TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC DÂN

VIỆN CÔNG NGHỆ THÔNG TIN VÀ KINH TẾ SỐ

BÀI THU HOẠCH NHÓM 1:

CÁC CÔNG NGHỆ HIỆN ĐẠI TRONG CÔNG NGHỆ THÔNG TIN

ĐỀ TÀI:

TÌM HIỂU VỀ TEXT MINING (CÀI ĐẶT VÀ CHẠY THỬ CÔNG CỤ)

Thành viên: Vũ Hoàng An 11218384
Bùi Quốc Anh 11210275
Nguyễn Thị Vân Anh 11218386
Trần Hoàng Kim Anh 11218387
Trần Tôn Anh 11218388

Lớp học phần: Các công nghệ hiện đại trong công nghệ thông tin_01

Giảng viên: TS. Nguyễn Trung Tuấn

Hà Nội 2023

TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC DÂN

VIỆN CÔNG NGHỆ THÔNG TIN VÀ KINH TẾ SỐ

BÀI THU HOẠCH NHÓM 1:

CÁC CÔNG NGHỆ HIỆN ĐẠI TRONG CÔNG NGHỆ THÔNG TIN

ĐỀ TÀI:

TÌM HIỂU VỀ TEXT MINING (CÀI ĐẶT VÀ CHẠY THỬ CÔNG CỤ)

Thành viên:	Vũ Hoàng An	11218384
	Bùi Quốc Anh	11210275
	Nguyễn Thị Vân Anh	11218386
	Trần Hoàng Kim Anh	11218387
	Trần Tôn Anh	11218388

Lớp học phần: Các công nghệ hiện đại trong công nghệ thông tin_01

Giảng viên: TS. Nguyễn Trung Tuấn

Hà Nội 2023

MỤC LỤC

MỞ ĐẦU 4

DANH MỤC THUẬT NGỮ 5

DANH MỤC HÌNH MINH HỌA, BẢNG BIỂU 7

TỔNG QUAN ĐỀ TÀI 8

1. Đề tài nghiên cứu 8
2. Phạm vi nghiên cứu 8
3. Đối tượng nghiên cứu 8
4. Mục đích nghiên cứu 9

NỘI DUNG 10

1. Cơ sở lý luận 10
   1.1. Khái niệm Text Mining 10
   1.2. Đặc điểm của Text Mining 13
   1.3. Cách thức Text Mining hoạt động 15
2. Ứng dụng của Text Mining 20
3. Một số kỹ thuật được sử dụng trong Text Mining 22
   3.1. Tổng quan các kỹ thuật 22
   3.2. Text Summarization – Tóm tắt văn bản 24
4. So sánh với các công nghệ khác 26
   4.1. Xử lý ngôn ngữ tự nhiên (Natural Language Processing) 26
   4.2. Khai thác dữ liệu (Data Mining) 29

CÀI ĐẶT VÀ ĐÁNH GIÁ 32

1. Dữ liệu thử nghiệm 32
2. Giới thiệu mô hình sử dụng 32

3. Cài đặt chương trình 33
3.1. Cài đặt và import thư viện 33
3.2. Tải bộ dữ liệu CNN/Daily Mail 33
3.3. Tiền xử lý dữ liệu 34
3.4. Xây dựng hàm đánh giá (Evaluate) 35
3.5. Tinh chỉnh mô hình 37
4. Đánh giá kết quả 38
KẾT LUẬN 40
TÀI LIỆU THAM KHẢO 41

MỞ ĐẦU

Trong thời đại kỹ thuật số hiện nay, dữ liệu văn bản tràn ngập từ mọi nguồn, từ các bài viết trên mạng xã hội đến các tài liệu kỹ thuật và email chúng ta nhận và gửi đi hàng ngày. Số lượng thông tin văn bản không ngừng gia tăng theo thời gian, đặt ra một thách thức cực lớn trong việc xử lý, phân tích và tận dụng tối đa tiềm tăng của nguồn dữ liệu này. Chính vấn đề này, Text Mining – một công nghệ nghiên cứu và xử lý chuyên về dữ liệu văn bản, ra đời.

Text Mining là quá trình sử dụng các phương pháp và công nghệ máy tính để trích xuất thông tin có giá trị từ dữ liệu văn bản. Text Mining không chỉ giúp tự động hóa quá trình tổng hợp và phân loại dữ liệu mà còn giúp hiểu sâu về ngữ nghĩa, ngôn ngữ và ngữ cảnh trong các tài liệu văn bản. Công nghệ này đã có sự phát triển đáng kể trong những năm gần đây, ngày càng trở thành một công cụ thiết yếu trong phân tích dữ liệu, nghiên cứu thị trường, quản lý văn bản và nhiều lĩnh vực khác.

Bài thu hoạch này của nhóm sẽ đi sâu vào chủ đề tìm hiểu Text Mining, cài đặt và chạy thử chương trình. Nhóm sẽ khám phá các khía cạnh quan trọng của công nghệ này, bao gồm khái niệm, cách thức hoạt động, ứng dụng thực tế,...; và cài đặt, chạy thử công nghệ này.

Nhóm 1 kỳ vọng bài thu hoạch sẽ cung cấp những tri thức cần thiết về Text Mining, cách thức Text Mining hoạt động và cải thiện quá trình ra quyết định, phân tích thông tin và đưa ra dự đoán trong nhiều lĩnh vực, từ ứng dụng trong các doanh nghiệp đến nghiên cứu khoa học. Từ những lý thuyết căn bản đó, nhóm sẽ cài đặt và chạy thử Text Mining, cụ thể là kỹ thuật Text Summarization – Tóm tắt văn bản trong công nghệ khai phá văn bản.

DANH MỤC THUẬT NGỮ

	Tiếng Anh	Tiếng Việt
A	Abstractive Summarization	Tóm tắt tóm lược ý
	AutoEncoder	Bộ tự mã hóa
	Automatic Machine Translation	Dịch máy tự động
B	Bag of Words	Mô hình túi từ
	Business Intelligence	Kinh doanh thông minh
C	Convolutional Neural Network	Mạng nơ-ron tích chập
	Customer Care Service	Dịch vụ chăm sóc khách hàng
D	Data Mining	Khai phá dữ liệu
	Deep Learning	Học sâu
	Dictionary-based	Dựa từ điển
	Document-Term-Matrix	Ma trận tài liệu-thuật ngữ
	Dynamic Padding	Bộ đệm động
E	Extractive Summarization	Tóm tắt trích chọn
F	Feature Selection	Lựa chọn tính năng
	Filtering	Lọc từ
	Fraud Detection	Phát hiện gian lận
I	Information Retrieval	Truy xuất thông tin
L	Lemmatization	Bổ đề ngôn ngữ
	Linguistic Processing	Xử lý ngôn ngữ học
	Logistic Regression	Hồi quy Logistic
	Longest Common Subsequence	Chuỗi con chung dài nhất
M	Machine Learning	Học máy
	Management Information System	Hệ thống thông tin quản lý
N	Naive Bayes Classification	Phân loại Naive Bayes
	Named-entity Recognition	Nhận diện thực thể được đặt tên
	Natural Language Generation	Sinh ngôn ngữ tự nhiên
	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
	Nonword	Từ không quan trọng

P	Part-of-speech Tagging	Gắn thẻ loại từ
	Pattern Analysis	Phân tích mẫu
	Pre-defined dictionary	Từ điển dựng sẵn
R	Random Forests	Rừng ngẫu nhiên
	Recurrent Neural Network	Mạng nơ-ron hồi quy
	Risk Management	Quản lý rủi ro
S	Sentiment Analysis	Phân tích cảm xúc
	Sequence and Path Analysis	Phân tích trình tự và đường dẫn
	Social Media Analysis	Phân tích truyền thông xã hội
	Stemming	Thu gọn từ
	Stopword	Từ dừng
	Supervised Learning	Học có giám sát
	Support Vector Machines	Máy vector hỗ trợ
T	Term Frequency-Inverse Document Frequency	Tần số từ - Độ quan trọng nghịch đảo của tài liệu
	Text Categorization	Phân loại văn bản
	Text Classification	Phân loại văn bản
	Text Cleanup	Làm sạch văn bản
	Text Clustering	Phân cụm văn bản
	Text Generation	Tạo lập văn bản
	Text Mining	Khai phá văn bản
	Text Pre-processing	Tiền xử lý văn bản
	Text Summarization	Tóm tắt văn bản
	Text Transformation	Biến đổi văn bản
	Tokenization	Mã hóa kỹ thuật số
U	Unsupervised Learning	Học không giám sát
V	Vector Space	Mô hình không gian vector
W	White Space Separation	Loại bỏ khoảng trắng
	Word Sense Disambigous	Định hướng nghĩa từ

DANH MỤC HÌNH MINH HỌA, BẢNG BIỂU

Danh mục hình minh họa

	Tên	Trang
H1.1	Các loại dữ liệu	10
H1.2	Quy trình xử lý văn bản Text Mining	12
H1.3	Quy trình phân tích Text Mining	13
H1.4	Quá trình hoạt động Text Mining	15
H1.5	Quy trình xử lý Text Mining	18
H2.1	Ứng dụng của Text Mining	21
H3.1	Tóm tắt trích chọn và Tóm tắt tóm lược ý	25
H4.1	Nhận dạng thực thể được đặt tên	27

Danh mục bảng biểu

	Tên	Trang
B3.1	Các công nghệ trong Text Mining	23
B4.1	Khác nhau giữa Narutal Language Processing và Text Mining	28
B4.2	Khác nhau giữa Data Mining và Text Mining	30

TỔNG QUAN ĐỀ TÀI

1. Đề tài nghiên cứu

Thế giới ngày càng phát triển, lượng thông tin được sinh ra và trao đổi ngày càng tăng lên, tạo ra một thách thức lớn trong vấn đề xử lý thông tin. Các thông tin được lưu trữ dưới nhiều hình thức khác nhau: có thể là các bài báo, tài liệu, thống kê hay các loại dữ liệu văn bản. Điều này đòi hỏi các tổ chức và cá nhân phải có giải pháp hiệu quả cho vấn đề quản lý và xử lý loại thông tin đa dạng này.

Không thể phủ nhận rằng lượng thông tin được chứa trong các loại văn bản không ngừng gia tăng và chiếm lượng lớn trong hệ thống thông tin, vì thế ta cần đặt ra vấn đề về việc xử lý cũng như khai thác nguồn thông tin gần như vô tận này. Và trong số các công nghệ xử lý thông tin dưới dạng văn bản, Text Mining có lẽ là "ứng cử viên sáng giá nhất" cho việc khám phá, truy xuất, trích xuất và phân tích thông tin từ các tài liệu văn bản.

2. Phạm vi nghiên cứu

Bài thu hoạch *Tìm hiểu về Text Mining* của nhóm 1 sẽ nghiên cứu về khái niệm, đặc điểm, ứng dụng của Text Mining, cách thức Text Mining hoạt động, các kỹ thuật nổi bật trong Text Mining và so sánh Text Mining với các phương pháp, kỹ thuật xử lý văn bản tương tự và đưa ra nhận xét.

Dựa trên những cơ sở lý thuyết đó, nhóm em đã tìm hiểu, cài đặt, và chạy thử một kỹ thuật khá nổi bật và phổ biến trong Text Mining là Text Summarization – kỹ thuật tóm tắt nội dung văn bản từ một đoạn văn bản dài cho trước, với đầu ra là những ý tóm tắt từ đoạn văn bản đó.

3. Đối tượng nghiên cứu

Bài thu hoạch sẽ tập trung vào nghiên cứu Text Mining, cài đặt và chạy thử. Nhờ sự phát triển của công nghệ và các cuộc cách mạng công nghiệp, Text Mining càng trở thành vấn đề đáng được nghiên cứu và quan tâm. Text Mining là quá trình kết hợp giữa các lĩnh vực khoa học máy tính, ngôn ngữ học và thống kê để phân tích và hiểu sâu hơn về thông tin chứa trong các tài liệu văn bản.

Bên cạnh nghiên cứu lý thuyết về Text Mining, nhóm đã cài đặt và chạy thử Text Summarization – một kỹ thuật trong Text Mining và đã có kết luận cùng đánh giá kết quả chạy thử chương trình.

4. Mục đích nghiên cứu

Text Mining là một kỹ thuật xử lý văn bản đầy tiềm năng, được áp dụng rộng rãi trong nhiều lĩnh vực. Text Mining có thể được tận dụng để trích xuất thông tin quan trọng từ các nguồn văn bản đa dạng, phân loại và phân đoạn văn bản để quản lý thông tin dễ dàng hơn, khám phá tri thức ẩn từ dữ liệu văn bản, dự đoán và phân tích xu hướng tương lai, nghiên cứu thị trường, quản lý tri thức tổ chức, và hỗ trợ quá trình ra quyết định. Đồng thời, việc nghiên cứu và chạy thử công nghệ Text Mining giúp chúng ta hiểu rõ hơn về cách xử lý thông tin từ các nguồn văn bản đa dạng, và tiến bộ trong việc tạo ra các phương pháp và công cụ mới. Với khả năng ứng dụng rộng rãi và tiềm năng không giới hạn, Text Mining đóng vai trò quan trọng trong việc biến thông tin từ các tài liệu văn bản thành tri thức và giá trị thực tiễn trong nhiều lĩnh vực khác nhau.

Không những thế, qua quá trình làm việc nhóm, đây còn là cơ hội để các thành viên trong nhóm nâng cao khả năng làm việc nhóm, khả năng giao tiếp và khả năng tiếp cận, giải quyết vấn đề một các khoa học, hợp lý, làm bước đệm vững chắc cho sự phát triển của từng cá nhân trong tương lai.

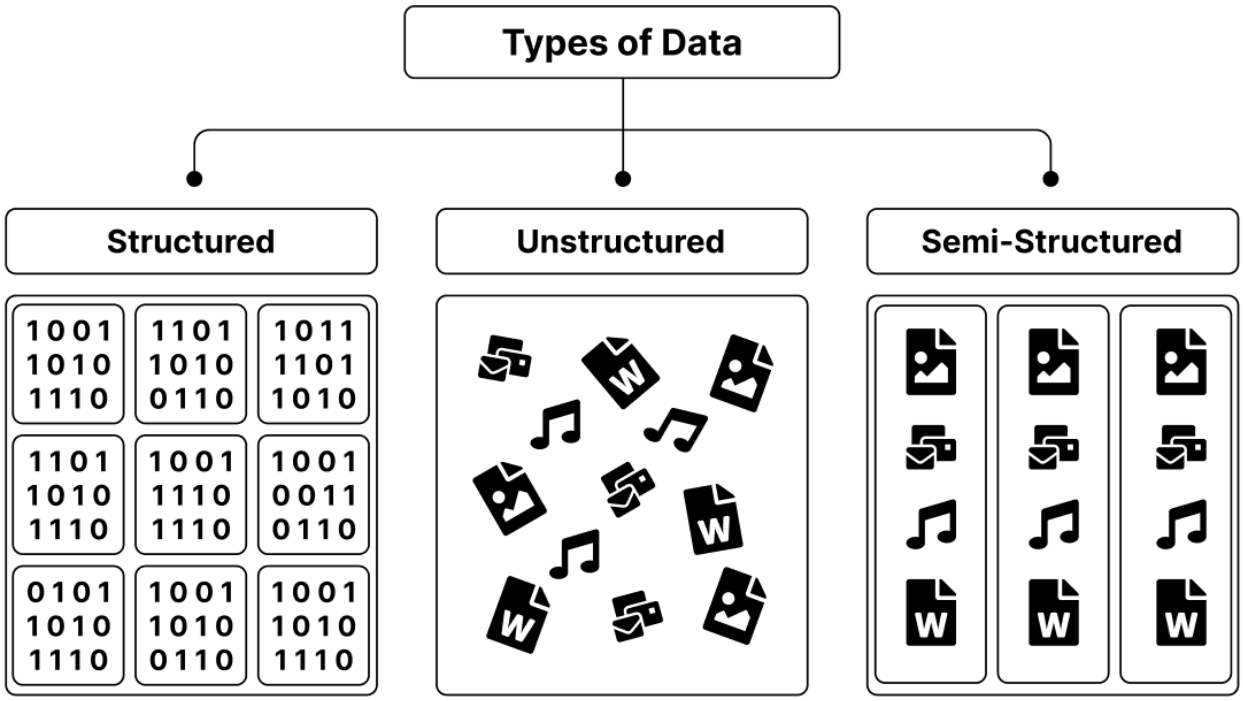
NỘI DUNG

1. Cơ sở lý luận

1.1. Khái niệm Text Mining

Text Mining là một thành phần của kỹ thuật khai thác dữ liệu, xử lý đặc biệt hiệu quả với dữ liệu văn bản phi cấu trúc. Text Mining sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên để trích xuất thông tin hữu ích và tri thức từ lượng lớn dữ liệu văn bản phi cấu trúc. Text Mining còn được sử dụng như một bước tiền xử lý cho quá trình khai thác dữ liệu hay như một quy trình độc lập cho các bài toán cụ thể.

Dữ liệu phi cấu trúc được hiểu là các thông tin ở dưới nhiều dạng khác nhau, chúng không tuân theo bất cứ quy ước nguyên mẫu dữ liệu nào và rất khó để lưu trữ và quản lý trong một cơ sở dữ liệu quan hệ chính thống. Một trong những loại dữ liệu phi cấu trúc phổ biến nhất là văn bản. Văn bản phi cấu trúc được tạo ra và thu thập dưới muôn hình vạn trạng, bao gồm các tài liệu văn bản, thư điện tử, tin nhắn, bản trình chiếu, phiếu trả lời khảo sát, bản ghi các cuộc gọi, bài đăng từ các blog và mạng xã hội,...



H1.1. Các loại dữ liệu

Text Mining được ứng dụng rộng rãi cho nhiều lĩnh vực và đã trở thành một công cụ quan trọng cho quá trình xử lý văn bản, trích xuất thông tin hữu ích từ dữ liệu văn bản và từ đó đưa ra phân tích và các quyết định. Có nhiều kỹ thuật và công cụ khác nhau để khai thác tri thức từ văn bản cũng như phân tích và tìm ra dữ liệu quan trọng cho quá trình dự đoán và ra quyết định. Việc lựa chọn đúng kỹ thuật và thuật toán trong Text Mining cho từng bài toán cụ thể góp phần lớn trong việc cải thiện tốc độ và cho ra kết quả chính xác trong từng trường hợp.

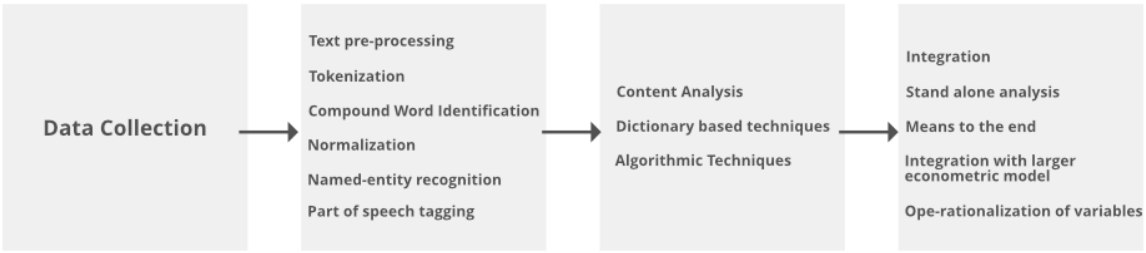
Hiện nay, hầu hết các công ty, tổ chức và dự án đều lưu trữ thông tin của họ dưới dạng số hóa như các tập dữ liệu lớn, các tài liệu số, thư viện số, các cơ sở dữ liệu, kho dữ liệu, bảng tính, website,... Việc xác định các mẫu và xu hướng thích hợp để trích xuất kiến thức từ lượng lớn dữ liệu này là một nhiệm vụ khó khăn. Text Mining là một phần thuộc công nghệ khai phá dữ liệu, nhằm trích xuất thông tin văn bản có giá trị từ kho dữ liệu văn bản. Text Mining là một công nghệ sử dụng các kỹ thuật tiên tiến như trí tuệ nhân tạo, thống kê số liệu, học máy,...

Quy trình xử lý văn bản của Text Mining thường bao gồm các bước:

- Thu thập dữ liệu văn bản phi cấu trúc từ nhiều nguồn khác nhau, có sẵn trong các định dạng tài liệu khác nhau như văn bản thô, trang web, tài liệu PDF,...
- Tiền xử lý và làm sạch dữ liệu: Các nhiệm vụ tiền xử lý và làm sạch dữ liệu được thực hiện để phát hiện và loại bỏ sự không đồng nhất trong dữ liệu. Quá trình làm sạch dữ liệu đảm bảo thu thập được dữ liệu văn bản thực sự và bao gồm một số phương pháp như:
    + Text Cleanup - Làm sạch văn bản: gồm nhiều nhiệm vụ khác nhau như loại bỏ quảng cáo trên các trang web phục vụ việc lấy dữ liệu văn bản.
    + Tokenization - Mã hóa kỹ thuật số: tạo ra phân đoạn của câu thành các từ bằng cách loại bỏ các khoảng trắng, dấu câu,...
    + Filtering - Lọc từ: loại bỏ các từ không liên quan đến nội dung như các mạo từ, liên từ, giới từ; thậm chí loại bỏ cả các từ lặp lại với tần suất cao.
    + Stemming - Thu gọn từ: loại bỏ các tiền tố, hậu tố của từ để đưa các từ về dạng cơ sở, dựa theo các quy tắc rút gọn. Kết quả của Stemming có thể là một từ không tồn tại hay không có nghĩa. Stemming thường xử lý nhanh và đơn giản, tuy nhiên kết quả lại tồn tại tỉ lệ sai lệch nhất định.

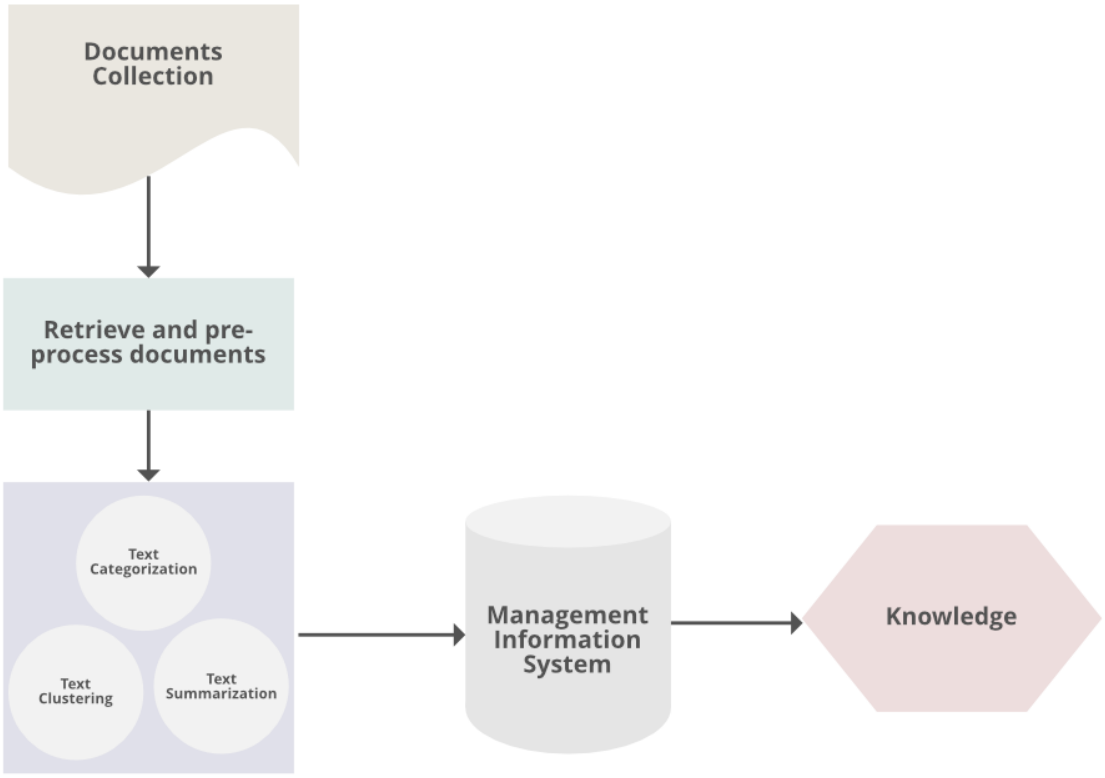
+ Lemmatization - Bổ đề ngôn ngữ: sử dụng kiến thức từ ngữ và từ điển để chuyển một từ về dạng cơ sở (lemma) có ý nghĩa trong ngôn ngữ. Kết quả trả về luôn là một từ có tồn tại và có ý nghĩa. Lemmatization chính xác hơn trong việc xác định từ gốc của một từ, nhưng lại đòi hỏi nhiều thời gian và tài nguyên hơn so với Stemming.
+ Linguistic Processing - Xử lý ngôn ngữ: gắn thẻ loại từ (Part-of-speech tagging), định hướng nghĩa từ (Word Sense Disambigous), xử lý cấu trúc ngữ nghĩa của từ.

- Chuyển đổi văn bản:
  + Xử lý và kiểm soát dữ liệu: Các nhiệm vụ xử lý và kiểm soát được áp dụng để xem xét và làm sạch sâu hơn tập dữ liệu.
  + Phân tích mẫu (Pattern Analysis): Quá trình này thường được thực hiện trong Hệ thống thông tin quản lý (Management Information System).
  + Thông tin được xử lý trong các bước trước được sử dụng để trích xuất thông tin quan trọng và thích hợp cho quy trình ra quyết định mạnh mẽ và thuận tiện cũng như phân tích xu hướng.



H1.2. Quy trình xử lý văn bản Text Mining

- Quy trình phân tích Text Mining:
  + Text Summarization – Tóm tắt văn bản: Tự động trích xuất một phần nội dung của văn bản và thể hiện toàn bộ nội dung của nó.
  + Text Categorization – Phân loại văn bản: Gán loại cho văn bản từ trong các loại đã được định sẵn bởi người dùng.
  + Text Clustering – Phân cụm văn bản: Phân đoạn văn bản thành một số nhóm dựa theo sự tương quan cơ bản.



H1.3. Quy trình phân tích Text Mining

1.2. Đặc điểm của Text Mining

1.2.1. Một số vấn đề khi ứng dụng Text Mining:

Quá trình triển khai Text Mining trong thực tế đôi khi gặp phải một số vấn đề, và dưới đây là những vấn đề quan trọng mà chúng ta cần quan tâm:

- Hiệu quả và hiệu suất của vấn đề đưa ra quyết định trong quá trình triển khai Text Mining trong thực tế, vấn đề này có thể bị ảnh hưởng từ cả phía máy tính và con người khi xử lý dữ liệu, nhất là đối với các tập dữ liệu lớn.
- Có khả năng xảy ra vấn đề không chắc chắn ở giai đoạn trung gian của quá trình triển khai Text Mining. Trong giai đoạn tiền xử lý, các quy tắc và hướng dẫn khác nhau được mô tả để chuẩn hóa văn bản, giúp quá trình khai thác văn bản trở nên hiệu quả. Vì thế, trước khi áp dụng phân tích mẫu vào tài liệu, cần phải thay đổi dữ liệu phi cấu trúc thành một loại dữ liệu có cấu trúc trung bình.
- Đôi khi thông điệp gốc hay ý nghĩa gốc của đầu vào có thể bị thay đổi trong quá trình khai thác văn bản.

- Hiện nay có khá nhiều thuật toán và kỹ thuật hỗ trợ văn bản đa ngôn ngữ. Điều này có thể tạo ra sự mơ hồ về ý nghĩa chính xác của văn bản, dẫn đến các kết quả "giả" sau quá trình chương trình hoạt động.
- Các từ đồng nghĩa, đa nghĩa và trái nghĩa khiến cho các công cụ Text Mining gặp khó khăn trong quá trình phân loại các văn bản .

1.2.2. Ưu điểm của Text Mining:

Như bất kỳ công nghệ nào, Text Mining cũng mang đến một loạt ưu điểm khi được áp dụng trong thực tế. Dưới đây là một số điểm mạnh mà ta có thể tận dụng khi sử dụng Text Mining:

- Xử lý lượng dữ liệu lớn: Text Mining cho phép các cá nhân và tổ chức trích xuất thông tin từ lượng lớn dữ liệu văn bản phi cấu trúc ở đầu vào, bao gồm rất nhiều loại văn bản như các phản hồi của khách hàng, bài đăng trên mạng xã hội và các bài báo.
- Có nhiều ứng dụng: Text Mining có rất nhiều ứng dụng, bao gồm phân tích cảm xúc (Sentiment Analysis), nhận dạng thực thể có tên (Named Entity Recognition), tóm tắt văn bản (Text summarization), v.v…. Phạm vi ứng dụng rộng lớn khiến Text Mining trở thành một công cụ linh hoạt để người dùng lấy được chính xác thông tin họ mong muốn từ các dữ liệu văn bản phi cấu trúc.
- Cải thiện việc ra quyết định: Text Mining trích xuất thông tin chi tiết từ dữ liệu văn bản phi cấu trúc để phục vụ cho quá trình ra quyết định dựa trên dữ liệu.

Không chỉ những lợi ích nêu trên, Text Mining giúp giảm chi phí, tăng năng suất trong các trung tâm dữ liệu của các tổ chức ứng dụng kỹ thuật này, góp phần tạo ra các dịch vụ mới và mô hình kinh doanh mới cho các công ty công nghệ khai thác Text Mining.

1.2.3. Nhược điểm của Text Mining:

Mặc dù Text Mining có nhiều ưu điểm, cũng có những hạn chế và nhược điểm cần được xem xét khi áp dụng công nghệ này:

- Độ phức tạp: Text Mining đòi hỏi các kỹ năng cao của người lập trình về các thuật toán xử lý ngôn ngữ tự nhiên và học máy.

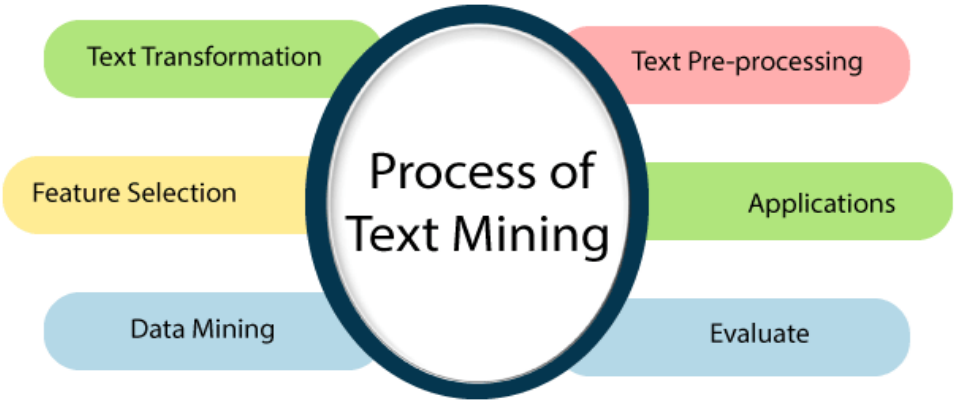
- Chất lượng dữ liệu: Chất lượng của dữ liệu văn bản đầu vào có thể không ổn định, ảnh hưởng đến độ chính xác của các kết quả đầu ra được trích xuất trong quá trình xử lý văn bản.
- Chi phí tính toán cao: Text Mining đòi hỏi tài nguyên tính toán và máy móc cao trong quá trình hoạt động. Các tổ chức nhỏ và các cá nhân có thể gặp khó khăn trong vấn đề triển khai Text Mining do hạn chế về máy móc và cấu hình máy tính trong quá trình hoạt động.
- Giới hạn kiểu dữ liệu đầu vào: Text Mining bị giới hạn ở kiểu dữ liệu đầu vào khi nó chỉ nhận kiểu dữ liệu văn bản phi cấu trúc và không thể sử dụng kỹ thuật này với các loại dữ liệu khác.
- Nhiễu trong kết quả: Text Mining sử dụng trong các tài liệu vẫn còn sai sót. Tuy nhiên trong hầu hết các trường hợp, nếu độ nhiễu của kết quả (hay tỉ lệ lỗi) đủ thấp thì lợi ích của tự động hóa sẽ lớn hơn khả năng xảy ra sai lầm của con người.

1.3. Cách thức Text Mining hoạt động

1.3.1. Quá trình Text Mining hoạt động

Text Mining là một quá trình tự động sử dụng xử lý ngôn ngữ tự nhiên để trích xuất những thông tin quý giá từ văn bản không có cấu trúc. Bằng cách biến đổi dữ liệu thành thông tin mà máy có thể hiểu, Text Mining tự động hóa quá trình phân loại văn bản theo cảm xúc, chủ đề và mục đích. Nó cải thiện theo thời gian khi ngữ cảnh của nội dung tăng lên, thay đổi hoặc khi cách tổ chức nội dung đó phát triển.

Quá trình Text Mining kết hợp các bước sau để trích xuất dữ liệu từ tài liệu:



H1.4. Quá trình hoạt động Text Mining

Text Pre-processing: Tiền xử lý là một nhiệm vụ quan trọng và là một bước quan trọng trong Text Mining. Bao gồm các bước sau:

+ clean up: loại bỏ mọi thông tin không cần thiết hoặc không mong muốn. Chẳng hạn như loại bỏ quảng cáo khỏi các trang web, chuẩn hóa văn bản được chuyển đổi từ định dạng nhị phân, ...

+ Tokenization: quá trình tách một cụm từ, câu, đoạn văn, một hoặc nhiều tài liệu văn bản thành các đơn vị nhỏ hơn.

+ part-of-speech tagging: qua trình đánh dấu một từ trong văn bản (ngữ liệu) tương ứng với một từ loại nào đó, dựa theo định nghĩa và bối cảnh văn phạm của từ đó.

Nêu ví dụ

Text Transformation – Biến đổi văn bản: Biến đổi dữ liệu văn bản để làm cho nó phù hợp với mục tiêu cụ thể, chẳng hạn như việc tách từ, chuẩn hóa từ viết hoa thành viết thường, xử lý các ký tự đặc biệt, và các biến đổi dữ liệu khác. Việc này làm cho dữ liệu dễ dàng xử lý hơn và chuẩn hóa nó trước khi thực hiện các tác vụ Text Mining, như phân tích cảm xúc, tạo biểu đồ từ dữ liệu văn bản, hay xây dựng mô hình dự đoán. Dưới đây là hai cách biểu diễn tài liệu chính thường được sử dụng:

1. Bag of words – Mô hình túi từ: "Bag of Words" (BoW) là một phương pháp quan trọng trong khai phá dữ liệu và xử lý ngôn ngữ tự nhiên. Nó được sử dụng để biểu diễn và xử lý văn bản hoặc dữ liệu văn bản trong các ứng dụng như phân loại văn bản, tìm kiếm thông tin, và khai thác dữ liệu. Các thức hoạt động của nó như sau:

- Thu thập dữ liệu văn bản: Trước tiên, bạn cần thu thập các tài liệu hoặc văn bản mà bạn muốn phân tích hoặc xử lý.
- Tiền xử lý dữ liệu: Bước này bao gồm việc loại bỏ các ký tự đặc biệt, chuyển đổi văn bản thành chữ thường (lowercase), và tách văn bản thành các từ (hoặc "tokens").
- Tạo Bag of Words: Bag of Words là một tập hợp (hoặc "túi") của tất cả các từ xuất hiện trong các văn bản của bạn, cùng với số lần mỗi từ xuất hiện trong mỗi văn bản cụ thể. Ví dụ: (1) John likes to watch movies. Mary likes movies too. (2) Mary also likes to watch football games.
- Sau đó ta sẽ có bag of words như sau:

BoW1 = {"John":1, "likes":2, "to":1, "watch":1, "movies":2, "Mary":1, "too":1};

BoW2 = {"Mary":1, "also":1, "likes":1, "to":1, "watch":1, "football":1, "games":1};

- Biểu diễn văn bản: Mỗi văn bản sau khi được biểu diễn bằng Bag of Words sẽ trở thành một vectơ số, trong đó mỗi thành phần của vectơ là số lần xuất hiện của từ tương ứng trong văn bản. Các vectơ này có thể được sử dụng để thực hiện các tác vụ như phân loại văn bản, tìm kiếm thông tin, hoặc tính toán sự tương đồng giữa các văn bản.

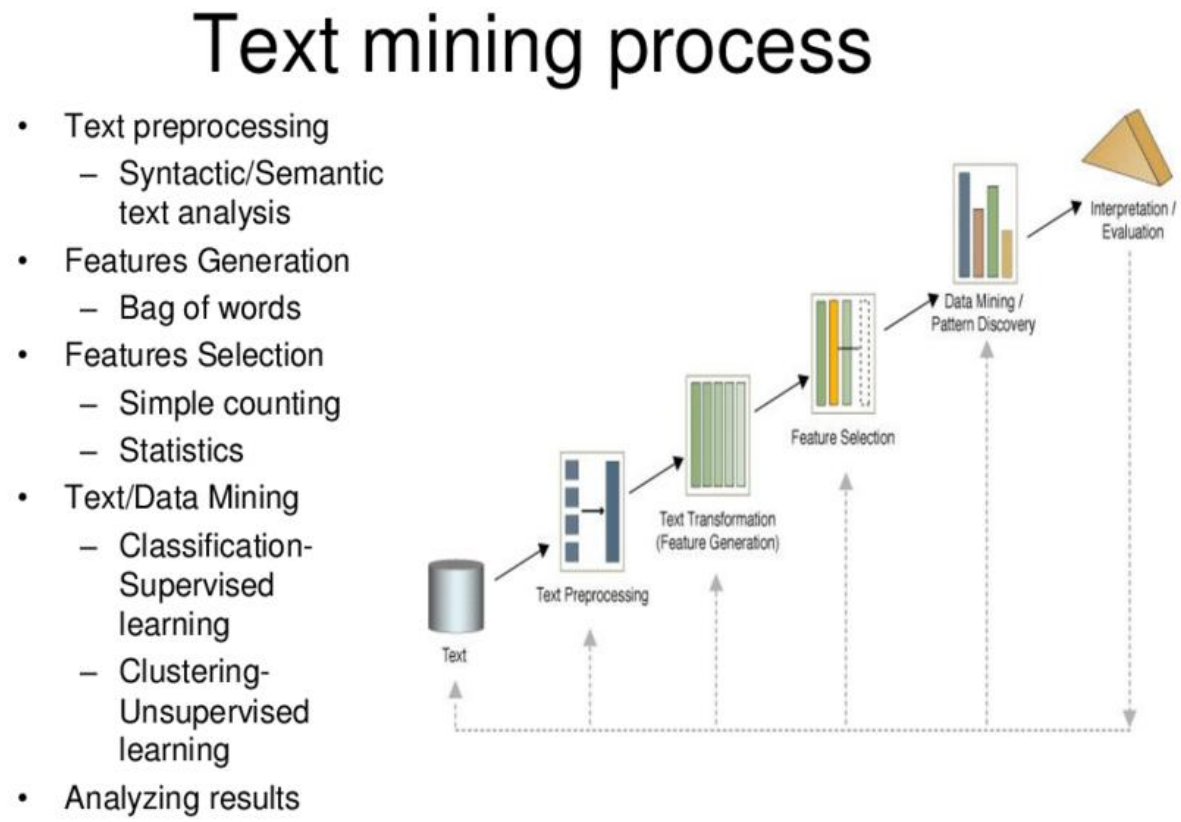
2. Vector Space – Mô hình không gian vector: vector space model hoặc term space model thuật ngữ là mô hình đại số để biểu diễn tài liệu văn bản (và bất kỳ đối tượng nào nói chung) dưới dạng vectơ của identifier (chẳng hạn như index terms). Nó được sử dụng trong việc lọc thông tin, truy xuất thông tin, lập index (chỉ mục) và xếp hạng mức độ liên quan. Cách hoạt động của vector space:

- Tài liệu và truy vấn được biểu diễn dưới dạng vectơ.
- Mỗi chiều tương ứng với một term riêng biệt. Nếu một thuật ngữ xuất hiện trong tài liệu thì giá trị của nó trong vectơ khác 0. Một số cách khác nhau để tính toán các giá trị này, còn được gọi là trọng số.
- Định nghĩa của term phụ thuộc vào ứng dụng. Thông thường, term là những từ đơn, từ khóa hoặc cụm từ dài hơn. Nếu các từ được chọn làm term thì số chiều của vectơ là số lượng từ trong từ vựng (số lượng các từ riêng biệt xuất hiện trong kho ngữ liệu).
- Các phép toán vectơ có thể được sử dụng để so sánh tài liệu với truy vấn.

Feature Selection: Lựa chọn tính năng là một phần quan trọng của khai phá dữ liệu. Đó là quá trình lựa chọn một tập hợp con các tính năng quan trọng để sử dụng trong việc tạo mô hình. Các tính năng dư thừa là những tính năng không cung cấp thêm thông tin. Các tính năng không liên quan không cung cấp thông tin hữu ích hoặc có liên quan trong bất kỳ bối cảnh nào. Lựa chọn tính năng có thể được định nghĩa là quá trình giảm đầu vào của quá trình xử lý hoặc tìm kiếm các nguồn thông tin thiết yếu. Lựa chọn đối tượng còn được gọi là lựa chọn biến.

Data Mining: Bây giờ, trong bước này, quy trình Text Mining kết hợp với quy trình thông thường. Các thủ tục Khai phá dữ liệu cổ điển được sử dụng trong cơ sở dữ liệu cấu trúc.

Evaluate: Sau đó, nó đánh giá kết quả. Thông qua việc đánh giá, có thể kiểm tra được tính chính xác của kết quả, đánh giá hiệu suất và loại bỏ các thông tin không cần thiết Mục tiêu của việc đánh giá là đảm bảo rằng các thông tin và mẫu dữ liệu đã được trích xuất hoặc tạo ra là chính xác và hữu ích cho mục đích cụ thể.



H1.5. Quy trình xử lý Text Mining

1.3.2. Cách thức triển khai Text Mining

Hiểu cách tổ chức dữ liệu văn bản: Dữ liệu có thể tồn tại trong các định dạng khác nhau, chẳng hạn như tệp tin văn bản, cơ sở dữ liệu, dữ liệu từ mạng xã hội, email, và nhiều nguồn khác. Việc hiểu cấu trúc này sẽ giúp xác định cách thu thập, trích xuất và lưu trữ dữ liệu một cách hiệu quả.

Tiền xử lý: Cần phải tiền xử lý dữ liệu trước khi áp dụng các kỹ thuật Text Mining. Điều này bao gồm việc loại bỏ ký tự không cần thiết, chuyển đổi văn bản thành chữ thường, tách từ, xử lý stopwords, nonwords (từ ngữ không quan trọng), và thực hiện stemming hoặc lemmatization để chuẩn hóa từ vựng.

Áp dụng các kỹ thuật Text Mining: Khi dữ liệu đã được tiền xử lý, có thể áp dụng các kỹ thuật Text Mining. Các phương pháp bao gồm

+ trích xuất thông tin (Information Retrieval): tìm thông tin liên quan từ một tập hợp các truy vấn hoặc cụm từ được xác định trước. Cách tiếp cận này thường được sử dụng trong hệ thống danh mục thư viện hoặc công cụ tìm kiếm trên web. Hệ thống IR sử dụng nhiều thuật toán khác nhau để theo dõi hành vi của người dùng và xác định dữ liệu liên quan. Tokenization liên quan đến việc chia một văn bản dài thành các câu hoặc từ được gọi là "tokens". Sau đó, các mã thông báo này được sử dụng trong các mô hình phân cụm văn bản hoặc các tác vụ liên quan đến liên kết tài liệu.

+ tóm tắt văn bản (Text Summarization): việc tạo ra một bản tóm tắt ngắn gọn, chính xác và trôi chảy của một tài liệu văn bản dài hơn. Các phương pháp tóm tắt văn bản tự động là rất cần thiết để giải quyết lượng dữ liệu văn bản ngày càng tăng. Điều này có thể giúp khám phá thông tin liên quan và sử dụng thông tin liên quan nhanh hơn.

+ phân loại văn bản (Text Classification)

+ phân tích cảm xúc (Sentiment Analysis)

và nhiều nhiệm vụ khác. Điều này thường đòi hỏi sử dụng các mô hình máy học hoặc mô hình xử lý ngôn ngữ tự nhiên, ví dụ điển hình như mô hình Transformer.

Phân tích kết quả: Sau khi áp dụng các kỹ thuật Text Mining, cần phân tích kết quả để hiểu thông tin mới và mẫu xuất hiện trong dữ liệu. Điều này có thể bao gồm việc tìm kiếm thông tin quan trọng, xác định các đặc điểm đáng chú ý, và tạo ra các báo cáo hoặc biểu đồ để thể hiện các phát hiện.

Trực quan hóa kết quả: Trực quan hóa dữ liệu là một phần quan trọng trong việc trình bày và diễn giải kết quả của Text Mining. Bằng cách sử dụng biểu đồ, đồ thị mạng, và các công cụ trực quan hóa khác, thông tin sẽ được hiển thị một cách rõ ràng và dễ hiểu cho người sử dụng cuối.

Triển khai hệ thống: Sau khi hệ thống Text Mining đã được xây dựng và kiểm tra, nó có thể được triển khai trong môi trường thực tế. Điều này có thể bao gồm tích hợp hệ thống vào ứng dụng hoặc trang web, cung cấp dịch vụ trực tiếp cho người dùng cuối, hoặc tích hợp vào quy trình công việc tự động trong doanh nghiệp.

2. Ứng dụng của Text Mining

Lượng dữ liệu sinh ra mỗi ngày trong quá trình hoạt động của các cá nhân và tổ chức là rất lớn, thống kê cho biết đến gần 80% lượng dữ liệu văn bản hiện nay là không có cấu trúc, hay nói các khác là chúng không được tổ chức theo một quy luật nào, khó để tìm kiếm, lưu trữ cũng như xử lý. Vì thế, ứng dụng của Text Mining trải dài trên nhiều lĩnh vực khác nhau có yêu cầu xử lý các loại dữ liệu từ văn bản:

- Risk Management (Quản lý rủi ro): Quản lý rủi ro là một quy trình có hệ thống và logic nhằm phân tích, xác định, xử lý và giám sát các rủi ro liên quan đến bất kỳ hành động hoặc quy trình nào trong tổ chức. Phân tích rủi ro không đầy đủ thường là nguyên nhân hàng đầu gây thất bại trong các dự án và quy trình của các doanh nghiệp, đặc biệt là các doanh nghiệp vừa và nhỏ còn thiếu kinh nghiệm trong giải quyết rủi ro. Text Mining sẽ tự động quét và phân tích các tài liệu để nắm bắt tổng quan về rủi ro có thể đối mặt. Sau đó, Text Mining này sẽ kết nối dữ liệu từ nhiều nguồn khác nhau để theo dõi xu hướng và mối quan hệ giữa các rủi ro, tạo ra các công cụ tự động hóa dự đoán và cảnh báo rủi ro tiềm ẩn. Sử dụng Text Mining giúp quản lý rủi ro một cách hiệu quả và thông minh hơn, giảm thiểu rủi ro và tối ưu hóa lợi ích cho doanh nghiệp.
- Customer Care Service (Dịch vụ chăm sóc khách hàng): Text Mining giúp doanh nghiệp hiểu rõ hơn về cảm nhận và ý kiến của khách hàng về sản phẩm và dịch vụ, cung cấp khả năng dự đoán nhu cầu của khách hàng trong tương lai, từ đó giúp tối ưu hóa chiến lược chăm sóc khách hàng. Công nghệ này giúp phát hiện sớm các vấn đề và khiếu nại từ khách hàng, giúp doanh nghiệp xử lý chúng kịp thời và tăng cường sự hài lòng của khách hàng. Ngoài ra, Text Mining còn giúp tạo ra phản hồi cá nhân hóa và nhanh chóng cho khách hàng, đồng thời quản lý dư luận và tương tác xã hội liên quan đến thương hiệu. Sử dụng Text Mining trong dịch vụ chăm sóc khách hàng giúp doanh nghiệp cải thiện chất lượng dịch vụ và xây dựng mối quan hệ tốt hơn với khách hàng.

- Business Intelligence (Kinh doanh thông minh): Trong lĩnh vực này, Text Mining tổ chức tổng hợp thông tin từ dữ liệu phi cấu trúc như bài viết trên mạng xã hội, email, và các nguồn dữ liệu khác để đưa ra quyết định và chiến lược kinh doanh thông minh. Text Mining còn giúp phát hiện xu hướng, mối quan hệ ẩn, dự đoán tương lai và quản lý các vấn đề cộng đồng liên quan đến thương hiệu. Chúng giúp các doanh nghiệp tổng hợp được các thông tin chi tiết, có chiều sâu và hiểu rõ hơn về khách hàng, môi trường kinh doanh của mình, từ đó tối ưu hóa các quyết định tạo ra lợi ích cạnh tranh.
- Social Media Analysis (Phân tích truyền thông xã hội): Text Mining tự động phân tích và trích xuất thông tin từ các dữ liệu văn bản trên các nền tảng mạng xã hội. Sử dụng Text Mining, doanh nghiệp có khả năng xác định xu hướng, cảm nhận, từ khóa quan trọng, và đánh giá sản phẩm và dịch vụ dựa trên phản hồi của người dùng trực tuyến. Text Mining cũng hỗ trợ giám sát dư luận trực tuyến và quản lý tình huống khủng hoảng, cung cấp thông tin để tự động hóa trả lời và tương tác với người dùng. Tất cả những điều này giúp doanh nghiệp hiểu rõ hơn về môi trường trực tuyến và xây dựng chiến lược dựa trên thông tin toàn diện hơn.
- Fraud Detection (Phát hiện gian lận): Text Mining sử dụng thông tin được và trích xuất từ các nguồn dữ liệu đầu vào, qua đó phân tích và xác định các dấu hiệu liên quan đến việc gian lận từ các dữ liệu văn bản phi cấu trúc như email, hồ sơ khách hàng, hay các bài viết trên mạng. Text Mining cung cấp khả năng phát hiện mô hình và biểu hiện của gian lận, xác định gian lận, phân tích ngữ cảnh và hệ thống các thông tin liên quan đến gian lận. Không chỉ thế, Text Mining còn có thể giám sát thời gian thực để ngăn chặn gian lận trước khi có bất kỳ thiệt hại nào bị gây ra, giúp các tổ chức và doanh nghiệp bảo vệ danh tiếng và tài sản trong cuộc chiến chống gian lận.

Không chỉ dừng lại ở những ứng dụng trên, về lợi ích mà nó mang lại, Text Mining dần trở thành một công cụ đắc lực cho các doanh nghiệp, tổ chức trong việc xử lý các vấn đề liên quan đến văn bản, đặc biệt là văn bản phi cấu trúc ở mọi lĩnh vực từ các lĩnh vực thuộc phạm trù kinh tế, chính trị cho đến các lĩnh vực xã hội như giáo dục, y tế,.... Những ứng dụng này cho thấy sự đa dạng và lợi ích to lớn của Text Mining trong việc nắm bắt và xử lý thông tin trong nhiều hoàn cảnh khác nhau để hỗ trợ các cá nhân và tổ chức trong việc phân tích, hỗ trợ ra quyết định hay thậm chí là giúp việc quản lý trở nên hiệu quả hơn.

H2.1. Ứng dụng của Text Mining

3. Một số kỹ thuật được sử dụng trong Text Mining

3.1. Tổng quan các kỹ thuật

Text Mining được chia làm 2 giai đoạn chính gồm: giai đoạn tiền xử lý văn bản và giai đoạn phân tích nội dung. Mỗi giai đoạn trong Text Mining đều gắn liền với những thuật toán và công nghệ riêng. Bảng sau sẽ tổng quan về các công nghệ sử dụng trong Text Mining:

	Thuật toán	**Vấn đề**	**Động cơ**	**Công nghệ**
Giai đoạn tiền xử lý văn bản	Mã hóa kỹ thuật số (Tokenization)	Làm thế nào để chuyển đổi văn bản thành từ hoặc định dạng văn bản?	Chuyển chuỗi thành một mã token văn bản duy nhất.	Loại bỏ khoảng trắng (White space separation)

	Nhận dạng từ ghép	Làm thế nào để nhận dạng các từ chung ý nghĩa?	Xác định các từ có nghĩa chung bị thiếu từ	n-grams
	Chuẩn hóa và giảm nhiễu âm	Làm thế nào để xử lý rất nhiều biến trong một ma trận tài liệu-thuật ngữ (Document-Term-Matrix)?	Giảm chiều của ma trận tài liệu-thuật ngữ	Thu gọn từ (Stemming), bổ đề ngôn ngữ (Lemmatization), xóa các từ ít quan trọng (stop words), từ ít phổ biến
	Xử lý ngôn ngữ học (Linguistic Processing)	Làm thế nào để xác định các từ có ý nghĩa đặc biệt hay mang chức năng ngữ pháp?	Gán thẻ/Đánh dấu từ	Nhận diện thực thể được đặt tên (Named-entity Recognition), gán thẻ loại từ (Part-of-speech Tagging)
Giai đoạn xử lý nội dung	Dựa từ điển (Dictionary-based)	Làm thể nào để xác định những đặc điểm và trạng thái xã hội học hoặc tâm lý tiềm ẩn được phản ánh trong ngôn ngữ tự nhiên?	Đo lường các khái niệm và cấu trúc ngữ cảnh, tâm lý, ngôn ngữ hoặc ngữ nghĩa.	Từ điển dựng sẵn (Pre-defined dictionaries), từ điển được tùy chỉnh (Customized dictionaries)
	Các kỹ thuật thuật toán	Làm thế nào để gán các văn bản vào các lớp đã được xác định trước?	Phân loại các thực thể văn bản vào các danh mục đã được xác định trước	Học có giám sát (Supervised Learning)

	Các kỹ thuật thuật toán	Làm thế nào để nhóm các tài liệu tương tự với nhau?	Gom nhóm các thực thể văn bản thành các nhóm chưa được xác định và không biết trước	Học không giám sát (Unsupervised Learning)

B3.1. Các công nghệ trong Text Mining

Một số kỹ thuật được dùng trong Text Mining có rất nhiều và chúng có thể được kể đến như: Trích xuất thông tin (Information Extraction), Phân loại văn bản (Text Categorization), Phân cụm văn bản (Text Clustering), Trực quan hóa văn bản (Text Visualization), Tóm tắt văn bản (Text Summarization), v.v…

Trong khuôn khổ bài thu hoạch, nhóm 1 đã tìm hiểu và cài đặt kỹ thuật Text Summarization, do đó sẽ giải thích kỹ hơn về kỹ thuật này.

3.2. Text Summarization – Tóm tắt văn bản

Tóm tắt văn bản là quá trình trích xuất thông tin quan trọng nhất từ một văn bản để tạo ra phiên bản ngắn gọn, xúc tích nhưng vẫn bảo toàn đầy đủ thông tin cơ bản và tuân theo ngữ pháp và chính tả. Bản tóm tắt cần duy trì thông tin quan trọng của toàn bộ văn bản gốc và có bố cục chặt chẽ, xem xét các yếu tố như độ dài câu, phong cách viết và cú pháp văn bản.

Phụ thuộc vào số lượng văn bản đầu vào, kỹ thuật tóm tắt có thể được chia thành hai loại chính: tóm tắt đơn văn bản và tóm tắt đa văn bản. Tóm tắt đơn văn bản đơn giản là việc rút gọn một văn bản thành một phiên bản ngắn gọn. Trong khi đó, tóm tắt đa văn bản đòi hỏi việc rút gọn một tập hợp các văn bản thành một phiên bản tóm tắt. Tóm tắt đa văn bản thường được sử dụng để tóm tắt thông tin từ nhiều văn bản để giúp người đọc hiểu nội dung cụ thể.

Tóm tắt văn bản thường được tiếp cận theo hai phương pháp chính: Tóm tắt trích chọn (Extractive Summarization) và Tóm tắt tóm lược ý (Abstractive Summarization).

Tóm tắt trích chọn là công việc chọn ra một tập con những từ đã có, những lời nói hoặc những câu của văn bản gốc để đưa vào bản tóm tắt. Những phần văn bản được cho là quan trọng được hình thành trên cơ sở một số thuật toán để xác định độ liên quan của chúng đối với ý nghĩa tổng thể của cả tài liệu. Những phần văn bản có mức độ liên quan cao nhất sẽ được chọn để đưa vào bản tóm tắt.

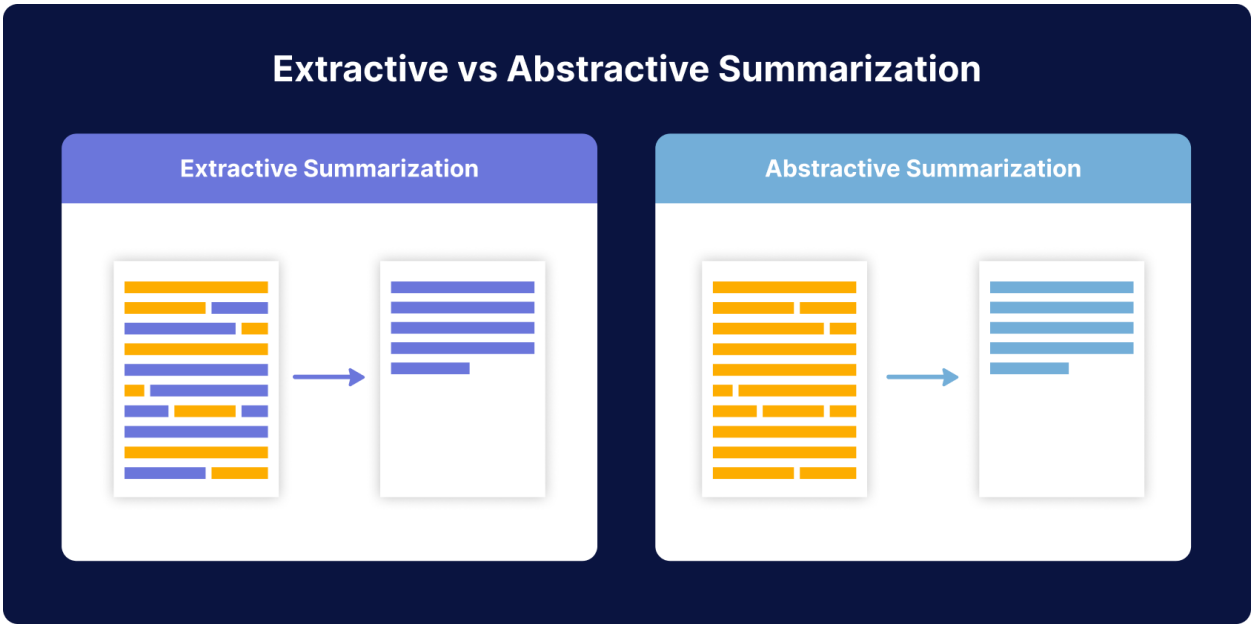
Có nhiều cách để thực hiện việc xếp hạng các phần văn bản như sử dụng TF-IDF (Term Frequency-Inverse Document Frequency), sử dụng các phương pháp dựa đồ thị như TextRank hay các phương pháp dựa trên mô hình học máy (Machine Learning) như Máy vector hỗ trợ (Support Vector Machines) hay Rừng ngẫu nhiên (Random Forests).

Mục tiêu chính của phương pháp này là duy trì ý nghĩa ban đầu của văn bản và bao gồm các từ đã từng xuất hiện trong văn bản. Phương pháp này hoạt động tốt khi đầu vào đã được tổ chức có cấu trúc, về vật lý và logic như nội dung trong các trang báo. Tuy nhiên, phương pháp này hoạt động kém hiệu quả khi phải xử lý các từ đồng âm hay các từ đồng nghĩa xuất hiện ở nội dung đầu vào.

Tóm tắt tóm lược ý phân tích văn bản đầu vào và tạo ra các cụm từ hoặc câu mới chứa bản chất của văn bản gốc và truyền đạt cùng ý nghĩa với văn bản gốc nhưng ngắn gọn và mạch lạc hơn.

Phương pháp này sẽ phân tích văn bảng đầu vào bằng một mô hình nơ-ron để tạo ra các cụm từ và câu mới chứa bản chất của văn bản gốc. Mô hình được huấn luyện trên lượng lớn dữ liệu văn bản và học cách hiểu về mối quan hệ giữa từ và câu, tạo ra văn bản mới truyền đạt cùng ý nghĩa với văn bản gốc một cách ngắn gọn, xúc tích và dễ hiểu. Phương pháp sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên tiên tiến như sinh ngôn ngữ tự nhiên (Natural Language Generation) và học sâu (Deep Learning) để hiểu ngữ cảnh và tạo ra bản tóm tắt.

Bản tóm tắt đầu ra của tóm tắt tóm lược ý thường ngắn và dễ hiểu, gần gũi với ngôn ngữ tự nhiên hơn so với phương pháp tóm tắt trích chọn. Tuy nhiên, phương pháp này vẫn còn nhiều hạn chế như các lỗi chính tả, ngữ pháp, hay thỉnh thoảng sẽ sinh ra những từ không tồn tại ở thế giới thực.



H3.1. Tóm tắt trích chọn và Tóm tắt tóm lược ý

Text Summarization là một kỹ thuật phổ biến trong Text Mining, nên nó cũng kế thừa những ứng dụng tuyệt vời của Text Mining trong thực tế đời sống. Tuy nhiên, tóm tắt văn bản sẽ có ứng dụng sâu hơn vào những tác vụ làm việc liên quan đến văn bản và cần đầu ra là một bản tóm tắt như ngành tài chính, chăm sóc sức khỏe, luật pháp,...

Một số ứng dụng cụ thể của Text Summarization có thể được kể đến như giám sát phương tiện truyền thông, viết báo, ứng dụng trong marketing và SEO, nghiên cứu tài chính, chatbot trả lời câu hỏi, e-learning, thực hiện các bản báo cáo sức khỏe, tóm tắt nộ dung, tự tạo nội dung mới, v.v…

4. So sánh với các công nghệ khác

4.1. Xử lý ngôn ngữ tự nhiên (Natural Language Processing)

4.1.1. Xử lý ngôn ngữ tự nhiên là gì?

Xử lý ngôn ngữ tự nhiên - Natural Language Processing, là một nhánh của khoa học máy tính, cụ thể hơn, một nhánh của trí tuệ nhân tạo liên quan đến việc cung cấp cho máy tính khả năng đọc hiểu văn bản và lời nói giống khả năng của con người.

Natural Language Processing kết hợp ngôn ngữ học tính toán - mô hình hóa ngôn ngữ của con người dựa trên quy tắc - với các mô hình thống kê, học máy và học sâu. Cùng

với nhau, những công nghệ này cho phép máy tính xử lý ngôn ngữ của con người dưới dạng dữ liệu văn bản hoặc giọng nói và 'hiểu' ý nghĩa đầy đủ, hoàn chỉnh với ý định và cảm xúc của người nói hoặc người viết.

Một số ứng dụng của Natural Language Processing:

- Phân tích cảm xúc (Sentiment Analysis) là quá trình phân loại cảm xúc được diễn đạt trong văn bản. Đầu vào của bài toán là một đoạn văn bản và đầu ra là xác suất mà cảm xúc thể hiện (tích cực, tiêu cực, trung tính). Phân tích quan điểm được áp dụng nhiều nhất trong phân loại đánh giá của khách hàng trên nhiều nền tảng trực tuyến khác nhau, xác định các dấu hiệu trong các bình luận trực tuyến.
- Dịch máy tự động (Automatic Machine Translation) nhiều ngôn ngữ khác nhau. Đầu vào là một văn bản bằng ngôn ngữ nguồn được chỉ định, đầu ra là văn bản bằng ngôn ngữ đích được chỉ định. Các ứng dụng của mô hình này được dùng để cải thiện khả năng giao tiếp giữa mọi người trên các nền tảng truyền thông. Một số hệ thống còn thực hiện nhận dạng ngôn ngữ; nghĩa là phân loại văn bản theo ngôn ngữ này hay ngôn ngữ khác.
- Nhận dạng thực thể được đặt tên (Named Entity Recognition): nhằm trích xuất thực thể trong một đoạn văn bản thành các danh mục được xác định trước như tên cá nhân, tổ chức, địa điểm và số lượng. Đầu vào là một văn bản, đầu ra là các thực thể được đặt tên khác nhau cùng với vị trí bắt đầu và kết thúc của chúng. Mô hình này được ứng dụng trong các tóm tắt bài báo và chống thông tin giả.

H4.1. Nhận dạng thực thể được đặt tên

- Tạo lập văn bản (Text Generation): tạo ra văn bản tương tự như văn bản do con người viết. Những mô hình đó có thể được tinh chỉnh để tạo ra văn bản ở nhiều thể loại và định dạng khác nhau như bài viết, blog hay mã máy tính. Phương pháp này hữu ích cho tính năng tự hoàn thành (dự đoán từ) và chatbots.

4.1.2. So sánh giữa Natural Language Processing và Text Mining

Cả Natural Language Processing và Text Mining có một vài điểm chung:

- Đầu vào: Đều làm việc với dữ liệu văn bản và là văn bản ngôn ngữ tự nhiên, cả với văn bản viết và văn bản nói. Cả hai đều xử lý và phân tích các đoạn văn bản, tài liệu, tập dữ liệu để trích xuất thông tin hoặc hiểu ngữ nghĩa.
- Công dụng: Khai phá thông tin từ văn bản. Dù Natural Language Processing tập trung vào hiểu và tạo ra các ứng dụng liên quan đến ngôn ngữ, Text Mining tập trung vào trích xuất thông tin cụ thể từ văn bản, nhưng cả hai công nghệ đều nhấn mạnh việc tìm kiếm, phân loại, và trích xuất dữ liệu hữu ích từ văn bản.
- Một số thuật toán: Đều dùng chung một số thuật toán được sử dụng rộng rãi trong lĩnh vực học máy và học sâu như: Hồi quy Logistic (Logistic Regression), Phân loại Naive Bayes (Naive Bayes Classification), Mạng nơ-ron tích chập (Convolutional Neural Network), Mạng nơ-ron hồi quy (Recurrent Neural Network), Bộ tự mã hóa (AutoEncoder), v.v…
- Natural Language Processing và Text Mining cùng có mối quan hệ mật thiết với hai lĩnh vực công nghệ hiện đại và nổi bật nhất bây giờ học máy và học sâu. Cả hai phần lớn thường dựa trên việc huấn luyện tập dữ liệu lớn để cho ra sản phẩm đầu ra.

Giữa Natural Language Processing và Text Mining có sự khác biệt:

	Natural Language Processing	**Text Mining**
Mục tiêu	Tạo ra các ứng dụng liên quan đến ngôn ngữ tự nhiên như chatbot, phân loại văn bản, dự đoán ý định, dịch máy, tổng hợp văn bản, học máy dựa trên văn bản.	Khai thác thông tin có trong dữ liệu văn bản để khám phá tri thức, mô hình hóa thông tin, tìm khiếm thông tin cụ thể.
Phương pháp	Sử dụng các kỹ thuật học máy và học sâu để hiểu và tạo ra ngôn ngữ tự nhiên, gồm sử dụng mô hình ngôn ngữ, mạng nơ-ron và thường cần dữ liệu huấn luyện lớn.	Sử dụng các phương pháp thống kê, khai phá dữ liệu, và gom cụm để trích xuất thông tin và tri thức từ dữ liệu văn bản. Các phương pháp thống kê như tần số từ (TF-IDF - Term Frequency-Inverse Document

		Frequency), và phân loại văn bản thường được sử dụng.
Ứng dụng	Các hệ thống yêu cầu đầu ra là đoạn văn bản ngôn ngữ giống với ngôn ngữ tự nhiên của con người hay phân tích văn bản ngôn ngữ tự nhiên.	Khai thác dữ liệu từ các nguồn văn bản lớn, tóm tắt văn bản, phân loại chủ đề trong các tập dữ liệu văn bản lớn.

B4.1. Khác nhau giữa Narutal Language Processing và Text Mining

Ngoài ra, Natural Language Processing và Text Mining còn được kết hợp với nhau để tạo ra một phương pháp mạnh mẽ trong các ứng dụng có khả năng hiểu và tri thức hóa dữ liệu văn bản. Kết hợp Natural Language Processing và Text Mining để giải quyết một số bài toán có độ phức tạp cao như bài toán phân tích ngữ cảnh, tóm tắt văn bản thông minh, phân tích ý kiến kết hợp phân loại văn bản, phân loại và gom cụm dữ liệu lớn.

4.2. Khai thác dữ liệu (Data Mining)

4.2.1. Data Mining là gì?

Khai thác dữ liệu là một kỹ thuật có sự hỗ trợ của máy tính được sử dụng trong phân tích để xử lý và khám phá các tập dữ liệu lớn. Với các công cụ và phương pháp khai thác dữ liệu, các tổ chức có thể khám phá các mẫu và mối quan hệ ẩn trong dữ liệu của họ. Data Mining biến dữ liệu thô thành kiến thức thực tế. Các công ty sử dụng kiến thức này để giải quyết vấn đề, phân tích tác động trong tương lai của các quyết định kinh doanh và tăng tỷ suất lợi nhuận.

Data Mining được ứng dụng trong rất nhiều các lĩnh vực: viễn thông, truyền thông, công nghệ, ngân hàng, bảo hiểm, giáo dục, sản xuất, bán lẻ, v.v…. Data Mining trong các lĩnh vực trên phân tích các dữ liệu thô có sẵn để đưa ra kết quả phân tích dữ liệu hay dự báo các xu hướng trong tương lai.

4.2.2. So sánh giữa Data Mining và Text Mining?

Data Mining và Text Mining đều có một số điêm chung sau đây:

- Mục tiêu: Khai phá thông tin. Cả Data Mining và Text Mining đều nhằm mục tiêu khai phá thông tin từ dữ liệu. Dù là thông tin từ dữ liệu số (Data Mining)

hoặc thông tin từ dữ liệu văn bản (Text Mining), cả hai lĩnh vực này đều hướng đến việc tìm hiểu tri thức, mô hình, hoặc mẫu từ dữ liệu.

- Quy trình xử lý dữ liệu: Đòi hỏi quy trình xử lý dữ liệu gồm các bước chính: thu thập dữ liệu, tiền xử lý dữ liệu, trích xuất đặc trưng, huấn luyện mô hình và kiểm tra mô hình.
- Ứng dụng trong thực tế: Có ứng dụng rộng rãi trong nhiều ngành nghề và lĩnh vực đòi hỏi phân tích thông tin từ một mẫu dữ liệu cụ thể và tối ưu hóa giá trị từ dữ liệu.

Tuy vậy, cả hai công nghệ này vẫn có một vài điểm khác biệt:

	Data Mining	**Text Mining**
Dữ liệu xử lý	Xử lý và khai phá dữ liệu từ nhiều nguồn, nhiều loại, cả có cấu trúc (cơ sở dữ liệu, bảng tính) và cả không có cấu trúc (hình ảnh, âm thanh, video).	Khai phá thông tin từ dữ liệu văn bản (bài viết, email, tin nhắn,...). Thường xử lý dữ liệu không có cấu trúc trong định dạng văn bản.
Mục tiêu	Tìm kiếm mô hình, quy tắc và cấu trúc trong dữ liệu để khám phá tri thức mới, dự báo kết quả, hỗ trợ ra quyết định	Khai phá thông tin từ dữ liệu văn bản. Nó bao gồm việc trích xuất tri thức, phân tích ý kiến, phân loại văn bản, và tạo ra cấu trúc dữ liệu từ văn bản.
Phương pháp	Khai phá luật kết hợp trong cơ sở dữ liệu, phân loại dữ liệu, phân cụm dữ liệu, phân tích trình tự và đường dẫn (Sequence and path analysis),...	Sử dụng các phương pháp thống kê, khai phá dữ liệu, và gom cụm để trích xuất thông tin và tri thức từ dữ liệu văn bản. Các phương pháp thống kê như tần số từ (TF-IDF - Term Frequency-Inverse Document Frequency), và phân loại văn bản thường được sử dụng.
Ứng dụng	Khai thác dữ liệu từ nhiều nguồn dữ liệu khác nhanh và nhiều loại dữ liệu khác nhau, từ đó đưa ra phân tích, dự báo xu hướng trong tương lai, tối ưu hóa quá trình ra quyết định.	Khai thác dữ liệu từ các nguồn văn bản lớn, tóm tắt văn bản, phân loại chủ đề trong các tập dữ liệu văn bản lớn.

Data Mining và Text Mining có thể kết hợp để tạo ra các ứng dụng mạnh mẽ hơn. Ví dụ, Data Mining có thể được sử dụng để xây dựng mô hình dự đoán dựa trên dữ liệu số, và sau đó kết quả từ mô hình này có thể được kết hợp với thông tin từ Text Mining để tạo ra ứng dụng tổng hợp thông tin toàn diện.

CÀI ĐẶT VÀ ĐÁNH GIÁ

1. Dữ liệu thử nghiệm

Nhóm sử dụng bộ dữ liệu CNN/Daily Mail để tiến hành thí nghiệm. CNN/Daily Mail là bộ dữ liệu tiếng Anh với hơn 300,000 bài báo khác nhau được đăng tải trên CNN và Daily Mail.

Thống kê về bộ dữ liệu:

- Bộ dữ liệu được chia làm 3 tập: Huấn luyện (training), phát triển (validation) và kiểm thử (test) với số lượng mẫu trong từng tập như sau:

Tập dữ liệu	**Số lượng mẫu trong tập dữ liệu**
Huấn luyện (training)	287,113
Phát triển (validation)	13,368
Kiểm thử (test)	11,490

- Mỗi mẫu dữ liệu bao gồm các trường dữ liệu sau:

Trường dữ liệu	**Mô tả**
`id`	Một chuỗi chứa hàm băm SHA1 (SHA1 hash) được định dạng thập lục phân biểu diễn cho URL của bài báo
`article`	Nội dung bài báo
`highlights`	Tóm tắt bài báo

2. Giới thiệu mô hình sử dụng

T5 (Text-To-Text Transfer Transformer), do Google phát triển dựa trên kiến trúc Transformer, là mô hình mã hóa-giải mã (encoder-decoder model) được đào tạo về nhiều tác vụ ngôn ngữ và có thể thực hiện chuyển đổi văn bản, như: Dịch văn bản sang ngôn ngữ khác, tóm tắt văn bản hoặc trả lời câu hỏi, v.v... Mô hình T5 yêu cầu một tiền tố nhất định tương ứng với từng tác vụ, ví dụ:

- Dịch từ tiếng Anh sang tiếng Đức: "`translate English to German: …`".
- Tóm tắt văn bản: "`summarize: …`".

T5 được chia thành nhiều phiên bản với số lượng tham số trong mỗi mô hình là khác nhau: `t5-small`, `t5-base`, `t5-large`, `t5-3b`, `t5-11b`. Trong bài báo cáo, nhóm ứng dụng mô hình `t5-small` với khoảng 60 triệu tham số để huấn luyện cho tác vụ tóm tắt trừu tượng (abstractive summarization) một đoạn tin tức ngắn.

3. Cài đặt chương trình

3.1. Cài đặt và import thư viện

```
!pip install accelerate -U
!pip install datasets
!pip install evaluate
!pip install rouge_score
!pip install transformers[torch]
```

```
from datasets import load_dataset
import pandas as pd
import numpy as np
import torch
from transformers import AutoModelForSeq2SeqLM,
Seq2SeqTrainingArguments, Seq2SeqTrainer
from transformers import AutoTokenizer, DataCollatorForSeq2Seq
from transformers import pipeline
import evaluate
```

Đăng nhập vào tài khoản Hugging Face để có thể tải lên và lưu trữ mô hình tại kho lưu trữ cá nhân.

```
from huggingface_hub import notebook_login

notebook_login()
```

3.2. Tải bộ dữ liệu CNN/Daily Mail

```
cnn_dailymail = load_dataset("cnn_dailymail", "3.0.0")
```

3.3. Tiền xử lý dữ liệu

Load tokenizer T5 để xử lý hai trường dữ liệu `article` và `highlights`. Tokenizer có chức năng tách một cụm từ, câu, đoạn văn, v.v… thành các đơn vị nhỏ hơn được gọi là Token. Các Tokens sau đó được so sánh với bộ từ điển có sẵn của T5 và được chuyển thành dạng số (ID):

```
checkpoint = "buianh0803/text-sum-2"
tokenizer = AutoTokenizer.from_pretrained(checkpoint)
```

Hàm tiền xử lý cần thực hiện những công việc sau:

- Thêm tiền tố vào đầu input để T5 biết đây là nhiệm vụ tóm tắt văn bản.
- Tokenize `article` và `highlights`.
- Cắt ngắn chuỗi để không dài hơn độ dài tối đa được gán cho tham số `max_length`.

```
def preprocess_function(data):
    prefix = "summarize: "
    inputs = [prefix + article for article in data["article"]]
    model_inputs = tokenizer(
         inputs,
         max_length=1024,
         truncation=True
    )

    labels = tokenizer(
         text_target=data["highlights"],
         max_length=128,
         truncation=True
    )

    model_inputs["labels"] = labels["input_ids"]
    return model_inputs
```

Áp dụng hàm tiền xử lý trên toàn bộ tập dữ liệu, việc đặt tham số `batched=True` giúp cho hàm `map` có thể xử lý nhiều mẫu cùng một lúc, dẫn đến tộc độ thực hiện của hàm sẽ nhanh hơn:

```
tokenized_datasets = cnn_dailymail.map(preprocess_function,
batched=True)
```

Tạo batch cho tập dữ liệu, đồng thời thêm bộ đệm động (Dynamic Padding) để các câu có độ dài bằng với độ dài của câu dài nhất trong batch với `DataCollatorForSeq2Seq`:

```
data_collator = DataCollatorForSeq2Seq(tokenizer=tokenizer,
model=checkpoint)
```

3.4. Xây dựng hàm đánh giá (Evaluate)

Để đánh giá hiệu suất của mô hình, nhóm sử dụng hệ thống độ đo ROUGE, cụ thể gồm 2 độ đo chính là: ROUGE-N, ROUGE-L.

- ROUGE-N: Ước lượng độ tương đồng N-grams giữa văn bản tóm tắt ứng cử và văn bản tóm tắt dẫn xuất [1].

$$\begin{cases} ROUGE-N_{recall} = \dfrac{Num\ N-gram\ matches}{Num\ N-grams\ in\ reference} \\ ROUGE-N_{precision} = \dfrac{Num\ N-gram\ matches}{Num\ N-grams\ in\ summary} \\ ROUGE-N_{F1-score} = 2\left(\dfrac{precision\ \cdot\ recall}{precision\ +\ recall}\right) \end{cases}$$

- ROUGE-L: Sử dụng độ đo chuỗi con chung dài nhất (LCS – Longest Common Subsequence) để ước lượng độ chính xác của văn bản tóm tắt ứng cử so với văn bản tóm tắt dẫn xuất [1].

$$\begin{cases} \underset{recall}{ROUGE-L} = \dfrac{LCS(gen,ref)}{Num\ words\ in\ reference} \\ \underset{precision}{ROUGE-L} = \dfrac{LCS(gen,ref)}{Num\ words\ in\ summary} \\ \underset{F1-score}{ROUGE-L} = 2\left(\dfrac{precision\ \cdot\ recall}{precision\ +\ recall}\right) \end{cases}$$

Trong đó:

- gen: Văn bản ứng cử – generated summary.
- ref: Văn bản dẫn xuất – reference summary.

```
rouge = evaluate.load("rouge")

def compute_metrics(eval_pred):
    predictions, labels = eval_pred

    decoded_preds = tokenizer.batch_decode(predictions,
skip_special_tokens=True)

    labels = np.where(labels != -100, labels, tokenizer.pad_token_id)
    decoded_labels = tokenizer.batch_decode(labels,
skip_special_tokens=True)

    result = rouge.compute(predictions=decoded_preds,
references=decoded_labels, use_stemmer=True)

    prediction_lens = [np.count_nonzero(pred !=
tokenizer.pad_token_id) for pred in predictions]

    result["gen_len"] = np.mean(prediction_lens)

    return {k: round(v, 4) for k, v in result.items()}
```

3.5. Tinh chỉnh mô hình

Load mô hình T5:

```
model = AutoModelForSeq2SeqLM.from_pretrained(checkpoint)
```

Định nghĩa tham số huấn luyện:

```
BATCH_SIZE = 16
training_args = Seq2SeqTrainingArguments(
    output_dir="text-sum-3",
    evaluation_strategy="epoch",
    learning_rate=2e-5,
    per_device_train_batch_size=BATCH_SIZE,
    per_device_eval_batch_size=BATCH_SIZE,
    weight_decay=0.01,
    save_total_limit=3,
    num_train_epochs=1,
    fp16=True,
    predict_with_generate=True,
    push_to_hub=True,
)
```

Truyền các tham số đã khởi tạo và bắt đầu quá trình tinh chỉnh mô hình:

```
trainer = Seq2SeqTrainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_datasets["train"],
    eval_dataset=tokenized_datasets["test"],
    tokenizer=tokenizer,
    data_collator=data_collator,
    compute_metrics=compute_metrics,
)

trainer.train()
```

Đẩy mô hình đã được huấn luyện xong lên Huggingface để có thể lưu trữ thông tin của mô hình và sử dụng cho lần tinh chỉnh mô hình sau:

```
trainer.push_to_hub()
```

4. Đánh giá kết quả

Chương trình được cài đặt bằng ngôn ngữ lập trình Python trên Google Colab Pro với cấu hình như sau:

GPU: NVIDIA T4 15GB
CPU: Intel(R) Xeon(R) CPU @ 2.00GHz, RAM 51GB
OS: Windows 11

Do cấu hình máy huấn luyện hạn chế nên nhóm dừng quá trình huấn luyện ở epoch 8 để kiểm tra kết quả.

Epoch	ROUGE-1	ROUGE-2	ROUGE-L
8	0.2485	0.1188	0.2056

Ví dụ đầu ra:

Nội dung văn bản	Tokyo (CNN)Police in Japan say they have arrested a 40-year-old man accused of fatally stabbing five neighbors in a farming community in Sumoto city. The man has admitted stabbing three women aged 59, 76 and 84, as well as two men aged 62 and 82, Deputy Police Chief Keizo Okumoto told CNN. He said the accused refused to comment further as he was awaiting his lawyer. The victims -- two couples and the 84-year-old woman -- lived within 100 meters (330 feet) of the suspect's home, police said. According to local media, the accused and the victims shared the same surname, but it is unclear if they are related. Sumoto city is on Awaji Island, Hyogo prefecture, in Japan. CNN's Susannah Cullinane contributed to this report from London.
Văn bản tóm tắt dẫn xuất	Police in Japan say they have arrested a man, 40, after five neighbors were fatally stabbed . The accused shares the same surname as the

	victims, aged 59 to 84, local media say . A police official says the man has admitted to the stabbings but refused to comment further .
Văn bản tóm tắt ứng cử	The 40-year-old man is accused of fatally stabbing five neighbors in a farming community in Sumoto city, Japan . The victims lived within 100 meters (330 feet) of the suspect's home .

Kết quả thử nghiệm cho thấy đoạn tóm tắt ứng cử không gặp phải các lỗi về chính tả và đã thể hiện được một vài nội dung giống với đoạn tóm tắt dẫn xuất trên tập dữ liệu CNN/Daily Mail.

KẾT LUẬN

Trong một thế giới số hóa ngày càng phát triển, và vai trò của công nghệ thông tin không ngừng gia tăng, việc hiểu và áp dụng Text Mining trở nên cực kỳ quan trọng và có giá trị.

Bài thu hoạch bắt đầu từ quá trình tìm hiểu cơ sở lý luận của Text Mining, xác định khái niệm, đặc điểm cũng như cách thức hoạt động của nó. Điều này giúp nắm bắt được bản chất của lĩnh vực này và nhận thức rằng Text Mining không chỉ đơn giản là một công nghệ, mà còn là một công cụ mạnh mẽ để khai phá thông tin từ văn bản.

Bên cạnh đó, bài thu hoạch đã để cập đến ứng dụng đa dạng của Text Mining, bao gồm các ứng dụng của công nghệ này trong đa dạng lĩnh vực khác nhau như phân tích dữ liệu, tin tức, hay xử lý ngôn ngữ tự nhiên, v.v. thể hiện tầm quan trọng và tính đa dạng của Text Mining trong thế giới hiện đại.

Tiếp đó, bài thu hoạch đã trình bày một số kỹ thuật được sử dụng trong Text Mining, tập trung vào Text Summarization, và nhóm cũng đã lấy kỹ thuật này làm nền tảng cho chương trình cài đặt và chạy thử ở cuối bài thu hoạch. Text Summarization là một kỹ thuật giúp rút gọn thông tin quan trọng từ một văn bản dài thành một bản tóm tắt nội dung ngắn gọn, súc tích giúp làm nổi bật ý chính và nội dung được đề cập.

Bài thu hoạch cũng đã so sánh Text Mining với một số công nghệ nổi bật trong ngành khai phá dữ liệu và làm nổi bật các điểm giống nhau, khác nhau giữa các kỹ thuật. Mỗi công nghệ đều có đặc điểm ưu việt và hạn chế riêng của mình, sự linh hoạt trong việc áp dụng các công nghệ này cho từng tình huống cụ thể giúp chúng ta làm chủ các công nghệ hiện đại trong một thế giới ngày càng tiến bộ.

Cuối cùng, nhóm 1 đã thực hiện cài đặt và chạy thử kỹ thuật Text Summarization trong Text Mining, thể hiện được sự thực tế và ứng dụng của Text Mining trong thực tế, biến lý thuyết trở thành cơ sở vững chắc trong quá trình thực hành.

Kết luận, bài luận này đã đưa ra cái nhìn tổng quan và cụ thể về Text Mining và Text Summarization, và hy vọng rằng giúp mọi người hiểu rõ hơn về lĩnh vực quan trọng này và cách công nghệ này đóng vai trò quan trọng trong việc xử lý và tóm tắt thông tin văn bản trong thời đại số hóa ngày càng phát triển.

TÀI LIỆU THAM KHẢO

[1] Đ. X. Dũng, "Tóm tắt văn bản sử dụng các kỹ thuật trong Deep Learning," Trường Đại học Công nghệ - Đại học quốc gia Hà Nội, Hà Nội, 2018.

[2] B. D. Vijay Kotu, Data Science (Second Edition), 2019.

[3] DataCamp, "Understanding Text Classification in Python," 11/2022. [Trực tuyến]. Available: https://www.datacamp.com/tutorial/text-classification-python.

[4] MonkeyLearn, "What Is Text Mining? A Beginner's Guide," [Trực tuyến]. Available: https://monkeylearn.com/text-mining/.

[5] GeeksforGeeks, "What is Information Retrieval?," 19/9/2023. [Trực tuyến]. Available: https://www.geeksforgeeks.org/what-is-information-retrieval/.

[6] O. Netsuite, "What Is Text Mining & How Does It Work?," 9/6/2022. [Trực tuyến]. Available: https://www.netsuite.com/portal/resource/articles/data-warehouse/text-mining.shtml.

[7] DeepLearning.AI, "Natural Language Processing (NLP) [A Complete Guide]," 11/1/2023. [Trực tuyến]. Available: https://www.deeplearning.ai/resources/natural-language-processing/.

[8] AnalyticsVidhya, "Exploring the Extractive Method of Text Summarization," 4/4/2023. [Trực tuyến]. Available: https://www.analyticsvidhya.com/blog/2023/03/exploring-the-extractive-method-of-text-summarization/.

[9] GeeksforGeeks, "Text Mining in Data Mining," 6/5/2023. [Trực tuyến]. Available: https://www.geeksforgeeks.org/text-mining-in-data-mining/.

[10] AnalyticsStep, "What is Text Mining? Process, Methods and Applications," 3/5/2021. [Trực tuyến]. Available: https://www.analyticssteps.com/blogs/what-text-mining-process-methods-and-applications.