

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CẦN THƠ
KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**



**NIÊN LUẬN NGÀNH
NGÀNH KHOA HỌC MÁY TÍNH**

Đề tài

**TÌM HIỂU XỬ LÝ NGÔN NGỮ TỰ NHIÊN VÀ
ỨNG DỤNG TÓM TẮT NỘI DUNG VĂN BẢN**

Sinh viên thực hiện: Nguyễn Hữu Tính

Mã số: B1710355

Khóa: 43

Cần Thơ, 6/2021

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CẦN THƠ
KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**



**NIÊN LUẬN NGÀNH
NGÀNH KHOA HỌC MÁY TÍNH**

Đề tài

**TÌM HIỂU XỬ LÝ NGÔN NGỮ TỰ NHIÊN VÀ
ỨNG DỤNG TÓM TẮT NỘI DUNG VĂN BẢN**

**Giáo viên hướng dẫn:
TS. Trần Nguyễn Dương Chi**

**Sinh viên thực hiện:
Nguyễn Hữu Tính
Mã số: B1710355
Khóa: 43**

Cần Thơ, 6/2021

[illegible]

i

LỜI CẢM ƠN

Để có được bài niên luận này, em xin được bày tỏ lòng biết ơn chân thành và sâu sắc đến Cô Trần Nguyễn Dương Chi – người đã trực tiếp tận tình hướng dẫn, giúp đỡ em. Trong suốt quá trình thực hiện niên luận, nhờ những sự chỉ bảo và hướng dẫn quý giá đó mà bài niên luận này được hoàn thành một cách tốt nhất.

Em cũng xin gửi lời cảm ơn chân thành đến các Thầy Cô Giảng viên Đại học Cần Thơ, đặc biệt là các Thầy Cô ở Khoa CNTT & TT, những người đã truyền đạt những kiến thức quý báu trong thời gian qua.

Em cũng xin chân thành cảm ơn bạn bè cùng với gia đình đã luôn động viên, khích lệ và tạo điều kiện giúp đỡ trong suốt quá trình thực hiện để em có thể hoàn thành bài niên luận một cách tốt nhất.

Tuy có nhiều cố gắng trong quá trình thực hiện niên luận, nhưng không thể tránh khỏi những sai sót. Em rất mong nhận được sự đóng góp ý kiến quý báu của quý Thầy Cô và các bạn để bài niên luận hoàn thiện hơn.

Cần Thơ, ngày tháng 6 năm 2021

Người viết

Nguyễn Hữu Tính

MỤC LỤC

PHẦN GIỚI THIỆU	1
1. Đặt vấn đề	1
2. Mục tiêu đề tài	1
3. Bố cục luận văn.....	1
PHẦN NỘI DUNG.....	2
CHƯƠNG 1	2
KHÁI QUÁT VỀ XỬ LÝ NGÔN NGỮ TỰ NHIÊN VÀ	2
TÓM TẮT VĂN BẢN.....	2
I. Xử lý ngôn ngữ tự nhiên:	2
1. Tổng quan:	2
II. Tóm tắt văn bản:	5
CHƯƠNG 2	6
MỘT SỐ NGHIÊN CỨU VỀ TÓM TẮT VĂN BẢN	6
I. Tóm tắt văn bản theo hướng trích chọn:	6
1. Phương pháp chủ đề đại diện dựa trên tần xuất:	7
2. Phương pháp đặc trưng đại diện:	9
II. Tóm tắt văn bản theo hướng tóm lược:	9
CHƯƠNG 3	11
I. Giới thiệu:	11
1. Tóm tắt văn bản với Spacy:	11
2. Thư viện NLTK:	12
II. Hệ thống tóm tắt văn bản:	13
1. Quy trình tóm tắt theo hướng trích suất sử dụng Spacy:	13
2. Quy trình tóm tắt theo hướng trích suất sử dụng Gensim:	15
3. Quy trình tóm tắt theo hướng trích suất sử dụng Suny:	15
4. Xây dựng bộ dữ liệu tóm tắt văn bản tiếng việt:	17
5. Tiền xử lý dữ liệu:	18
6. Mã hóa bài viết bằng mô hình ngôn ngữ của spaCy:	19
7. Trích xuất các từ khóa quan trọng và tính toán trọng lượng chuẩn hóa:	19
8. Tính mức độ quan trọng của từng câu trong bài viết dựa trên sự xuất hiện của từ khóa: 21	
9. Lọc các câu dựa trên mức độ quan trọng được tính toán:	22
10. Tạo bản tóm tắt:	22

CHƯƠNG 4:	23
ĐÁNH GIÁ TÓM TẮT VĂN BẢN	23
I. Môi trường thử nghiệm:	23
II. Đánh giá tóm tắt văn bản:	24
1. Các phương pháp đánh giá tóm tắt văn bản:	24
2. Đánh giá tóm tắt văn bản:	27
PHẦN KẾT LUẬN	35
1. Kết quả đạt được:	35
Xây dựng được phần mềm tóm tắt văn bản với chức năng cơ bản như:	35
2. Hạn chế:	35
3. Hướng phát triển:	35
CHƯƠNG 5:	36
DEMO CHƯƠNG TRÌNH	36
TÀI LIỆU THAM KHẢO	39

DANH MỤC HÌNH

Hình 1. Mô hình sequence-to-sequence với cơ chế attention	10
Hình 2. Mô hình bài toán tóm tắt văn bản	13
Hình 3. Mô hình ngôn ngữ Spacy	14
Hình 4. Quy trình thực hiện tóm tắt văn bản tiếng việt với Spacy	16
Hình 5. Dữ liệu cào từ trang web	17
Hình 6. Giao diện chính của chương trình	36
Hình 7. Giao diện chức năng thông qua url	37
Hình 8. Giao diện so sánh các phương pháp	38

ABSTRACT

TÓM TẮT

Bài báo trình bày cách thức rút trích các câu có nội dung quan trọng trong các văn bản khoa học tiếng Việt dựa trên cấu trúc. Hệ thống rút trích được xây dựng dựa trên một quy trình chặt chẽ mà bài báo đề xuất với việc áp dụng nhiều phương pháp khác nhau trong việc tính toán độ quan trọng thông tin của câu. Kết quả thử nghiệm cho thấy kết hợp phương pháp độ đo cục bộ và toàn cục (TF.IDF) với cách đánh giá câu theo cách cộng dồn trọng số từ cho kết quả tốt nhất. Bước đầu thử nghiệm trên các bài báo khoa học và toàn văn báo cáo thuộc lĩnh vực Công nghệ thông tin đã cho những kết quả có độ chính xác cao so với yêu cầu.

Đối với những người làm nghiên cứu thì việc tìm kiếm tài liệu để tham khảo là một vấn đề vô cùng quan trọng, trong khi đó không phải chỉ đọc lướt qua là người ta có thể nắm hết các ý mà tác giả muốn nêu trong tài liệu. Có khi mất khá nhiều thời gian để đọc hết một tài liệu rồi nhận ra tài liệu đó không phù hợp với mục tiêu tìm kiếm của mình. Khác với việc chúng ta đọc rồi tự rút ra cho mình những ý chính trong toàn bộ văn bản như lâu nay mọi người thường làm, điều đó không tránh khỏi sự chủ quan trong chọn lựa ý chính vì mỗi người có những trình độ khác nhau, có chuyên môn khác nhau. Trong khi đặc điểm của văn bản khoa học là trong mỗi văn bản, tác giả – nhà khoa học – luôn mong muốn trình bày, thậm chí là khẳng định một ý tưởng khoa học cụ thể.

Với mục đích giúp con người tiết kiệm thời gian hơn trong việc tìm kiếm, sàng lọc và tổng hợp các thông tin một cách khách quan trong kho tri thức khổng lồ của nhân loại – Internet, bài báo muốn đề cập đến một quy trình cho phép máy tính có thể tự động rút trích ý chính từ văn bản tương đối chính xác nhất mà cụ thể là các văn bản khoa học trong ngành công nghệ thông tin như bài báo khoa học và toàn văn báo cáo. Bên cạnh đó bài báo trình bày nhiều phương pháp thực hiện khác nhau trong việc tính độ quan trọng thông tin của câu để đưa ra nhận xét đánh giá phương pháp nào là tối ưu, từ đó đưa vào quy trình thực hiện việc rút trích.

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

PHẦN GIỚI THIỆU

1. Đặt vấn đề

Với sự phát triển mạnh mẽ của công nghệ thông tin và mạng máy tính, lượng tài liệu văn bản khổng lồ được tạo ra với nhiều mục đích sử dụng khác nhau khiến cho việc đọc hiểu và trích lược các thông tin cần thiết trong khối tri thức đồ sộ này tốn rất nhiều thời gian và chi phí (đặc biệt là chi phí cho hạ tầng và truyền dẫn thông tin đáp ứng yêu cầu cho một số lượng ngày càng nhiều các thiết bị cầm tay). Để tăng hiệu quả cũng như dễ dàng hơn trong việc tiếp nhận thông tin của người dùng, nhiều nghiên cứu về khai phá dữ liệu và xử lý ngôn ngữ tự nhiên đã được thực hiện. Một trong những nghiên cứu quan trọng đóng vai trò then chốt đó tóm tắt văn bản tự động.

Bài toán tóm tắt văn bản tiếng Việt cũng được nghiên cứu và áp dụng nhiều kỹ thuật như đối với tiếng Anh; tuy nhiên, tóm tắt văn bản nói riêng và xử lý ngôn ngữ tự nhiên nói chung áp dụng cho tiếng Việt gặp nhiều thách thức hơn. Sở dĩ là vì tiếng Việt với đặc trưng là tiếng đơn âm và có thanh điệu nên việc tách từ, tách các thành phần ngữ nghĩa trong câu tiếng Việt đòi hỏi xử lý phức tạp hơn so với xử lý câu tiếng Anh, thêm vào đó, không có nhiều kho dữ liệu tiếng Việt được chuẩn hóa và công bố.

2. Mục tiêu đề tài

Tìm hiểu ngôn ngữ tự nhiên và tạo ứng dụng tóm tắt văn bản theo hướng trích suất

3. Bố cục luận văn

Phần giới thiệu

Giới thiệu tổng quát về đề tài.

Phần nội dung

Chương 1: Khái quát về xử lý ngôn ngữ tự nhiên và bài toán tóm tắt văn bản.

Chương 2: Một số nghiên cứu về tóm tắt văn bản.

Chương 3: Xây dựng hệ thống tóm tắt văn bản.

Chương 4: Thử nghiệm và đánh giá.

Chương 5: Demo chương trình.

Phần kết luận

Trình bày kết quả đạt được và hướng phát triển hệ thống.

PHẦN NỘI DUNG

CHƯƠNG 1

KHÁI QUÁT VỀ XỬ LÝ NGÔN NGỮ TỰ NHIÊN VÀ TÓM TẮT VĂN BẢN

I. Xử lý ngôn ngữ tự nhiên:

1. Tổng quan:

Xử lý ngôn ngữ tự nhiên (natural language processing - NLP) là một nhánh của trí tuệ nhân tạo tập trung vào các ứng dụng trên ngôn ngữ của con người. Trong trí tuệ nhân tạo thì xử lý ngôn ngữ tự nhiên là một trong những phần khó nhất vì nó liên quan đến việc phải hiểu ý nghĩa ngôn ngữ-công cụ hoàn hảo nhất của tư duy và giao tiếp.

Xử lý ngôn ngữ chính là xử lý thông tin khi đầu vào là “dữ liệu ngôn ngữ” (dữ liệu cần biến đổi), tức dữ liệu “văn bản” hay “tiếng nói”. Các dữ liệu liên quan đến ngôn ngữ viết (văn bản) và nói (tiếng nói) đang dần trở nên kiểu dữ liệu chính con người có và lưu trữ dưới dạng điện tử. Đặc điểm chính của các kiểu dữ liệu này là không có cấu trúc hoặc nửa cấu trúc và chúng không thể lưu trữ trong các khuôn dạng cố định như các bảng biểu. Theo đánh giá của công ty Oracle, hiện có đến 80% dữ liệu không cấu trúc trong lượng dữ liệu của loài người đang có [Oracle Text]. Với sự ra đời và phổ biến của Internet, của sách báo điện tử, của máy tính cá nhân, của viễn thông, của thiết bị âm thanh, ... người người ai cũng có thể tạo ra dữ liệu văn bản hay tiếng nói. Vấn đề là làm sao ta có thể xử lý chúng, tức chuyển chúng từ các dạng ta chưa hiểu được thành các dạng ta có thể hiểu và giải thích được, tức là ta có thể tìm ra thông tin, tri thức hữu ích cho mình.

Tuy nhiên, một văn bản thật sự (một bài báo khoa học chẳng hạn) có thể có đến hàng nghìn câu, và ta không phải có một mà hàng triệu văn bản. Web là một nguồn dữ liệu văn bản khổng lồ, và cùng với các thư viện điện tử – khi trong một tương lai gần các sách báo xưa nay và các nguồn âm thanh được chuyển hết vào máy tính (chẳng hạn bằng các chương trình nhận dạng chữ, thu nhập âm thanh, hoặc gõ thẳng vào máy) – sẽ sớm chứa hầu như toàn bộ kiến thức của nhân loại. Vấn đề là làm sao “xử lý” (chuyển đổi) được khối dữ liệu văn bản và tiếng nói khổng lồ này qua dạng khác để mỗi người có được thông tin và tri thức cần thiết từ chúng. Xử lý ngôn ngữ tự nhiên đã được ứng dụng trong thực tế để giải quyết các bài toán như: nhận dạng chữ viết, nhận dạng tiếng nói, tổng hợp tiếng nói, dịch tự động, tìm kiếm thông tin, tóm tắt văn bản, khai phá dữ liệu và phát hiện tri thức.

1.1. Các bước xử lý ngôn ngữ tự nhiên:

Phân tích hình thái - Trong bước này từng từ sẽ được phân tích và các ký tự không phải chữ (như các dấu câu) sẽ được tách ra khỏi các từ. Trong tiếng Anh và nhiều ngôn ngữ khác, các từ được phân tách với nhau bằng dấu cách. Tuy nhiên trong tiếng Việt, dấu cách được dùng để phân tách các tiếng (âm tiết) chứ không phải từ. Cùng với các ngôn ngữ như tiếng Trung, tiếng Hàn, tiếng Nhật, phân tách từ trong tiếng Việt là một công việc không hề đơn giản.

Phân tích cú pháp - Dãy các từ sẽ được biến đổi thành các cấu trúc thể hiện sự liên kết giữa các từ này. Sẽ có những dãy từ bị loại do vi phạm các luật văn phạm.

Phân tích ngữ nghĩa - Thêm ngữ nghĩa vào các cấu trúc được tạo ra bởi bộ phân tích cú pháp.

Tích hợp văn bản - Ngữ nghĩa của một câu riêng biệt có thể phụ thuộc vào những câu đứng trước, đồng thời nó cũng có thể ảnh hưởng đến các câu phía sau.

Phân tích thực nghĩa - Cấu trúc thể hiện điều được phát ngôn sẽ được thông dịch lại để xác định nó thật sự có nghĩa là gì.

1.2. Các bài toán và ứng dụng:

Nhận dạng chữ viết: Có hai kiểu nhận dạng, thứ nhất là nhận dạng chữ in, ví dụ nhận dạng chữ trên sách giáo khoa rồi chuyển nó thành dạng văn bản điện tử như dưới định dạng doc của Microsoft Word chẳng hạn. Phức tạp hơn là nhận dạng chữ viết tay, có khó khăn bởi vì chữ viết tay không có khuôn dạng rõ ràng và thay đổi từ người này sang người khác. Với chương trình nhận dạng chữ viết in có thể chuyển hàng ngàn đầu sách trong thư viện thành văn bản điện tử trong thời gian ngắn. Nhận dạng chữ viết của con người có ứng dụng trong khoa học hình sự và bảo mật thông tin (nhận dạng chữ ký điện tử).

Nhận dạng tiếng nói: Nhận dạng tiếng nói rồi chuyển chúng thành văn bản tương ứng. Giúp thao tác của con người trên các thiết bị nhanh hơn và đơn giản hơn, chẳng hạn thay vì gõ một tài liệu nào đó bạn đọc nó lên và trình soạn thảo sẽ tự ghi nó ra. Đây cũng là bước đầu tiên cần phải thực hiện trong ước mơ thực hiện giao tiếp giữa con người với robot. Nhận dạng tiếng nói có khả năng trợ giúp người khiếm thị rất nhiều.

Tổng hợp tiếng nói: Từ một văn bản tự động tổng hợp thành tiếng nói. Thay vì phải tự đọc một cuốn sách hay nội dung một trang web, nó tự động đọc cho chúng ta. Giống như nhận dạng tiếng nói, tổng hợp tiếng nói là sự trợ giúp tốt cho người khiếm thị, nhưng ngược lại nó là bước cuối cùng trong giao tiếp giữa robot với người.

Dịch tự động (machine translate): Như tên gọi đây là chương trình dịch tự động từ ngôn ngữ này sang ngôn ngữ khác. Một phần mềm điển hình về tiếng

Việt của chương trình này là Evtrans của Softex, dịch tự động từ tiếng Anh sang tiếng Việt và ngược lại, phần mềm từng được trang web vdict.com mua bản quyền, đây cũng là trang đầu tiên đưa ứng dụng này lên mạng. Tháng 10 năm 2008 có hai công ty tham gia vào lĩnh vực này cho ngôn ngữ tiếng Việt là công ty Lạc Việt (công ty phát hành từ điển Lạc Việt) và Google, một thời gian sau đó Xalo.vn cũng đưa ra dịch vụ tương tự.

Tìm kiếm thông tin (information retrieval): Đặt câu hỏi và chương trình tự tìm ra nội dung phù hợp nhất. Thông tin ngày càng đầy lên theo cấp số nhân, đặc biệt với sự trợ giúp của internet việc tiếp cận thông tin trở lên dễ dàng hơn bao giờ hết. Việc khó khăn lúc này là tìm đúng nhất thông tin mình cần giữa bề bộn tri thức và đặc biệt thông tin đó phải đáng tin cậy. Các máy tìm kiếm dựa trên giao diện web như Google hay Yahoo hiện nay chỉ phân tích nội dung rất đơn giản dựa trên tần suất của từ khoá và thứ hạng của trang và một số tiêu chí đánh giá khác để đưa ra kết luận, kết quả là rất nhiều tìm kiếm không nhận được câu trả lời phù hợp, thậm chí bị dẫn tới một liên kết không liên quan gì do thủ thuật đánh lừa của các trang web nhằm giới thiệu sản phẩm (có tên tiếng Anh là SEO viết tắt của từ search engine optimization). Thực tế cho đến bây giờ chưa có máy tìm kiếm nào hiểu được ngôn ngữ tự nhiên của con người trừ trang www.ask.com được đánh giá là "hiểu" được những câu hỏi có cấu trúc ở dạng đơn giản nhất. Mới đây cộng đồng mạng đang xôn xao về trang Wolfram Alpha, được hứa hẹn là có khả năng hiểu ngôn ngữ tự nhiên của con người và đưa ra câu trả lời chính xác. Lĩnh vực này hứa hẹn tạo ra bước nhảy trong cách thức tiếp nhận tri thức của cả cộng đồng.

Tóm tắt văn bản: Từ một văn bản dài tóm tắt thành một văn bản ngắn hơn theo mong muốn nhưng vẫn chứa những nội dung thiết yếu nhất.

Khai phá dữ liệu (data mining) và phát hiện tri thức: Từ rất nhiều tài liệu khác nhau phát hiện ra tri thức mới. Thực tế để làm được điều này rất khó, nó gần như là mô phỏng quá trình học tập, khám phá khoa học của con người, đây là lĩnh vực đang trong giai đoạn đầu phát triển. Ở mức độ đơn giản khi kết hợp với máy tìm kiếm nó cho phép đặt câu hỏi để từ đó công cụ tự tìm ra câu trả lời dựa trên các thông tin trên web mặc cho việc trước đó có câu trả lời lưu trên web hay không (giống như trang Yahoo! hỏi và đáp, nơi chuyên đặt các câu hỏi để người khác trả lời), nói một cách nôm na là nó đã biết xử lý dữ liệu để trả lời câu hỏi của người sử dụng, thay vì máy móc đáp trả những gì chỉ có sẵn trong bộ nhớ.

II. Tóm tắt văn bản:

1. Bài toán tóm tắt văn bản:

Tóm tắt văn bản tự động là tác vụ để tạo ra một tóm tắt chính xác và hợp ngữ pháp trong khi vẫn giữ được các thông tin chính và ý nghĩa của văn bản gốc. Trong các năm gần đây, có rất nhiều hướng tiếp cận đã được nghiên cứu cho tóm tắt văn bản tự động và đã được áp dụng rộng rãi trong nhiều lĩnh vực. Ví dụ, máy tìm kiếm sinh ra các trích đoạn như là các bản xem trước của tài liệu, các website tin tức sinh ra các đoạn mô tả ngắn gọn cho bài viết (thường là tiêu đề của bài viết).

Mục tiêu của tóm tắt văn bản là tạo ra bản tóm tắt giống như cách con người tóm tắt, đây là bài toán đầy thách thức, bởi vì khi con người thực hiện tóm tắt một văn bản, chúng ta thường đọc toàn bộ nội dung rồi dựa trên sự hiểu biết và cảm thụ của mình để viết lại một đoạn tóm tắt nhằm làm nổi bật các ý chính của văn bản gốc. Nhưng vì máy tính khó có thể có được tri thức và khả năng ngôn ngữ như của con người, nên việc thực hiện tóm tắt văn bản tự động là một công việc phức tạp.

2. Các hướng tiếp cận tóm tắt văn bản:

Nhìn chung, có hai hướng tiếp cận cho tóm tắt văn bản tự động là trích chọn (extraction) và tóm lược (abstraction). Theo [32], tóm tắt văn bản có thể được phân loại dựa trên đầu vào (đơn hay đa văn bản), mục đích (tổng quát, theo lĩnh vực cụ thể, hay dựa trên truy vấn) và loại đầu ra (trích chọn hay tóm lược).

Phương pháp tóm tắt trích chọn thực hiện đánh giá các phần quan trọng của văn bản và đưa chúng một cách nguyên bản vào bản tóm tắt, do đó, phương pháp này chỉ phụ thuộc vào việc trích chọn các câu từ văn bản gốc dựa trên việc xếp hạng mức độ liên quan của các cụm từ để chỉ chọn những cụm từ liên quan nhất tới nội dung của tài liệu gốc. Trong khi đó, phương pháp tóm tắt tóm lược nhằm tạo ra văn bản tóm tắt mới có thể không gồm các từ hay các cụm từ trong văn bản gốc. Nó cố gắng hiểu và đánh giá văn bản sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên tiên tiến để tạo ra một văn bản ngắn hơn, truyền đạt được những thông tin quan trọng nhất từ văn bản gốc. Mặc dù các tóm tắt được con người thực hiện thường không giống như trích chọn, song hầu hết các nghiên cứu về tóm tắt văn bản hiện tại vẫn tập trung vào tóm tắt bằng phương pháp trích chọn vì về cơ bản các tóm tắt sinh bởi phương pháp trích chọn cho kết quả tốt hơn so với tóm tắt bằng phương pháp tóm lược. Điều này là bởi vì phương pháp tóm tắt bằng tóm lược phải đối mặt với các vấn đề như thể hệ ngữ nghĩa, suy luận và sinh ngôn ngữ tự nhiên, các vấn đề này phức tạp hơn nhiều lần so với việc trích chọn câu. Hướng tiếp cận tóm tắt bằng tóm lược khó hơn so với tóm tắt bằng trích chọn, song phương pháp này được kỳ vọng có thể tạo ra được các văn bản tóm tắt giống như cách con người thực hiện.

CHƯƠNG 2

MỘT SỐ NGHIÊN CỨU VỀ TÓM TẮT VĂN BẢN

I. Tóm tắt văn bản theo hướng trích chọn:

Như đã đề cập trong chương 1, các kỹ thuật tóm tắt bằng trích chọn sinh ra các đoạn tóm tắt bằng cách chọn một tập các câu trong văn bản gốc. Các đoạn tóm tắt này chứa các câu quan trọng nhất của đầu vào. Đầu vào có thể là đơn văn bản hoặc đa văn bản. Trong khuôn khổ của luận văn này, đầu vào của bài toán tóm tắt văn bản là đơn văn bản. Các hệ thống tóm tắt văn bản theo hướng trích chọn thường gồm các tác vụ: xây dựng một đại diện trung gian (intermediate representation) của văn bản đầu vào thể hiện các đặc điểm chính của văn bản; tính điểm (xếp hạng) các câu dựa trên đại diện trung gian đã xây dựng; chọn các câu đưa vào tóm tắt. Mỗi hệ thống tóm tắt văn bản tạo ra một số đại diện trung gian của văn bản mà nó sẽ thực hiện tóm tắt và tìm các nội dung nổi bật dựa trên đại diện trung gian này. Có hai hướng tiếp cận dựa trên đại diện trung gian là chủ đề đại diện (topic representation) và các đặc trưng đại diện (indicator representation). Các phương pháp dựa trên chủ đề đại diện biến đổi văn bản đầu vào thành một đại diện trung gian và tìm kiếm các chủ đề được thảo luận trong văn bản. Kỹ thuật tóm tắt dựa trên chủ đề đại diện tiêu biểu là phương pháp tiếp cận dựa trên tần xuất (frequency). Phương pháp dựa trên các đặc trưng đại diện thực hiện mô tả các câu trong văn bản như một danh sách các đặc trưng quan trọng chẳng hạn như độ dài câu, vị trí của câu trong tài liệu hay câu có chứa những cụm từ nhất định. Khi các đại diện trung gian đã được tạo ra, một điểm số thể hiện mức độ quan trọng sẽ được gán cho mỗi câu. Đối với phương pháp dựa trên chủ đề đại diện, điểm số của một câu thể hiện mức độ giải thích của câu đối với một vài chủ đề quan trọng nhất của văn bản. Trong hầu hết các phương pháp dựa trên đặc trưng đại diện, điểm số được tính bằng tổng hợp các dấu hiệu từ các đặc trưng khác nhau. Các kỹ thuật học máy thường được sử dụng để tìm trọng số cho các đặc trưng. Cuối cùng hệ thống tóm tắt sẽ lựa chọn các câu quan trọng nhất để tạo ra bản tóm tắt. Có thể áp dụng các thuật toán tham lam để chọn các câu quan trọng nhất từ văn bản gốc, hoặc biến việc lựa chọn câu thành một bài toán tối ưu trong đó xem xét ràng buộc tối đa hóa tầm quan trọng tổng thể và sự gắn kết ngữ nghĩa trong khi tối thiểu hóa sự dư thừa. Có nhiều yếu tố khác cần được cân nhắc khi lựa chọn các câu quan trọng, ví dụ ngữ cảnh của bản tóm tắt hay loại tài liệu cần tóm tắt (bài báo tin tức, email, báo cáo khoa học). Các tiêu chí này có thể trở thành các trọng số bổ sung cho việc lựa chọn các câu quan trọng đưa vào bản tóm tắt.

1. Phương pháp chủ đề đại diện dựa trên tần xuất:

1.1. Word probability:

Xác suất của từ (word probability) là dạng đơn giản nhất sử dụng tần xuất trên văn bản đầu vào như là một chỉ số quan trọng. Phương pháp này khá phụ thuộc vào độ dài của văn bản đầu vào, ví dụ, một từ xuất hiện ba lần trong một văn bản 10 từ có thể là từ quan trọng song có thể nó là một từ bình thường trong văn bản 1000 từ. Xác suất của một từ w : $p(w)$ được tính dựa trên số lần xuất hiện của từ w , $n(w)$, trong toàn bộ các từ thuộc văn bản đầu vào N .

$$P(w) = n(w)/N$$

Hệ thống SumBasic phát triển dựa trên ý tưởng sử dụng xác suất của từ để tính toán câu quan trọng. Với mỗi câu S_j trong văn bản đầu vào, nó gán một trọng số bằng xác suất trung bình của các từ chứa nội dung trong câu (một danh sách các từ không mang thông tin – stop words – sẽ bị loại khỏi quá trình đánh trọng số):

$$\text{Weight}(S_j) = \frac{\sum_{w_i \in S_j} p(w_i)}{|\{w_i | w_i \in S_j\}|}$$

Tiếp theo nó sẽ chọn các câu có điểm số tốt nhất gồm những từ có xác suất cao nhất. Bước này đảm bảo rằng các từ có xác suất cao nhất đại diện cho chủ đề của văn bản đầu vào sẽ được đưa vào bản tóm tắt. Sau khi chọn một câu đưa vào tóm tắt, xác suất của mỗi từ trong câu được hiệu chỉnh:

$$p_{\text{new}}(w_i) = p_{\text{old}}(w_i)^2$$

Việc hiệu chỉnh này thể hiện rằng xác suất một từ xuất hiện hai lần trong bản tóm tắt là thấp hơn so với xác suất từ xuất hiện chỉ một lần. Quá trình lặp lại cho đến khi đạt được độ dài cần thiết của văn bản tóm tắt.

1.2. Phương pháp TF-IDF:

TF-IDF (Term Frequency – Inverse Document Frequency) là 1 kỹ thuật sử dụng trong khai phá dữ liệu văn bản. Trọng số này được sử dụng để đánh giá tầm quan trọng của một từ trong một văn bản. Giá trị cao thể hiện độ quan trọng cao và nó phụ thuộc vào số lần từ xuất hiện trong văn bản nhưng bù lại bởi tần suất của từ đó trong tập dữ liệu. Một vài biến thể của tf-idf thường được sử dụng trong các hệ thống tìm kiếm như một công cụ chính để đánh giá và sắp xếp văn bản dựa vào truy vấn của người dùng. Tf-idf cũng được sử dụng để lọc những từ stopwords trong các bài toán như tóm tắt văn bản và phân loại văn bản.

TF: Term Frequency (Tần suất xuất hiện của từ) là số lần từ xuất hiện trong văn bản. Vì các văn bản có thể có độ dài ngắn khác nhau nên một số từ có thể xuất hiện nhiều lần trong một văn bản dài hơn là một văn bản ngắn. Như vậy, term frequency thường được chia cho độ dài văn bản (tổng số từ trong một văn bản).

$$tf(t, d) = \frac{f(t, d)}{\max\{f(w, d) : w \in d\}}$$

Trong đó:

- $tf(t, d)$: tần suất xuất hiện của từ t trong văn bản d
- $f(t, d)$: Số lần xuất hiện của từ t trong văn bản d
- $\max(\{f(w, d) : w \in d\})$: Số lần xuất hiện của từ có số lần xuất hiện nhiều nhất trong văn bản d

IDF: Inverse Document Frequency (Nghịch đảo tần suất của văn bản), giúp đánh giá tầm quan trọng của một từ. Khi tính toán TF, tất cả các từ được coi như có độ quan trọng bằng nhau. Nhưng một số từ như “is”, “of” và “that” thường xuất hiện rất nhiều lần nhưng độ quan trọng là không cao. Như thế chúng ta cần giảm độ quan trọng của những từ này xuống.

$$idf(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

Trong đó:

- $idf(t, D)$: giá trị idf của từ t trong tập văn bản
- $|D|$: Tổng số văn bản trong tập D
- $|\{d \in D : t \in d\}|$: thể hiện số văn bản trong tập D có chứa từ t .

Cơ số logarit trong công thức này không thay đổi giá trị idf của từ mà chỉ thu hẹp khoảng giá trị của từ đó. Vì thay đổi cơ số sẽ dẫn đến việc giá trị của các từ thay đổi bởi một số nhất định và tỷ lệ giữa các trọng lượng với nhau sẽ không thay đổi. (nói cách khác, thay đổi cơ số sẽ không ảnh hưởng đến tỷ lệ giữa các giá trị IDF). Việc sử dụng logarit nhằm giúp giá trị tf-idf của một từ nhỏ hơn, do có công thức tính tf-idf của một từ trong 1 văn bản là tích của tf và idf của từ đó.

Cụ thể, công thức tính tf-idf hoàn chỉnh như sau:

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D)$$

Khi đó: Những từ có giá trị TF-IDF cao là những từ xuất hiện nhiều trong văn bản này, và xuất hiện ít trong các văn bản khác. Việc này giúp lọc ra những từ phổ biến và giữ lại những từ có giá trị cao (từ khoá của văn bản đó).

2. Phương pháp đặc trưng đại diện:

2.1. Phương pháp đồ thị cho tóm tắt văn bản:

Phương pháp dựa trên đồ thị thể hiện văn bản như là một đồ thị liên thông. Các câu tạo thành các đỉnh của đồ thị và các cạnh giữa các câu thể hiện sự liên quan giữa hai câu với nhau. Một kỹ thuật thường được sử dụng để nối hai đỉnh đó là đo lường sự tương đồng giữa hai câu và nếu nó lớn hơn một ngưỡng nhất định thì chúng liên thông nhau. Đồ thị này thể hiện kết quả ở hai phần: thứ nhất, một phần đồ thị con được tạo bảo các chủ đề rời rạc trong văn bản; thứ hai, các câu được kết nối tới nhiều câu khác trong đồ thị là các câu quan trọng có thể lựa chọn đưa vào văn bản tóm tắt. Một phương pháp dựa trên đồ thị tiêu biểu đó là TextRank. Phương pháp dựa trên đồ thị không cần các kỹ thuật xử lý ngôn ngữ tự nhiên đặc thù cho từng ngôn ngữ ngoài việc tách câu và từ, nên nó có thể áp dụng cho nhiều ngôn ngữ khác nhau.

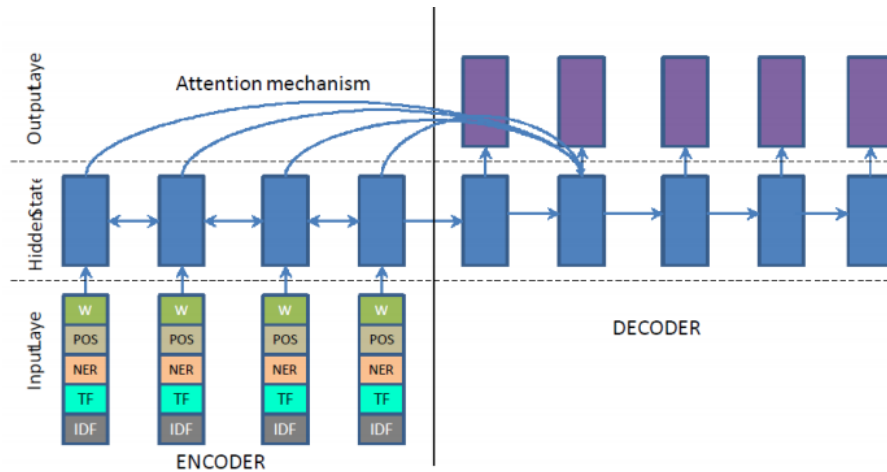
2.2. Kỹ thuật học máy cho tóm tắt văn bản:

Phương pháp áp dụng học máy cho tóm tắt văn bản thực hiện giải bài toán phân loại nhị phân. Tư tưởng của chúng là phân loại các câu trong văn bản đầu vào thành hai tập là tập các câu tóm tắt và tập các câu không là tóm tắt dựa vào các đặc trưng mà chúng có. Tập dữ liệu huấn luyện gồm các văn bản và các bản tóm tắt trích chọn tương ứng. Xác suất một câu được chọn vào văn bản tóm tắt là điểm số của câu. Việc lựa chọn các hàm phân loại đóng vai trò quan trọng trong việc tính điểm cho các câu. Một số đặc trưng phân loại thường được sử dụng trong tóm tắt văn bản gồm có vị trí của câu trong văn bản, độ dài của câu, tồn tại của các từ viết hoa, độ tương đồng của câu với tiêu đề của văn bản... Có nhiều kỹ thuật học máy được áp dụng trong tóm tắt văn bản, tiêu biểu là áp dụng của mô hình Markov ẩn (Hidden Markov Model).

II. Tóm tắt văn bản theo hướng tóm lược:

Những năm gần đây với sự phát triển của phần cứng máy tính, cùng với nhiều kỹ thuật tiên tiến dựa trên mạng nơ ron nhân tạo và kiến trúc mạng học sâu, một số nghiên cứu về tóm tắt văn bản bằng tóm lược đã được thực hiện với mục tiêu tạo được văn bản tóm tắt giống như cách con người thực hiện.

Nallapati và cộng sự [22] áp dụng mô hình chuỗi sang chuỗi (sequence-to-sequence) với cơ chế attention kết hợp với các đặc trưng ngôn ngữ (part-of-speech, name-entity và TF-IDF) để thực hiện tóm tắt văn bản theo hướng tóm lược (hình 2.1). Kết quả cho thấy mô hình có khả năng sinh ra các từ không có trong văn bản đầu vào, nhiều ví dụ cho thấy mô hình có thể sinh ra được đoạn tóm tắt gần giống với con người viết.



Hình 1. Mô hình *sequence-to-sequence* với cơ chế *attention*

Tác giả See và cộng sự trong đề xuất cải tiến mạng pointer-generator trên mô hình chuỗi sang chuỗi cho phép thực hiện sao chép một (các từ) từ văn bản gốc vào văn bản tóm tắt trong trường hợp mô hình sinh ra một từ không có trong tập từ vựng (unknown word). Mô hình được thử nghiệm trên bộ dữ liệu tiếng anh các bài báo của CNN/DailyMail cho kết quả khá khả quan. Hình 2.2. minh họa ví dụ chạy thử nghiệm được tác giả công bố.

CHƯƠNG 3

XÂY DỰNG HỆ THỐNG TÓM TẮT VĂN BẢN

I. Giới thiệu:

Kỹ thuật tóm tắt bằng trích chọn sinh ra các đoạn tóm tắt bằng cách chọn một tập các câu trong văn bản gốc. Các đoạn tóm tắt này chứa các câu quan trọng nhất của đầu vào. Đầu vào có thể là đơn văn bản hoặc đa văn bản. Trong khuôn khổ của luận văn này, đầu vào của bài toán tóm tắt văn bản là đơn văn bản. Các hệ thống tóm tắt văn bản theo hướng trích chọn thường gồm các tác vụ: xây dựng một đại diện trung gian (intermediate representation) của văn bản đầu vào thể hiện các đặc điểm chính của văn bản; tính điểm (xếp hạng) các câu dựa trên đại diện trung gian đã xây dựng; chọn các câu đưa vào tóm tắt.

1. Tóm tắt văn bản với Spacy:

Spacy là một thư viện phần mềm mã nguồn mở dành cho xử lý ngôn ngữ tự nhiên nâng cao, được viết bằng hai ngôn ngữ Python và Cython. Thư viện này được xuất bản với giấy phép MIT và các nhà phát triển chính là Matthew Honnibal và Ines Montani, cũng là những người sáng lập công ty phần mềm Explosion.

Không giống như NLTK, được sử dụng rộng rãi trong giảng dạy và nghiên cứu, spaCy tập trung vào việc cung cấp phần mềm để sử dụng trong sản xuất. spaCy cũng hỗ trợ các quy trình làm việc học sâu cho phép nối kết với các mô hình thống kê được huấn luyện bởi các thư viện máy học phổ biến như TensorFlow, PyTorch hay Apache MXNet thông qua thư viện học máy Thinc của riêng nó. Sử dụng Thinc làm chương trình phụ trợ (backend) của nó, spaCy làm nổi bật các mô hình mạng thần kinh tích chập cho các tác vụ gán nhãn từ loại (part-of-speech tagging), cây phân tích cú pháp, phân loại tài liệu và nhận dạng thực thể có tên (NER). Các mô hình thống kê mạng thần kinh nhân tạo được tích hợp trước để thực hiện các tác vụ này sẵn có ở 17 ngôn ngữ, bao gồm tiếng Anh, Bồ Đào Nha, Tây Ban Nha, Nga, Trung Quốc, và cũng có một mô hình NER đa ngữ. Thêm nữa, spaCy cũng hỗ trợ token hóa cho hơn 65 ngôn ngữ, cho phép người dùng huấn luyện mô hình tùy chỉnh trên các tập dữ liệu của riêng mình.

1.1. Các mô hình thống kê của Spacy:

Các mô hình này là động cơ năng lượng của spaCy. Các mô hình này cho phép spaCy thực hiện một số tác vụ liên quan đến NLP, chẳng hạn như gán thẻ một phần giọng nói, nhận dạng thực thể được đặt tên và phân tích cú pháp phụ thuộc.

- **en_core_web_sm:** CNN đa tác vụ bằng tiếng Anh được đào tạo trên OntoNotes. Kích thước - 11 MB
- **en_core_web_md:** CNN đa tác vụ bằng tiếng Anh được đào tạo trên OntoNotes, với các vector GloVe được đào tạo về Common Crawl. Kích thước - 91 MB

- **en_core_web_lg:** CNN đa tác vụ bằng tiếng Anh được đào tạo trên OntoNotes, với các vector GloVe được đào tạo về Common Crawl. Kích thước - 789 MB

1.2. Các chức năng chính của Spacy:

Các tính năng chính của spaCy:

- Hỗ trợ khoảng 60 ngôn ngữ.
- Các mô hình đã được đào tạo có sẵn cho các ngôn ngữ và ứng dụng khác nhau.
- Học đa nhiệm bằng cách sử dụng các máy biến áp đã được đào tạo trước đó như BERT (Kết xuất bộ mã hóa hai chiều của Máy biến áp).
- Hỗ trợ các vector được đào tạo trước và nhúng từ.
- Hiệu suất cao.
- Mô hình hệ thống đào tạo tại chỗ sẵn sàng sử dụng.
- Mã hóa có động cơ ngôn ngữ.
- Các thành phần sẵn sàng sử dụng có sẵn để liên kết các thực thể được đặt tên, đánh dấu các phần của giọng nói, phân loại văn bản, phân tích sự phụ thuộc dựa trên thể, chia câu, đánh dấu các phần của giọng nói, phân tích hình thái, tạo gốc, v.v.
- Hỗ trợ mở rộng chức năng với các thành phần và thuộc tính tùy chỉnh.
- Hỗ trợ tạo mô hình của riêng bạn dựa trên PyTorch, TensorFlow và các khuôn khổ khác.
- Các công cụ tích hợp để Liên kết đối tượng được đặt tên và Trực quan hóa cú pháp (NER, Nhận dạng đối tượng được đặt tên).
- Quy trình đóng gói và triển khai mô hình đơn giản và quản lý quy trình làm việc.
- Độ chính xác cao.

Thư viện được viết bằng Python với các phần tử trong Cython, một phần mở rộng Python cho phép gọi hàm trực tiếp bằng ngôn ngữ C.

Mã dự án được phân phối theo giấy phép MIT. Các mô hình ngôn ngữ đã sẵn sàng cho 58 ngôn ngữ.

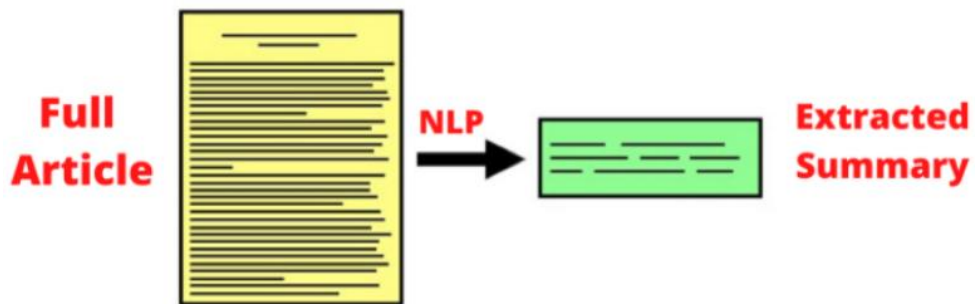
2. Thư viện NLTK:

Thư viện Natural Language Toolkit (tạm dịch là Bộ công cụ Ngôn ngữ Tự nhiên, viết tắt NLTK) là một nền tảng dẫn đầu để xây dựng các chương trình Python làm việc với dữ liệu ngôn ngữ của con người. Thư viện cung cấp giao diện để sử dụng với hơn 50 tài nguyên từ vựng và ngữ liệu (corpora), điển hình là WordNet cùng với các thư viện thích hợp để xử lý bài toán phân loại, token hóa (tokenization), tìm từ gốc (stemming),... Thư viện NLTK còn là công cụ tuyệt vời để giảng dạy

và giải quyết các bài toán về tính toán ngôn ngữ sử dụng Python. Ở các khóa học đào tạo Thạc sĩ và Tiến sĩ chuyên ngành, NLTK là một thư viện bắt buộc các học viên phải nắm vững và sử dụng thường xuyên.

II. Hệ thống tóm tắt văn bản:

Bài toán tóm tắt văn bản theo hướng trích suất sinh ra các đoạn tóm tắt bằng cách chọn một tập các câu trong văn bản gốc. Các đoạn tóm tắt này chứa các câu quan trọng nhất của đầu vào.



Hình 2. Mô hình bài toán tóm tắt văn bản

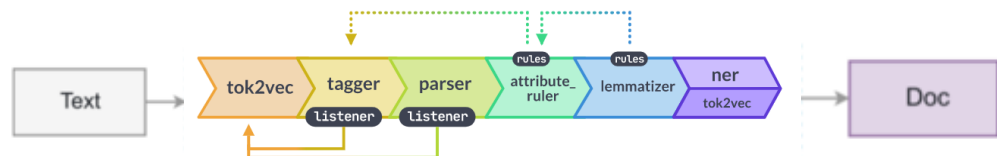
1. Quy trình tóm tắt theo hướng trích suất sử dụng Spacy:

Các bước thực hiện tóm tắt văn bản theo hướng trích xuất:

- Thu thập dữ liệu phù hợp: dữ liệu phù hợp cho bài toán tóm tắt văn bản tiếng Việt: văn bản đầy đủ được lấy từ website bằng phương pháp Cào dữ liệu (DrawData) và văn bản tóm tắt mẫu (do con người thực hiện tóm tắt).
- Xử lý dữ liệu: Khi dữ liệu được lấy về sẽ không được tốt, từ đó cho ra kết quả cuối cùng không đạt kết quả cao. Do đó, cần làm sạch dữ liệu, ở đây đa phần là xóa các thẻ HTML, JavaScip và cũng có thể là xóa bỏ các từ không cần thiết hay các ký tự không có ý nghĩa (\$%&##"). Với Python, BeautifulSoup và lxml là hai thư viện được cộng đồng sử dụng nhiều nhất và vô cùng mạnh mẽ, tiện lợi. Sau khi làm sạch dữ liệu, bước tiếp theo sẽ tách từ, Trong tiếng Việt, dấu cách (space) không được sử dụng như 1 kí hiệu phân tách từ, nó chỉ có ý nghĩa phân tách các âm tiết với nhau. Vì thế, để xử lý tiếng Việt, công đoạn tách từ (word segmentation) là 1 trong những bài toán cơ bản và quan trọng bậc nhất. Ví dụ: từ “đất nước” được tạo ra từ 2 âm tiết “đất” và “nước”, cả 2 âm tiết này đều có nghĩa riêng khi đứng độc lập, nhưng khi ghép lại sẽ mang một nghĩa khác. Vì đặc điểm này, bài toán tách từ trở thành 1 bài toán tiền đề cho các ứng dụng xử lý ngôn ngữ tự nhiên khác như phân loại văn bản, tóm tắt văn bản, máy dịch tự động, ... Tách từ chính xác hay không là công việc rất quan trọng, nếu không chính xác rất có thể dẫn đến việc ý nghĩa của câu sai, ảnh hưởng đến tính chính xác của chương trình. Về phương pháp, hiện nay cũng có khá nhiều mã nguồn nghiên cứu được public, có thể tham khảo tại word-segmentation.

Bước cuối của giai đoạn này là loại bỏ các StopWords, StopWords là những từ xuất hiện nhiều trong ngôn ngữ tự nhiên, tuy nhiên lại không mang nhiều ý nghĩa. Ở tiếng việt StopWords là những từ như: để, này, kia... Tiếng anh là những từ như: is, that, this... Có rất nhiều cách để loại bỏ StopWords nhưng có 2 cách chính là: Dùng từ điển và dựa theo tần suất xuất hiện của từ.

- Mã hóa bài viết bằng mô hình ngôn ngữ của spaCy. Những gì diễn ra đầu tiên là một nhiệm vụ được gọi là mã hóa , nó sẽ chia văn bản thành các đơn vị phân tích cần xử lý thêm. Trong hầu hết các trường hợp, mã thông báo tương ứng với các từ được phân tách bằng khoảng trắng, nhưng các dấu chấm câu cũng được coi là mã thông báo độc lập. Bởi vì máy tính coi các từ là chuỗi ký tự, việc gán dấu câu cho mã thông báo của riêng chúng sẽ ngăn không cho dấu câu theo sau gắn vào các từ đứng trước chúng. Sơ đồ bên dưới phác thảo các tác vụ mà spaCy có thể thực hiện sau khi văn bản đã được mã hóa:



Hình 3. Mô hình ngôn ngữ Spacy

Khi đó:

- Tokenizer: Phân đoạn văn bản thành các câu đơn.
 - Tagger: Gán thẻ cho đoạn văn bản.
 - Parser: Gán nhãn phụ thuộc
 - Ner: Phát hiện và gán nhãn các thực thể được đặt tên.
- Trích xuất các từ khóa quan trọng và tính toán trọng lượng chuẩn hóa. Từ các câu được tách ra, đếm số lần xuất hiện của các từ với từ phổ biến sẽ được gán 1 theo TFIDF.
 - Tính mức độ quan trọng của từng câu trong bài viết dựa trên sự xuất hiện của từ khóa. Dựa vào trọng số của từ được tính ở bước trên, trọng số của câu sẽ được tính bằng tổng trọng số của từ xuất hiện trong câu đó.
 - Sắp xếp các câu dựa trên mức độ quan trọng được tính toán. Khi có được trọng số các câu hoàn chỉnh, hệ thống sắp xếp lại trọng số các câu và chọn các câu có trọng số cao nhất làm thành đoạn văn bản tóm tắt.
 - Thử nghiệm và đánh giá với bộ dữ liệu đã được xử lý phía trên.

2. Quy trình tóm tắt theo hướng trích suất sử dụng Gensim:

Quy trình giống nhau ở các bước cào dữ liệu và xử lý dữ liệu và khác nhau ở các phần còn lại của quy trình.

Gensim là một thư viện python rất tiện dụng để thực hiện các tác vụ NLP. Quá trình tóm tắt văn bản sử dụng gensim thư viện dựa trên Thuật toán TextRank.

TextRank là một kỹ thuật tóm tắt khai thác. Nó dựa trên khái niệm rằng những từ xuất hiện thường xuyên hơn là đáng kể. Do đó, những câu có chứa các từ thường xuyên là rất quan trọng.

Dựa trên điều này, thuật toán chỉ định điểm số cho từng câu trong văn bản. Các câu được xếp hạng cao nhất sẽ lọt vào phần tóm tắt. Sau khi nhập gensimgói, bước đầu tiên là nhập summarize từ. Nó là một chức năng được xây dựng sẵn để thực hiện TextRank.gensim.summarization. Tiếp theo, chuyển ngữ liệu văn bản làm đầu vào cho summarize:

```
short_summary = summarize(original_text)
print(short_summary)
```

Các thông số là:

- ratio: Nó có thể nhận các giá trị từ 0 đến 1. Nó thể hiện tỷ lệ của bản tóm tắt so với văn bản gốc.
- word_count: Nó quyết định số từ trong bản tóm tắt.

3. Quy trình tóm tắt theo hướng trích suất sử dụng Suny:

Quy trình giống nhau ở các bước cào dữ liệu và xử lý dữ liệu và khác nhau ở các phần còn lại của quy trình.

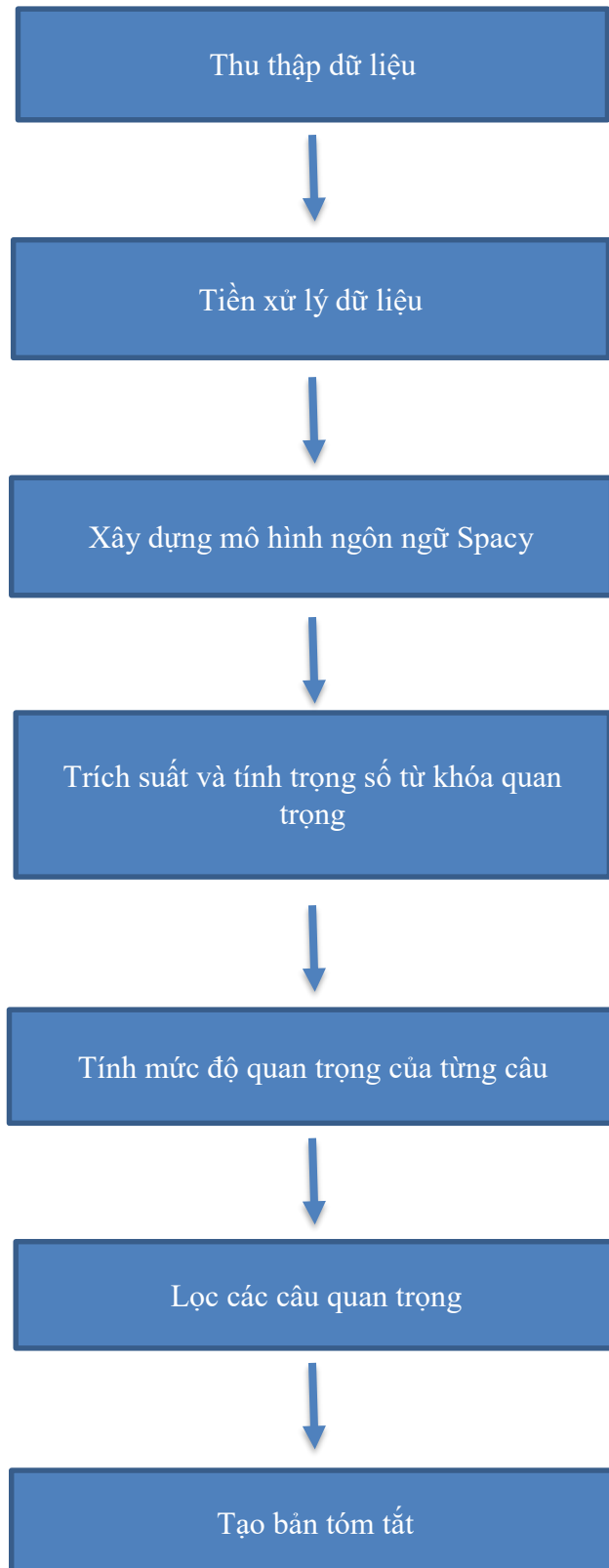
sumyLibraryray cung cấp cho bạn một số thuật toán để triển khai tính năng Tổng hợp Văn bản. Chỉ cần nhập thuật toán mong muốn của bạn thay vì phải tự viết mã nó.

Trong phần này, tôi sẽ thảo luận về việc triển khai các thuật toán dưới đây để tóm tắt bằng cách sử dụng sumy:

LexRank: Thứ hạng cao hơn, cao hơn là ưu tiên được đưa vào văn bản tóm tắt. Vì nguồn văn bản ở đây là một chuỗi, bạn cần sử dụng hàm để khởi tạo trình phân tích cú pháp. Bạn có thể chỉ định ngôn ngữ được sử dụng làm đầu vào cho. PlainTextParser.from_string()Tokenizer. Tiếp theo, tạo một mô hình tóm tắt lex_rank_summarizer để phù hợp với văn bản của bạn. lex_rank_summarizer (document, sentences_count). Bạn có thể quyết định số lượng câu bạn muốn trong phần tóm tắt thông qua tham số sentences_count.

KL-Sum: chọn các câu dựa trên sự giống nhau về cách phân bố từ như văn bản gốc. Nó nhằm mục đích hạ thấp tiêu chí phân kỳ KL. Nó sử dụng phương pháp tối ưu hóa tham lam và tiếp tục thêm các câu cho đến khi sự phân kỳ KL giảm xuống.

Các bước được tiến hành như thể hiện trong hình, chi tiết các bước được thể hiện trong các mục tiếp theo của niên luận.



Hình 4. Quy trình thực hiện tóm tắt văn bản tiếng việt với Spacy

4. Xây dựng bộ dữ liệu tóm tắt văn bản tiếng Việt:

Bài toán tóm tắt văn được đã được rất nhiều tác giả nghiên cứu, đặc biệt là đối với tóm tắt văn bản tiếng Anh. Với tóm tắt văn bản tiếng Anh, bộ dữ liệu kinh điển được sử dụng là bộ dữ liệu Gigaword với khoảng bốn triệu bài báo (Graff và các 31 cộng sự, 2003), chi phí mua giấy phép sử dụng bộ dữ liệu này là 6,000 USD nên chỉ có những tổ chức lớn mới có khả năng tiếp cận kho dữ liệu này. Một kho dữ liệu khác thường được sử dụng cho tóm tắt văn bản tiếng Anh đó là bộ dữ liệu các bài báo của CNN/Daily Mail với hơn 90,000 bài báo CNN và hơn 200,000 bài báo Daily Mail.

Tuy nhiên, đối với tóm tắt văn bản tiếng Việt, hiện tại chưa có kho dữ liệu chính thức nào được công bố, đây là thách thức lớn đối với chúng tôi. Vì vậy, để chuẩn bị dữ liệu thực hiện bài toán tóm tắt văn bản tiếng Việt, chúng tôi tiến hành thu thập dữ liệu là các bài báo trên một số website tin tức của Việt Nam. Dữ liệu mà chúng tôi quan tâm đó là phần tóm tắt dưới tiêu đề của bài báo, và nội dung văn bản của bài báo.

Dữ liệu bao gồm các bài báo về các chủ đề: *Công nghệ thông tin, giải trí, giáo dục, pháp luật, sức khỏe, thể giới, thể thao.*



Thủ tướng Phạm Minh Chính: Tiếp tục đổi mới đồng bộ, toàn diện công tác xây dựng pháp luật

VTV



Thủ môn được HLV Park Hang-seo bổ sung phút chót vào danh sách đi UAE là ai?

NGƯỜI LAO ĐỘNG



Bộ Y tế: Chặn dịch ở tỉnh Bắc Giang phải nhanh gấp 10 lần Đà Nẵng

Vietnam+



Mỹ cam kết hỗ trợ tái thiết Gaza

tin tức 39 liên quan



Sửa đổi một số mức phạt vi phạm về môi trường

Chính Phủ 1 liên quan

Hình 5. Dữ liệu cào từ trang web

5. Tiền xử lý dữ liệu:

Với dữ liệu thu được từ các website tin tức trực tuyến của Báo Cần Thơ, tiến hành tiền xử lý để làm sạch dữ liệu và loại bỏ các ký tự nhiễu trong văn bản như sau:

- Khi dữ liệu được lấy về sẽ không được tốt, từ đó cho ra kết quả cuối cùng không đạt kết quả cao. Do đó, cần làm sạch dữ liệu, ở đây đa phần là xóa các thẻ HTML, JavaScript, cũng có thể là xóa bỏ các từ không cần thiết hay các ký tự không có ý nghĩa (\$%&##"), loại bỏ các dấu gạch đầu dòng, các dấu gạch ngang trong văn bản và loại bỏ các dấu hai chấm ":" trước mỗi danh sách liệt kê. Với Python, BeautifulSoup và lxml là hai thư viện được cộng đồng sử dụng nhiều nhất và vô cùng mạnh mẽ, tiện lợi.
- Thay thế các dấu chấm phẩy ";" phân tách ý thành dấu chấm ngắt câu "."
- Thêm dấu chấm kết thúc câu cho những chú thích dưới ảnh không có dấu kết thúc câu.
- Tách các câu trong phần tóm tắt của bài báo bằng phân tách các câu dựa trên kết thúc câu bởi dấu chấm, dấu chấm hỏi và dấu chấm than.
- Tách văn bản thành các token. Trong tiếng Việt, dấu cách (space) không được sử dụng như 1 kí hiệu phân tách từ, nó chỉ có ý nghĩa phân tách các âm tiết với nhau. Vì thế, để xử lý tiếng Việt, công đoạn tách từ (word segmentation) là 1 trong những bài toán cơ bản và quan trọng bậc nhất. Ví dụ: từ "đất nước" được tạo ra từ 2 âm tiết "đất" và "nước", cả 2 âm tiết này đều có nghĩa riêng khi đứng độc lập, nhưng khi ghép lại sẽ mang một nghĩa khác. Vì đặc điểm này, bài toán tách từ trở thành 1 bài toán tiền đề cho các ứng dụng xử lý ngôn ngữ tự nhiên khác như phân loại văn bản, tóm tắt văn bản, máy dịch tự động, ... Tách từ chính xác hay không là công việc rất quan trọng, nếu không chính xác rất có thể dẫn đến việc ý nghĩa của câu sai, ảnh hưởng đến tính chính xác của chương trình. Về phương pháp, hiện nay cũng có khá nhiều mã nguồn nghiên cứu được public, có thể tham khảo tại word-segmentation. Bước cuối của giai đoạn này là loại bỏ các StopWords, StopWords là những từ xuất hiện nhiều trong ngôn ngữ tự nhiên, tuy nhiên lại không mang nhiều ý nghĩa. Ở tiếng việt StopWords là những từ như: để, này, kia... Tiếng anh là những từ như: is, that, this... Có rất nhiều cách để loại bỏ StopWords nhưng có 2 cách chính là: Dùng từ điển và dựa theo tần suất xuất hiện của từ.

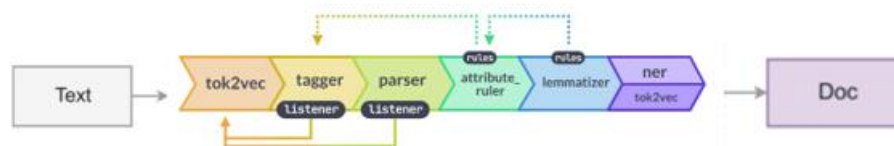
6. Mã hóa bài viết bằng mô hình ngôn ngữ của spaCy:

Mô hình Spacy được sử dụng thông qua chức năng tải tích hợp.

```
nlp = spacy.load("en_core_web_lg")
import en_core_web_lg
nlp = en_core_web_lg.load()
```

Mô hình Spacy thực hiện chuyển văn bản đầu vào thành chữ thường và mã hóa nó bằng mô hình ngôn ngữ Spacy. Nó lặp qua từng từ xuất hiện trong văn bản và bỏ qua nếu nó là một từ khóa hay là dấu chấm câu.

Những gì diễn ra đầu tiên là một nhiệm vụ được gọi là mã hóa, nó sẽ chia văn bản thành các đơn vị phân tích cần xử lý thêm. Trong hầu hết các trường hợp, mã thông báo tương ứng với các từ được phân tách bằng khoảng trắng, nhưng các dấu chấm câu cũng được coi là mã thông báo độc lập. Bởi vì máy tính coi các từ là chuỗi ký tự, việc gán dấu câu cho mã thông báo của riêng chúng sẽ ngăn không cho dấu câu theo sau gắn vào các từ đứng trước chúng. Sơ đồ bên dưới phác thảo các tác vụ mà spaCy có thể thực hiện sau khi văn bản đã được mã hóa:



Khi

đó:

- Tokenizer: Phân đoạn văn bản thành các câu đơn.
- Tagger: Gán thẻ cho đoạn văn bản.
- Parser: Gán nhãn phụ thuộc
- Ner: Phát hiện và gắn nhãn các thực thể được đặt tên.

7. Trích xuất các từ khóa quan trọng và tính toán trọng lượng chuẩn hóa:

TFIDF - Dạng viết tắt cho Tần suất thuật ngữ - Tần suất tài liệu nghịch đảo, nó được sử dụng để biểu thị mức độ quan trọng của một từ nhất định đối với một tài liệu trên một bộ sưu tập hoàn chỉnh và được tính qua 3 bước lần lượt theo 3 công thức sau:

Tần suất xuất hiện của từ:

$$tf(t, d) = \frac{f(t, d)}{\max\{f(w, d) : w \in d\}} \quad (1)$$

Trong đó:

- $tf(t, d)$: tần suất xuất hiện của từ t trong văn bản d
- $f(t, d)$: Số lần xuất hiện của từ t trong văn bản d

- $\max(\{f(w, d) : w \in d\})$: Số lần xuất hiện của từ có số lần xuất hiện nhiều nhất trong văn bản d

Nghịch đảo tần suất của văn bản:

$$\text{idf}(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \quad (2)$$

Trong đó:

- $\text{idf}(t, D)$: giá trị idf của từ t trong tập văn bản
- $|D|$: Tổng số văn bản trong tập D
- $|\{d \in D : t \in d\}|$: thể hiện số văn bản trong tập D có chứa từ t.

Tích TFIDF:

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D) \quad (3)$$

Ví dụ: xét đoạn văn bản nhỏ sau, ký hiệu là data:

Bé 18 tháng tuổi ở TP.HCM dương tính với nCoV. Bệnh nhi là cháu ngoại của bà chủ quán bánh canh ở hẻm 287, đường Nguyễn Đình Chiểu, quận 3, TP.HCM. Giám đốc Sở Y tế Nguyễn Tấn Bình cho biết bé ở cùng gia đình tại quận Tân Bình. Bệnh nhi được điều trị tại Bệnh viện Nhi đồng Thành phố.

Đầu tiên, theo công thức sẽ đếm xem từng từ trong văn bản xuất hiện bao nhiêu lần. điển hình từ “bé” xuất hiện 2 lần nên “bé”=2. Tương tự ta được các từ còn lại:

'biết': 1, 'bà': 1, 'bánh': 1, 'bé': 2, 'bình': 1, 'bệnh': 3, 'bình': 1, 'canh': 1, 'chiểu': 1, 'cho': 1, 'cháu': 1, 'chủ': 1, 'cùng': 1, 'của': 1, 'dương': 1, 'gia': 1, 'giám': 1, 'hẻm': 1, 'là': 1, 'ngoại': 1, 'nguyễn': 2, 'nhi': 3, 'phố': 1, 'quán': 1, 'quận': 2, 'sở': 1, 'thành': 1, 'tháng': 1, 'trị': 1, 'tuổi': 1, 'tân': 1, 'tính': 1, 'tại': 2, 'tân': 1, 'tế': 1, 'viện': 1, 'với': 1, 'y': 1, 'điều': 1, 'đình': 2, 'đường': 1, 'được': 1, 'độc': 1, 'đồng': 1, 'ở': 3

Sau khi có được số lần xuất hiện của các từ, theo công thức (1) xét thấy từ “nhi” có số lần xuất hiện nhiều nhất nên áp dụng công thức (1) cho từ “biết” được:

$$\text{Tf}(\text{biết}, \text{data}) = \frac{1}{3} = 0.333333333.$$

Công thức (2) dùng để giảm mức độ quan trọng của các từ không cần thiết trong đoạn văn bản. ví dụ như từ “ở”,...

Sau khi áp dụng lần lượt các công thức như trên được các trọng số quan trọng của từ như sau:

'biết': 0.3333333333333333,	'bà': 0.3333333333333333,	'bánh': 0.3333333333333333,
0.3333333333333333,	'bé': 0.6666666666666666,	'bình': 0.3333333333333333,
0.3333333333333333,	'bệnh': 1.0,	'bình': 0.3333333333333333,
0.3333333333333333,	'chiều': 0.3333333333333333,	'cho': 0.3333333333333333,
0.3333333333333333,	'cháu': 0.3333333333333333,	'chủ': 0.3333333333333333,
0.3333333333333333,	'cùng': 0.3333333333333333,	'của': 0.3333333333333333,
0.3333333333333333,	'đương': 0.3333333333333333,	'gia': 0.3333333333333333,
0.3333333333333333,	'giám': 0.3333333333333333,	'hẻm': 0.3333333333333333,
0.3333333333333333,	'là': 0.3333333333333333,	'ngoại': 0.3333333333333333,
0.3333333333333333,	'nguyên': 0.6666666666666666,	'nhì': 1.0,
0.3333333333333333,	'quán': 0.3333333333333333,	'phố': 0.3333333333333333,
0.6666666666666666,	'sở': 0.3333333333333333,	'quận': 0.3333333333333333,
0.3333333333333333,	'tháng': 0.3333333333333333,	'thành': 0.3333333333333333,
0.3333333333333333,	'tuổi': 0.3333333333333333,	'trị': 0.3333333333333333,
0.3333333333333333,	'tính': 0.3333333333333333,	'tân': 0.3333333333333333,
0.3333333333333333,	'tính': 0.3333333333333333,	'tính': 0.3333333333333333,
'tấn': 0.3333333333333333,	'tế': 0.3333333333333333,	'tính': 0.3333333333333333,
0.3333333333333333,	'vớ': 0.3333333333333333,	'y': 0.3333333333333333,
'điều': 0.3333333333333333,	'đình': 0.6666666666666666,	'đường': 0.3333333333333333,
0.3333333333333333,	'được': 0.3333333333333333,	'độc': 0.3333333333333333,
0.3333333333333333,	'đồng': 0.3333333333333333,	'ở': 1.0

8. Tính mức độ quan trọng của từng câu trong bài viết dựa trên sự xuất hiện của từ khóa:

Sau khi nhận được trọng số của từng từ riêng lẻ, suy ra trọng số mỗi câu của mỗi câu. Bằng cách này, biết được tầm quan trọng của từng câu để có thể loại bỏ những câu không quan trọng khỏi phần tóm tắt.

Trọng số mỗi câu được tính theo công thức:

$$\text{Trọng số câu} = \sum \text{Trọng số các từ xuất hiện trong câu}$$

Ví dụ:

Bé 18 tháng tuổi ở TP.HCM dương tính với nCoV = “bé” + “tháng” + “tuổi” + “dương” + “tính” + “vớ” + “ở” = 0.6666666666666666 + 0.3333333333333333 + 0.3333333333333333 + 0.3333333333333333 + 0.3333333333333333 + 0.3333333333333333 + 0.3333333333333333 + 1 = 3.3333333333333335

Tính tương tự với các câu còn lại sẽ được kết quả như sau:

Bé 18 tháng tuổi ở TP.HCM dương tính với nCoV.: 3.333333333333335,
Bệnh nhi là cháu ngoại của bà chủ quán bánh canh ở hẻm 287, đường Nguyễn
Đình Chiểu, quận 3, TP.HCM.: 9.0,
Sở Y tế: 0.6666666666666666,
cho biết bé ở cùng gia đình tại quận: 5.0,
Tân Bình.: 0.6666666666666666,
Bệnh nhi được điều trị: 3.0000000000000004,
tại Bệnh viện: 1.9999999999999998,
Nhi đồng Thành phố.: 1.9999999999999998

9. Lọc các câu dựa trên mức độ quan trọng được tính toán:

Sắp xếp các câu theo cách câu có những từ quan trọng nhất sẽ quan trọng hơn.

10. Tạo bản tóm tắt:

Bản tóm tắt cuối cùng đang được tạo chỉ sử dụng thông tin có giá trị. Tất cả nội dung ít quan trọng hơn hoặc không quan trọng sẽ bị xóa khỏi nội dung.

Bé 18 tháng tuổi ở TP.HCM dương tính với nCoV. Bệnh nhi là cháu ngoại của
bà chủ quán bánh canh ở hẻm 287, đường Nguyễn Đình Chiểu, quận 3,
TP.HCM.

CHƯƠNG 4

ĐÁNH GIÁ TÓM TẮT VĂN BẢN

I. Môi trường thử nghiệm:

Mô hình tóm tắt văn bản tiếng việt được xây dựng và thử nghiệm trên máy Tính có cấu hình như sau:

- CPU: Intel(R) Core(TM) i5-8265U CPU @ 1.60GHz 1.80 GHz
- RAM: 12GB.
- GPU: NVIDIA MX230.
- Hệ điều hành Windows 10.
- Ngôn ngữ lập trình: Python trên trình biên dịch Python 3.9
- IDE: Google Colab + Python 3.9 + Visual studio code.

Các công cụ chính sử dụng:

Google Colab: Cho phép xây dựng mô hình một cách trực quan. Nó cung cấp các thư viện tích hợp cho phép cấu hình tham số trong quá trình huấn luyện, áp dụng các công thức tính toán trên số học và ma trận, đồng thời hiển thị các kết quả bằng biểu đồ, đồ thị.

NLTK: NLTK là viết tắt của Natural Language Toolkit, đây là công cụ xử lý ngôn ngữ tự nhiên mạnh trên môi trường Python. Luận văn sử dụng NLTK để thực hiện tách từ đơn, phục vụ cho việc chuyển văn bản từ dạng thông thường (text) sang dạng nhị phân (binary).

Pyvi: Thư viện Python để tách từ Tiếng Việt [31]. Luận văn sử dụng Pyvi để xây dựng tập từ điển và tách từ từ văn bản đầu vào.

SpaCy: spaCy là một thư viện phần mềm mã nguồn mở để xử lý ngôn ngữ tự nhiên nâng cao. spaCy được thiết kế đặc biệt để sử dụng trong sản xuất và giúp bạn xây dựng các ứng dụng xử lý và “hiểu” khối lượng lớn văn bản. Nó có thể được sử dụng để xây dựng hệ thống khai thác thông tin hoặc hiểu ngôn ngữ tự nhiên hoặc để xử lý trước văn bản để học sâu.

II. Đánh giá tóm tắt văn bản:

1. Các phương pháp đánh giá tóm tắt văn bản:

1.1. 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.

1.2. 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.

$$Precision = \frac{|SH \cap SM|}{|SM|}$$

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.

$$Recall = \frac{|SH \cap SM|}{|SH|}$$

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 được trong đoạn $[0,1]$, hiển nhiên giá trị tốt nhất là 1.

$$f - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

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 là một số không âm, nghĩa là precision quan trọng hơn, nghĩa là recall quan trọng hơn.

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

1.3. Đánh giá dựa trên nội dung:

1.3.1. 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.

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

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.

1.3.2. Phương pháp ROUGE:

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 bản 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:

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

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.

1.3.2. Phương pháp đánh giá BLEU (Bilingual Evaluation Understudy):

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.

2. Đánh giá tóm tắt văn bản:

Đánh giá tóm tắt văn bản được thử nghiệm trên các bài báo ở các lĩnh vực khác nhau. Tiêu biểu điển hình là bài báo về y tế và giáo dục được lấy từ trang Baocantho và sử dụng phương pháp đánh giá thủ công.

Văn bản tóm tắt mẫu và văn bản do hệ thống thực hiện tóm tắt được xây dựng trên văn bản gốc và văn bản có số câu bằng nhau.

❖ Văn bản gốc: bài báo về y tế.

Trung tâm Sàng lọc - Chẩn đoán trước sinh và sơ sinh tại BV Phụ sản TP Cần Thơ được Tổng cục Dân số - Kế hoạch hóa gia đình giao trọng trách thực hiện Đề án nâng cao chất lượng dân số 12 tỉnh ĐBSCL. Theo đó, Trung tâm thực hiện tầm soát các bệnh, tật bẩm sinh giai đoạn trước sinh cho thai phụ đến khám thai tại BV trong 3 tháng đầu và 3 tháng giữa thai kỳ. Trong 48 giờ đầu sau sinh, trẻ sinh ra tại viện đều được tầm soát bệnh lý bẩm sinh thông qua xét nghiệm lấy máu gót chân. Trung tâm còn tiếp nhận mẫu từ các BV có chuyên khoa sản trên địa bàn thành phố và các tỉnh trong vùng gửi đến. Trước đây, khi có kết quả chẩn đoán bệnh, tất cả các trường hợp nghi ngờ nguy cơ cao đều phải chuyển lên tuyến trên. Các gia đình có con mắc bệnh phải thường xuyên đưa con lên TP Hồ Chí Minh tái khám, trong khi các BV tuyến trên quá tải, cùng với việc di chuyển tới lui, tốn kém nhiều chi phí..., khiến một số gia đình nản chí, ảnh hưởng đến hiệu quả điều trị cho con.

Từ thực tế đó, các bác sĩ chuyên khoa Nhi - Sơ sinh BV Phụ sản TP Cần Thơ quyết tâm triển khai các kỹ thuật trong lĩnh vực này để tiếp nhận điều trị cho bệnh nhi. BS CKI Thạch Thị Ngọc Yến, Phó Trưởng Khoa Nhi - Sơ sinh, chia sẻ: “Với chỉ đạo và ủng hộ của Ban Giám đốc BV, đội ngũ bác sĩ của khoa được luân phiên lên tuyến trên, ở BV Nhi Trung ương Hà Nội học về lĩnh vực này. Bước đầu khó khăn nhiều, vì đây là lĩnh vực mới đối với Khoa Nhi - Sơ sinh; thứ hai là đây là lĩnh vực khó. Thật sự phải cảm ơn các anh chị đồng nghiệp ở tuyến trên, ngay cả bậc tiền bối đã nhiệt tình hỗ trợ cho đơn vị. Từ năm 2019 đến thời điểm này, BV Phụ sản TP Cần Thơ đã phát hiện, điều trị 4 ca tăng sản, 11 ca suy giáp và theo dõi 2 bệnh về rối loạn chuyển hóa bẩm sinh”.

Bệnh tăng sản tuyến thượng thận bẩm sinh là bệnh không phổ biến, tần suất 1/10.000-1/15.000 trẻ sinh ra. Bệnh suy giáp bẩm sinh có tỷ lệ gặp phải từ 1/3.000 - 1/4.000. Bệnh thiếu hụt citrin có tỷ lệ gặp từ 1/17.000 - 1/34.000. Bệnh thiếu hụt Enzyme Beta-Ketothiolase là bệnh hiếm gặp. Việt Nam đã có 35 ca đã được chẩn đoán trong đó có 1 trường hợp do BV Phụ sản TP Cần Thơ sàng lọc

phát hiện cuối năm 2019. Trẻ mắc các bệnh lý này, nếu không được sàng lọc, phát hiện sớm, chậm trễ can thiệp, điều trị, chắc chắn ảnh hưởng đến sự phát triển thể chất và tinh thần của trẻ, thậm chí khiến trẻ tử vong. Một số trẻ mắc bệnh lý về nội tiết hay rối loạn chuyển hóa bẩm sinh nếu không điều trị sớm, sẽ trở nên ngờ nghệch.

Cô N.T.T (ngụ ở quận Ninh Kiều) đưa cháu nội 7 tháng tuổi đến tái khám tại Khoa Nhi - Sơ sinh BV Phụ sản TP Cần Thơ, chia sẻ: “Các bác sĩ, điều dưỡng ở Khoa Nhi - Sơ sinh là những người mẹ đã hồi sinh cho bé một cuộc đời”. Cô T kể, khi có kết quả chẩn đoán bé có nguy cơ cao mắc bệnh tăng sản thượng thận bẩm sinh, gia đình không biết đó là bệnh gì. Sau đó, bé nhập viện và diễn biến nặng dần, tưởng đâu không qua khỏi. Suốt 39 ngày nằm viện, cháu phải thở máy, tất cả nhờ công chăm sóc tận tình của cán bộ y tế nơi đây để giúp cháu qua cơn nguy kịch. BS Ngọc Yến và đồng nghiệp cũng theo dõi, điều trị bệnh tăng sản cho cháu khỏe, ổn định.

Không chỉ trường hợp của cháu cô T, mà hầu hết các trường hợp mắc bệnh bẩm sinh, BS Ngọc Yến và tập thể Khoa Nhi - Sơ sinh đều theo dõi chặt chẽ, chăm sóc chu đáo. Vì là lĩnh vực mới, lại là lĩnh vực khó, nên BS Ngọc Yến thường xuyên mài mò học hỏi chuyên môn từ sách vở, kết nối chặt chẽ với các đồng nghiệp, các chuyên gia đầu ngành trong lĩnh vực bệnh lý bẩm sinh ở tuyến trên. TS Vũ Chí Dũng, Trưởng khoa Nội tiết - Chuyển hóa - Di truyền, BV Nhi Trung ương Hà Nội, là người luôn đồng hành, hỗ trợ Khoa Nhi - Sơ sinh BV Phụ sản trong việc chẩn đoán, theo dõi và can thiệp điều trị kịp thời cho các bé bị bệnh, để sống khỏe mạnh như bao trẻ khác.

Văn bản tóm tắt mẫu:

Trung tâm Sàng lọc - Chẩn đoán trước sinh và sơ sinh tại BV Phụ sản TP Cần Thơ được Tổng cục Dân số - Kế hoạch hóa gia đình giao trọng trách thực hiện án nâng cao chất lượng dân số 12 tỉnh ĐBSCL. Theo đó, Trung tâm thực hiện tầm soát các bệnh, tật bẩm sinh giai đoạn trước sinh cho thai phụ đến khám thai tại BV trong 3 tháng đầu và 3 tháng giữa thai kỳ. **Trong 48 giờ đầu sau sinh, trẻ sinh ra tại viện đều được tầm soát bệnh lý bẩm sinh thông qua xét nghiệm lấy máu gót chân.** Cô N.T.T đưa cháu nội 7 tháng tuổi đến tái khám tại Khoa Nhi - Sơ sinh BV Phụ sản TP Cần Thơ, chia sẻ: “**Các bác sĩ, điều dưỡng Khoa Nhi - Sơ sinh là những người mẹ hồi sinh cho bé một cuộc đời**”. Cô T kể, khi có kết quả chẩn đoán bé có nguy cơ cao mắc bệnh tăng sản thượng thận bẩm sinh, gia đình không biết là bệnh gì. Sau đó, bé nhập viện và diễn biến nặng dần, tưởng đâu không qua khỏi. Không chỉ trường hợp

của cháu cô T, mà hầu hết các trường hợp mắc bệnh bẩm sinh, BS Ngọc Yến và tập thể Khoa Nhi - Sơ sinh đều theo dõi chặt chẽ, chăm sóc chu đáo.

Link: <https://baocantho.com.vn/-ia-chi-dieu-tri-benh-bam-sinh-hiem-gap-tai-can-tho-a134027.html>

Văn bản tóm tắt với Spacy:

TS Vũ Chí Dũng, Trưởng khoa Nội tiết - Chuyển hóa - Di truyền, BV Nhi Trung ương Hà Nội, là người luôn đồng hành, hỗ trợ Khoa Thờ đã phát hiện, điều trị 4 ca tăng sản, 11 ca suy giáp và theo dõi 2 bệnh về rối loạn chuyển hóa bẩm sinh”. Thờ, chia sẻ: **“Các bác sĩ, điều dưỡng ở Khoa Nhi - Sơ sinh là những người mẹ đã hồi sinh cho bé một cuộc đời”**. Cô T kể, khi có kết quả chẩn đoán bé có nguy cơ cao mắc bệnh tăng sản thượng thận bẩm sinh, gia đình không biết là bệnh gì. BS CKI Thạch Thị Ngọc Yến, Phó Trưởng Khoa Nhi - Sơ sinh, chia sẻ: “Với chỉ đạo và ủng hộ của Ban Giám đốc BV, đội ngũ Trước đây, khi có kết quả chẩn đoán bệnh, tất cả các trường hợp nghi ngờ nguy cơ cao đều phải chuyển lên tuyến trên. **Trong 48 giờ đầu sau sinh, trẻ sinh ra tại viện đều được tầm soát bệnh lý bẩm sinh thông qua xét nghiệm lấy máu gót chân.**

⇒ Văn bản sau khi tóm tắt rúc trích các câu có trọng số cao để làm thành bài báo tóm tắt. Sau khi tóm tắt có thể thấy, văn bản có số câu ít hơn văn bản gốc về mặt số lượng và vẫn giữ được nội dung đầy đủ mà bài báo gốc mang đến về mặt nội dung.

⇒ Khi đối chiếu với văn bản tóm tắt mẫu cho thấy:

- Về nội dung: văn bản hệ thống tóm tắt vẫn giữ được nội dung gần giống với văn bản tóm tắt mẫu, không có thông tin dư thừa và không thiếu đi thông tin quan trọng.
- Về mặt ngữ pháp: văn bản tóm tắt sinh ra đoạn văn gồm các câu có đầy đủ chủ ngữ vị ngữ, không xuất hiện lỗi sai từ. Các câu liên kết với nhau thông qua dấu chấm câu – dấu “.” Tuy nhiên, kết cấu câu chữ vẫn chưa được hoàn thiện như ở văn bản tóm tắt mẫu.

So sánh với các phiên bản tóm tắt khác – Gensum, Sumy:

Spacy: TS Vũ Chí Dũng, Trưởng khoa Nội tiết - Chuyển hóa - Di truyền, BV Nhi Trung ương Hà Nội, là người luôn đồng hành, hỗ trợ Khoa Thờ đã phát hiện, điều trị 4 ca tăng sản, 11 ca suy giáp và theo dõi 2 bệnh về rối loạn chuyển hóa bẩm sinh”. Thờ, chia sẻ: “Các bác sĩ, điều dưỡng ở Khoa Nhi - Sơ

sinh là những người mẹ đã hồi sinh cho bé một cuộc đời”. Cô T kể, khi có kết quả chẩn đoán bé có nguy cơ cao mắc bệnh tăng sản thượng thận bẩm sinh, gia đình không biết BS CKI Thạch Thị Ngọc Yến, Phó Trưởng Khoa Nhi - Sơ sinh, chia sẻ: “Với chỉ đạo và ủng hộ của Ban Giám đốc BV, đội ngũ Trước đây, khi có kết quả chẩn đoán bệnh, tất cả các trường hợp nghi ngờ nguy cơ cao đều phải chuyển lên tuyến trên. Trong 48 giờ đầu sau sinh, trẻ sinh ra tại viện đều được tầm soát bệnh lý bẩm sinh thông qua xét nghiệm lấy máu gót chân.

Gensum: Trung tâm Sàng lọc - Chẩn đoán trước sinh và sơ sinh tại BV Phụ sản TP Cần Thơ được Tổng cục Dân số - Kế hoạch hóa gia đình giao trọng trách thực hiện Đề án nâng cao chất lượng dân số 12 tỉnh ĐBSCL. Các gia đình có con mắc bệnh phải thường xuyên đưa con lên TP Hồ Chí Minh tái khám, trong khi các BV tuyến trên quá tải, cùng với việc di chuyển tới lui, tốn kém nhiều chi phí..., khiến một số gia đình nản chí, ảnh hưởng đến hiệu quả điều trị cho con. Từ thực tế đó, các bác sĩ chuyên khoa Nhi - Sơ sinh BV Phụ sản TP Cần Thơ quyết tâm triển khai các kỹ thuật trong lĩnh vực này để tiếp nhận điều trị cho bệnh nhi. Từ năm 2019 đến thời điểm này, BV Phụ sản TP Cần Thơ đã phát hiện, điều trị 4 ca tăng sản, 11 ca suy giáp và theo dõi 2 bệnh về rối loạn chuyển hóa bẩm sinh”. TS Vũ Chí Dũng, Trưởng khoa Nội tiết - Chuyển hóa - Di truyền, BV Nhi Trung ương Hà Nội, là người luôn đồng hành, hỗ trợ Khoa Nhi - Sơ sinh BV Phụ sản trong việc chẩn đoán, theo dõi và can thiệp điều trị kịp thời cho các bé bị bệnh, để sống khỏe mạnh như bao trẻ khác.

Summary: TS Vũ Chí Dũng, Trưởng khoa Nội tiết - Chuyển hóa - Di truyền, BV Nhi Trung ương Hà Nội, là người luôn đồng hành, hỗ trợ Khoa Thơ đã phát hiện, điều trị 4 ca tăng sản, 11 ca suy giáp và theo dõi 2 bệnh về rối loạn chuyển hóa bẩm sinh”. Thơ, chia sẻ: “Các bác sĩ, điều dưỡng ở Khoa Nhi - Sơ sinh là những người mẹ đã hồi sinh cho bé một cuộc đời”. Cô T kể, khi có kết quả chẩn đoán bé có nguy cơ cao mắc bệnh tăng sản thượng thận bẩm sinh, gia đình không biết BS CKI Thạch Thị Ngọc Yến, Phó Trưởng Khoa Nhi - Sơ sinh, chia sẻ: “Với chỉ đạo và ủng hộ của Ban Giám đốc BV, đội ngũ Trước đây, khi có kết quả chẩn đoán bệnh, tất cả các trường hợp nghi ngờ nguy cơ cao đều phải chuyển lên tuyến trên. Trong 48 giờ đầu sau sinh, trẻ sinh ra tại viện đều được tầm soát bệnh lý bẩm sinh thông qua xét nghiệm lấy máu gót chân.

❖ **Văn bản gốc: bài báo về giáo dục.**

Thời điểm này, các trường học, thầy cô, phụ huynh và học sinh lớp 9 cơ bản đã chuẩn bị tốt nhất cho kỳ thi. Trương Gia Hân - học sinh lớp 9 Trường THCS An Thái, quận Bình Thủy, đã có 9 năm đạt học sinh Giỏi - cho biết, em được thầy cô ôn tập tại trường, đồng thời tự rèn luyện thêm, nên khá tự tin. Chị Nguyễn Kiều Dung, phường An Hòa, quận Ninh Kiều, có con gái đang học lớp 9, chia sẻ: “Thời điểm kỳ thi cận kề, tôi khuyên con nên học và nghỉ ngơi khoa học, ăn uống lành mạnh và ngủ đủ giấc”.

Phía các trường THCS trên địa bàn thành phố đã tổ chức củng cố kiến thức cho học sinh ngay từ đầu năm học; bắt đầu ôn tập từ đầu học kỳ II và sau đó tăng tốc ôn thi từ tháng 4. Ông Nguyễn Văn Chi, Trưởng Phòng Giáo dục và Đào tạo (GD&ĐT) huyện Thới Lai, cho biết từ đầu năm học ngành đã chỉ đạo các trường chọn giáo viên có năng lực chuyên môn tốt, nhiệt tình trong giảng dạy và có kinh nghiệm về công tác chủ nhiệm để chủ nhiệm và dạy bộ môn lớp 9. Đồng thời họp cha mẹ học sinh triển khai kế hoạch dạy học và bồi dưỡng học sinh cuối cấp; nhất là ba môn Ngữ văn, Toán, Ngoại ngữ.

Năm học 2020-2021, TP Cần Thơ có khoảng 15.360 học sinh lớp 9. Qua tư vấn, hướng dẫn của các trường THCS, đến nay có khoảng 13.370 học sinh lớp 9 có nguyện vọng thi vào 27 trường THPT công lập. Trong khi đó, tổng chỉ tiêu tuyển của các trường THPT công lập khoảng 11.000 học sinh. Như vậy sẽ có các em không trúng tuyển, với các hướng đi khác như vào học các trường ngoài công lập; Trung tâm Giáo dục nghề nghiệp - Giáo dục thường xuyên (GDNN-GDTX) quận, huyện; các trường nghề... Dù đây là thực tế mỗi kỳ tuyển sinh lớp 10 hằng năm, nhưng vẫn không tránh khỏi việc nhiều phụ huynh, học sinh lo lắng: nếu không vào lớp 10 trường công, thì lựa chọn trường ngoài công lập, các trung tâm và trường nghề... sao cho tốt nhất?

Tình hình những năm qua cho thấy, vẫn còn nhiều cánh cửa rộng mở với học sinh không trúng tuyển lớp 10 các trường THPT công lập. Theo thầy Nguyễn Văn Chi, học sinh có học lực trung bình - yếu cần mạnh dạn đăng ký học lớp 10 tại các trung tâm GDNN-GDTX hoặc các trường trung cấp nghề tại địa phương, theo hệ vừa học nghề vừa học văn hóa... Ngành đã chỉ đạo các trường THCS phải làm tốt công tác tư vấn hướng nghiệp, phân luồng để tất cả học sinh đều chọn lựa loại hình học tập cho phù hợp; tuyệt đối không để học sinh bỏ học sau tốt nghiệp THCS.

Một hướng đi rất khả thi mà học sinh không trúng tuyển lớp 10 có thể chọn là học song song văn hóa và học nghề tại các cơ sở giáo dục nghề nghiệp. TP Cần Thơ hiện có nhiều cơ sở giáo dục nghề nghiệp với các chương trình đào tạo đa dạng, phong phú, phát triển toàn diện kiến thức và kỹ năng cho học sinh. Người tốt nghiệp THCS đi học nghề được miễn 100% học phí (trừ các môn học văn hóa phải đóng học phí theo quy định); được liên thông lên các bậc học cao hơn. Đơn cử như theo Ban giám hiệu Trường Cao đẳng Kinh tế - Kỹ thuật Cần Thơ, bên cạnh bậc cao đẳng, trường xét điểm tổng kết của học bạ cuối năm lớp 9 cho các ngành nghề trung cấp. Thời gian học cao đẳng hoặc trung cấp là 3 năm. Trong đó học sinh học trung cấp vừa học văn hóa THPT, vừa học chuyên môn. Trường cũng đã áp dụng phân luồng nghề đào tạo kép theo tiêu chuẩn Nhật Bản, với mô hình KOSEN (mô hình 9+), cho phép người học tốt nghiệp có thể học liên thông lên cao đẳng, đại học. Chương trình đào tạo được xây dựng kết hợp giữa hai phần văn hóa và chuyên môn; với phần chuyên môn tăng dần theo từng năm. Việc đào tạo song hành này giúp học sinh được tiếp tục học liên thông từ trung cấp lên cao đẳng, rút ngắn thời gian đào tạo.

Con đường học tập của học sinh không trúng tuyển lớp 10 THPT công lập vẫn có nhiều hướng để đi đến thành công, nếu như có sự lựa chọn phù hợp năng lực, điều kiện gia đình.

Link: <https://baocantho.com.vn/nhieu-co-hoi-cho-hoc-sinh-khong-trung-tuyen-lop-10-thpt-cong-lap-a133825.html>

Đoạn văn tóm tắt mẫu:

Chị Nguyễn Kiều Dung, phường An Hòa, quận Ninh Kiều, có con gái đang học lớp 9, chia sẻ: "**Thời điểm kỳ thi cận kề, tôi khuyên con nên học và nghỉ ngơi khoa học, ăn uống lành mạnh và ngủ giấc**". Phía các trường THCS trên địa bàn thành phố tổ chức củng cố kiến thức cho học sinh ngay từ đầu năm học; bắt đầu ôn tập từ đầu học kỳ II và sau tăng tốc ôn thi từ tháng 4. Đồng thời hợp cha mẹ học sinh triển khai kế hoạch dạy học và bồi dưỡng học sinh cuối cấp; **nhất là ba môn Ngữ văn, Toán, Ngoại ngữ**. Trong học sinh học trung cấp vừa học văn hóa THPT, vừa học chuyên môn. Trường cũng áp dụng phân luồng nghề đào tạo kép theo tiêu chuẩn Nhật Bản, với mô hình KOSEN, **cho phép người học tốt nghiệp có thể học liên thông lên cao đẳng, đại học**.

Theo

website

Summarize:

https://smmry.com/4486213242#&SM_LENGTH=7

Văn bản tóm tắt với Spacy:

Học lớp 9, chia sẻ: **“Thời điểm kỳ thi cận kề, tôi khuyên con nên học và nghỉ ngơi khoa học, ăn uống lành mạnh và ngủ đủ giấc”**. Thời điểm này, các trường học, thầy cô, phụ huynh và học sinh lớp 9 cơ bản đã chuẩn bị tốt nhất cho kỳ thi. học sinh triển khai kế hoạch dạy học và bồi dưỡng học sinh cuối cấp; **nhất là ba môn Ngữ văn, Toán, Ngoại ngữ**. (mô hình 9+), **cho phép người học tốt nghiệp có thể học liên thông lên cao đẳng, đại học**. học sinh Giỏi - cho biết, em được thầy cô ôn tập tại trường, đồng thời tự rèn luyện thêm, nên khá tự tin. Việc đào tạo song hành này giúp học sinh được tiếp tục học liên thông từ trung cấp lên cao đẳng, rút ngắn thời gian đào tạo. học lớp 10 tại các trung tâm GDNN-GDTX hoặc các trường trung cấp nghề tại địa phương, theo hệ vừa học nghề.

So sánh với hai phiên bản tóm tắt bằng Gensim và Suny:

Spacy: Học lớp 9, chia sẻ: “Thời điểm kỳ thi cận kề, tôi khuyên con nên học và nghỉ ngơi khoa học, ăn uống lành mạnh và ngủ đủ giấc”. Thời điểm này, các trường học, thầy cô, phụ huynh và học sinh lớp 9 cơ bản đã chuẩn bị tốt nhất cho kỳ thi. học sinh triển khai kế hoạch dạy học và bồi dưỡng học sinh cuối cấp; nhất là ba môn Ngữ văn, Toán, Ngoại ngữ. (mô hình 9+), cho phép người học tốt nghiệp có thể học liên thông lên cao đẳng, đại học. học sinh Giỏi - cho biết, em được thầy cô ôn tập tại trường, đồng thời tự rèn luyện thêm, nên khá tự tin. Việc đào tạo song hành này giúp học sinh được tiếp tục học liên thông từ trung cấp lên cao đẳng, rút ngắn thời gian đào tạo. học lớp 10 tại các trung tâm GDNN-GDTX hoặc các trường trung cấp nghề tại địa phương, theo hệ vừa học nghề.

Gensim: Ông Nguyễn Văn Chi, Trưởng Phòng Giáo dục và Đào tạo (GD&ĐT) huyện Thới Lai, cho biết từ đầu năm học ngành đã chỉ đạo các trường chọn giáo viên có năng lực chuyên môn tốt, nhiệt tình trong giảng dạy và có kinh nghiệm về công tác chủ nhiệm để chủ nhiệm và dạy bộ môn lớp 9.

Theo thầy Nguyễn Văn Chi, học sinh có học lực trung bình - yếu cần mạnh dạn đăng ký học lớp 10 tại các trung tâm GDNN-GDTX hoặc các trường trung cấp nghề tại địa phương, theo hệ vừa học nghề vừa học văn hóa... Ngành đã chỉ đạo các trường THCS phải làm tốt công tác tư vấn hướng nghiệp, phân

luồng để tất cả học sinh đều chọn lựa loại hình học tập cho phù hợp; tuyệt đối không để học sinh bỏ học sau tốt nghiệp THCS.

Trường cũng đã áp dụng phân luồng nghề đào tạo kép theo tiêu chuẩn Nhật Bản, với mô hình KOSEN (mô hình 9+), cho phép người học tốt nghiệp có thể học liên thông lên cao đẳng, đại học.

Con đường học tập của học sinh không trùng tuyến lớp 10 THPT công lập vẫn có nhiều hướng để đi đến thành công, nếu như có sự lựa chọn phù hợp năng lực, điều kiện gia đình.

Summy: Học lớp 9, chia sẻ: “Thời điểm kỳ thi cận kề, tôi khuyên con nên học và nghỉ ngơi khoa học, ăn uống lành mạnh và ngủ đủ giấc”. Thời điểm này, các trường học, thầy cô, phụ huynh và học sinh lớp 9 cơ bản đã chuẩn bị tốt nhất cho kỳ thi. học sinh triển khai kế hoạch dạy học và bồi dưỡng học sinh cuối cấp; nhất là ba môn Ngữ văn, Toán, Ngoại ngữ. (mô hình 9+), cho phép người học tốt nghiệp có thể học liên thông lên cao đẳng, đại học. học sinh Giỏi - cho biết, em được thầy cô ôn tập tại trường, đồng thời tự rèn luyện thêm, nên khá tự tin. Việc đào tạo song hành này giúp học sinh được tiếp tục học liên thông từ trung cấp lên cao đẳng, rút ngắn thời gian đào tạo. học lớp 10 tại các trung tâm GDNN-GDTX hoặc các trường trung cấp nghề tại địa phương, theo hệ vừa học vừa học nghề.

⇒ Từ kết quả sinh tóm tắt cho thấy rằng, cả ba phương pháp đều cho đoạn văn tóm tắt khá tốt, có logic về mặt ngữ nghĩa, liên kết câu và không có lỗi sai về mặt từ vựng. Tuy nhiên, so về nội dung, Spacy lại cho nội dung tóm tắt đầy đủ và không mất mát nội dung của văn bản gốc.

Bài toán tóm tắt văn bản theo hướng trích xuất có thể cho kết quả khả quan, có khả năng tạo ra văn bản tóm tắt gần giống với cách con người thực hiện tóm tắt.

PHẦN KẾT LUẬN

1. Kết quả đạt được:

Xây dựng được phần mềm tóm tắt văn bản với chức năng cơ bản như:

- Tóm tắt văn bản từ khung giao diện được người dùng dán văn bản vào
- Tóm tắt văn bản từ link url của các trang internet nguồn mở
- So sánh với các phiên bản khác nhau của tóm tắt văn bản

Niên luận xây dựng tập dữ liệu cho tóm tắt văn bản tiếng Việt, sẵn sàng chia sẻ cho mục đích nghiên cứu và áp dụng trong tóm tắt văn bản tiếng Việt.

Niên luận cũng đã thử nghiệm mô hình đã xây dựng với dữ liệu tiếng Việt. Thử nghiệm với dữ liệu tiếng Việt về tin tức từ báo cần thơ (<https://baocantho.com.vn/>) và một số báo khác cho kết quả khả quan.

2. Hạn chế:

Mặc dù cơ bản đã hoàn thành được bài niên luận về tìm hiểu ngôn ngữ tự nhiên và ứng dụng vào tóm tắt văn bản nhưng vẫn còn một vài hạn chế nhất định:

- Chưa cho phép người dùng chọn số câu tóm tắt thích hợp.
- Không có chức năng upload văn bản bằng file (.txt, .doc) do chỉ upload được file tiếng anh, chưa xử lý được tiếng việt ở chức năng này

3. Hướng phát triển:

- ❖ Hoàn thiện thêm hai chức năng nói trên để ứng dụng được hoàn chỉnh và phong phú hơn về mặt chức năng
- ❖ Để tăng độ chính xác cho tóm tắt văn bản, điều kiện quan trọng là xây dựng tập dữ liệu đầu vào chất lượng hơn, thể hiện chính xác hơn độ tương quan, mối liên hệ giữa các từ, các token. Do đó, việc xây dựng tập dữ liệu lớn và phong phú về chủ đề, đa dạng về mặt từ vựng là rất cần thiết cho quá trình tóm tắt văn bản.
- ❖ Chức năng so sánh các phiên bản khác nhau: phát triển thành 3 khung kết quả đầu ra, mỗi khung chứa một bản tóm tắt văn bản của một phiên bản giúp người dùng dễ so sánh hơn.

CHƯƠNG 5 DEMO CHƯƠNG TRÌNH

Giao diện chính của chương trình:

The screenshot displays the main interface of the program. On the left is a sidebar menu with links: Home, URL, Comparer, and About. The main content area is titled 'CHƯƠNG TRÌNH TÓM TẮT VĂN BẢN' and contains a text input field with the following text: 'Trung tâm Sàng lọc - Chẩn đoán trước sinh và sơ sinh tại BV Phụ sản TP Cần Thơ được Tổng cục Dân số - Kế hoạch hóa gia đình giao trọng trách thực hiện Đề án nâng cao chất lượng dân số 12 tỉnh ĐBSCL. Theo đó, Trung tâm thực hiện tầm soát các bệnh, tật bẩm sinh giai đoạn trước sinh cho thai phụ đến khám thai tại BV trong 3 tháng đầu và 3 tháng giữa thai kỳ. Trong 48 giờ đầu sau sinh, trẻ sinh ra tại viện đều được tầm soát bệnh lý bẩm sinh thông qua xét nghiệm lấy máu gót chân. Trung tâm còn tiếp nhận mẫu từ các BV có chuyên khoa sản trên địa bàn thành phố và các tỉnh trong vùng gửi đến. Trước đây, khi có kết quả chẩn đoán bệnh, tật cả các trường hợp nghi ngờ nguy cơ cao đều phải chuyển lên tuyến trên. Các gia đình có con mắc bệnh phải thường xuyên đưa con lên TP Hồ Chí Minh tái khám, trong k'. Below the input field are four buttons: 'Reset', 'Summarize', 'Clear Result', and 'Main Points'. At the bottom, there is a 'Summary' section containing a detailed summary of the text input.

CHƯƠNG TRÌNH TÓM TẮT VĂN BẢN

Cho văn bản cần tóm tắt vào khung

Trung tâm Sàng lọc - Chẩn đoán trước sinh và sơ sinh tại BV Phụ sản TP Cần Thơ được Tổng cục Dân số - Kế hoạch hóa gia đình giao trọng trách thực hiện Đề án nâng cao chất lượng dân số 12 tỉnh ĐBSCL. Theo đó, Trung tâm thực hiện tầm soát các bệnh, tật bẩm sinh giai đoạn trước sinh cho thai phụ đến khám thai tại BV trong 3 tháng đầu và 3 tháng giữa thai kỳ. Trong 48 giờ đầu sau sinh, trẻ sinh ra tại viện đều được tầm soát bệnh lý bẩm sinh thông qua xét nghiệm lấy máu gót chân. Trung tâm còn tiếp nhận mẫu từ các BV có chuyên khoa sản trên địa bàn thành phố và các tỉnh trong vùng gửi đến. Trước đây, khi có kết quả chẩn đoán bệnh, tật cả các trường hợp nghi ngờ nguy cơ cao đều phải chuyển lên tuyến trên. Các gia đình có con mắc bệnh phải thường xuyên đưa con lên TP Hồ Chí Minh tái khám, trong k

Reset Summarize

Clear Result Main Points

Summary: TS Vũ Chí Dũng, Trưởng khoa Nội tiết - Chuyển hóa - Di truyền, BV Nhi Tr ung ương Hà Nội, là người luôn đồng hành, hỗ trợ Khoa Tho, chia sẻ: "Các bác sĩ, điều dưỡng ở Khoa Nhi - Sơ sinh là những người mẹ đã hồi sinh cho bé một cuộc đ ời". Tho đã phát hiện, điều trị 4 ca tăng sản, 11 ca suy giáp và theo dõi 2 bệnh về rối loạn chuyển hóa bẩm sinh". Theo đó, Trung tâm thực hiện tầm soát các bện h, tật bẩm sinh giai đoạn trước sinh cho thai phụ đến khám thai tại Trước đây, k hi có kết quả chẩn đoán bệnh, tật cả các trường hợp nghi ngờ nguy cơ cao đều phả i chuyển lên tuyến trên. Có T kể, khi có kết quả chẩn đoán bé có nguy cơ cao mắc bệnh tăng sản thượng thận bẩm sinh, gia đình không biết BS CKI Thạch Thị Ngọc Y ến, Phó Trưởng Khoa Nhi - Sơ sinh, chia sẻ: "Với chỉ đạo và ủng hộ của Ban Giám đốc BV

Hình 6. Giao diện chính của chương trình

Giao diện chính của chương trình gồm các phần như sau:

- Bảng phân chia chức năng: Tóm tắt bằng văn bản, link url, so sánh các phương pháp,...
- Khung để nhập văn bản đầu vào
- Các nút Bottom để gọi hàm tóm tắt văn bản
- Khung trả kết quả đầu ra là văn bản tóm tắt

Các chức năng chính:

- Reset: Cho phép người dùng thực thi lại chương trình sau khi đã làm một văn bản tóm tắt. Nút reset sẽ tự động xóa phần khung nhập văn bản để người dùng có thể nhập văn bản mới vào.
- Summarize: Nút thực thi lệnh tóm tắt văn bản và hiển thị văn bản được tóm tắt ở khung phía dưới.
- Clear Result: Thực thi xóa khung văn bản tóm tắt – khung phía dưới

Chức năng tóm tắt thông qua link url từ internet

The screenshot shows a web application with a sidebar on the left containing links: Home, URL (selected), Comparer, and About. The main area is titled 'URL' and contains a text input field with the URL '-bao-ho-giup-ha-nhiệt-cho-cac-y-bac-si-a133867.html'. Below the input field are four buttons: 'Reset', 'Get Text', 'Clear Result', and 'Summarize'. The 'Get Text' button has been clicked, and the original text from the URL is displayed in a text area. Below this, a 'Summary' section shows a summarized version of the text.

Home
URL
Comparer
About

URL

Cho link bài báo cần tóm tắt

-bao-ho-giup-ha-nhiệt-cho-cac-y-bac-si-a133867.html

Reset

Get Text

Clear Result

Summarize

triệu USD tiền ăn cấp của tội phạm mạng châu Á Những tính năng mới trong iOS 14.6 và iPadOS 14.6 4 cách đơn giản để in danh sách tập tin của thư mục 3 tiện ích diệt virus mã độc đáng tin cậy Tổng biên tập: TRƯƠNG VĂN CHUYỀN Giấy phép số 320/GP-BTTTT, do Bộ Thông Tin và Truyền Thông cấp ngày 04-07-2017 24 Trần Văn Hoài, P.Xuân Khánh, Q.Ninh Kiều, Tp.Cần Thơ - Điện thoại: (0292) 3830098 - Fax: (0292) 3830561

Email: toasoan@baocantho.com.vn Liên hệ giao dịch quảng cáo, rao vặt: quangcao@baocantho.com.vn Ghi rõ nguồn "Báo điện tử Cần Thơ" khi phát hành lại thông tin từ website này Ph

át triển bởi:

Summary: Do bộ đồ này rất bí nên chúng ta chỉ có thể tạo ra không khí đối lưu, giúp "hạ nhiệt" cho các y, bác sĩ. Để bảo đảm sức khỏe cho nhân viên y tế ở các điểm nóng của dịch, các chuyên gia cho rằng người trực tiếp tại tâm Bắc Giang để dùng cho nhân viên y tế tuyến đầu và sẽ sớm được phổ biến cho tất cả các nhân viên y tế. Chúng ta có thể tăng giảm tốc độ, vì khi nghỉ ngơi có thể dùng quạt tốc độ nhẹ hơn", TS. dịch không nên làm việc trong điều kiện phải mang bảo hộ kín liên tục nhiều giờ mà cần được thay ca sau 2-3 tiếng, tránh để kiệt sức. Đoàn Ngọc Hải, Viện trưởng Viện Sức khỏe Nghề nghiệp và Môi trường (Bộ Y tế) cho biết đã có giải pháp chống nóng cho nhân viên y tế khi phải mặc bộ đồ bảo hộ để chống d

Hình 7. Giao diện chức năng thông qua url

Các chức năng:

- Reset: Cho phép người dùng thực thi lại chương trình sau khi đã làm một văn bản tóm tắt bằng link url. Nút reset sẽ tự động xóa phần khung nhập link url để người dùng có thể nhập một link url mới vào.
- Get Text: Thực thi lấy văn bản về (văn bản gốc) cho người dùng xem trước đúng link url chưa
- Summarize: Nút thực thi lệnh tóm tắt văn bản và hiển thị văn bản được tóm tắt ở khung phía dưới.
- Clear Result: Thực thi xóa khung văn bản tóm tắt – khung phía dưới

Chức năng so sánh các phương pháp khác nhau (Spacy, Gensum, Sumy):

Home
URL
Comparer
About

Các phương pháp

Cho văn bản cần tóm tắt vào khung

Trung tâm Sàng lọc - Chẩn đoán trước sinh và sơ sinh tại BV Phụ sản TP Cần Thơ được Tổng cục Dân số - Kế hoạch hóa gia đình giao trọng trách thực hiện Đề án nâng cao chất lượng dân số 12 tỉnh ĐBSCL. Theo đó, Trung tâm thực hiện tầm soát các bệnh, tật bẩm sinh giai đoạn trước sinh cho thai phụ đến khám thai tại BV trong 3 tháng đầu và 3 tháng giữa thai kỳ. Trong 48 giờ đầu sau sinh, trẻ sinh ra tại viện đều được tầm soát bệnh lý bẩm sinh thông qua xét nghiệm lấy máu gót chân. Trung tâm còn tiếp nhận mẫu từ các BV có chuyên khoa sản trên địa bàn thành phố và các tỉnh trong vùng gửi đến. Trước đây, khi có kết quả chẩn đoán bệnh, tất cả các trường hợp nghi ngờ nguy cơ cao đều phải chuyển lên tuyến trên. Các gia đình có con mắc bệnh phải thường xuyên đưa con lên TP Hồ Chí Minh tái khám, trong k

Reset SpaCy Main Points

Clear Result Gensim Sumy

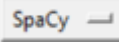
SpaCy

c chẩn đoán, theo dõi và can thiệp điều trị kịp thời cho các bé bị bệnh, để sống khỏe mạnh như bao trẻ khác.

Sumy Summary: TS Vũ Chí Dũng, Trưởng khoa Nội tiết - Chuyển hóa - Di truyền, BV Nhi Trung ương Hà Nội, là người luôn đồng hành, hỗ trợ Khoa Thờ, chia sẻ: "Các bác sĩ, điều dưỡng ở Khoa Nhi - Sơ sinh là những người mẹ đã hồi sinh cho bé một cuộc đời". Thờ đã phát hiện, điều trị 4 ca tăng sản, 11 ca suy giáp và theo dõi 2 bệnh về rối loạn chuyển hóa bẩm sinh". Theo đó, Trung tâm thực hiện tầm soát các bệnh, tật bẩm sinh giai đoạn trước sinh cho thai phụ đến khám thai tại Trước đây, khi có kết quả chẩn đoán bệnh, tất cả các trường hợp nghi ngờ nguy cơ cao đều phải chuyển lên tuyến trên. Có T kể, khi có kết quả chẩn đoán bé có nguy cơ cao mắc bệnh tăng sản thượng thận bẩm sinh, gia đình không biết BS CKI Thạch Thị Ngọc Yến, Phó Trưởng Khoa Nhi - Sơ sinh, chia sẻ: "Với chỉ đạo và ủng hộ của Ban Giám đốc BV, đội ngũ

Hình 8. Giao diện so sánh các phương pháp

Các nút chức năng:

- Spacy: Thực hiện lệnh tóm tắt văn bản với Spacy
- Gensim: Thực hiện lệnh tóm tắt văn bản với Gensim
- Suny: Thực hiện lệnh tóm tắt văn bản với Suny
- Reset: Cho phép người dùng thực thi lại chương trình sau khi đã làm một văn bản tóm tắt. Nút reset sẽ tự động xóa phần khung nhập văn bản để người dùng có thể nhập văn bản mới vào.
- Clear Result: Xóa khung văn bản tóm tắt
- Chức năng mở rộng : Khi người dùng không muốn so sánh các phiên bản và cũng không thích tóm tắt bằng phương pháp Spacy ở hai giao diện chức năng trên, người dùng có thể chọn duy nhất một trong ba phiên bản ở chức năng mở rộng này. Khi đó hệ thống chỉ tóm tắt với phiên bản được người dùng lựa chọn.

TÀI LIỆU THAM KHẢO

- [1] <https://www.machinelearningplus.com/nlp/text-summarization-approaches-nlp-example/>
- [2] https://www.tutorialspoint.com/python_text_processing/index.htm
- [3] <https://stackabuse.com/text-summarization-with-nltk-in-python>
- [4] <https://www.analyticsvidhya.com/blog/2019/06/comprehensive-guide-text-summarization-using-deep-learning-python/>
- [5] ["Introducing spaCy"](#). explosion.ai. Truy cập ngày 18 tháng 12 năm 2016.
- [6] [^ "Release v3.0.0: Transformer-based pipelines, new training system, project templates, custom models, improved component API, type hints & lots more · explosion/spaCy"](#). GitHub (bằng tiếng Anh). Truy cập ngày 2 tháng 2 năm 2021.
- [7] [^ "Models & Languages | spaCy Usage Documentation"](#). spacy.io. Truy cập ngày 10 tháng 3 năm 2020.
- [8] [Deep learning with word2vec and Gensim](#)
- [9] Řehůřek, Radim (2011). ["Scalability of Semantic Analysis in Natural Language Processing"](#) (PDF). Retrieved 27 January 2015. my open-source gensim software package that accompanies this thesis