, hành vi của sự vật. Động từ thường kết hợp với các từ "đã, đang, hãy, đừng…". Trong đa phần các câu, động từ thường đóng vai trò là vị ngữ, nhưng trong một số trường hợp động từ lại đóng vai trò là chủ ngữ. Động từ có thể chia thành hai loại là động từ tình thái, và động từ chỉ hành động.Ví dụ: Một số động từ tình thái là: "sẽ", "có thể", "nên"… Một số động từ chỉ hành động như: "đi", "học", "nói".
- **Tính từ**: Là những từ chỉ tính chất của sự vật, sự việc. Ví dụ như: "xinh đẹp", "lung linh"…

2.3.1.2 Đặc điểm về câu

Hai thành phần chính trong tiếng Việt là chủ ngữ và vị ngữ [8]. Chủ ngữ trả lời cho câu hỏi là ai, cái gì, con gì… Vị ngữ trả lời cho câu hỏi làm gì, đi đâu, như thế nào…Ví dụ:

- "Tôi làm luận văn thạc sĩ". Chủ ngữ trong câu là "tôi", vị ngữ trong câu là "làm luận văn thạc sĩ".
- "Làm nông nghiệp cần quan tâm tới thời tiết". Chủ ngữ là "làm nông nghiệp", vị ngữ là "cần quan tâm tới thời tiết".

Ngoài chủ ngữ và vị ngữ trong câu còn có thêm trạng ngữ, định ngữ, và bổ ngữ.

Câu bao gồm có câu đơn và câu ghép.

Câu ghép là câu có lớn hơn hoặc bằng hai vế, mỗi vế mang câu trúc tương tự câu đơn. Câu ghép gồm hai loại là câu ghép đẳng lập và câu ghép chính phụ:

- Câu ghép đẳng lập là câu mà các vế trong câu độc lập về nghĩa. Ví dụ: "Tôi học đại học còn em tôi học trung học" hay "Bầu trời quang đãng và gió trong lành", "Mùa hè nắng nóng còn mùa thu không khí mát mẻ".
- Câu ghép chính phụ là câu bao gồm hai vế, một vế chính và một vế phụ, hai vế có quan hệ về mặt nghĩa, và được kết nối với nhau bằng các cặp quan hệ từ "vì-nên", "nếu-thì", "mặc dù-nhưng". Ví dụ câu: "Mặc dù thời gian ngắn nhưng anh ấy vẫn hoàn thành nhiệm vụ

được giao".

2.3.2 Một số nghiên cứu tóm tắt văn bản tiếng Việt

2.3.2.1. Trích rút

Đối với tóm tắt tiếng Việt, hầu hết các nghiên cứu về tóm tắt văn bản tiếng Việt đi theo hướng trích rút câu.

Một số công trình công bố điển hình dạng này như:

Nghiên cứu của Nguyễn Thị Thu Hà [5] đề xuất xây dựng hệ thống tóm tắt văn bản tiếng Việt dựa trên trích xuất câu và rút gọn câu. Việc trích rút câu được thực hiện theo hai phương pháp: 1) dựa trên lý thuyết tập mờ và mô hình chủ đề; và 2) dựa trên lượng thông tin và độ ngôn ngữ. Việc rút gọn câu được thực hiện theo hai cách: 1) xác định chuỗi phù hợp và 2) kết nối các chuỗi con phù hợp nhất.

Đỗ Phúc và các cộng sự rút trích nội dung chính của khối thông điệp bằng phương pháp gom cụm đồ thị [6].

Nghiên cứu của nhóm tác giả Nguyễn Thị Ngọc Tú, xây dựng mô hình đồ thị trong tóm tắt văn bản tiếng Việt với nghiên cứu "ứng dụng đồ thị trong tóm tắt đa văn bản tiếng Việt" [9].

Ngoài ra còn có sự góp mặt của nhóm tác giả Nguyễn Trọng Phúc và Lê Thanh Hương [7] sử dụng cấu trúc diễn ngôn tiếng Việt đối với hệ thống tóm tắt tự động. Cấu trúc diễn ngôn là một phương tiện cho phép biểu diễn mối quan hệ diễn ngôn giữa các đoạn văn bản. Cây cấu trúc diễn ngôn cho phép đánh giá được tầm quan trọng của các mệnh đề và các câu. Trên cơ sở đó có thể trích rút các câu quan trọng đưa vào văn bản tóm tắt.

Nghiên cứu [1] của Nguyễn Nhật An đề xuất phương pháp tóm tắt văn bản tiếng Việt theo hướng trích rút dựa trên bộ hệ số đặc trưng.

Nhóm tác giả Trương Quốc Định và Nguyễn Quang Dũng cũng đã đề cập đến phương pháp dựa trên mô hình đồ thị có trọng số [3]. Mỗi đỉnh của đồ thị biểu diễn một câu, cạnh nối hai câu có gán trọng số thể hiện độ tương đồng ngữ nghĩa của chúng và cuối cùng một giải thuật PageRank dựa trên đồ thị được tùy biến để tích hợp độ tương tự câu. Sau cùng các câu quan trọng nhất sẽ được trích rút trong văn bản tóm tắt .

2.3.2.2. Tóm lược

Học sâu là phương pháp học máy được nghiên cứu và sử dụng rộng rãi trong những năm gần đây, mở ra hướng đi mới cho các bài toán như xử lý ảnh, xử lý tiếng nói và xử lý ngôn ngữ tự nhiên... Với tiếng Việt, chưa có nhiều nghiên cứu về hướng tiếp cận này nên việc áp dụng bài toán này trong thực tế là một điều thú vị và mới mẻ và hứa hẹn nhiều khả năng phát triển.

Nghiên cứu [10] của nhóm tác giả Lâm Quang Tường, đã sử dụng học sâu cho bài toán tóm tắt văn bản tự động đối với tiếng Việt. Đây được coi như một nghiên cứu xuất bản chính thức đầu tiên theo hướng tóm lược cho bài toàn tóm tắt văn bản tiếng Việt. Nhóm tác giả đã sử dụng mô hình Word2vec để rút trích những đặc trưng riêng của văn bản tiếng Việt, phục vụ cho mô hình Sequence to sequence with Attention nhằm tạo kết quả đầu ra là chuỗi các từ. Tuy kết quả còn chưa cao nhưng mô hình đã giải quyết thành công mục tiêu của bài toán.

Đề tài "Tóm tắt văn bản sử dụng các kỹ thuật trong deep learning" [2] của tác giả Đoàn Xuân Dũng, tác giả đã sử dụng mạng nơ-ron tích chập với mạng GRU (Gated Recurrent Units) kết hợp với cơ chế Attention để giải quyết bài toán tóm tắt tóm lược văn bản tiếng Việt. Tác giả đã tiến hành thực nghiệm trên hai bộ dữ liệu khác nhau, với các cấu hình mạng CNN (Convolution Neural Network) khác nhau, kết quả cho thấy nghiên cứu cho kết quả khá khả quan.

CHƯƠNG 3. XÂY DỰNG MÔ HÌNH TÓM TẮT VĂN BẢN TIẾNG VIỆT THEO PHƯƠNG PHÁP ĐỒ THỊ.

Trong chương 3, luận văn tập trung trình bày ba vấn đề. Vấn đề thứ nhất là thuật toán iSpreadRank, vấn đề thứ hai là đưa ra mô hình tóm tắt văn bản tiếng Việt dựa theo phương pháp đồ thị áp dụng thuật toán iSpreadRank, vấn đề thứ ba là xây dựng mô hình, trong đó trình bày chi tiết các bước, và các thuật toán dùng trong từng bước.

3.1. Thuật toán iSpreadRank

iSpreadRank [29] được Jen-Yuan Yeh và cộng sự đề xuất áp dụng cho bài toán tóm tắt văn bản theo hướng tiếp cận trích xuất câu.

Đầu vào của thuật toán iSpreadRank:

- Ma trận biểu diễn sự liên kết của các câu trong văn bản, ma trận này được suy ra từ đồ thị có trọng số thể hiện sự tương đồng giữa các câu.
- Trọng số (độ quan trọng) ban đầu của các câu.

Về bản chất iSpreadRank là một dạng của thuật toán lan truyền kích hoạt, đối tượng kích hoạt lan truyền là trọng số của các câu, iSpreadRank cho rằng trọng số của một câu phụ thuộc vào 3 yếu tố: 1) số lượng câu mà có sự kết nối với nó; 2) trọng số của các câu kết nối với nó; 3) sức mạnh liên kết của câu đó với các câu khác, "sức mạnh liên kết" được đo bằng độ tương đồng, nghĩa là hai câu càng tương đồng thì "sức mạnh liên kết" càng lớn và ngược lại. Trọng số của các câu được cập nhật và điều chỉnh lặp đi lặp lại trên toàn mạng, bảng xếp hạng câu được suy ra theo thứ tự tầm quan trọng của các câu. Thuật toán iSpreadRank chia làm ba bước: 1) Khởi tạo; 2) Suy diễn; 3) Dự đoán.

- Bước khởi tạo: Biến đổi đồ thị có trọng số ban đầu thành ma trận kề để tính toán.
- Bước suy diễn: Tính toán độ quan trọng của các câu.
- Bước dự đoán: Đưa ra bảng xếp hạng các câu dựa trên kết quả của bước suy diễn.

3.1.1. Khởi tạo

Gọi $G = (V, E)$ là đồ thị có trọng số biểu diễn sự tương đồng giữa các câu, trong đó $V = \{s_1, s_2, \dots s_n\}$ là tập các đỉnh của đồ thị, E là tập cạnh của đồ

thị. Đồ thị G sẽ được biến đổi thành một ma trận kề A, các hàng và các cột được gán nhãn bởi chỉ số của các đỉnh, giá trị của các phần tử trong ma trận kề được tính bởi công thức:

$$a_{ij} = \begin{cases} 0 & if\ i = j \\ sim(s_i, s_j) & if\ i \neq j \end{cases} \quad (3.1)$$

Vì G là đồ thị vô hướng nên A sẽ là ma trận đối xứng , nên $a_{ij} = a_{ji}$

3.1.2. Suy luận

Mỗi đỉnh trong đồ thị có một trọng số đại diện cho tầm quan trọng của một câu, thuật toán được lặp đi lặp lại để cập nhật liên tục các trọng số này cho đến khi đạt được trạng thái ổn định hoặc gặp điều kiện dừng, trong luận văn tôi đặt điều kiện dừng là số lần lặp bằng 500. Trong mỗi lần lặp, các đỉnh sẽ có một trọng số mới bằng cách "thu thập" trọng số của các đỉnh liên kết với nó, sau đó lại truyền đi trọng số mới đó đến các đỉnh liên kết lân cận.

Việc lan truyền này, được biểu diễn bằng công thức đại số sau đây:

$$X(t) = X(0) + M \times X(t-1), \quad M = \sigma R^T \quad (3.2)$$

Trong đó:

X: Là vector n chiều lưu giữ trọng số của n đỉnh $\{s_1, s_2, \dots s_n\}$ của đồ thị.

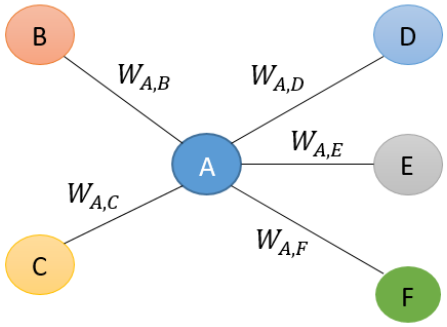
σ: Ý nghĩa của tham số này chỉ ra tỉ lệ trọng số của một đỉnh truyền cho các đỉnh khác, trong luận văn $\sigma = 0.8$

R: Là ma trận thu được từ ma trận kề A bằng công thức: $r_{ij} = \frac{a_{ij}}{\sum_k a_{ik}}$

Trạng thái cân bằng của đồ thị đạt được khi:

$$\sum_i |X_i(t) - X_I(t-1)| \leq \varepsilon, \qquad \varepsilon = 0.0001 \quad (3.3)$$

Hình 4 minh hoạ việc thu thập và lan truyền trọng số của đỉnh A trong một lần lặp:



Hình 4.Minh hoạ quá trình lan truyền kích hoạt

Bước 1. Thu thập trọng số từ các đỉnh lân cận trong lần lặp t

$$X_A(t) = X_A(0) + \sigma\frac{W_{B,A}}{\sum_K W_{B,K}}X_B(t-1) + \sigma\frac{W_{C,A}}{\sum_K W_{C,K}}X_C(t-1) + \sigma\frac{W_{D,A}}{\sum_K W_{D,K}}X_D(t-1) + \sigma\frac{W_{E,A}}{\sum_K W_{E,K}}X_E(t-1) + \sigma\frac{W_{F,A}}{\sum_K W_{F,K}}X_F(t-1) \quad (3.4)$$

Bước 2. Lan truyền trọng số hiện tại cho các đỉnh lân cận trong lần lặp kế tiếp (t+1)

$$X_B(t+1) = X_B(0) + \sigma\frac{W_{A,B}}{\sum_K W_{A,K}}X_A(t) + \cdots \quad (3.5)$$

$$X_C(t+1) = X_C(0) + \sigma\frac{W_{A,C}}{\sum_K W_{A,K}}X_A(t) + \cdots \quad (3.6)$$

$$X_D(t+1) = X_D(0) + \sigma\frac{W_{A,D}}{\sum_K W_{A,K}}X_A(t) + \cdots \quad (3.7)$$

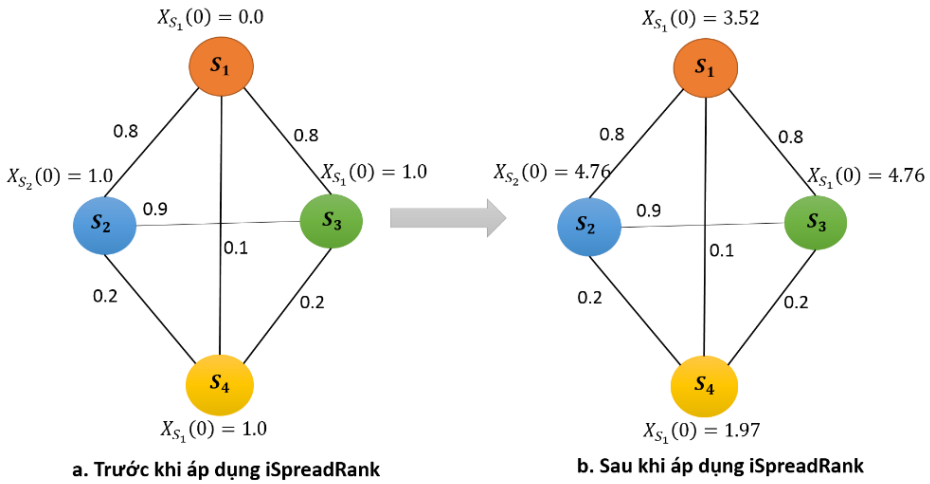
$$X_E(t+1) = X_E(0) + \sigma\frac{W_{A,E}}{\sum_K W_{A,K}}X_A(t) + \cdots \quad (3.8)$$

$$X_F(t+1) = X_F(0) + \sigma\frac{W_{A,F}}{\sum_K W_{A,K}}X_A(t) + \cdots \quad (3.9)$$

3.1.3. Dự đoán

Khi thuật toán kết thúc, đồ thị đạt trạng thái ổn định với mỗi đỉnh được gán nhãn có trọng số là độ quan trọng của nó.

Quan sát ví dụ dưới đây:



Hình 5.Trọng số đỉnh của đồ thị trước và sau áp dụng thuật toán iSpreadRank

Hình 5 giải thích cơ chế hoạt động của thuật toán iSpreadRank. Hình 5a biểu diễn trạng thái ban đầu của đồ thị, hạng ban đầu là: Hạng (S2) = hạng(S3) = hạng(S4) > hạng(S1). Hình 5b biểu diễn trạng thái hội tụ của iSpreadRank, chúng ta có một mạng mới, trong đó: Hạng(S2) = hạng(S3) > hạng(S1) > hạng(S4). Có thể thấy sự thay đổi khi S1 được thăng hạng lên vị trí trước S4 trong bảng xếp hạng mới. Bảng 1 sẽ mô tả chi tiết giá trị các tham số tương ứng với ví dụ trên.

Bảng 1.Chi tiết các tham số trong thuật toán iSpreadRank

A Ma trận kề của đồ thị có trọng số biểu diễn sự tương đồng của các câu	$\begin{bmatrix} 0.0 & 0.8 & 0.8 & 0.1 \\ 0.8 & 0.0 & 0.9 & 0.2 \\ 0.8 & 0.9 & 0.0 & 0.2 \\ 0.1 & 0.2 & 0.2 & 0.0 \end{bmatrix}$
R Ma trận thu được từ ma trận A bởi công thức: $r_{ij} = \frac{a_{ij}}{\sum_k a_{ik}}$	$\begin{bmatrix} 0.0 & \frac{0.8}{1.7} & \frac{0.8}{1.7} & \frac{0.1}{1.7} \\ \frac{0.8}{1.9} & 0.0 & \frac{0.9}{1.9} & \frac{0.2}{1.9} \\ \frac{0.8}{1.9} & \frac{0.9}{1.9} & 0.0 & \frac{0.2}{1.9} \\ \frac{0.1}{0.5} & \frac{0.2}{0.5} & \frac{0.2}{0.5} & 0.0 \end{bmatrix}$

R^T Ma trận chuyển vị của ma trận R	$\begin{bmatrix} 0.0 & \frac{0.8}{1.9} & \frac{0.8}{1.9} & \frac{0.1}{0.5} \\ \frac{0.8}{1.7} & 0.0 & \frac{0.9}{1.9} & \frac{0.2}{0.5} \\ \frac{0.8}{1.7} & \frac{0.9}{1.9} & 0.0 & \frac{0.2}{0.5} \\ \frac{0.1}{1.7} & \frac{0.2}{1.9} & \frac{0.2}{1.9} & 0.0 \end{bmatrix}$
X(0) Vector chứa thông tin về trọng số khởi tạo ban đầu của các đỉnh trong câu	$\begin{bmatrix} x1 \\ x2 \\ \dots \\ xn \end{bmatrix}$ Trọng số khởi tạo được tính bằng nhiều phương pháp: PageRank, Đặc trưng của câu…
$X_A(t)$ Vector chứa thông tin về trọng số của đỉnh A trong lần lặp t	$X(t) = X(0) + \sigma R^T \times X(t-1)$

Giả sử $X(0) = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$, X(1) được tính như sau:

$$X(1) = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} + 0.8 \times \begin{bmatrix} 0.0 & \frac{0.8}{1.9} & \frac{0.8}{1.9} & \frac{0.1}{0.5} \\ \frac{0.8}{1.7} & 0.0 & \frac{0.9}{1.9} & \frac{0.2}{0.5} \\ \frac{0.8}{1.7} & \frac{0.9}{1.9} & 0.0 & \frac{0.2}{0.5} \\ \frac{0.1}{1.7} & \frac{0.2}{1.9} & \frac{0.2}{1.9} & 0.0 \end{bmatrix} \times \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.8337 \\ 1.6989 \\ 1.6989 \\ 1.1684 \end{bmatrix} \quad (3.10)$$

Sau 20 lần lặp, mạng đạt được trạng thái ổn định:

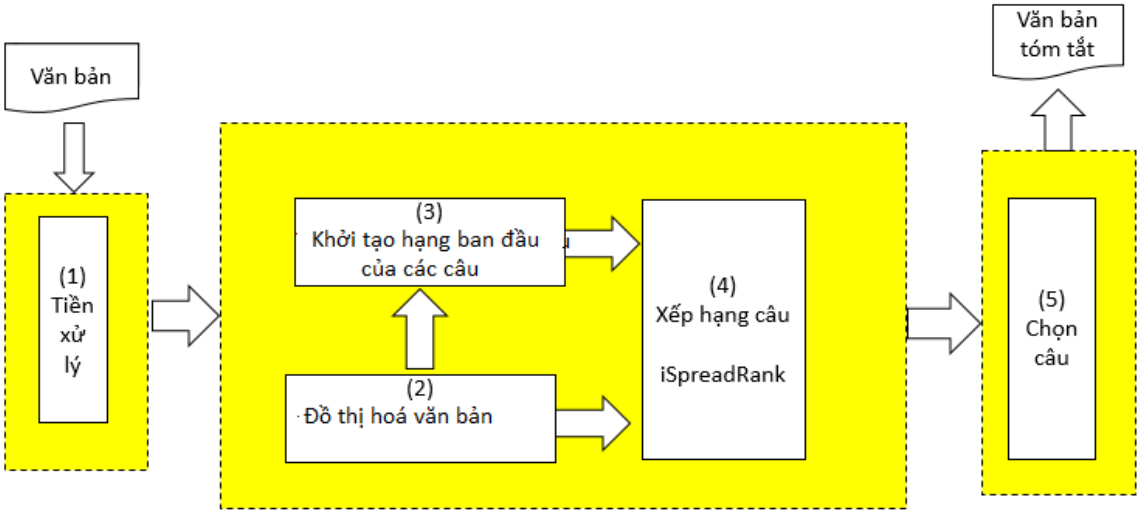
Bảng 2.Kết quả thực hiện thuật toán sau 20 lần lặp

Số lần lặp	S_1	S_2	S_3	S_4
0	0.0000	1.0000	1.0000	1.0000
1	0.8337	1.6989	1.6989	1.1684
5	2.4058	3.5114	3.5114	1.6392
10	3.1543	4.3489	4.3489	1.8594
20	3.4802	4.7131	4.7131	1.9552
Hội tụ	3.5193	4.7568	4.7568	1.9667

Theo bảng 2 trọng số của câu 1 tăng nhanh hơn trọng số của câu 4. Điều này do câu 1 liên kết mạnh với câu 2 và câu 3, do đó nhận nhiều "năng lượng" hơn từ câu 2 và câu 3. Ngược lại, câu 2 và câu 3 truyền ít năng lượng hơn cho câu 4 vì câu 4 có liên kết yếu với câu 2 và câu 3.

3.2. Thiết kế mô hình

Dựa vào thuật toán iSpreadRank, tôi xây dựng mô hình tóm tắt văn bản tự động cho tiếng Việt như sau:



Hình 6.Mô hình tóm tắt văn bản tiếng Việt áp dụng thuật toán iSpreadRank

Quá trình tóm tắt được chia thành 3 pha chính: 1) tiền xử lý; 2) xếp hạng câu; 3) sinh văn bản tóm tắt. Chi tiết các bước thực hiện, có thể chia thành 5 bước: 1) tiền xử lý; 2) đồ thị hoá văn bản; 3) khởi tạo hạng ban đầu; 4) cập nhật hạng của các câu ; 5) chọn câu vào văn bản tóm tắt.

3.2.1. Tiền xử lý

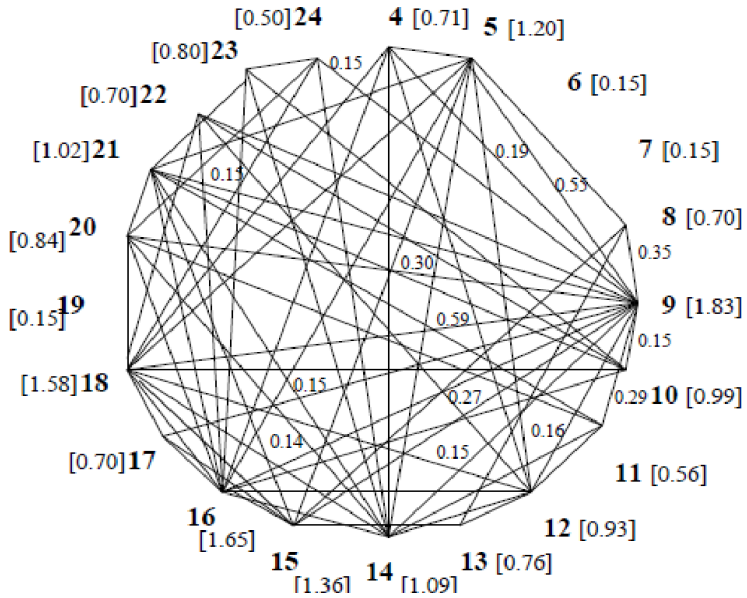
Chuẩn hóa văn bản đầu vào, bao gồm các công việc sau:

- Loại bỏ các ký tự thừa, các thành phần không thuộc văn bản tiếng Việt.
- Loại bỏ từ dừng.
    - Từ dừng là những từ không có ý nghĩa từ vựng, nó được dùng đi kèm với thực từ để thiết lập mối quan hệ giữa các sự vật hiện tượng do thực từ diễn đạt. Ví dụ: "Tôi đang viết luận văn thạc sĩ". Trong câu này, "đang " là một hư từ chỉ thời gian. Nó đi kèm với từ "viết", làm thành phụ tố cho từ đó và tạo thành cụm từ "đang viết luận văn" (cả cụm này làm vị ngữ của câu). Ta thấy, khi bỏ từ "đang" câu vẫn mang đủ ý nghĩa chính "tôi viết luận văn".
- Tách từ. Tách từ là việc xác định danh giới của các từ, vì như trình bày tại phần 2.3.1.1, từ trong tiếng Việt có thể là từ đơn hoặc từ ghép.
- Tách câu. Xác định các câu có trong văn bản.

Trong bước này, tôi sử dụng thư viện vnTokenizer [20] để thực hiện tách từ và tách câu. Liên kết tải thư viện: http://mim.hus.vnu.edu.vn/dsl/tools/tokenizer

3.2.2. Đồ thị hoá văn bản.

Văn bản đầu vào được mô hình hoá thành một đồ thị vô hướng và có trọng số. Trong đó mỗi đỉnh đại diện cho một câu, trọng số của câu được xác định bởi đại lượng độ tương đồng giữa hai câu. Để tính giá trị tương đồng này, mỗi câu được sẽ được biểu diễn thành một vector. Hình 7 minh hoạ của một đồ thị tương đồng:



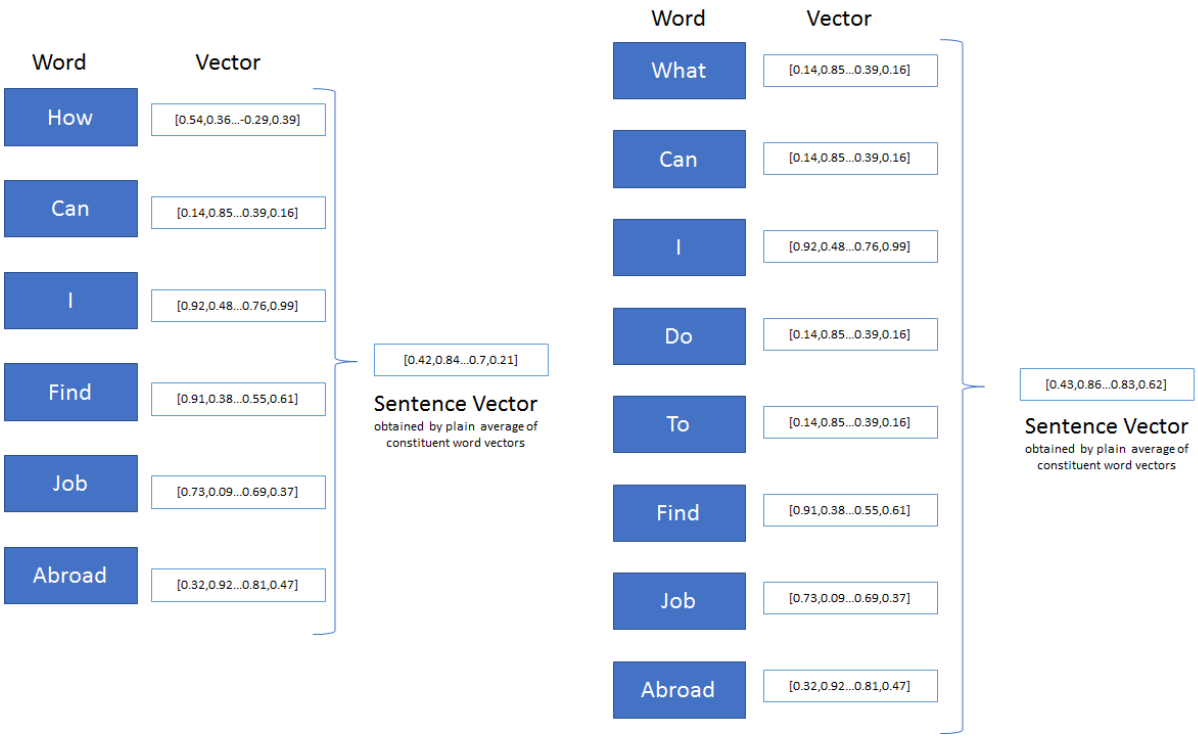
Hình 7.Đồ thị mạng tương đồng của các câu trong văn bản

(Nguồn ảnh: [14])

Hai đỉnh được gọi là có liên kết khi và chỉ khi chúng có tính tương đồng nhau, và giá trị tương đồng ấy phải lớn hơn một ngưỡng cho trước. Do vậy, trọng số cạnh giữa hai đỉnh của đồ thị thể hiện mối quan hệ về mặt ngữ nghĩa . Để tính độ tương đồng giữa hai câu, có hai nhiệm vụ cần hoàn thành đó là:

- Chuẩn hoá câu thành vector.
- Sau khi có vector câu, tiến hành tính cosin góc giữa hai vector, giá trị này chính là giá trị tương đồng mà ta cần tính. Giá trị cosin càng lớn, góc giữa hai vector càng nhỏ, hai câu càng tương đồng.

3.2.2.1. Chuẩn hoá câu thành vector dựa vào Word2Vec.



Hình 8.Ví dụ về chuyển đổi vector từ sang vector câu

(Nguồn ảnh: https://images.viblo.asia)

Ý tưởng của phương pháp này là biểu diễn một từ thành một vector, vector câu sẽ được tính bằng tổng hoặc trung bình của tất cả các vector từ.

Cách đơn giản nhất để biểu diễn từ thành vector là one-hot-vector bằng 2 bước: 1) Tạo từ điển từ tất cả các tập từ trong văn bản; 2) Tạo vector có số chiều bằng kích thước từ điển, các phần tử của vector chỉ có giá trị 0 và duy nhất một phần tử bằng 1 tại ví trí xuất hiện của từ đó trong từ điển. Tuy nhiên, phương pháp này có những hạn chế như sau:

- Với những bộ dữ liệu lớn, số chiều của vector quá lớn.
- Không xác định được mối tương quan ngữ nghĩa giữa các từ.

Word2Vec [19] có thể giải quyết nhược điểm của One-hot-vector. Word2Vec là kỹ thuật chuyển đổi một từ sang một vector trong không gian vector với số chiều thấp mà vẫn giữ được thông tin ngữ cảnh của từ đó.

Chúng ta có thể dễ dàng training một mô hình Word2Vec bằng thư viện gensim trong python:

```
model = Word2Vec(train_data, size=100, window=10, min_count=3, workers=
4, sg=1)
```

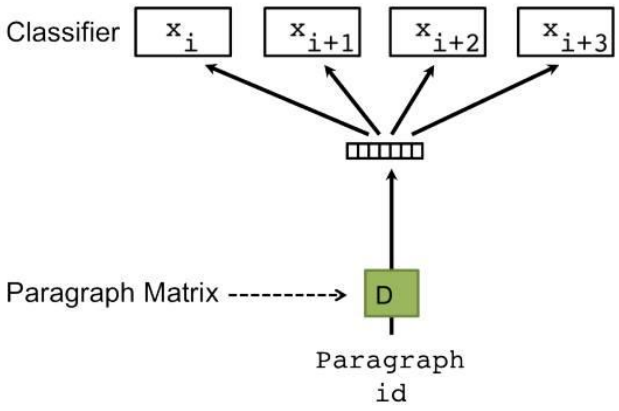
Trong đó:

- train_data: Tập dữ liệu.
- size: Kích thước vector.
- window: Khoảng cách tối đa của một từ với từ dự đoán.
- min_count: Từ nào xuất hiện với số lần nhỏ hơn min_count sẽ bị loại bỏ.

3.2.2.2. Chuẩn hoá câu thành vector dựa vào doc2Vec.

Trong nghiên cứu [21] của Quốc Lê và Tomas Mikolov, nhóm tác giả đã trình bày một phương pháp có khả năng biểu diễn các câu văn, đoạn văn hay cả văn bản thành một vector. Tôi mặc định vai trò của Doc2vec là biểu diễn câu thành vector. Doc2vec bao gồm 2 phần chính là: DBOW(distributed bag of words) và DM (distributed memory).

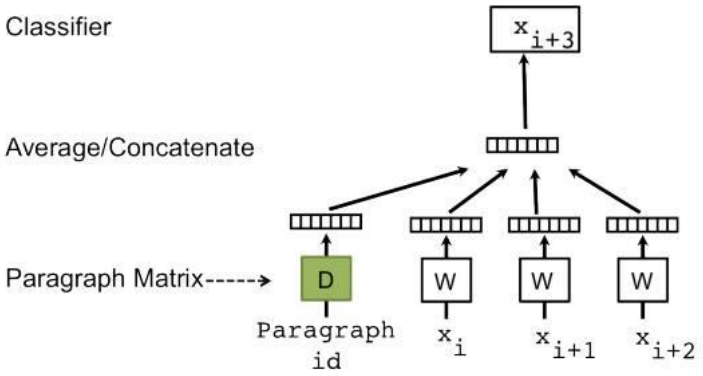
DBOW: ParagraphID đại diện cho câu được training. Sau khi training vector ParagraphID chính là vector câu mà chúng ta muốn tìm.



Hình 9.Phân phối Bag of Words của vector câu.

(Nguồn ảnh: [21])

DM: Coi các câu chỉ là các "từ", sau đó nối "từ " này vào tập các từ trong câu. Trong quá trình training, các phần tử trong vector câu và vector từ đều được cập nhật.



Hình 10.Mô hình cập nhật vector câu

(Nguồn: [21])

Chúng ta có thể tạo một mô hình Doc2vec bằng thư viện gensim trong python.

3.2.2.3. Chuẩn hoá câu thành vector dựa vào TF.ISF.

TF.ISF(Term frequency-inverse sentence frequency) bắt nguồn từ TF.IDF, TF.IDF được sử dụng để xác định mức độ quan trọng của một từ trong một văn bản, mà văn bản đó nằm trong một tập các văn bản. Luận văn này, tiếp cận bài toán tóm tắt đơn văn bản nên tôi sử dụng trọng số TF.ISF.

$$TF.ISF(w) = TF(w,s) \times ISF(w) \quad (3.11)$$

Trong đó:

- $TF(w,s)$: Là tần số xuất hiện của từ w trong câu s
- $ISF(w) = \log\frac{N_s}{SF(w)}$: Là tần số nghịch của từ w trong tập câu thuộc văn bản. N_s là tổng số câu trong văn bản, $SF(w)$ là tổng số câu chứa từ w.

Trong luận văn tôi sử dụng phương pháp này để chuẩn hoá câu thành vector bởi sự ổn định, xử lý nhanh và không cần dữ liệu training.

3.2.2.4. Tính độ tương đồng của hai câu bằng cosine.

Độ tương đồng của hai câu được tính bằng công thức cosine như sau:

$$sim(s_i, s_j) = \frac{\overrightarrow{s_i} \times \overrightarrow{s_j}}{|\overrightarrow{s_i}| \times |\overrightarrow{s_j}|} \quad (3.12)$$

Trong đó s_i, s_j là các vector câu.

$$\overrightarrow{s_i} = (w_{w_{i1}}, w_{w_{i2}}, w_{w_{i3}}, \dots, w_{w_{in}})$$

$$\overrightarrow{s_j} = (w_{w_{j1}}, w_{w_{j1}}, w_{w_{j1}}, \dots, w_{w_{jn}})$$

3.2.3. Khởi tạo hạng ban đầu của các câu

Để xác định hạng (trọng số) ban đầu của các câu, luận văn này tôi thử nghiệm hai cách với hai phương pháp là tính trọng số ban đầu của các câu bằng điểm PageRank và bằng điểm đặc trưng (feature score).

3.2.3.1. Điểm PageRank

PageRank được phát triển tại đại học Stanford bởi Lary Page, và sau đó Sergey Brin như một phần dự án xây dựng công cụ tìm kiếm.

$$PR(T_1) = 1 - d + d \times (\frac{PR(T_2)}{C(T_2)} + \cdots + \frac{PR(T_n)}{C(T_n)}) \quad (3.13)$$

Trong đó:

$PR(T_1)$: PageRank của đỉnh T_1.

d: Damping factor. Trong luận văn d = 0.85

$PR(T_2)$: PageRank của đỉnh T_2. Đỉnh T_2 có liên kết với T_1.

$C(T_2)$: Tổng số liên kết của đỉnh T_2 với các đỉnh khác.

3.2.3.2. Điểm đặc trưng

Cách xác định điểm đặc trưng được trình bày trong nghiên cứu [29]. Trong đó quan tâm tới ba đặc trưng là vị trí câu, độ trọng tâm (centroid) và độ tương đồng với câu đầu tiên. Điểm đặc trưng được tính bằng công thức sau:

$$SCORE(s) = w_c C_s + w_p P_s + w_f F_s \quad (3.14)$$

Trong đó:

w_c, w_p, w_f lần lượt là trọng số của từng đặc trưng về trọng tâm, vị trí, và tương đồng với câu đầu tiên. Luận văn này $w_c = w_p = w_f = 1$.

$C_s = \sum_i^n f(w_i) \times \frac{N_{w_i}}{N} \times ISF(w_i)$ (n là tổng số từ khác nhau trong câu s,

$f(w_i)$ số lần xuất hiện của từ w_i trong câu s, $\frac{N_{w_i}}{N}$ tỉ số giữa số lần xuất hiện của từ w_i trong toàn văn bản và số câu trong văn bản, $ISF(w_i)$ tần số nghịch của từ w_i.)

$P_s = \frac{N-i+1}{N}$ trong đó N là số câu trong văn bản, i là chỉ số vị trí của câu .

$F_s = sim(s, s_1)$ độ tương đồng của câu s với câu đầu tiên trong văn bản. sim được tính theo công thức 3.12.

3.2.4. Xếp hạng câu

Luận văn này tôi sử dụng thuật toán iSpreadRank để xếp hạng câu, chi tiết thuật toán đã được trình bày tại mục 3.1.

3.2.5. Trích chọn câu

Sau khi có bảng xếp hạng câu, chúng ta sẽ chọn lấy những câu theo thứ tự xếp hạng từ cao xuống thấp. Để tránh dư thừa thông tin, các câu được thêm vào văn bản tóm tắt nếu nó không quá giống với các câu được chọn trước đó, mức độ dư thừa được xác định bởi một ngưỡng, trong luận văn tôi đặt ngưỡng này bằng 0.4. Vì vậy chỉ những câu có điểm xếp hạng cao và thông tin ít dư thừa mới được chọn vào văn bản tóm tắt.

Đầu vào: Danh sách các câu đã được xếp hạng

$R = \left\{S_{S_1}, \dots, S_{S_m} \middle| i_j \in \{1, \dots, n\}\right\}$ trong đó $r(S_{S_1}) > \cdots > r(S_{S_n})$

$boolean\ visited[]$ // Mảng lưu vết các câu đã duyệt

Bước 1

$Sum = \{S_1\}$ // Câu có trọng số cao nhất được chọn đầu tiên

$visited[1] = true$

Bước 2

Duyệt tất cả các câu trong tập câu R ($j = i_2$ đến i_n) thoả mãn $visited[j] = false$

Nếu $sim(s_j, s_i) \geq \theta\ ;\ s_i \in Sum$

Bỏ qua s_j

Nếu không thì

Thêm s_j vào Sum

$visited[j] = true$

Bước 3
Lặp lại bước 2 cho đến khi kích thước của S đạt yêu cầu

Bước 4
Xắp sếp các câu trong tập *Sum* theo thứ tự xuất hiện trong văn bản

Đầu ra: Văn bản tóm tắt trong đó các câu được sắp xếp theo thứ tự xuất hiện trong văn bản gốc.

CHƯƠNG 4. ĐÁNH GIÁ KẾT QUẢ ĐẠT ĐƯỢC

Theo nghiên cứu [28], khi đánh giá trên bộ dữ liệu DUC 2004. Kết quả thực nghiệm cho thấy, khi xem xét số liệu ROUGE, hiệu suất xếp hạng của **iSpreadRank** tốt và **tốt** hơn PageRank.

iSpreadRank > PageRank > Normalized Similarity-based Degree > HITS

Bảng 3.So sánh hiệu suất tóm tắt của iSpreadRank với một số thuật toán khác

Models	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L
Rand. baseline	0.31549	0.04634	0.01031	0.00368	0.33878
Normalized Similarity-based Degree	0.34976	0.06391	0.01724	0.00559	0.35814
HITS	0.34485	0.06597	0.01880	0.00697	0.34972
PageRank	0.35366	**0.07499**	**0.02288**	**0.00978**	0.35977

iSpreadRank	**0.36047**	0.07385	0.02141	0.00778	**0.36444**

Để chứng minh sự hiệu quả của iSpreadRank cho văn bản tiếng Việt, tôi tiến hành cài đặt với 3 cách khác nhau:

- **SYS1:** Thuần PageRank.
- **SYS2:** Mô hình tóm tắt sử dụng thuật toán iSpreadRank với đầu vào là điểm PageRank.
- **SYS3**: Mô hình tóm tắt sử dụng thuật toán iSpreadRank với đầu vào là điểm đặc trưng (feature score).

4.1. Môi trường thực nghiệm

4.1.1. Môi trường phần cứng

Chương trình được thử nghiệm trên PC có cấu hình như sau:

- CPU: Intel Core i5 2.3Ghz
- Ram 4Gb
- Hệ điều hành: Windows 8

4.1.2. Môi trường phần mềm

Mô hình được cài đặt bằng ngôn ngữ lập trình Java với công cụ Eclipse.

Đánh giá kết quả thực nghiệm bằng tool ROUGE:
https://github.com/kavgan/ROUGE-2.0

4.2. Dữ liệu thực nghiệm

Dữ liệu thực nghiệm tác giả sử dụng trong luận văn được lấy từ tập dữ liệu trong đề tài "Nghiên cứu một số phương pháp tóm tắt văn bản tự động trên máy tính áp dụng cho Tiếng Việt", do PGS.TS. Lê Thanh Hương làm chủ nhiệm [4]. Tập dữ liệu bao gồm 200 văn bản gốc và 200 văn bản tóm tắt mẫu tương ứng, được chia thành 06 chủ đề.

Liên kết tải bộ dữ liệu: http://is.hust.edu.vn/~huonglt/donvanban.rar

Bảng 4.Danh sách chủ đề và số lượng văn bản tương ứng

STT	**Chủ đề**	**Số văn bản**
1	Chính trị	31
2	Khoa học công nghệ	25
3	Khoa học – giáo dục	22

4	Kinh tế	53
5	Văn hóa	34
6	Xã hội	35

Bảng 5.Danh sách các văn bản được sử dụng

Chính trị	**Khoa học - CN**	KHGD3	KT12	KT44	VH22	XH19
CT01	KHCN1	KHGD4	KT13	KT45	VH23	XH20
CT02	KHCN2	KHGD5	KT14	KT46	VH24	XH21
CT03	KHCN3	KHGD6	KT15	KT47	VH25	XH22
CT04	KHCN4	KHGD7	KT16	KT48	VH26	XH23
CT05	KHCN5	KHGD8	KT17	KT49	VH27	XH24
CT06	KHCN6	KHGD9	KT18	KT50	VH28	XH25
CT07	KHCN7	KHGD10	KT19	KT51	VH29	XH26
CT08	KHCN8	KHGD11	KT20	KT52	VH30	XH27
CT09	KHCN9	KHGD12	KT21	KT53	VH31	XH28
CT10	KHCN10	KHGD13	KT22	**Văn hóa**	VH32	XH29
CT11	KHCN11	KHGD14	KT23	VH01	VH33	XH30
CT12	KHCN12	KHGD15	KT24	VH02	VH34	XH31
CT13	KHCN13	KHGD16	KT25	VH03	**Xã hội**	XH32
CT14	KHCN14	KHGD17	KT26	VH04	XH01	XH33
CT15	KHCN15	KHGD18	KT27	VH05	XH02	XH34
CT16	KHCN16	KHGD19	KT28	VH06	XH03	XH35
CT17	KHCN17	KHGD20	KT29	VH07	XH04	
CT18	KHCN18	KHGD21	KT30	VH08	XH05	
CT19	KHCN19	KHGD22	KT31	VH09	XH06	

CT20	KHCN20	**Kinh tế**	KT32	VH10	XH07	
CT21	KHCN21	KT1	KT33	VH11	XH08	
CT22	KHCN22	KT2	KT34	VH12	XH09	
CT23	KHCN23	KT3	KT35	VH13	XH10	
CT24	KHCN24	KT4	KT36	VH14	XH11	
CT25	KHCN25	KT5	KT37	VH15	XH12	
CT26		KT6	KT38	VH16	XH13	
CT27		KT7	KT39	VH17	XH14	
CT28		KT8	KT40	VH18	XH15	
CT29	**Khoa học–GD**	KT9	KT41	VH19	XH16	
CT30	KHGD1	KT10	KT42	VH20	XH17	
CT31	KHGD2	KT11	KT43	VH21	XH18	

Độ dài văn bản tóm tắt được giới hạn là 3 câu. Độ dài này gần tương đương với độ dài văn bản mẫu do người tóm tắt. 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 (n=1, n=2, n=3, n=4).
- 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.

4.3. Tiến hành thực nghiệm

Trong nghiên cứu [4] của PGS.TS. Lê Thanh Hương, tác giả đã sử dụng thuật toán PageRank cải tiến để trích rút ra những câu quan trọng dựa trên đặc trưng TF.ISF, đặc biệt độ quan trọng của từ còn phụ thuộc vào việc từ đó có xuất hiện trong tiêu đề của văn bản không. Kết quả thực nghiệm của nghiên cứu là:

Bảng 6.Kết quả tóm tắt của nghiên cứu [4]

ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4
0.5939	0.389	0.337	0.311

Tiến hành thực nghiệm trên SYS1, SYS2, SYS3, kết quả thu được như sau:

Bảng 7.Kết quả tóm tắt của SYS1

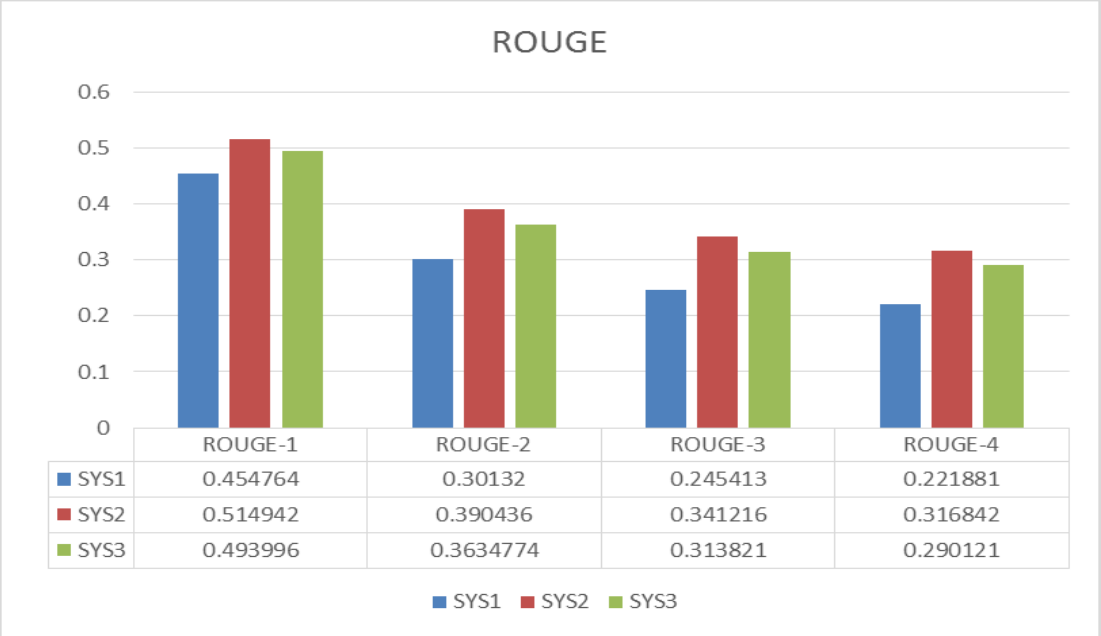
N-gram	**Re-call**	**Precision**	**F-score**
ROUGE-1	0.47501	0.463355	0.454764
ROUGE-2	0.316588	0.305373	0.30132
ROUGE-3	0.257127	0.249568	0.245413
ROUGE-4	0.222037	0.22632	0.221881

Bảng 8.Kết quả tóm tắt của SYS2.

N-gram	**Re-call**	**Precision**	**F-score**
ROUGE-1	0.523513	0.542311	0.514942
ROUGE-2	0.399859	0.408042	0.390436
ROUGE-3	0.349411	0.357209	0.341261
ROUGE-4	0.324173	0.332662	0.316842

Bảng 9.Kết quả tóm tắt của SYS3.

N-gram	**Re-call**	**Precision**	**F-score**
ROUGE-1	0.537141	0.489039	0.493996
ROUGE-2	0.399594	0.356491	0.363474
ROUGE-3	0.345448	0.307761	0.313821
ROUGE-4	0.319658	0.284788	0.290121



Hình 11.Biểu đồ so sánh độ chính xác sử dụng ROUGE tính trên F-score

Từ biểu đồ của hình 11 ta thấy khi xem xét độ chính xác dựa trên độ đo F-score:

- SYS2 và SYS3 tốt hơn SYS1, nghĩa là thuật toán iSpreadRank với tham số đầu vào là điểm PageRank hay điểm đặc trưng đều cho kết quả tốt hơn thuần PageRank.
- SYS2 cho kết quả tốt nhất. nghĩa là với tham số đầu vào là điểm PageRank cho kết quả tốt hơn tham số đầu vào là điểm đặc trưng.

So sánh kết quả của SYS2 với kết quả [4] của PGS.TS Lê Thanh Hương, cho thấy ROUGE -2,3,4 luận văn cho kết quả tốt hơn, ROUGE-1 luận văn cho kết quả thấp hơn.

Một số ví dụ về kết quả tóm tắt đầu ra của SYS2 được trình bày tại bảng 10.

Bảng 10. Một số ví dụ về kết quả tóm tắt của SYS2

Ví dụ 1	Hoàng Anh Gia Lai tái cấu trúc Công ty của bầu Đức sẽ bán các dự án thuộc lĩnh vực thủy điện và bất động sản nhằm dự trữ tiền mặt và giảm nợ vay. Công ty cổ phần Hoàng Anh Gia Lai (Mã CK: HAG) vừa công bố nghị quyết HĐQT. Theo đó, công ty thông qua chủ trương tái cấu trúc các đơn vị thuộc ngành thủy điện bằng cách bán các dự án đã đi vào hoạt động và đang trong giai đoạn đầu tư. Đồng thời, Hoàng Anh Gia Lai cũng thống nhất thông qua chủ trương tái cấu trúc các đơn vị thuộc ngành bất động sản bằng hình thức bán sỉ căn hộ và bán các dự án hoặc cổ phiếu của các công ty con đang sở hữu dự án. Mục đích bán, theo doanh nghiệp, nhằm thu tiền mặt về dự trữ và giảm nợ vay. Trước đó, trong báo cáo tài chính hợp nhất quý I, công ty do ông Đoàn Nguyên Đức làm Chủ tịch có 2.444 tỷ đồng tiền và các khoản tương đương tiền, nợ ngắn hạn là 3.446 tỷ đồng. Tại ĐHCĐ, bầu Đức cho biết, tỷ trọng doanh thu từ bất động sản trong năm 2013 sẽ giảm từ 64% xuống còn 14%.
Văn bản tóm tắt	*Công ty của bầu Đức sẽ bán các dự án thuộc lĩnh vực thủy điện và bất động sản nhằm dự trữ tiền mặt và giảm nợ vay.* *Theo đó, công ty thông qua chủ trương tái cấu trúc các đơn vị thuộc ngành thủy điện bằng cách bán các dự án đã đi vào hoạt động và đang trong giai đoạn đầu tư.* *Đồng thời, Hoàng Anh Gia Lai cũng thống nhất thông qua chủ trương tái cấu trúc các đơn vị thuộc ngành bất động sản bằng hình thức bán sỉ căn hộ và bán các dự án hoặc cổ phiếu của các công ty con đang sở hữu dự án.*
Ví dụ 2	Mỹ lại tăng thuế chống bán phá giá cá tra Việt Nam. Bộ Thương mại Mỹ (DOC) vừa có quyết định tăng thuế chống bán phá giá cá tra trong đợt xem xét hành chính lần 8 (POR 8) lên 1,29 USD/kg, từ mức 0,77 USD/kg đã đưa ra trong tháng 3. Thông tin trên được ông Trương Đình Hòe - Tổng thư ký Hiệp hội Chế biến và Xuất khẩu thủy sản Việt Nam (VASEP) cho biết. Theo vị này, DOC quyết định tăng thuế do cho rằng việc tính toán trước

	đó có sai sót, nên đã tiến hành điều chỉnh trong lần áp thuế tiếp theo. Trong thông cáo báo chí của DOC cũng chỉ rõ, cơ quan này bị nhầm toàn bộ tỷ lệ sử dụng cá của Việt An, việc tiêu thụ dầu diesel của Docifish và đưa doanh số hàng bán bị trả lại vào quá trình tính toán biên độ phá giá của các công ty Việt An và Vĩnh Hoàn", thông báo của DOC cho hay. Sau khi điều chỉnh, mức thuế chống bán phá giá cá tra với 12 doanh nghiệp bị tăng lên 1,29 USD/kg, tương đương tăng 67% so với mức thuế mới công bố cách đây hai tháng. Riêng Vĩnh Hoàn được giữ nguyên mức thuế 0,19 USD/kg, Việt An bị áp mức cao nhất là 2,39 USD/kg, so với 1,34 USD/kg trước đó. Trước câu hỏi doanh nghiệp xuất khẩu cá tra sẽ bị ảnh hưởng ra sao với quyết định tăng thuế của DOC, ông Hòe chia sẻ: "Với mức cũ doanh nghiệp đã không có đường xuất khẩu sang Mỹ, huống chi là với mức tăng thêm lên 1,29 USD một kg". Trong phán quyết hồi tháng 3, DOC đã bất ngờ đổi quốc gia thay thế từ Bangladesh sang Indonesia - nơi có yếu tố đầu vào nuôi cá tra cao hơn Việt Nam, khiến mức thuế chống bán phá giá tăng lên hàng chục lần. Bởi theo POR7, mức thuế trung bình áp dụng cho Việt An chỉ là 0,02 USD/kg, Vĩnh Hoàn và các công ty khác là 0 cent. Để phản đối quyết định trên, VASEP và hầu hết các doanh nghiệp bị áp thuế cùng gửi đơn kiện DOC lên Tòa án Thương mại Quốc tế (CIT). Được biết, CIT đã chấp nhận đơn khởi kiện và yêu cầu Hải quan Mỹ tạm dừng không thu thuế chống bán phá giá của các doanh nghiệp theo kết luận cuối cùng của POR8 cho tới khi có phán quyết cuối cùng của Tòa án này.
Văn bản tóm tắt	*Sau khi điều chỉnh, mức thuế chống bán phá giá cá tra với 12 doanh nghiệp bị tăng lên 1,29 USD/kg, tương đương tăng 67% so với mức thuế mới công bố cách đây hai tháng.* *Trước câu hỏi doanh nghiệp xuất khẩu cá tra sẽ bị ảnh hưởng ra sao với quyết định tăng thuế của DOC, ông Hòe chia sẻ: "Với mức cũ doanh nghiệp đã không có đường xuất khẩu sang Mỹ, huống chi là với mức tăng thêm lên 1,29 USD một kg".* *Được biết, CIT đã chấp nhận đơn khởi kiện và yêu cầu Hải quan Mỹ tạm dừng không thu thuế chống bán phá giá của các doanh nghiệp theo kết luận cuối cùng của POR8 cho tới khi có phán quyết cuối cùng của Tòa án này.*

Ví dụ 3	Hà Nội tháo dỡ hai cầu bộ hành để xây cầu vượt Mới được đưa vào sử dụng chưa lâu, hai cây cầu vượt dành cho người đi bộ trên đường Nguyễn Chí Thanh và Trần Khát Chân đã bị tháo dỡ để dành không gian cho cầu vượt dành cho xe cơ giới. Để giải quyết tình trạng ùn tắc giao thông vào giờ cao điểm tại nút giao Đại Cồ Việt - Trần Khát Chân, đầu tháng 2/2013, Hà Nội đã khởi công cây cầu vượt dài hơn 350 m, rộng 11 m. Cùng với đó, cây cầu dành cho người đi bộ trên đường gần Trần Khát Chân mới được đưa vào sử dụng đã phải tháo dỡ. Phần thân cầu được dùng lại, dự kiến sẽ lắp trên đường Giải Phóng. Một cây cầu vượt dài 276m, rộng 17m, dành cho 4 làn xe cơ giới cũng mới được khởi công tại nút giao Nguyễn Chí Thanh - Liễu Giai. Cây cầu vượt dành cho người đi bộ trên đường Nguyễn Chí Thanh (nằm ngay đầu cầu vượt cho xe cơ giới) cũng sẽ phải tháo dỡ, lắp đặt lại cách vị trí cũ 100m. Đại diện Sở Giao thông vận tải Hà Nội cho biết, việc tháo dỡ cầu dành cho người đi bộ để xây dựng cầu vượt đã được tính toán kỹ. "Cầu dành cho người đi bộ có thể tháo dỡ lắp đặt sang vị trí khác. Do vậy, việc tháo dỡ cầu bộ hành để xây dựng cầu vượt dành cho xe cơ giới đem lại hiệu quả cao hơn", đại diện Sở Giao thông vận tải nói.
Văn bản tóm tắt	*Mới được đưa vào sử dụng chưa lâu, hai cây cầu vượt dành cho người đi bộ trên đường Nguyễn Chí Thanh và Trần Khát Chân đã bị tháo dỡ để dành không gian cho cầu vượt dành cho xe cơ giới.* *Cùng với đó, cây cầu dành cho người đi bộ trên đường gần Trần Khát Chân mới được đưa vào sử dụng đã phải tháo dỡ.* *Đại diện Sở Giao thông vận tải Hà Nội cho biết, việc tháo dỡ cầu dành cho người đi bộ để xây dựng cầu vượt đã được tính toán kỹ.*

Để có thêm kết luận về hiệu quả tóm tắt của mô hình này với các chủ đề khác nhau, trên SYS2 tôi tiến hành đánh độ chính xác trên 6 chủ đề của tập dữ liệu đầu vào. Kết quả đạt được như số liệu trong bảng 11.

Bảng 11.Kết quả tóm tắt trên từng chủ đề

CHÍNH TRỊ (CT)			
N-gram	Re-call	Precision	F-score
ROUGE-1	0.638618	0.506263	0.552175
ROUGE-2	0.520317	0.415658	0.451557
ROUGE-3	0.465483	0.375312	0.405767
ROUGE-4	0.437183	0.354376	0.381907
KHOA HỌC CÔNG NGHỆ (KHCN)			
N-gram	Re-call	Precision	F-score
ROUGE-1	0.568510	0.528414	0.533281
ROUGE-2	0.449245	0.401547	0.413094
ROUGE-3	0.386475	0.344224	0.354706
ROUGE-4	0.353400	0.313636	0.323624
KHOA HỌC GIÁO DỤC (KHDG)			
N-gram	Re-call	Precision	F-score
ROUGE-1	0.433021	0.533314	0.463474
ROUGE-2	0.304992	0.362280	0.319804
ROUGE-3	0.256379	0.304124	0.267644
ROUGE-4	0.234632	0.278876	0.244296
KINH TẾ (KT)			
N-gram	Re-call	Precision	F-score
ROUGE-1	0.493498	0.579375	0.519330
ROUGE-2	0.379894	0.442192	0.397567
ROUGE-3	0.334185	0.390893	0.350196
ROUGE-4	0.308180	0.362371	0.323328
VĂN HOÁ (VH)			
N-gram	Re-call	Precision	F-score
ROUGE-1	0.444837	0.472404	0.434046
ROUGE-2	0.289870	0.306672	0.282316
ROUGE-3	0.244330	0.256426	0.236787

ROUGE-4	0.222449	0.233966	0.215395
XÃ HỘI (XH)			
N-gram	Re-call	Precision	F-score
ROUGE-1	0.559191	0.593613	0.563908
ROUGE-2	0.492032	0.469877	0.446388
ROUGE-3	0.391730	0.419197	0.396817
ROUGE-4	0.370897	0.40048	0.377103

Dựa trên số liệu của bảng 11, tôi thấy rằng kết quả tóm tắt có sự khác nhau giữa các chủ đề, cụ thể với chủ đề Chính trị (CT) mô hình cho kết quả tốt nhất, chủ đề Văn hoá (VH) mô hình cho kết quả thấp nhất. Qua đó thấy rằng, đặc trưng về chủ đề cũng là một đặc trưng quan trọng ảnh hưởng tới độ chính xác của bài toán tóm tắt văn bản.

Trên SYS2, tôi tiến hành chọn ra 54 văn bản, đây là những văn bản tóm tắt cho điểm F-score nhỏ hơn 0.45 trên ROUGE-1. Kết quả thu được như bảng 12.

Bảng 12.Danh sách văn bản có kết quả tóm tắt thấp

STT	**Tên file**	**F-Score**		**STT**	**Tên file**	**F-score**
1	CT09.TXT	0.4357		28	KT48.TXT	0.36545
2	CT10.TXT	0.44186		29	KT49.TXT	0.42466
3	CT17.TXT	0.0125		30	KT50.TXT	0.42997
4	CT29.TXT	0.43066		31	KT6.TXT	0.34746
5	KHCN14.TXT	0.35897		32	VH01.TXT	0.38647
6	KHCN17.TXT	0.43515		33	VH05.TXT	0.35172
7	KHCN19.TXT	0.43333		34	VH06.TXT	0.39496
8	KHCN7.TXT	0.32957		35	VH07.TXT	0.2844
9	KHGD1.TXT	0.25279		36	VH08.TXT	0.38
10	KHGD15.TXT	0.27397		37	VH16.TXT	0.3609
11	KHGD19.TXT	0.31111		38	VH17.TXT	0.23602
12	KHGD20.TXT	0.44882		39	VH21.TXT	0.42623
13	KHGD22.TXT	0.35176		40	VH22.TXT	0.24299
14	KHGD4.TXT	0.37019		41	VH23.TXT	0.18039
15	KT10.TXT	0.44706		42	VH24.TXT	0.38806
16	KT14.TXT	0.35088		43	VH26.TXT	0.43294
17	KT15.TXT	0.12709		44	VH28.TXT	0.37073
18	KT19.TXT	0.37433		45	VH32.TXT	0.36364
19	KT2.TXT	0.41791		46	VH33.TXT	0.4
20	KT20.TXT	0.40876		47	XH06.TXT	0.43902

21	KT30.TXT	0.35754		48	XH07.TXT	0.39739
22	KT33.TXT	0.39844		49	XH08.TXT	0.42985
23	KT35.TXT	0.37722		50	XH12.TXT	0.41475
24	KT39.TXT	0.40876		51	XH23.TXT	0.33452
25	KT44.TXT	0.42997		52	XH25.TXT	0.31746
26	KT45.TXT	0.36545		53	XH29.TXT	0.40667
27	KT46.TXT	0.42466		54	XH30.TXT	0.35918

Qua phân tích và kiểm tra lại nội dung văn bản tóm tắt mẫu và văn bản tóm tắt sinh ra từ hệ thống của các văn bản trong bảng 12, ngoài việc độ chính xác có sự phân bố khác nhau giữa các chủ đề, tôi thấy rằng một số văn bản trên có kết quả tóm tắt thấp còn do những nguyên nhân sau:

- Lỗi không đồng bộ về định dạng encoding (mã hoá) giữa văn bản tóm tắt hệ thống và văn bản tóm tắt mẫu. Cụ thể văn bản "CT17.TXT", trong văn bản mẫu encoding là "Encode ucs-2 le bom", trong khi đó văn bản tóm tắt hệ thống là "Encode UTF-8". Việc này dẫn đến tool rouge đọc đầu vào sai với văn bản này, và cho ra kết quả rất thấp F-score là 0.0125.
- Lỗi văn bản bản tóm tắt mẫu không khớp với văn bản gốc. Cụ thể văn bản "KT15.TXT", nội dung văn bản gốc liên quan đến kinh tế, nhưng tại văn bản mẫu nội dung lại nói về giáo dục. Vì vậy nội dung văn bản tóm tắt hệ thống sinh ra sẽ khác nội dung với văn bản mẫu. Kết quả đánh giá tại văn bản này F-score là 0.12709.
- Độ dài chênh lệch giữa văn bản tóm tắt mẫu và văn bản tóm tắt hệ thống. Cụ thể văn bản "VH32.TXT" số câu trong văn bản tóm tắt mẫu là 5, trong khi số câu trong văn bản tóm tắt hệ thống là 3. Kết quả đánh giá F-score là 0.36364.

Kết quả phân tích này bổ sung thêm những điểm cần chú ý, để tôi phát triển và cải tiến mô hình tóm tắt văn bản tiếng Việt sau này.

KẾT LUẬN

Những vấn đề đã giải quyết được trong luận văn

- Luận văn đã trình bày tổng quan về cơ sở lý thuyết về tóm tắt văn bản bao gồm khái niệm, phân loại, các hướng tiếp cận, các phương pháp đánh giá tóm tắt văn bản.
- Luận văn đã trình bày chi tiết thuật toán iSpreadRank bao gồm dữ liệu đầu vào, đầu ra, các bước thực hiện thuật toán.
- Luận văn đã xây dựng hoàn chỉnh và cài đặt thành công mô hình tóm tắt văn bản Tiếng Việt tự động áp dụng thuật toán iSpreadRank. Mô hình có những ưu điểm nổi bật như sau:
  - Không cần dữ liệu training, thích hợp với những ngôn ngữ ít tài nguyên (bộ dữ liệu chuẩn) như tiếng Việt.
  - Thuật toán rõ ràng, dễ tích hợp thêm tri thức, có thể tính trọng số đầu vào của các câu bằng nhiều phương pháp khác nhau. Hiện tại trong luận văn, tôi trình bày hai phương pháp là PageRank, và điểm đặc trưng với 3 đặc trưng, tuy nhiên chúng ta có thể thử nghiệm với 4 hoặc 5 đặc trưng, hoặc nhiều phương pháp khác.
  - Có thể tóm tắt các văn bản lớn. Đây cũng là một ưu điểm so với tóm tắt tóm lược, bởi như đã biết mô hình tóm tắt tóm lược như mô hình Sequence-to-Sequence gặp nhiều khó khăn trong việc tóm tắt văn bản lớn.
  - Dễ cài đặt. Khi xây dựng hệ thống tóm tắt văn bản dựa theo mô hình này, lập trình viên không cần nhiều những kiến thức chuyên sâu về ngôn ngữ học cũng như xử lý ngôn ngữ tự nhiên vẫn có thể xây dựng được ứng dụng tóm tắt văn bản.
- Kết quả bước đầu cho thấy mô hình cho kết quả tốt.

Công việc tương lai cần làm

- Nghiên cứu, áp dụng các phương pháp mới giúp nâng cao chất lượng văn bản tóm tắt bằng việc rút gọn các câu trong văn bản tóm tắt. Trên cơ sở các kiến thức về tóm tắt văn bản đã tìm hiểu, nghiên cứu và xây dựng hệ thống tóm tắt văn bản theo kiểu tóm lược.
- Một trong những hạn chế của mô hình hiện tại là việc coi các câu là độc lập với nhau, vì vậy đặc trưng chủ đề trong văn bản bị coi nhẹ, trong tương lai, khi xây dựng mô hình, tôi sẽ nghiên cứu, áp dụng thêm một số thuật toán như Naïve-Bayes, để giải quyết vấn đề này.

- Thu thập dữ liệu mẫu để phục vụ cho việc đánh giá được chính xác và khách quan hơn.
- Tích hợp mô hình vào xây dựng ứng dụng tóm tắt tin tức cho điện thoại di động.

TÀI LIỆU THAM KHẢO

Tiếng Việt

[1] Nguyễn Nhật An (2015), "*Nghiên cứu phát triển các kỹ thuật tự động tóm tắt văn bản tiếng Việt*", Luận án tiến sĩ, Viện Khoa học và Công nghệ quân sự.

[2] Đoàn Xuân Dũng (2018), *"Tóm tắt văn bản sử dụng các kỹ thuật trong deep learning"*, Luận văn thạc sĩ, Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội.

[3] Trương Quốc Định, Nguyễn Quang Dũng (2012), "Một giải pháp tóm tắt văn bản tiếng Việt tự động", *Hội thảo quốc gia lần thứ XV: một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông Hà Nội*, 03-04/12/2012.

[4] Lê Thanh Hương (2014). "Nghiên cứu một số phương pháp tóm tắt văn bản tự động trên máy tính áp dụng cho Tiếng Việt", *Báo cáo tổng kết đề tài B2012 - 01 – 24*, Trường Đại học Bách Khoa Hà Nội.

[5] Nguyễn Thị Thu Hà (2012), "*Phát triển một số thuật toán tóm tắt văn bản tiếng Việt sử dụng phương pháp học bán giám sát*"**,** Luận án tiến sĩ, Học viện kỹ thuật quân sự.

[6] Đỗ Phúc, Mai Xuân Hùng, Nguyễn Thị Kim Phụng (2008) "Gom cụm đồ thị và ứng dụng vào việc rút trích nội dung chính của khối thông điệp trên diễn đàn thảo luận", *Tạp chí Phát triển Khoa học Công nghệ*, Tập 11, Số 05 - 2008, tr. 21-32.

[7] Nguyễn Trọng Phúc, Lê Thanh Hương (2008), "Tóm tắt văn bản sử dụng cấu trúc diễn ngôn", *Proc of ICTrda08*.

[8] Trịnh Văn Quỳnh, Hoàng Thị Khánh, Đỗ Thị Lan Hương, Nguyễn Thị Hà (2017). "*Chiến thuật ôn tập Ngữ Văn lớp 9 luyện thi vào 10 Bằng sơ đồ tư duy*", Nhà xuất bản Đại học Quốc gia Hà Nội.

[9] Nguyễn Thị Ngọc Tú, Nguyễn Thị Thu Hà , Lê Thanh Hương , Hồ Ngọc Vinh, Đào Thanh Tĩnh, Nguyễn Ngọc Cương (2015), "ứng dụng đồ thị trong tóm tắt đa văn bản tiếng Việt". *Kỷ yếu Hội nghị Quốc gia lần thứ VIII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR).*

[10] Lâm Quang Tường , Phạm Thế Phi, và Đỗ Đức Hào (2017), "Tóm tắt văn bản tiếng Việt tự động với mô hình SEQUENCE-TO-SEQUENCE". *Tạp chí Khoa học Trường Đại học Cần Thơ*, Số chuyên đề: Công nghệ thông tin (2017), tr.125-132.

Tiếng Anh

[11] Mehdi Allahyari , Seyedamin Pouriyeh, Mehdi Assef, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez and Krys Kochut (2017), "Text Summarization Techniques: A Brief Survey", *arXiv*, July 2017, USA.

[12] John M Conroy and Dianne P. O'leary (2001), "Text summarization via hidden Markov models", *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*.

[13] Hal Daum III and Daniel Marcu (2006), "Bayesian Query Focused Summarization", *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pp.305–312, Sydney.

[14] Mangesh Dahale (2014), *"Text Summarization for Compressed Inverted Indexes and Snippets"*, Master's Theses and Graduate Research, San Jose State University.

[15] Rafael Ferreira, Frederico Freitas, Luciano de Souza Cabral, Rafael Dueire Lins, Rinaldo Lima. (2013), "A Four Dimension Graph Model for Automatic Text Summarization", *IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT*).

[16] Vishal Gupta (2010), "A Survey of Text Summarization Extractive Techniques". *JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE*, VOL. 2, NO. 3.

[17] Xu Han, Tao Lv, Zhirui Hu, Xinyan Wang, and Cong Wang (2016), "Text Summarization Using FrameNet-Based Semantic Graph Model". *Scientific Programming Volume 2016, Article ID 5130603*.

[18] Simon Kemp (2019), "Digital 2019: Global internet use accelerates", *Wearesocial.com, Global Digital 2019 reports*, 30 January 2019.

[19] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean (2013), "Distributed Representations of Words and Phrasesand their Compositionality", *arXiv:1310.4546v1*.

[20] Hong Phuong Le, Thi Minh Huyen Nguyen, Roussanaly Azim, Vinh H.T (2008), "A Hybrid Approach to Word Segmentation of Vietnamese Texts", *In: Martín-Vide C., Otto F., Fernau H. (eds) Language and Automata Theory and Applications. LATA 2008. Lecture Notes in Computer Science*, vol 5196, Springer, Berlin, Heidelberg.

[21] Quoc V. Le, Tomas Mikolov (2014), "Distributed Representations of Sentences and Documents", *arXiv:1405.4053v2*.

[22] Lin, Chin-Yew (2004), "ROUGE: a Package for Automatic Evaluation of Summaries", *In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, Barcelona, Spain, July 25 - 26, 2004.

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002) "BLEU: a Method for Automatic Evaluation of Machine Translation", *Computational Linguistics (ACL)*, Philadelphia, July 2002, pp. 311-318.

[24] Abigail See, Peter J. Liu, Christopher D. Manning (2017). "Get To The Point: Summarization with Pointer-Generator Networks", *arXiv:1704.04368*.

[25] Xinghao Song, Chunming Yang, Hui Zhang and Xujian Zhao (2018), "The Algorithm of Automatic Text Summarization Based on Network Representation Learning", *Springer Nature Switzerland AG 2018 M. Zhang et al. (Eds.): NLPCC 2018, LNAI 11109*, pp.362–371.

[26] Dingding Wang, Shenghuo Zhu, Tao Li, and Yihong Gong (2009), "Multidocument summarization using sentence-based topic models", *In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. Association for Computationa*.

[27] Kang Yang , Kamal Al-Sabahi , Yanmin Xiang and Zuping Zhang (2018), "An Integrated Graph Model for Document Summarization". *Information 2018, 9(9), 232; https://doi.org/10.3390/info9090232*.

[28] Jen-Yuan Yeh, Wei-Pang Yang, Hao-Ren Ke, Pei-Cheng Cheng.

(2014), "Extraction-based News Summarization Using Sentence Centrality in the Sentence Similarity Network", *Journal of Information Management,* Vol. 21, No. 3, pp. 271-304**.**

[29] Jen-Yuan Yeh, Hao-Ren Ke, Wei-Pang Yang (2008), *"*iSpreadRank: Ranking sentences for extraction-based summarization using feature weight propagation in the sentence similarity network", *Expert Systems with Applications 35 (2008),* pp.1451–1462.