**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**ĐỒ ÁN CUỐI KÌ**

**MÁY HỌC**

*Người thực hiện*: **Trần Thái Hà – 22636801**

**Trần Minh Hiếu – 22689931**

Lớp **: 420300371102**

Khoá  **: 18**

*Người hướng dẫn*: **TS. BÙI THANH HÙNG**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2025**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**ĐỒ ÁN CUỐI KÌ**

**MÁY HỌC**

*Người thực hiện*: **Trần Thái Hà – 22636801**

**Trần Minh Hiếu – 22689931**

Lớp **: 420300371102**

Khoá  **: 18**

Người hướng dẫn: **TS. BÙI THANH HÙNG**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2025**

LỜI CẢM ƠN

Bên cạnh việc báo cáo nội dung cuối kì môn Máy học thì nhóm chúng em xin gửi lời cảm ơn sâu sắc đến với thầy Hùng vì đã tạo cho chúng em điều kiện thuận lợi để phát triển trong ngành Khoa học dữ liệu và hướng dẫn tận tình chúng em trong các bài giảng trên lớp và các bài tập về nhà chất lượng. Ngoài việc đó ra, thì chúng em còn được thử sức với việc làm đồ án, cách làm việc nhóm hiệu quả, trong việc bắt cặp thì chúng em đã có những thấu hiểu, cách tương tác, lắng nghe và trình bày ý kiến của bản thân. Lời cuối cùng nhóm em xin chúc Thầy thật nhiều sức khỏe để có thể dẫn dắt thêm nhiều thế hệ sinh viên trong thời gian sắp tới và chúc cho đề tài cuối kì của các nhóm thành công rực rỡ.

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH**

Tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được sự hướng dẫn của TS. Bùi Thanh Hùng. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Công nghiệp TP Hồ Chí Minh không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày tháng năm*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

Trần Thái Hà

Trần Minh Hiếu

PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

TÓM TẮT

Bài toán tập trung vào phát triển một hệ thống hỏi đáp tự động bằng tiếng Việt trong lĩnh vực y tế, nhằm giải quyết nhu cầu cung cấp thông tin chính xác và nhanh chóng về các bệnh lý. Hệ thống cần nhận diện ý định (intent) của câu hỏi, bao gồm các khía cạnh như triệu chứng, nguyên nhân, và phương pháp điều trị, đồng thời xác định tên bệnh được đề cập trong câu hỏi để trả lời phù hợp. Thách thức nằm ở việc xử lý ngôn ngữ tự nhiên tiếng Việt với sự phức tạp về ngữ nghĩa, cấu trúc câu, và các biến thể ngôn ngữ. Ngoài ra, hệ thống phải khai thác hiệu quả dữ liệu từ ontology y tế để cung cấp câu trả lời đáng tin cậy, phù hợp ngữ cảnh, và có khả năng xử lý các câu hỏi đa ý định (kết hợp nhiều khía cạnh như triệu chứng và nguyên nhân). Mục tiêu là xây dựng một hệ thống không chỉ chính xác mà còn có khả năng mở rộng, hỗ trợ người dùng trong việc tìm hiểu thông tin y tế một cách dễ dàng và hiệu quả.

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc195481827)

[PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc195481828)

[TÓM TẮT iv](#_Toc195481829)

[MỤC LỤC 1](#_Toc195481830)

[DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT 3](#_Toc195481831)

[DANH MỤC CÁC HÌNH VẼ 4](#_Toc195481832)

[DANH MỤC CÁC BẢNG 5](#_Toc195481833)

[XÂY DỰNG HỆ THỐNG GIẢI THÍCH TRÍ TUỆ NHÂN TẠO BẰNG CÁC PHƯƠNG PHÁP HỌC MÁY 6](#_Toc195481834)

[1 Giới thiệu về bài toán 6](#_Toc195481835)

[2 Phân tích yêu cầu của bài toán 6](#_Toc195481836)

[2.1 Yêu cầu của bài toán 6](#_Toc195481837)

[2.2 Các phương pháp giải quyết bài toán 7](#_Toc195481838)

[2.3 Phương pháp đề xuất giải quyết bài toán 8](#_Toc195481839)

[3 Phương pháp giải quyết bài toán 8](#_Toc195481840)

[3.1 Mô hình tổng quát 8](#_Toc195481841)

[3.2 Đặc trưng của mô hình đề xuất 9](#_Toc195481842)

[4 Thực nghiệm 9](#_Toc195481843)

[4.1 Dữ liệu 9](#_Toc195481844)

[4.2 Xử lý dữ liệu 9](#_Toc195481845)

[4.3 Công nghệ sử dụng 9](#_Toc195481846)

[4.4 Cách đánh giá 10](#_Toc195481847)

[5 Kết quả đạt được 10](#_Toc195481848)

[6 Kết luận 10](#_Toc195481849)

[LÀM VIỆC NHÓM 12](#_Toc195481850)

[TÀI LIỆU THAM KHẢO 13](#_Toc195481851)

[PHỤ LỤC 15](#_Toc195481852)

[TỰ ĐÁNH GIÁ (Bài nhóm) 17](#_Toc195481853)

**DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT**

**CÁC KÝ HIỆU**

*f Tần số của dòng điện và điện áp (Hz)*

*p Mật độ điện tích khối (C/m3)*

*N: số mẫu*

*L: số nhãn*

*: nhãn dự đoán*

*: nhãn thực tế*

*1: hàm chỉ báo*

**CÁC CHỮ VIẾT TẮT**

CSTD: Công suất tác dụng

MF: Máy phát điện

BER: Tỷ lệ bít lỗi

RF: Random Forest

SVM: Support Vector Machine

TF-IDF: Term Frequency-inverse Document Frequency

NLP: Natural Language Processing

OWL: Ontology Web Language

RDFlib: Resource Description Framework ( thư viện python để truy vấn và xử lý dữ liệu RDF )

DANH MỤC CÁC HÌNH VẼ

[Hình 1: Kiến Trúc FTP 7](#_Toc9880)

[Hình 2: Kiến trúc TP 8](#_Toc7961)

[Hình 3: Mô hình tổng quát 11](#_Toc14982)

[Hình 4: Dữ liệu 13](#_Toc15843)

[Hình 5: Ma trận nhầm lẫn của Random Forest 16](#_Toc17806)

[Hình 6: Precison, Recall, F1-score của Random Forest 17](#_Toc313)

[Hình 7: Biểu đồ nhiệt Accuracy của Random Forest 17](#_Toc32356)

[Hình 8: Ma trận nhầm lẫn của SVM 18](#_Toc15824)

[Hình 9: Precison, Recall, F1-score của SVM 18](#_Toc17521)

[Hình 10: Biểu đồ nhiệt Accuracy của SVM 19](#_Toc25809)

DANH MỤC CÁC BẢNG

[Bảng 1: Kết quả đạt được 16](#_Toc20582)

Xây dựng hệ thống hỏi đáp tự động tiếng Việt theo hướng tiếp cận Ontology

1.1 Giới thiệu về bài toán

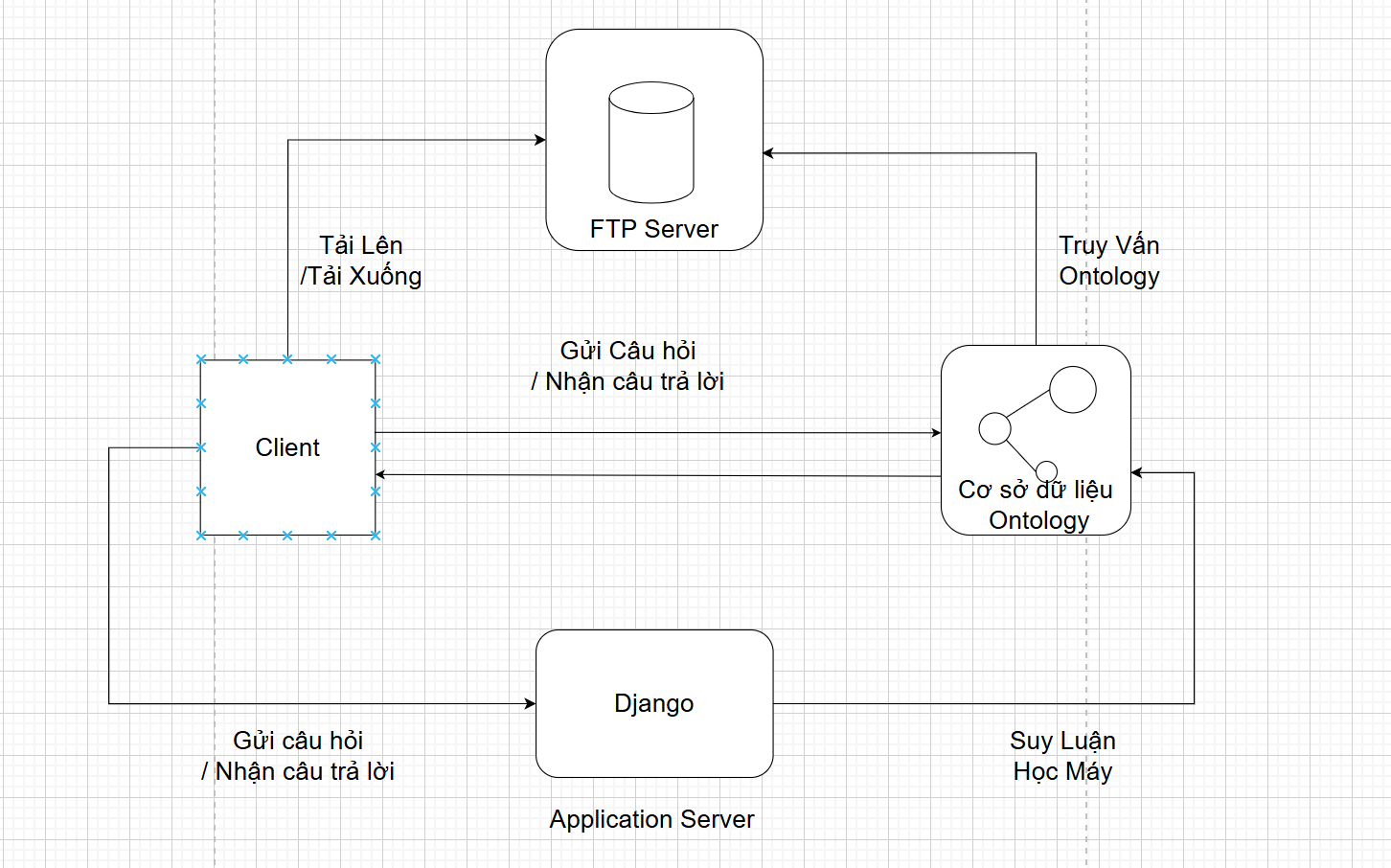
* Bài toán tập trung vào xây dựng một hệ thống hỏi đáp tự động bằng tiếng Việt trong lĩnh vực y tế, sử dụng cách tiếp cận dựa trên ontology để cung cấp thông tin về các bệnh lý. Hệ thống cần nhận diện ý định (intent) của câu hỏi, bao gồm triệu chứng, nguyên nhân, và phương pháp điều trị, đồng thời xác định tên bệnh được đề cập để trả lời chính xác. Hai mô hình học máy được sử dụng để phân loại đa nhãn là Random Forest và Support Vector Machine (SVM), kết hợp với kỹ thuật biểu diễn văn bản TF-IDF và truy vấn ontology y tế. Mục tiêu là xử lý ngôn ngữ tự nhiên tiếng Việt, phân loại chính xác các ý định (đơn hoặc đa ý định), và trích xuất thông tin từ ontology để hỗ trợ người dùng tìm hiểu thông tin y tế một cách hiệu quả.
* Ý nghĩa bài toán:
* Ý nghĩa thực tiễn: Hệ thống giúp người dùng, đặc biệt là những người không có chuyên môn y tế, dễ dàng tiếp cận thông tin về bệnh lý, triệu chứng, nguyên nhân, và cách điều trị bằng ngôn ngữ tự nhiên tiếng Việt. Điều này đặc biệt hữu ích trong bối cảnh nhu cầu tư vấn y tế trực tuyến ngày càng tăng, giúp giảm tải cho nhân viên y tế và nâng cao nhận thức sức khỏe cộng đồng.
* Ý nghĩa khoa học: Bài toán góp phần phát triển các hệ thống xử lý ngôn ngữ tự nhiên tiếng Việt, vốn còn hạn chế so với các ngôn ngữ phổ biến. Việc sử dụng cả Random Forest và SVM để phân loại đa nhãn, kết hợp ontology, mang lại một cách tiếp cận mới cho xử lý câu hỏi phức tạp, mở ra tiềm năng ứng dụng trong các lĩnh vực như giáo dục, pháp luật.
* Ý nghĩa xã hội: Hệ thống có thể hỗ trợ người dân ở vùng sâu, vùng xa hoặc những người gặp khó khăn trong việc tiếp cận dịch vụ y tế, từ đó góp phần cải thiện chất lượng cuộc sống và thúc đẩy bình đẳng trong tiếp cận thông tin y tế.

2 Phân tích yêu cầu của bài toán

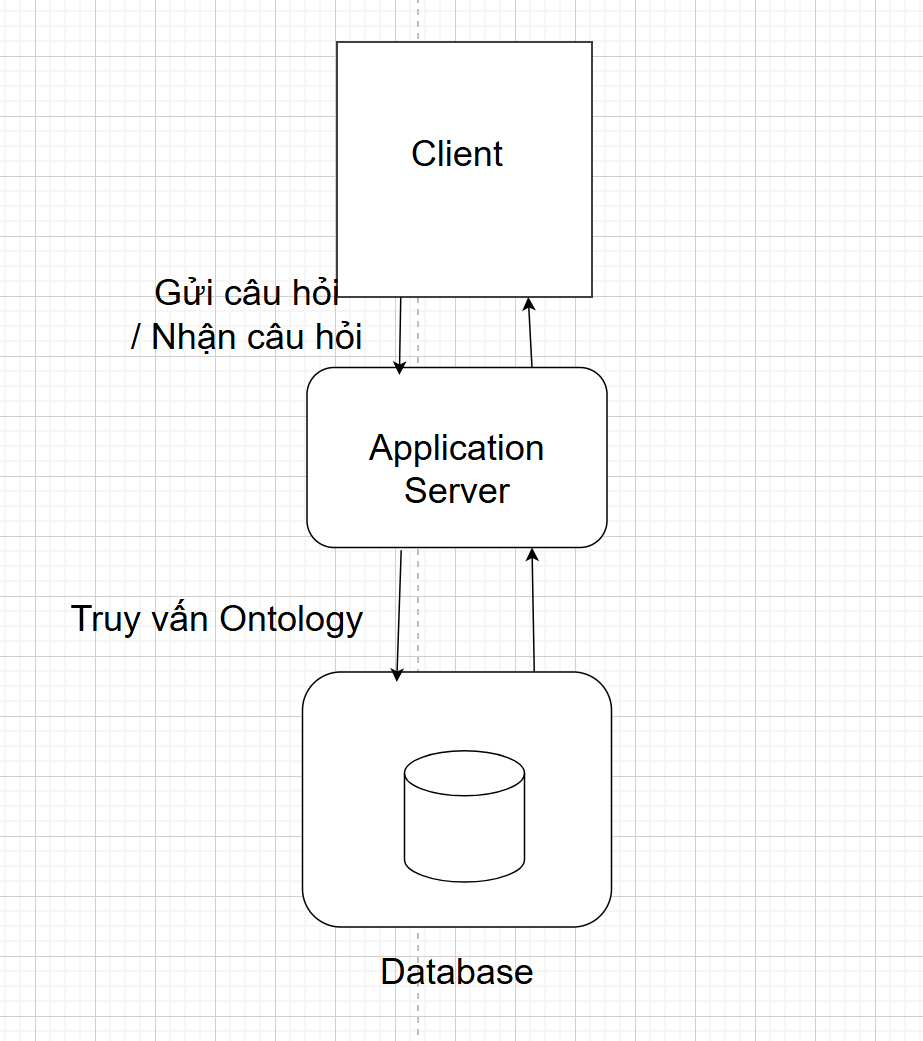
2.1 Yêu cầu của bài toán

Trong bài toán xây dựng hệ thống hỏi đáp tự động tiếng Việt theo hướng tiếp cận Ontology, yêu cầu xây dựng một hệ thống có khả năng:

* Cung cấp câu trả lời chính xác và dễ hiểu dựa trên việc nhận diện ý định (triệu chứng, nguyên nhân, điều trị) và tên bệnh trong các câu hỏi tiếng Việt, sử dụng ontology như cơ sở tri thức.
* Nhận diện và phân loại ý định đa nhãn một cách hiệu quả, bao gồm các câu hỏi có nhiều ý định đồng thời, đồng thời trích xuất chính xác tên bệnh để truy vấn ontology.
* So sánh hiệu quả của các phương pháp phân loại (như Random Forest và SVM) nhằm xác định phương pháp tối ưu cho việc nhận diện ý định và đảm bảo độ chính xác cao trong trả lời.
* Đảm bảo tính minh bạch và dễ sử dụng cho người dùng không chuyên, với các câu trả lời được trình bày rõ ràng, dễ hiểu, và phù hợp với ngữ cảnh y tế tiếng Việt.



Hình 1: Kiến Trúc FTP



Hình 2: Kiến trúc TP

2.2 Các phương pháp giải quyết bài toán

Sử dụng SVM để nhận diện ý định:

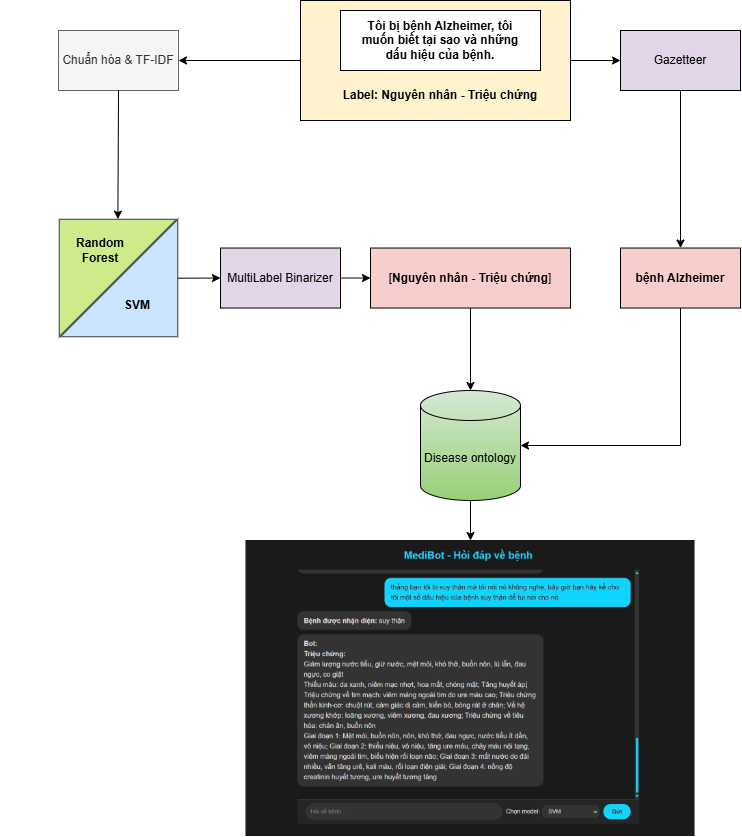
* Tiêu đề bài báo: “TRÍCH CHỌN THÔNG TIN Y TẾ TIẾNG VIỆT CHO BÀI TOÁN TÌM KIẾM NGỮ NGHĨA”
* Tác giả: Trần Thị Ngân
* Năm xuất bản: 2009
* Phương pháp giải quyết:
* Nhận diện thực thể bằng Gazetteer: Sử dụng tập Gazetteer (danh sách từ vựng thực thể y tế như tên bệnh, triệu chứng, thuốc) để so khớp và nhận diện thực thể trong văn bản tiếng Việt.
* Phân loại quan hệ ngữ nghĩa bằng SVM: Đề xuất sử dụng Support Vector Machine (SVM) để gán nhãn 6 quan hệ ngữ nghĩa (LÀ, CÓ, GÂY\_RA, LIÊN\_QUAN, ĐIỀU\_TRỊ, TÁC\_ĐỘNG) cho câu văn bản y tế, nhằm hỗ trợ khai thác thông tin ngữ nghĩa.
* Xây dựng ontology y tế: Xây dựng ontology tiếng Việt với các lớp thực thể (bệnh, triệu chứng, thuốc) và quan hệ ngữ nghĩa, sử dụng công cụ Protégé và ngôn ngữ OWL, để lưu trữ thông tin y tế và hỗ trợ tìm kiếm ngữ nghĩa .
* Dữ liệu thực nghiệm:
* Tổng quan: Văn bản y tế tiếng Việt thu thập từ các nguồn trực tuyến, chú thích thủ công khoảng 1000 câu.
* Nhận diện thực thể: Dữ liệu chú thích chứa các thực thể như tên bệnh, triệu chứng, thuốc, dùng để xây dựng tập Gazetteer.
* Phân loại quan hệ: 1000 câu được chú thích với 6 quan hệ ngữ nghĩa, dự kiến chia 500 câu huấn luyện và 500 câu kiểm thử (chưa thực hiện đầy đủ).
* Ontology: Dữ liệu từ văn bản y tế để xây dựng các lớp thực thể và quan hệ, lưu trữ trong ontology.
* Kết quả đạt được:
* Nhận diện thực thể: Qua 10 lần thực nghiệm, đạt Precision 56.25%-77.06%, Recall 50%-76.92%, F1-Measure 52.94%-71.49% (Bảng 4, Chương 5.4).
* Phân loại quan hệ bằng SVM: Chưa có kết quả cụ thể do chưa thực hiện đầy đủ thí nghiệm, nhưng kỳ vọng hiệu quả cao với dữ liệu chú thích chất lượng.
* Xây dựng ontology: Ontology y tế tiếng Việt được xây dựng với các lớp thực thể và 6 quan hệ, hỗ trợ nhận diện thực thể hiệu quả (F1-Measure như trên), nhưng chưa đánh giá khả năng truy vấn.
* Hạn chế:
* Nhận diện thực thể: Hiệu quả phụ thuộc vào sự tương đồng giữa dữ liệu huấn luyện và kiểm thử; không nhận diện tốt các thực thể mới hoặc biến thể ngoài Gazetteer.
* Phân loại quan hệ bằng SVM: Chưa hoàn thiện thí nghiệm do hạn chế thời gian, dữ liệu tiếng Việt hạn chế về quy mô và độ đa dạng, chi phí tính toán của SVM cao.

2.3 Phương pháp đề xuất giải quyết bài toán

* Trong dự án này, chúng tôi đề xuất sử dụng phương pháp Random Forest để xây dựng hệ thống hỏi đáp tự động tiếng Việt trong lĩnh vực y tế. Cụ thể, Random Forest cung cấp khả năng phân loại đa nhãn nhanh chóng và hiệu quả, tận dụng tập hợp nhiều cây quyết định để xử lý các câu hỏi phức tạp với dữ liệu không đồng nhất, đồng thời tránh overfitting. Phương pháp này mang lại sự linh hoạt và độ chính xác cao trong việc nhận diện ý định (triệu chứng, nguyên nhân, điều trị) từ các đặc trưng TF-IDF.
* Bên cạnh đó, chúng tôi cũng sử dụng phương pháp SVM trong một thực nghiệm riêng biệt để nhận diện ý định. SVM tối ưu hóa ranh giới phân cách, mang lại độ chính xác cao với các đặc trưng TF-IDF, đặc biệt phù hợp cho các câu hỏi yêu cầu phân loại chính xác.

3 Phương pháp giải quyết bài toán

3.1 Mô hình tổng quát



Hình 3: Mô hình tổng quát

3.2 Đặc trưng của mô hình đề xuất

* Bài toán xây dựng hệ thống hỏi đáp tự động tiếng Việt theo hướng tiếp cận ontology, sử dụng cả hai mô hình Random Forest và SVM để phân loại đa nhãn:

1. Nhận diện ý định (intent) của câu hỏi:

* Hệ thống phải phân loại chính xác ý định của câu hỏi, bao gồm ý định đơn (triệu chứng, nguyên nhân, điều trị) và đa ý định (kết hợp các khía cạnh như triệu chứng và nguyên nhân, hoặc triệu chứng, nguyên nhân, điều trị).
* Đảm bảo xử lý các câu hỏi phức tạp với nhiều ý định đồng thời, sử dụng cả Random Forest và SVM để so sánh hiệu quả phân loại đa nhãn.

1. Xác định tên bệnh trong câu hỏi:

* Hệ thống cần trích xuất chính xác tên bệnh được đề cập trong câu hỏi, bất kể cách diễn đạt đa dạng của người dùng (ví dụ: "bệnh Alzheimer", "teo não").
* Xử lý các biến thể ngôn ngữ tiếng Việt, bao gồm dấu, cách viết có hoặc không có từ khóa như "bệnh".

1. Truy vấn ontology để cung cấp câu trả lời:

* Sử dụng ontology y tế để trích xuất thông tin về triệu chứng, nguyên nhân, hoặc điều trị dựa trên tên bệnh và ý định.
* Đảm bảo câu trả lời chính xác, đầy đủ, và phù hợp với ngữ cảnh câu hỏi thông qua truy vấn RDFlib.

1. Xử lý ngôn ngữ tự nhiên tiếng Việt:

* Hiểu và xử lý các câu hỏi tiếng Việt với cấu trúc ngữ pháp đa dạng, từ câu đơn giản đến phức tạp.
* Chuẩn hóa dữ liệu đầu vào (loại bỏ dấu ngoặc, ký tự không cần thiết) để đảm bảo tính thống nhất.

1. Xây dựng và huấn luyện mô hình phân loại ý định:

* Phát triển hai mô hình học máy:
* Random Forest: Sử dụng MultiOutputClassifier với Random Forest để phân loại đa nhãn, tối ưu hóa tham số (n\_estimators, max\_depth) bằng GridSearchCV.
* SVM: Sử dụng MultiOutputClassifier với SVM (kernel rbf/linear), tối ưu hóa tham số (C, gamma, kernel) bằng GridSearchCV.
* Chuyển đổi câu hỏi thành vector bằng TfidfVectorizer, mã hóa nhãn đa ý định bằng MultiLabelBinarizer.

1. Đánh giá hiệu suất hệ thống:

* Đánh giá cả hai mô hình qua các chỉ số: Accuracy, Hamming Loss, Subset Accuracy, Precision, Recall, và F1-score.
* Kiểm tra khả năng nhận diện ý định và trả lời đúng trên danh sách câu hỏi test, so sánh hiệu quả giữa Random Forest và SVM.
* Trực quan hóa kết quả bằng ma trận nhầm lẫn và biểu đồ Precision, Recall, F1-score.

4 Thực nghiệm

4.1 Dữ liệu

* Dữ liệu về disease ontology được thư thập từ các website về y tế như: vinmec, nutrihome, vnvc, nhathuoclongchau, hugging face, …
* Dữ liệu gồm có các cột:
* Question: Chứa câu hỏi về bệnh.
* Name: Chứa tên bệnh.
* Intent: Label mục đích của câu hỏi.
* Scientific\_name: Chứa tên khoa học của bệnh.
* Answer: Câu trả lời cho câu hỏi.



Hình 4: Dữ liệu

4.2 Xử lý dữ liệu

Tiền xử lý dữ liệu, bao gồm:

* Chuyển đổi định dạng: Chuyển cột Intent từ chuỗi thành danh sách.
* Mã hóa nhãn: Sử dụng MultiLabelBinarizer để mã hóa nhãn đa nhãn thành vector nhị phân.
* Biểu diễn văn bản: Chuyển câu hỏi thành vector TF-IDF bằng TfidfVectorizer.
* Chia dữ liệu: Phân chia tập huấn luyện và kiểm thử.
* Chuẩn hóa dữ liệu: Chuẩn hóa tên bệnh và ý định (loại bỏ khoảng trắng, ánh xạ nhãn) để nhận diện tên bệnh và truy vấn ontology.

4.3 Công nghệ sử dụng

Dự án sử dụng ngôn ngữ lập trình Python. Các thư viện chính bao gồm scikit-learn (Random Forest, SVM, TfidfVectorizer, MultiLabelBinarizer), pandas (xử lý dữ liệu), rdflib (truy vấn ontology), và unidecode (chuẩn hóa văn bản). Công cụ Jupyter Notebook được dùng để phát triển và thử nghiệm, cùng với Protégé để xây dựng ontology.

4.4 Cách đánh giá

Dự án xây dựng hệ thống hỏi đáp tự động tiếng Việt trong lĩnh vực y tế sử dụng các độ đo sau để đánh giá hiệu suất mô hình Random Forest và SVM trong phân loại đa nhãn ý định (triệu chứng, nguyên nhân, điều trị):

1. Accuracy (Độ chính xác):

* Ý nghĩa: Tỷ lệ mẫu dự đoán đúng toàn bộ nhãn ý định so với tổng số mẫu.
* Công thức:

1. Hamming loss:

* Ý nghĩa: Tỷ lệ nhãn ý định bị dự đoán sai so với tổng số nhãn.
* Công thức:

(Với N: số mẫu, L: số nhãn, : nhãn dự đoán, : nhãn thực tế, 1: hàm chỉ báo).

1. Subnet Accuracy:

* Ý nghĩa: Tỷ lệ các mẫu được dự đoán đúng toàn bộ tập hợp nhãn ý định.
* Công thức:

1. Precision, Recall, F1-score:

* Ý nghĩa:
* Precision: Tỷ lệ nhãn dự đoán đúng trong số nhãn được dự đoán là Positive.
* Recall: Tỷ lệ nhãn Positive được dự đoán đúng trong số nhãn Positive thực tế.
* F1-score: Trung bình điều hòa của Precision và Recall.
* Công thức:

**5 Kết quả đạt được**

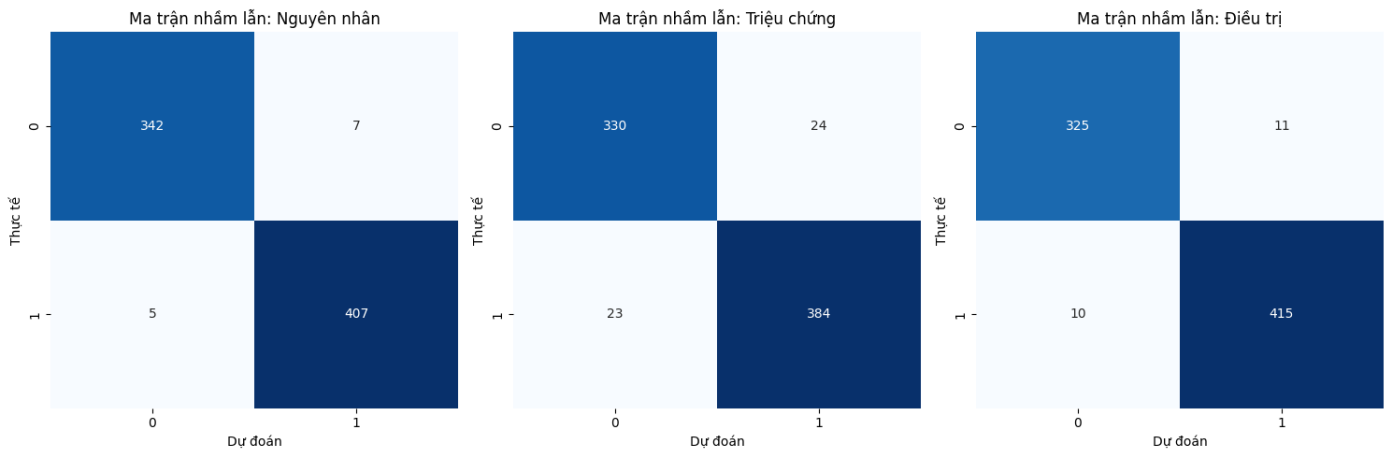
Tham số thực nghiệm:

* Random Forest: n\_estimators=[10, 20, 25], max\_depth=[5, 10, 15], tối ưu: n\_estimators=25, max\_depth=15; dữ liệu TF-IDF (1000 đặc trưng), 3 lớp (Triệu chứng, Nguyên nhân, Điều trị).
* SVM: C=[0.1, 1, 10], gamma=[0.01, 0.1, 1], kernel=[rbf, linear], 18 tổ hợp tham số; dữ liệu TF-IDF, 3 lớp. Cả hai chia dữ liệu 80% huấn luyện, 20% kiểm thử, không dùng epoch hay tầng.

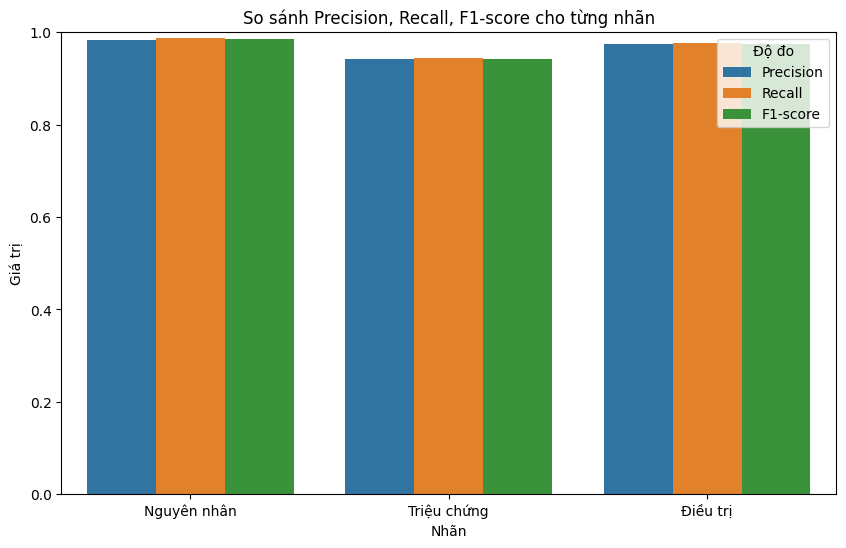
Kết quả đạt được:

|  |  |  |
| --- | --- | --- |
| Phương pháp | Random Forest | SVM |
| Accuracy | 92.38% | 95.27% |
| Hamming Loss | 3.5% | 2.5% |
| Subnet Accuracy | 92.38% | 95.27% |
| Precision | 96.61% | 97.76% |
| Recall | 96.93% | 97.48% |
| F1-Score | 96.76% | 97.61% |

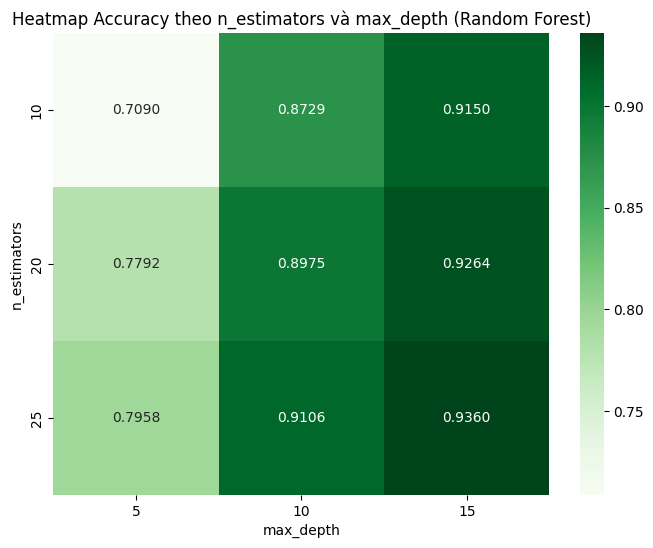
Bảng 1: Kết quả đạt được



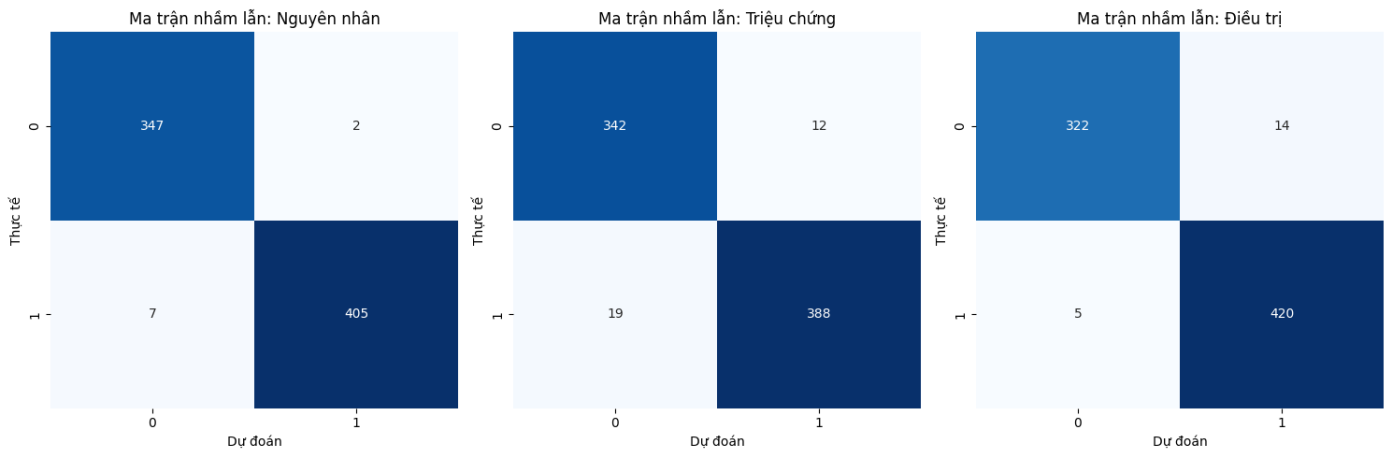
Hình 5: Ma trận nhầm lẫn của Random Forest



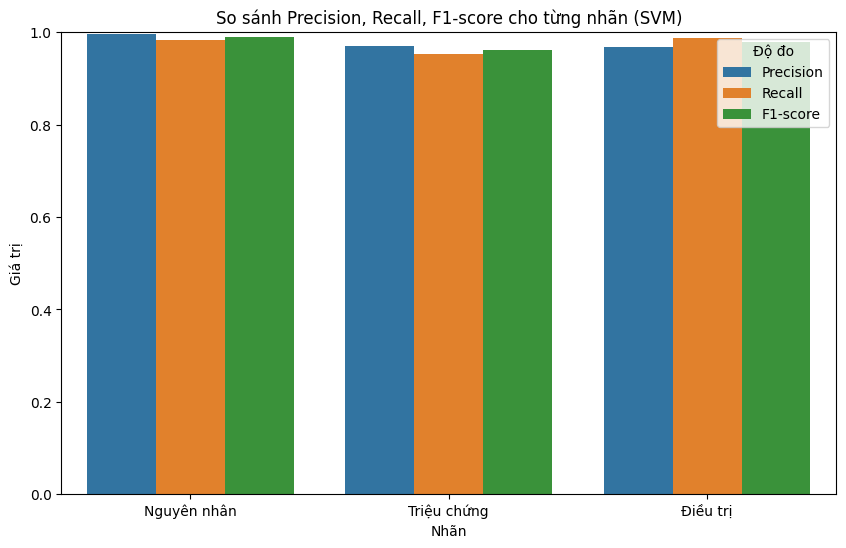
Hình 6: Precison, Recall, F1-score của Random Forest



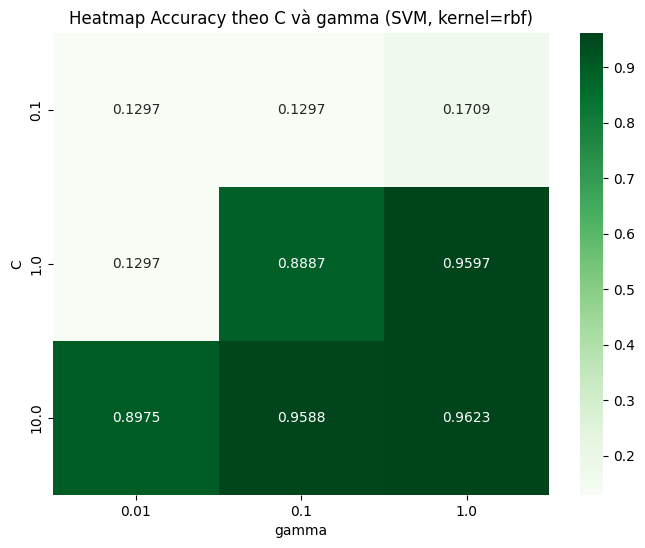
Hình 7: Biểu đồ nhiệt Accuracy của Random Forest



Hình 8: Ma trận nhầm lẫn của SVM



Hình 9: Precison, Recall, F1-score của SVM



Hình 10: Biểu đồ nhiệt Accuracy của SVM

SVM đạt hiệu suất cao nhất (Accuracy 95.27%, F1-score 97.61%):

* Lý do:
* SVM sử dụng kernel rbf (hoặc linear) để ánh xạ dữ liệu TF-IDF vào không gian chiều cao, tối ưu hóa ranh giới phân cách (maximum margin hyperplane), giúp phân biệt tốt các ý định phức tạp, kể cả trong bài toán đa nhãn.
* Chiến lược one-vs-rest trong MultiOutputClassifier đảm bảo mỗi nhãn được huấn luyện độc lập, giảm thiểu lỗi khi dự đoán nhiều ý định đồng thời.
* GridSearchCV với các tham số C, gamma, và kernel được tối ưu hóa, giúp mô hình đạt hiệu suất cao trên dữ liệu kiểm thử.

Random Forest đạt hiệu suất tốt (Accuracy 92.38%, F1-score 96.76%):

* Lý do:
* Random Forest là một mô hình ensemble, sử dụng nhiều cây quyết định để giảm overfitting và xử lý tốt dữ liệu không đồng nhất, như các câu hỏi tiếng Việt với từ vựng và cấu trúc đa dạng.
* Tham số tối ưu (n\_estimators=25, max\_depth=15) được tìm qua GridSearchCV, đảm bảo mô hình đủ sâu và đa dạng để nhận diện ý định.
* Tuy nhiên, Random Forest kém SVM vì không tối ưu hóa ranh giới phân cách chặt chẽ như SVM, dẫn đến một số lỗi nhãn nhỏ (Hamming Loss 3.5% so với 2.5% của SVM).

6 Kết luận

Kết quả đạt được:

* Dự án xây dựng hệ thống hỏi đáp tự động tiếng Việt trong lĩnh vực y tế, sử dụng ontology và phân loại đa nhãn, đạt được kết quả khả quan. Hai mô hình Random Forest và SVM được triển khai để nhận diện ý định (triệu chứng, nguyên nhân, điều trị) và trích xuất tên bệnh. SVM vượt trội với Accuracy 95.27%, Hamming Loss 2.5%, Subset Accuracy 95.27%, và Precision/Recall/F1-score >97%, trong khi Random Forest đạt Accuracy 92.38%, Hamming Loss 3.5%, và Precision/Recall/F1-score>96%. Hệ thống truy vấn ontology trả lời chính xác dựa trên ý định và tên bệnh, đáp ứng nhu cầu cung cấp thông tin y tế dễ hiểu.

Hạn chế của phương pháp:

* Đặc trưng đơn giản: Chỉ sử dụng TF-IDF, thiếu các đặc trưng ngữ nghĩa nâng cao như word embedding, dẫn đến khó xử lý các câu hỏi có cách diễn đạt phức tạp.
* Nhận diện tên bệnh: Chỉ dựa trên so khớp chuỗi, dễ bỏ sót tên bệnh mới hoặc biệt ngữ.
* Ontology giới hạn: Ontology y tế có quy mô nhỏ, chưa được mở rộng tự động, hạn chế khả năng trả lời các câu hỏi đa dạng.
* Chi phí tính toán: SVM có hiệu suất cao nhưng tốn tài nguyên tính toán, gây khó khăn khi xử lý dữ liệu lớn.

Hướng phát triển trong tương lai:

* Mở rộng dữ liệu: Thu thập dữ liệu câu hỏi thực tế từ người dùng và tích hợp các nguồn y tế tiếng Việt để tăng tính đa dạng và thực tế.
* Tích hợp học sâu: Áp dụng các mô hình học sâu như BERT hoặc LSTM để khai thác ngữ nghĩa sâu hơn, cải thiện nhận diện ý định và tên bệnh.
* Cải tiến đặc trưng: Sử dụng word embedding (như Word2vec, fastText) để bổ sung đặc trưng ngữ nghĩa, tăng khả năng xử lý các câu hỏi phức tạp.
* Mở rộng ontology: Phát triển phương pháp tự động hóa mở rộng ontology, bổ sung các khái niệm và quan hệ mới để đáp ứng các câu hỏi đa dạng hơn.
* Tối ưu hóa hiệu suất: Giảm chi phí tính toán của SVM bằng cách sử dụng các kỹ thuật như giảm chiều dữ liệu hoặc thử nghiệm các mô hình nhẹ hơn.

**LÀM VIỆC NHÓM**

Nhóm làm việc theo hình thức online và offline

Phân chia công việc: Bạn Hà sẽ tìm hiểu về hướng giải quyết bài toán và thực hiện với Random Forest. Bạn Hiếu sẽ tìm hiểu về cách sử dụng ontology cho việc truy vấn dữ liệu và hiện thực bài toán với SVM.

Tổng số lần gặp nhau 6(tính theo buổi)

Tổng thời gian gặp nhau 40(tính theo giờ)

**TÀI LIỆU THAM KHẢO**

1. Trần Thị Ngân (2009), “Trích chọn thông tin y tế Tiếng Việt cho bài toán tìm kiếm ngữ nghĩa”.
2. Nguyễn Cẩm Tú. Nhận biết các loại thực thể trong văn bản tiếng Việt nhằm hỗ trợ Web ngữ nghĩa và tìm kiếm hướng thực thể. Khóa luận tốt nghiệp ĐHCN 5/2005, tr. 3, tr.
3. Nguyễn Minh Tuấn. Phân lớp câu hỏi hướng tới tìm kiếm ngữ nghĩa tiếng Việt trong lĩnh vực y tế. Khóa luận tốt nghiệp ĐHCN 5/2008, tr. 2-26.
4. Andreas Vlachos. Evaluating and combining biomedical named entity recognition systems,Computer Laboratory ,University of Cambridge, 2007. [4]. Brandon Beamer, Alla Rozovskaya, Roxana Girju. Automatic Semantic Relation Extraction with Multiple Boundary Generation. University of Illinois at Urbana-Champaign, 2008, tr. 3-4.
5. Brandon Beamer, Alla Rozovskaya, Roxana Girju. Automatic Semantic Relation Extraction with Multiple Boundary Generation. University of Illinois at Urbana-Champaign, 2008, tr. 3-4.

**PHỤ LỤC**

**MỘT SỐ CHÚ Ý KHI VIẾT BÁO CÁO**

1. Thống nhất kích cỡ chữ, kiểu chữ trong toàn bộ báo cáo. Không tô màu chữ, chỉ dùng màu đen
2. Các công thức phải tự gõ và đánh số theo Chương, ví dụ 1.1, 2.1, 2.2, 2.3
3. Các hình và Bảng phải đánh số theo chương, ví dụ Hình 1.1, Hình 2.1, Bảng 3.1, Bảng 3.2
4. Các hình nếu lấy ở ngoài phải đề footnote chú thích nguồn ở dưới
5. Hình mô hình tổng quát phải tự vẽ bằng Word, không dán hình
6. Các tài liệu tham khảo phải đính vào luận văn theo thứ tự từ nhỏ tới lớn, bắt đầu từ 1, ít nhất phải từ 5-15 tài liệu tham khảo, lựa chọn các tài liệu tham khảo mới
7. Tóm tắt trình bày được các nội dung sau: giới thiệu, phương pháp làm, kết quả, nhận xét (không dùng hình, bảng ở mục này)

TỰ ĐÁNH GIÁ (Bài nhóm)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Nội dung | Điểm chuẩn | Tự chấm | Ghi chú |
| 1  (8.5đ) | 1.1 Giới thiệu về bài toán | 0.5 | 0.5 |  |
| 1.2 Phân tích yêu cầu của bài toán | 1 | 1 |  |
| 1.3 Phương pháp giải quyết bài toán | 1.5 | 1.25 |  |
| 1.4 Thực nghiệm | 4 | 3.5 |  |
| **1.5 Kết quả đạt được** | 1 | 1 |  |
| 1.6 Kết luận | 0.5 | 0.5 |  |
| 2  (1đ) | Báo cáo (chú ý các chú ý 2,3,4,6 ở trang trước, nếu sai sẽ bị trừ điểm nặng) | 1đ | 1 |  |
| 3  (0.5đ) | Điểm nhóm (chú ý trả lời các câu hỏi trong mục làm việc nhóm) | 0.5đ | 0.5 |  |
| Tổng điểm | | | 9.25 |  |

TỰ ĐÁNH GIÁ (Bài cá nhân)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Nội dung | Điểm chuẩn | Tự chấm | Ghi chú |
| 1  (9đ) | 1.1 Giới thiệu về bài toán | 0.5 |  |  |
| 1.2 Phân tích yêu cầu của bài toán | 1 |  |  |
| 1.3 Phương pháp giải quyết bài toán | 1.5 |  |  |
| 1.4 Thực nghiệm | 4.5 |  |  |
| **1.5 Kết quả đạt được** | 1 |  |  |
| 1.6 Kết luận | 0.5 |  |  |
| 2  (1đ) | Báo cáo (chú ý các chú ý 2,3,4,6 ở trang trước, nếu sai sẽ bị trừ điểm nặng) | 1đ |  |  |
| Tổng điểm | | |  |  |