

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

ĐỒ ÁN TỐT NGHIỆP

Xây dựng chatbot hỗ trợ du lịch lễ hội tại Việt Nam ứng dụng Web ngữ nghĩa

NGUYỄN THỊ KIỀU THƯƠNG

thuong.ntk153728@sis.hust.edu.vn

Ngành Công nghệ thông tin

Giảng viên hướng dẫn: TS. Đỗ Bá Lâm

Chữ ký của GVHD

Bộ môn: Hệ thống thông tin

Viện: Công nghệ thông tin và truyền thông

HÀ NỘI, 6/2020

ĐỀ TÀI TỐT NGHIỆP

1. Thông tin về sinh viên

Họ và tên sinh viên: Nguyễn Thị Kiều Thương

Điện thoại liên lạc: 0773390954 Email: thuong.ntk153728@sis.hust.edu.vn

Lớp: Việt Nhật - IS2

Hệ đào tạo: Chính quy

Đề án tốt nghiệp được thực hiện tại: Đại học Bách Khoa Hà Nội

Thời gian làm ĐATN: Từ ngày 01/03/2020 đến 20/06/2020

2. Mục đích nội dung của ĐATN

Xây dựng Chatbot hỗ trợ du lịch lễ hội sử dụng công nghệ Web ngữ nghĩa.

3. Các nhiệm vụ cụ thể của ĐATN

- Tìm hiểu, nghiên cứu về công nghệ Web ngữ nghĩa.
- Tìm hiểu và áp dụng bài toán NER cho việc thu thập và xử lý dữ liệu đầu vào.
- Ứng dụng công nghệ Web ngữ nghĩa trong xây dựng Chatbot.

4. Lời cam đoan của sinh viên

Tôi – *Nguyễn Thị Kiều Thương* - cam kết ĐATN là công trình nghiên cứu của bản thân tôi dưới sự hướng dẫn của TS. *Đỗ Bá Lâm*.

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

Hà Nội, ngày 26 tháng 06 năm 2020

Tác giả ĐATN

Nguyễn Thị Kiều Thương

5. Xác nhận của giáo viên hướng dẫn về mức độ hoàn thành của ĐATN và cho phép bảo vệ:

Hà Nội, ngày 26 tháng 06 năm 2020

Giáo viên hướng dẫn

TS. Đỗ Bá Lâm

Lời cảm ơn

Em xin gửi lời cảm ơn đến thầy Đỗ Bá Lâm người trực tiếp hướng dẫn em thực hiện đồ án này cùng toàn thể thầy cô trong trường và tập thể Việt Nhật K60.

Tóm tắt nội dung đồ án

Mục tiêu nghiên cứu của đồ án tốt nghiệp này bao gồm (1) xây dựng kho tri thức ngữ nghĩa về du lịch lễ hội ở Việt Nam; và (2) xây dựng Chatbot sử dụng kho tri thức đã xây dựng ở trên.

Đồ án tốt nghiệp này tập trung nghiên cứu và thực hiện xây dựng kho tri thức ngữ nghĩa về du lịch lễ hội ở Việt Nam. Nguồn dữ liệu được thu thập từ các websites liên quan đến lễ hội ở Việt Nam, đặc biệt là từ Wikipedia. Do nguồn dữ liệu này lớn cũng như có cấu trúc phức tạp, đa dạng nên tôi đã nghiên cứu và sử dụng bài toán NER (Named-entity recognition) để tiền xử lý, làm sạch, và tích hợp các nguồn dữ liệu này. Sau đó tôi tiến hành chuyển đổi dữ liệu tổng hợp ở trên sang dạng dữ liệu ngữ nghĩa theo ontology đã xây dựng. Công cụ xây dựng ontology là Protégé và sử dụng ngôn ngữ truy vấn là SPRAQL.

Dựa trên kho dữ liệu ngữ nghĩa được xây dựng, tôi đã thiết kế và phát triển một ứng dụng chatbot hỗ trợ du lịch lễ hội tại Việt Nam. Công nghệ xây dựng và xử lý chatbot gồm khung hỗ trợ Chatbot Microsoft Bot Framework và dịch vụ xử lý ngôn ngữ tự nhiên LUIS. Chatbot sử dụng công nghệ web ngữ nghĩa thể hiện ưu điểm của web ngữ nghĩa như khả năng suy diễn và tính linh hoạt trong sử dụng dữ liệu. Cấu trúc đồ án gồm 6 chương:

CHƯƠNG 1. GIỚI THIỆU: Trình bày nhiệm vụ đồ án, tóm tắt các nội dung được thực hiện trong đồ án.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT: Trình bày cơ sở lý thuyết và phương pháp áp dụng những lý thuyết đó vào việc giải quyết bài toán.

CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG: Phân tích thiết kế hệ thống, biểu đồ usecase, biểu đồ lớp, biểu đồ hoạt động, biểu đồ trình tự và thiết kế cơ sở dữ liệu.

CHƯƠNG 4. KẾT QUẢ THỰC NGHIỆM: Trình bày các kết quả đạt được khi thực hiện chương trình cũng như đánh giá ưu nhược điểm của hệ thống.

CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN: Tổng kết các kết quả đạt được của đồ án và định hướng phát triển trong tương lai.

Sinh viên thực hiện

Nguyễn Thị Kiều Thương

MỤC LỤC

| | |
|---|-----------|
| CHƯƠNG 1. GIỚI THIỆU | 1 |
| 1.1 Đặt vấn đề | 1 |
| 1.2 Nhiệm vụ được tiến hành thực hiện trong khuôn khổ đề án tốt nghiệp..... | 1 |
| 1.3 Đối tượng sử dụng..... | 2 |
| 1.4 Công nghệ sử dụng..... | 2 |
| CHƯƠNG 2. CƠ SỞ LÝ THUYẾT..... | 3 |
| 2.1 Web ngữ nghĩa (Semantic Web) | 3 |
| 2.1.1 Khái niệm về Web ngữ nghĩa | 3 |
| 2.1.2 Kiến trúc Web ngữ nghĩa | 3 |
| 2.1.3 RDF (Resource Description Language)..... | 4 |
| 2.1.4 Khái niệm về Ontology | 7 |
| 2.1.5 Ngôn ngữ truy vấn SPRAQL..... | 8 |
| 2.2 Nhận dạng thực thể có tên (Name Entity Recognition – NER) | 9 |
| 2.2.1 Khái niệm NER..... | 9 |
| 2.2.2 Các nhãn thực thể và từ loại..... | 10 |
| 2.2.3 Trường điều kiện ngẫu nhiên CRFs..... | 11 |
| 2.3 Chatbot | 11 |
| 2.3.1 Khái niệm Chatbot | 12 |
| 2.3.2 Các thành phần của Chatbot | 12 |
| 2.3.3 Xây dựng Chatbot | 13 |
| CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG..... | 14 |
| 3.1 Mô tả bài toán..... | 14 |
| 3.2 Kiến trúc hệ thống | 14 |
| 3.2.1 Các thành phần xây dựng nên cơ sở tri thức..... | 14 |
| 3.2.2 Các thành phần xây dựng nên Chatbot | 15 |
| 3.3 Xây dựng cơ sở tri thức về lễ hội ở Việt Nam | 15 |
| 3.3.1 Thu thập nguồn dữ liệu | 16 |
| 3.3.2 Định nghĩa danh sách thực thể và quan hệ cần trích rút. | 16 |
| 3.3.3 Mô hình huấn luyện CRFs | 17 |
| 3.3.4 Xây dựng Ontology..... | 19 |
| 3.3.5 Tạo cơ sở tri thức | 23 |
| 3.4 Xây dựng hệ thống Chatbot | 25 |

| | | |
|--|--|-----------|
| 3.4.1 | Đặc tả Use Case | 25 |
| 3.4.2 | Mô tả kịch bản với người dùng | 27 |
| 3.4.3 | Cài đặt và xử lý ngôn ngữ tự nhiên..... | 27 |
| 3.5 | Truy vấn dữ liệu bằng SPARQL..... | 29 |
| CHƯƠNG 4. KẾT QUẢ THỰC NGHIỆM | | 31 |
| 4.1 | Kết quả gán nhãn, trích xuất và thu thập dữ liệu liên quan đến lễ hội..... | 31 |
| 4.1.1 | Kịch bản thử nghiệm..... | 31 |
| 4.1.2 | Kết quả kiểm thử | 31 |
| 4.2 | Kết quả xây dựng cơ sở tri thức | 34 |
| 4.2.1 | Kết quả xây dựng lớp và thuộc tính của Ontology | 34 |
| 4.2.2 | Kết quả chuyển đổi dữ liệu sang dạng có ngữ nghĩa | 35 |
| 4.2.3 | Tổng kết về cơ sở tri thức | 36 |
| 4.3 | Kết quả cài đặt Chatbot hỗ trợ du lịch | 37 |
| 4.3.1 | Lời chào hệ thống..... | 37 |
| 4.3.2 | Chức năng Tìm hiểu lễ hội Việt Nam | 37 |
| 4.3.3 | Chức năng Tìm kiếm các lễ hội | 40 |
| 4.3.4 | Chức năng Gợi ý du lịch lễ hội | 44 |
| 4.3.5 | Đánh giá về Chatbot trên cơ sở Web ngữ nghĩa | 46 |
| CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN..... | | 49 |
| 5.1 | Đánh giá kết quả thực hiện..... | 49 |
| 5.2 | Hướng phát triển trong tương lai | 49 |
| TÀI LIỆU THAM KHẢO | | 51 |

DANH MỤC HÌNH VẼ

| | |
|---|----|
| Hình 1: Kiến trúc phân tầng Web ngữ nghĩa | 3 |
| Hình 2: Mô hình bộ ba RDF | 4 |
| Hình 3: Đồ thị miêu tả mối quan hệ của một bộ ba RDF | 5 |
| Hình 4: Đồ thị miêu tả mối quan hệ sử dụng plain literal..... | 6 |
| Hình 5: Đồ thị miêu tả mối quan hệ sử dụng typed literal..... | 6 |
| Hình 6: Ví dụ về một Ontology..... | 7 |
| Hình 7: Sơ đồ xây dựng cơ sở tri thức | 14 |
| Hình 8: Sơ đồ kiến trúc hệ thống | 15 |
| Hình 9: Sự phân cấp lớp trong Ontology lễ hội..... | 21 |
| Hình 10: Mối quan hệ giữa các lớp trong Ontology lễ hội | 22 |
| Hình 11: Giao diện Cellfie của Protégé | 25 |
| Hình 12: Mô hình kịch bản giữa Chatbot và người dùng | 27 |
| Hình 13: Nhận diện ý định (intent) của người dùng với LUIS | 28 |
| Hình 14: Nhận diện thông tin (entity) từ người dùng với LUIS | 28 |
| Hình 15: Kết quả về một ví dụ truy vấn SPRAQL | 30 |
| Hình 16: Ví dụ một bài báo trên Wikipedia..... | 31 |
| Hình 17: Kết quả sau khi crawl dữ liệu từ một bài viết..... | 32 |
| Hình 18: Kết quả thư mục lưu trữ các bài viết sau khi crawl | 32 |
| Hình 19: Kết quả của mô hình CRFs | 33 |
| Hình 20: Kết quả trích xuất và lưu trữ dữ liệu | 34 |
| Hình 21: Cài đặt các lớp của Ontology | 34 |
| Hình 22: Cài đặt các Object properties của Ontology | 35 |
| Hình 23: Cài đặt các Data property của Ontology | 35 |
| Hình 24: Các thể hiện của lớp lễ hội..... | 36 |
| Hình 25: Các thuộc tính và quan hệ của một thể hiện trong lớp lễ hội..... | 36 |
| Hình 26: Lời chào của Chatbot | 37 |
| Hình 27: Minh họa chức năng Tìm hiểu lễ hội Việt Nam (1)..... | 37 |
| Hình 28: Minh họa chức năng Tìm hiểu lễ hội Việt Nam (2)..... | 38 |
| Hình 29: Ví dụ về tìm kiếm lễ hội theo tên (1) | 39 |
| Hình 30: Ví dụ về tìm kiếm lễ hội theo tên (2) | 39 |
| Hình 31: Chức năng tìm kiếm các lễ hội..... | 40 |
| Hình 32: Ví dụ tìm kiếm lễ hội theo dân tộc..... | 40 |
| Hình 33: Ví dụ tìm kiếm lễ hội thông qua địa điểm | 41 |
| Hình 34: Ví dụ về tìm kiếm lễ hội thông qua mục đích của lễ hội (1) | 41 |
| Hình 35: Ví dụ về tìm kiếm lễ hội thông qua mục đích của lễ hội (2) | 42 |
| Hình 36: Ví dụ về tìm kiếm lễ hội thông qua mục đích (3) | 42 |

| | |
|--|----|
| Hình 37: Ví dụ về tìm kiếm lễ hội thông qua mục đích (4) | 43 |
| Hình 38: Ví dụ về tìm kiếm lễ hội thông qua hoạt động trong lễ hội | 44 |
| Hình 39: : Chức năng gợi ý du lịch lễ hội (1) | 45 |
| Hình 40: Chức năng gợi ý du lịch lễ hội (2) | 46 |
| Hình 41: Minh họa về khả năng suy diễn | 48 |

DANH MỤC HÌNH VẼ

| | |
|---|----|
| Bảng 1: Ví dụ về gán nhãn thực thể có tên | 10 |
| Bảng 2: Danh sách thực thể và nhãn tương ứng | 16 |
| Bảng 3: Danh sách nhãn quan hệ | 17 |
| Bảng 4: Ví dụ về gán nhãn thực thể | 18 |
| Bảng 5: Ví dụ về gán nhãn quan hệ | 18 |
| Bảng 6: Ví dụ một số cột trong file saveAllFes.xlsx | 24 |

CHƯƠNG 1. GIỚI THIỆU

1.1 Đặt vấn đề

Ngày nay, sự phát triển nhanh chóng của Internet đã tạo điều kiện cho mọi người có thể dễ dàng chia sẻ thông tin, tri thức thông qua các bài viết trên các websites về rất nhiều lĩnh vực khác nhau. Từ đó đã góp phần tạo nên một nguồn dữ liệu lớn, phong phú và hết sức hữu ích cho nhân loại. Như một hệ quả, người dùng khi tìm hiểu về một lĩnh vực từ các nguồn tin trên Internet, sẽ thu nhận được một số lượng lớn thông tin trả về. Người dùng, tiếp đó cần phải tự đọc và tự chọn lọc những thông tin có ích đối với bản thân mình, một công việc rất tốn thời gian và công sức. Do vậy, nảy sinh nhu cầu về việc xây dựng các hệ thống có thể hiểu dữ liệu, cung cấp cho người dùng những thông tin phù hợp nhất với nhu cầu tìm kiếm của họ. Web ngữ nghĩa đã ra đời nhằm góp phần giải quyết thách thức này, thông qua việc gắn ngữ nghĩa vào dữ liệu.

Du lịch Việt Nam hiện đang là một ngành rất có tiềm năng phát triển lớn, và đang thu hút được sự quan tâm rộng rãi của người dân. Việc tìm hiểu và xây dựng kho dữ liệu du lịch lễ hội ở Việt Nam sử dụng công nghệ Web ngữ nghĩa, sẽ giúp tạo ra một cơ sở tri thức giàu ý nghĩa, có khả năng suy diễn và hiểu bởi máy tính. Thêm vào đó, đây cũng là bước tiền đề cho những dự án lớn hơn như xây dựng kho tri thức dữ liệu về mọi lĩnh vực trong du lịch Việt Nam, bao gồm không chỉ lễ hội mà còn cả khách sạn, khu nghỉ dưỡng, ẩm thực, giao thông,... góp phần tạo ra trải nghiệm du lịch thông minh, thúc đẩy du lịch, quảng bá văn hóa, đất nước và con người Việt Nam.

Hiện nay chatbot đang là một ứng dụng thu hút sự quan tâm rất lớn của các công ty, tổ chức và chính phủ trong việc cung cấp sự chăm sóc, trải nghiệm khác biệt cho người dùng, nhờ khả năng hoạt động liên tục 24/7. Đây là giải pháp giúp giải quyết gánh nặng cho các tổ chức khi muốn hỗ trợ người dùng tìm hiểu về các thông tin hoạt động của mình. Việc phát triển chatbot hỗ trợ du lịch lễ hội, do vậy cũng sẽ rất cần thiết để cung cấp các thông tin lễ hội tới người dùng thông qua hình thức trao đổi thân thiện và thuận tiện.

Do vậy, trong Đồ án tốt nghiệp này, tôi sẽ thực hiện hai công việc. Đầu tiên, tôi xây dựng cơ sở tri thức về lễ hội ở Việt Nam dựa trên nguồn thông tin trên Internet. Tiếp đó, một ứng dụng Chatbot sẽ được phát triển, sử dụng cơ sở tri thức đã xây dựng, để cung cấp thông tin một cách thông minh đến người dùng.

1.2 Nhiệm vụ được tiến hành thực hiện trong khuôn khổ đồ án tốt nghiệp

1. Nghiên cứu tìm hiểu về web ngữ nghĩa (Semantic Web) cũng như các công cụ để có thể xây dựng và truy vấn cơ sở tri thức ngữ nghĩa.
2. Tìm kiếm thu thập nguồn dữ liệu về lễ hội ở Việt Nam trên Internet. Nghiên cứu kỹ thuật tiền xử lý, làm sạch và lưu trữ dữ liệu.

3. Thực hiện chuyển đổi dữ liệu sau khi đã xử lý ở bước 2 sang dạng dữ liệu ngữ nghĩa dựa trên một Ontology đề xuất.
4. Tìm hiểu các công nghệ để xây dựng chatbot.
5. Tiến hành cài đặt ứng dụng.
6. Đánh giá ưu, nhược điểm của sản phẩm.

1.3 Đối tượng sử dụng

Dành cho những người muốn tìm hiểu về các lễ hội ở Việt Nam hoặc có mong muốn du lịch lễ hội phù hợp với sở thích của bản thân.

1.4 Công nghệ sử dụng

1. Thiết kế Ontology: Ứng dụng Protégé, ngôn ngữ OWL và RDF(s).
2. Truy vấn Ontology: SPRAQL.
3. Xử lý gán nhãn dữ liệu: Thuật toán CRFs.
4. Công nghệ Chatbot: Microsoft Bot Framework.
5. Xử lý ngôn ngữ tự nhiên trong Chatbot: Language Understanding Intelligent Service (LUIS).
6. Ứng dụng xây dựng trên ngôn ngữ: Python.

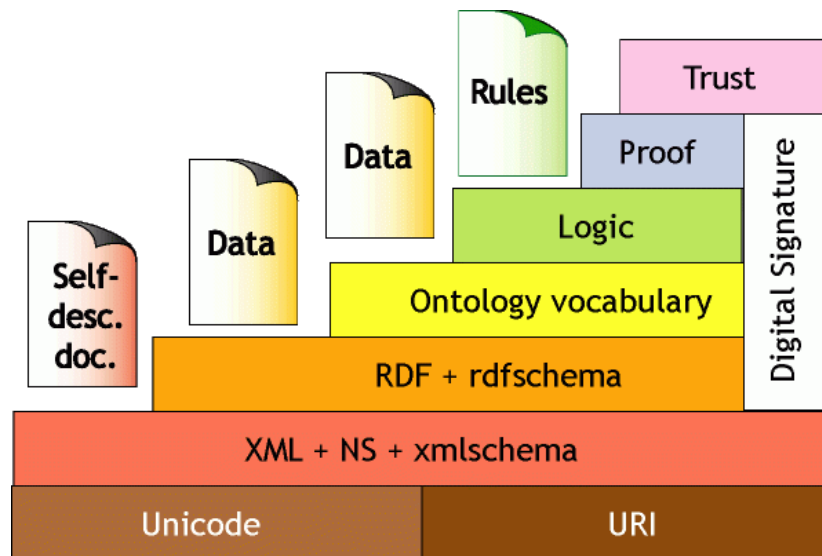
CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1 Web ngữ nghĩa (Semantic Web)

2.1.1 Khái niệm về Web ngữ nghĩa

Khái niệm về Web ngữ nghĩa được đưa ra bởi Tim Berners-Lee - người phát minh ra WWW (World Wide Web), URI, HTTP, và HTML. Web ngữ nghĩa là sự mở rộng Web hiện tại, ở đó thông tin luôn được gắn với một “ngữ nghĩa” xác định – cho phép máy tính và con người cộng tác tốt hơn. (Theo: [1]) Chính vì vậy Web ngữ nghĩa giúp tạo ra một mạng lưới thông tin được liên kết giúp cho máy tính có thể dễ dàng xử lý thông tin. Trong khi đó, kỹ thuật công nghệ web hiện tại chỉ tập trung vào việc lưu trữ và tìm kiếm thông tin đã được lưu trữ. Chính vì vậy sự ra đời của Web ngữ nghĩa đã tạo ra một bước tiến vượt bậc so với Web hiện tại dựa vào khả năng làm việc với thông tin thay vì chỉ đơn thuần là lưu trữ thông tin. Mục tiêu của web có ngữ nghĩa là làm cho dữ liệu trên Web được định nghĩa và liên kết theo một cách thức nào đó để chúng có thể được sử dụng bởi máy tính không chỉ với mục đích hiển thị, mà còn với mục đích tự động hóa, tích hợp và tái sử dụng dữ liệu giữa nhiều ứng dụng khác biệt.

2.1.2 Kiến trúc Web ngữ nghĩa



Hình 1: Kiến trúc phân tầng Web ngữ nghĩa

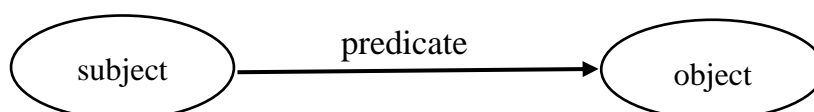
Kiến trúc Web ngữ nghĩa là tập hợp bao gồm các lớp như sau :

- Lớp Unicode & URI: Là lớp đầu tiên của kiến trúc Web ngữ nghĩa, nhằm định danh tài nguyên dựa trên tập kí tự quốc tế.
 - Unicode: là một bảng mã chuẩn quốc tế nhằm thống nhất sự giao tiếp giữa các quốc gia.
 - URI (Uniform Resource Identifier): cung cấp cách thức để định danh các tài nguyên.

- Lớp XML, NS và XMLSchema: cung cấp khả năng tích hợp các thành phần trong web ngữ nghĩa với XML và các khái niệm có liên quan (không gian tên, lược đồ XML).
- Lớp RDF và RDFSchema: là lớp cho phép tạo ra các mô tả ngữ nghĩa về dữ liệu và mối quan hệ giữa chúng. RDF là viết tắt của Resource Description Language là một ngôn ngữ, mô tả các đối tượng và quan hệ giữa chúng. RDF cho phép gán kiểu cho các tài nguyên và làm nền tảng cho Ontology. Dữ liệu được biểu diễn ở dạng bộ ba <subject, predicate, object>. Do vậy, dữ liệu được biểu diễn dưới dạng một đồ thị có hướng và có nhãn. Trong đó, predicate là nhãn trên cạnh, có hướng đi từ subject tới object. RDF SCHEMA cung cấp sự mô tả về các lớp và thuộc tính, mà đối tượng được biểu diễn có liên quan tới.
- Lớp Ontology: là ngôn ngữ được thiết kế để cung cấp khả năng biểu diễn tri thức phong phú và đa dạng hơn so với RDF và RDFS. Cụ thể hơn, RDF bị giới hạn tới việc biểu diễn mối quan hệ giữa hai đối tượng, trong khi RDFS tập trung vào kiến trúc phân tầng của lớp và thuộc tính. Ontology được xây dựng dựa trên RDF, và cung cấp một bộ từ vựng lớn hơn, cũng như có cú pháp đa dạng hơn RDF.
- Lớp Logic: cung cấp các luật suy diễn. Nhờ đó, máy tính có thể thực hiện các suy luận thông qua những nguyên tắc được đưa ra, để thu được thêm những thông tin mới.
- Lớp Proof: sử dụng các luật của lớp Logic để kiểm tra tính đúng đắn của một suy diễn nào đó.
- Lớp Digital Signature: được sử dụng để định danh nguồn của một tài liệu cụ thể, cho phép kiểm tra nội dung của tài liệu có bị thay đổi bởi một bên thứ ba hay không.
- Lớp Trust: nhằm đánh giá mức độ tin cậy và quyết định có nên tin tưởng các nguồn thông tin hay không.

2.1.3 RDF (Resource Description Language)

RDF (Resource Description Language) là một ngôn ngữ, mô tả các đối tượng và quan hệ giữa chúng, đây chính là nền tảng của Web ngữ nghĩa. Dưới đây là mô tả về mô hình của RDF với 3 thành phần chính là subject, predicate và object.



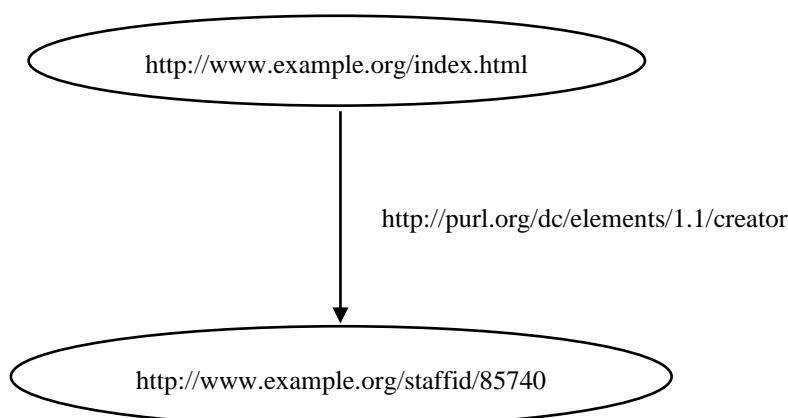
Hình 2: Mô hình bộ ba RDF

- Subject là một tài nguyên mà chúng ta đề cập tới. Nó thường được biểu diễn bởi một URI.
- Predicate mô tả thuộc tính của một tài nguyên. Ví dụ như tiêu đề, tác giả, ... Nó có thể được biểu diễn bởi một URI.
- Object là giá trị của thuộc tính đó (ví dụ: một tác giả có tên Xuân Diệu).

Ví dụ một bộ ba RDF trên <https://www.w3.org/TR/rdf-primer/> (Theo:[10])

“<http://www.example.org/index.html> has a creator whose value is John Smith”

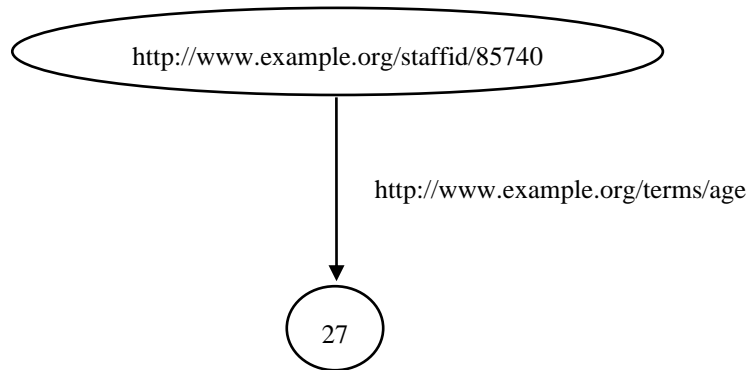
| | |
|-----------|---|
| Subject | http://www.example.org/index.html |
| Predicate | http://purl.org/dc/elements/1.1/creator |
| Object | http://www.example.org/staffid/85740 |



Hình 3: Đồ thị miêu tả mối quan hệ của một bộ ba RDF

- Literal sử dụng để biểu diễn các giá trị như ngày tháng, chữ số... Bất cứ cái gì có thể biểu diễn bởi một giá trị Literal thì đều có thể biểu diễn dưới dạng một URI. Một Literal có thể là object của một phát biểu nhưng không thể là subject hay là predicate. Literal có hai kiểu là plain literal và typed literal.
- Plain literal là kiểu chuỗi, trong ngôn ngữ tự nhiên có thể gọi là kiểu text. Ví dụ về plain literal lấy từ <https://www.w3.org/TR/rdf-primer/> (Theo:[10]) mô tả tuổi của John Smith là 27. Ở đây giá trị tuổi là một kiểu plain literal với hai giá trị nối vào chuỗi là ‘2’ và ‘7’.

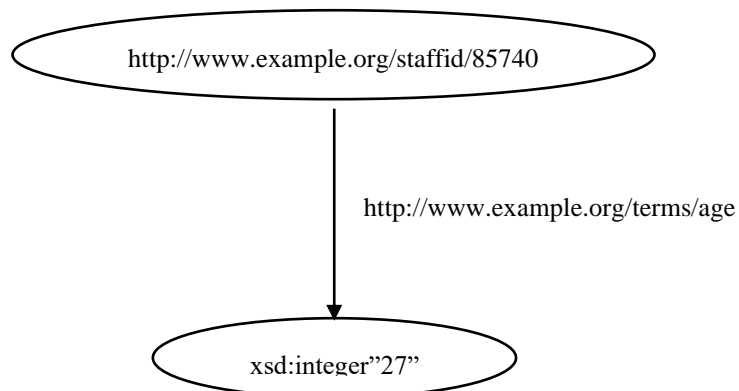
| | |
|-----------|--------------------------------------|
| Subject | http://www.example.org/staffid/85740 |
| Predicate | http://www.example.org/terms/age |
| Object | "27" |



Hình 4: Đồ thị miêu tả mối quan hệ sử dụng plain literal

- Typed literal kết hợp chuỗi với một định danh URI nào đó tạo ra kiểu dữ liệu đặc biệt. Ví dụ về typed literal lấy từ <https://www.w3.org/TR/rdf-primer/> (Theo:[10]) mô tả tuổi của John Smith là 27. Ở đây giá trị tuổi muốn biểu diễn là một con số ta cần dùng kiểu typed literal.

| | |
|-----------|--|
| Subject | http://www.example.org/staffid/85740 |
| Predicate | http://www.example.org/terms/age |
| Object | "27"^^ <http://www.w3.org/2001/XMLSchema#integer> |



Hình 5: Đồ thị miêu tả mối quan hệ sử dụng typed literal

2.1.4 Khái niệm về Ontology

Ontology là một tập các khái niệm và quan hệ giữa các khái niệm được định nghĩa cho một lĩnh vực nhất định nhằm vào việc biểu diễn dữ liệu và máy tính có thể hiểu được. Đây cũng là một hướng tiếp cận để xây dựng lên các bộ cơ sở tri thức trong Web ngữ nghĩa.

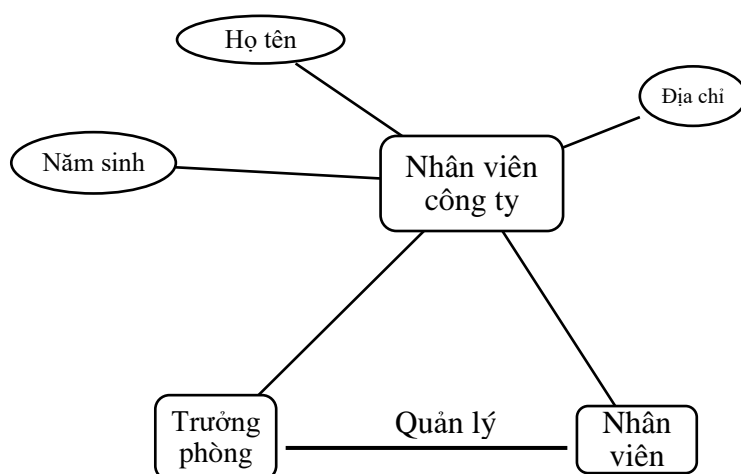
Ontology bao gồm các thành phần chính:

- Một bộ từ vựng mô tả các khái niệm và quan hệ giữa chúng.
- Đặc tả ý nghĩa từ vựng.
- Các ràng buộc mô tả các tri thức bổ sung về lĩnh vực nhất định.

Các khái niệm cơ bản trong Ontology:

- Mệnh đề (Axioms): Các mệnh đề được biểu diễn bởi OWL.
- Thể hiện (Individuals): Là những đối tượng cụ thể đại diện cho lớp.
- Lớp (Classes): Định nghĩa một loại thực thể nhất định, bao gồm các thuộc tính có cùng tính chất chung nào đó. (Ví dụ: giảng viên, sinh viên, lớp học)
- Thuộc tính (Properties): Mô tả đặc điểm của các đối tượng
- Ràng buộc (Restrictions): Mô tả các ràng buộc giữa các thuộc tính trong Ontology. Bao gồm ràng buộc về phạm vi (range) và ràng buộc về miền (domain).
- Quan hệ (Relations): Là mối liên hệ giữa các đối tượng trong Ontology.

Ví dụ về một Ontology miêu tả quan hệ giữa các chức vụ trong một công ty. Khái niệm nhân viên công ty có hai khái niệm con là nhân viên và trưởng phòng.



Hình 6: Ví dụ về một Ontology

Năm sinh, Họ tên, Địa chỉ là các thuộc tính của lớp Nhân viên công ty. Trưởng phòng và Nhân viên là lớp con của lớp Nhân viên công ty. Quản lý là quan hệ giữa 2 lớp Trưởng phòng và Nhân viên

Trong số các phương pháp xây dựng Ontology ta có thể tham khảo phương pháp của Noy và McGuinness đã đề nghị vào năm 2003 với một tập các bước để xây dựng Ontology như sau:

- Bước 1: Quyết định miền và phạm vi của Ontology bằng cách xác định rõ mục đích của ontology và thông tin sẽ được lưu trong ontology.
- Bước 2: Xem xét sự dùng lại của các Ontology đã có: Sự dùng lại những Ontology đã được xây dựng cùng miền tri thức vừa tối thiểu thời gian và công sức vừa có thể đem đến chất lượng cao hơn nếu các ontology đã được kiểm thử kỹ càng.
- Bước 3: Xác định các thuật ngữ quan trọng sẽ sử dụng trong Ontology.
- Bước 4: Định nghĩa các lớp và phân cấp các lớp.
- Bước 5: Định nghĩa các thuộc tính của các lớp.
- Bước 6: Định nghĩa đặc điểm của thuộc tính.
- Bước 7: Bổ sung các thể hiện cho Ontology

Ngôn ngữ sử dụng trong Ontology là OWL (Ontology Web Language) là ngôn ngữ gần như XML dùng để mô tả các hệ cơ sở tri thức, cung cấp các chuẩn nhằm tạo ra một nền tảng quản lý, chia sẻ và tái sử dụng dữ liệu. Dưới đây là một vài cú pháp trong Ontology:

Namespaces: Đây phần dùng để khai báo về các thuật ngữ sẽ được sử dụng. Ví dụ khai báo namespaces trong file festivalVietNam.owl:

```
<rdf:RDF
xmlns:owl= "http://www.w3.org/2002/07/owl#"
xmlns:rdf= "http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs= "http://www.w3.org/2000/01/rdf-schema#"
xmlns:xsd= "http://www.w3.org/2001/XMLSchema#">
```

Ontology Headers: Các thẻ này hỗ trợ các nhiệm vụ quan trọng như nhận xét (rdfs:comment), gán nhãn (rdfs:label), kiểm soát phiên bản (owl:PriorVersion) hay thông báo sự bao gồm các Ontology khác (owl:imports). Ví dụ:

```
<owl:Ontology rdf:about = "">
<rdfs:comment>An example OWL Ontology</rdfs:comment>
<owl:priorVersion rdf:resource = http://www.w3.org/TR/2003/PR-owl-guide-20031215/wine/>
<owl:imports rdf:resource= "http://www.w3.org/TR/2004/REC-owl-guide-20040210/food"/>
<rdfs:label>Wine Ontology</rdfs:label>
```

2.1.5 Ngôn ngữ truy vấn SPRAQL

SPRAQL là một ngôn ngữ truy vấn tài liệu RDF. Giống như SQL cho phép truy xuất, sửa đổi dữ liệu trong cơ sở dữ liệu quan hệ SPRAQL là một ngôn ngữ thực hiện các truy vấn phức tạp trên dữ liệu dạng đồ thị RDF được khuyến nghị

bởi W3C (2008). Truy vấn SPARQL có cấu tạo gồm một tập hợp 3 mẫu (subject, predicate, object). Ví dụ: Nam (a subject) có con là (a predicate) Khoa (a object) là một bộ ba biểu diễn trong đồ thị RDF. Khi đó kết quả của câu truy vấn SPARQL "SELECT ?x ?y ?z" sẽ là "Nam có con là Khoa".

Sức mạnh lớn nhất của SPARQL là khả năng truy vấn thông qua các mối quan hệ trong dữ liệu đồ thị RDF. Trong quá trình này, các mẫu đơn giản có thể được kết hợp thành các mẫu phức tạp hơn, giúp khám phá các mối quan hệ phức tạp hơn trong dữ liệu. Ví dụ: "Tìm cặp tài nguyên (a, b), sao cho tồn tại x là cha của a và b là anh của x". Kết quả mong đợi: cặp bác – cháu.

Cú pháp tổng quát của SPARQL-SELECT có dạng như sau:

PREFIX ns: <namespaceURI>

PREFIX: <.>

SELECT variables

[FROM <dataURI>]

[FROM NAMED <dataURI>]

WHERE {constraints [FILTER] [OPTIONAL]}

[ORDER BY variables] [OFFSET/LIMIT n] [DISTINCT]

Trong đó:

PREFIX: chỉ định tên cho một URI.

SELECT: trả về tất cả hoặc vài giá trị biến theo mệnh đề.

WHERE CONSTRUCT: trả về một đồ thị RDF với các biến liên quan.

DESCRIBE: trả về một "mô tả" của tài nguyên tìm được

ASK: trả về kết quả tìm một mẫu đồ thị có hay không.

WHERE: danh sách, kết nối các mẫu (đồ thị) truy vấn.

OPTIONAL: danh sách, kết nối các mẫu (đồ thị) truy vấn tùy chọn.

AND: biểu thức logic (để lọc các giá trị)

2.2 Nhận dạng thực thể có tên (Name Entity Recognition – NER)

2.2.1 Khái niệm NER

Nhận dạng thực thể có tên (Name Entity Recognition – NER) còn gọi là nhận dạng thực thể định danh hay trích xuất thực thể là một nhiệm vụ con của trích xuất thông tin. Đây là bài toán có nhiệm vụ tìm kiếm, nhận biết và phân loại các thành phần nguyên tử trong một văn bản như địa điểm, thời gian, nhân vật ... Ví dụ: Với đoạn văn bản sau: "*Dịch bệnh này bắt đầu xuất hiện từ tháng 12 năm 2019, với tâm dịch đầu tiên được ghi nhận là thành phố Vũ Hán thuộc tỉnh Hồ Bắc Trung Quốc.*" (Theo: Đại dịch Covid 19 - vi.wikipedia.org)

NER sẽ giúp nhận biết "thành phố Vũ Hán", "tỉnh Hồ Bắc", "Trung Quốc" là các thực thể thuộc nhóm địa điểm, "*tháng 12 năm 2019*" là thời gian.

NER giúp các ứng dụng xử lý ngôn ngữ tự nhiên biết cách tự động trích xuất thực thể định danh trong văn bản. Hiện nay một số phương pháp học máy hoặc học sâu để thực hiện bài toán NER. Trong đề án này tôi sử dụng phương pháp trường điều kiện ngẫu nhiên (Conditional random fields - CRFs).

2.2.2 Các nhãn thực thể và từ loại

Trong quá trình thực hiện bài toán gán nhãn thực thể có tên, bước đầu tiên ta phải tiến hành tách từ và gán nhãn từ loại (danh từ, động từ, tính từ ...) cho từ đó. Sau đó các từ sẽ được xác định thuộc một trong các nhãn thực thể hay quan hệ đã định nghĩa để gán nhãn tương ứng.

Nhãn thực thể được gán theo cấu trúc BIO. Trong đó B là Begin gán cho từ đầu tiên của thực thể, I là Inside gán cho các từ còn lại của thực thể, O là Other gán cho các từ không nằm trong bất kì thực thể nào. Các nhãn loại từ bao gồm có: N (Noun) gán cho danh từ, NP (Proper Noun) gán cho danh từ riêng, V (Ver) gán cho động từ, A (adjective) gán cho tính từ, ... Ví dụ : “*Hội Gióng Phù Đổng chính thống được tổ chức hàng năm vào ba ngày mùng 7, mùng 8 và mùng 9 tháng 4 Âm lịch tại xã Phù Đổng, huyện Gia Lâm, thành phố Hà Nội.*” (Theo: *Hội Gióng* - vi.wikipedia.org)

Ta có B-FES và I-FES là nhãn thực thể tên lễ hội, B-TIM và I-TIM là nhãn thời gian, B-LOC và I-LOC là nhãn địa điểm, ... cùng các nhãn loại từ như sau:

Bảng 1: Ví dụ về gán nhãn thực thể có tên

| Từ | Từ loại | Nhãn |
|--------------------|-----------|--------------|
| <i>Hội</i> | <i>N</i> | <i>B-FES</i> |
| <i>Gióng</i> | <i>Np</i> | <i>I-FES</i> |
| <i>Phù Đổng</i> | <i>Np</i> | <i>I-FES</i> |
| <i>Chính thống</i> | <i>N</i> | <i>O</i> |
| <i>được</i> | <i>V</i> | <i>O</i> |
| <i>tổ chức</i> | <i>V</i> | <i>B-FIT</i> |
| <i>hàng</i> | <i>N</i> | <i>B-TIM</i> |
| <i>năm</i> | <i>N</i> | <i>I-TIM</i> |
| <i>vào</i> | <i>E</i> | <i>I-TIM</i> |
| <i>ba</i> | <i>M</i> | <i>I-TIM</i> |
| <i>ngày</i> | <i>N</i> | <i>I-TIM</i> |
| <i>mùng</i> | <i>V</i> | <i>I-TIM</i> |
| <i>7</i> | <i>M</i> | <i>I-TIM</i> |
| <i>,</i> | <i>CH</i> | <i>I-TIM</i> |
| <i>mùng</i> | <i>N</i> | <i>I-TIM</i> |
| <i>8</i> | <i>M</i> | <i>I-TIM</i> |

| | | |
|-----------|-----------|--------------|
| và | <i>C</i> | <i>I-TIM</i> |
| mùng | <i>N</i> | <i>I-TIM</i> |
| 9 | <i>M</i> | <i>I-TIM</i> |
| tháng | <i>N</i> | <i>I-TIM</i> |
| 4 | <i>M</i> | <i>I-TIM</i> |
| Âm lịch | <i>N</i> | <i>I-TIM</i> |
| tại | <i>E</i> | <i>B-FIL</i> |
| xã | <i>N</i> | <i>B-LOC</i> |
| Phù Đổng | <i>Np</i> | <i>I-LOC</i> |
| , | <i>CH</i> | <i>I-LOC</i> |
| huyện | <i>N</i> | <i>I-LOC</i> |
| Gia Lâm | <i>Np</i> | <i>I-LOC</i> |
| , | <i>CH</i> | <i>I-LOC</i> |
| thành phố | <i>N</i> | <i>O</i> |
| Hà Nội | <i>Np</i> | <i>B-LOC</i> |

2.2.3 Trường điều kiện ngẫu nhiên CRFs

Trường điều kiện ngẫu nhiên CRFs (Conditional random fields) là một dạng của mô hình xác suất thường áp dụng cho dự đoán cấu trúc trong nhận diện mẫu và học máy. CRFs là một kiểu mô hình đồ thị vô hướng xác suất. Nó dùng để mã hóa những mối quan hệ của những mẫu quan sát được và xây dựng nên những đặc tả phù hợp. Đặc biệt CRFs được ứng dụng trong phân tích cấu trúc câu, nhận diện thực thể trong văn bản. (Theo : [2])

Giả sử $G = (V, E)$ là một đồ thị vô hướng, E là tập các cạnh vô hướng và V là tập đỉnh của đồ thị sao cho $Y = \{Y_v | v \in V\}$. Ta có (X, Y) là một trường ngẫu nhiên có điều kiện khi biến Y_v , điều kiện X tuân theo tính chất Markov đối với đồ thị G .

$p(Y_v | X, Y_u, u \neq v)$ là xác suất của biến ngẫu nhiên Y_v ứng với điều kiện X và tất cả các biến ngẫu nhiên khác, u và v là hai đỉnh thuộc đồ thị.

$p(Y_v | X, Y_u, u \sim v)$ là xác suất của biến ngẫu nhiên Y_v ứng với điều kiện X và các biến ngẫu nhiên u, v tương ứng là hai đỉnh kề trong đồ thị là:

Ta có nếu $p(Y_v | X, Y_u, u \neq v) = p(Y_v | X, Y_u, u \sim v)$ thì (X, Y) sẽ là một trường ngẫu nhiên điều kiện.

Quá trình huấn luyện mô hình CRF thực chất là bài toán đi tìm tập tham số của mô hình. Kỹ thuật sử dụng khi học tham số θ là dùng maximum likelihood để học $p(Y_i | X_i; \theta)$. Maximum likelihood là phương pháp ước lượng giá trị tham số bởi những giá trị làm cực đại hóa hàm likehooh. Vì vậy có thể nói việc huấn luyện mô hình CRFs chính là việc đi tìm cực đại của hàm likehooh. (Theo : [2])

2.3 Chatbot

2.3.1 Khái niệm Chatbot

Chatbot là một phần mềm trí tuệ nhân tạo mô phỏng một cuộc trò chuyện với người dùng bằng ngôn ngữ tự nhiên của con người thông qua giao diện của các ứng dụng nhắn tin trên máy tính hay điện thoại.

Chatbot là hình thức để con người và máy tính tương tác một cách tự nhiên nhất. Đây là hình thức thường được các tổ chức sử dụng để cải tiến quy trình phục vụ, nâng cao trải nghiệm cho khách hàng đặc biệt trong thời đại bùng nổ của tin nhắn và các dịch vụ online trên Internet. Một vài ứng dụng của chatbot có thể kể đến như trợ lý cá nhân, tìm kiếm thông tin, giới thiệu sản phẩm, chăm sóc khách hàng, đặt chỗ, thanh toán trực tuyến, ... Khi sử dụng chatbot sẽ mang lại những lợi ích cụ thể như chatbot hoạt động liên tục nên có thể truy cập bất kỳ lúc nào, công suất xử lý lớn, có thể thay đổi linh hoạt, các thao tác được tự động hóa, không dễ xảy ra tình trạng bỏ sót tin nhắn...(Theo: [3])

Hiện nay có nhiều cách tiếp cận cũng như các công cụ khác nhau để phát triển chatbot. Tùy thuộc vào lĩnh vực và mục đích sử dụng chung ta có thể lựa chọn các công nghệ phù hợp

2.3.2 Các thành phần của Chatbot

Chatbot được cấu tạo từ 5 thành phần chính đó là hiểu ngôn ngữ tự nhiên NLU (Natural Language Understanding), hệ thống câu hỏi và câu trả lời, nền tảng điện toán đám mây, các công cụ, ngôn ngữ lập trình Chatbot và hệ thống đầu cuối

NLU là thành phần quan trọng của chatbot đảm nhiệm công việc mô phỏng suy nghĩ của con người để hiểu được ngôn ngữ tự nhiên của con người. Nhờ thành phần này chatbot sẽ diễn giải được những gì người dùng nói, chuyển đổi thông tin từ người dùng thành đầu vào để hệ thống tiếp tục xử lý. NLU bao gồm xử lý ngôn ngữ tự nhiên (Natural Learning Process - NLP) để xác định ý định của câu hỏi và trích xuất thông tin. Từ câu hỏi của người hỏi 3 thành phần chính cần được trích xuất:

- Phân loại tên miền (Domain Classification)
- Phân loại ý định (Intent classification)
- Trích xuất thông tin (Entity extraction)

Hệ thống câu hỏi và câu trả lời là thành phần chính để chatbot có thể đưa ra câu trả lời. Quá trình sinh câu trả lời bao gồm sinh thủ công và tự động. Sinh thủ công là tạo ra danh sách các câu hỏi thường gặp và ánh xạ các câu trả lời của nó. Điều này giúp chatbot nhanh chóng xác định câu trả lời cho những câu hỏi quan trọng. Sinh tự động bao gồm việc tải các tài liệu có sẵn của công ty, tổ chức như các tài liệu liên quan đến chính sách và các loại tài liệu hỏi đáp khác lên bot và yêu cầu nó tự đào tạo. Công cụ đưa ra một danh sách các câu hỏi và các câu trả lời liên quan đến các loại tài liệu này. Sau đó, bot có thể trả lời chúng một cách tự tin. (Theo: [3])

Nền tảng điện toán đám mây (Cloud Platform): Đây là nơi lưu trữ và phát triển các hệ thống AI được coi là phần “cơ thể vật lý” của Chatbot. Hiện nay các platform phổ biến nhất là Google Cloud Platform, Messenger Platform, Microsoft Azure Learning, ...

Các công cụ, ngôn ngữ lập trình Chatbot: Các nền tảng API.AI, WIT.AI, ... hỗ trợ xây dựng bởi nhiều ngôn ngữ lập trình như Python, C++, Android, IOS, Node.js, C#, .NET, Ruby, Java, Rất nhiều công ty cung cấp nền tảng điện toán đám mây như Google, Facebook, Amazon cho phép bạn sử dụng công cụ của họ để xây dựng hệ thống Chatbot và phát triển hệ thống AI. (Theo: [4])

Hệ thống đầu cuối có thể là bất cứ nền tảng nào có giao diện tiếp cận người dùng. Ngoài những chatbot được tích hợp ngay trên sản phẩm, ta có thể kết nối chatbot trên các ứng dụng như Facebook, Slack, Skype, ...

2.3.3 Xây dựng Chatbot

a. Hiểu ý định người dùng

Việc xác định ý định người dùng (Intent) từ những câu giao tiếp đầu vào rất quan trọng. Để xây dựng được các lớp ý định ta cần đưa ra tập các câu hỏi với các cách diễn đạt khác nhau thể hiện ý định đó để huấn luyện. Ví dụ: Người dùng có ý định “*Hỏi số ca nhiễm mới Covid19 hàng ngày*”, Intent này có thể thể hiện qua những câu hỏi sau của người dùng:

Hôm nay có bao nhiêu ca nhiễm mới ở Việt Nam?

Số ca nhiễm Covid ngày hôm nay trên thế giới?

Số người dương tính với Covid19 trong ngày tại Mỹ?

Đến hết 24h ngày hôm nay thế giới có ghi nhận thêm ca nhiễm mới nào không?

b. Trích xuất thông tin

Từ câu đầu vào của người dùng ta có thể trích xuất được thông tin qua việc nhận biết các lớp Entity đã được định nghĩa và huấn luyện trước đó. Ví dụ Entity “*phạm vi lãnh thổ*” bao gồm các thể hiện như *Thế giới, Châu Á, Châu Âu, Mỹ, Việt Nam, Trung Quốc, Pháp ...* Khi đó với câu đầu vào “*Hôm nay có bao nhiêu ca nhiễm mới ở Việt Nam?*” ta xác định được một Entity “*phạm vi lãnh thổ*” là *Việt Nam*. Từ đó hệ thống sẽ lấy ra những thông tin liên quan đến lãnh thổ này để trả câu trả lời cho người dùng.

c. Quản lý hội thoại

Quản lý hội thoại để đảm bảo việc trao đổi giữa người và máy được thông suốt. Trong quá trình trao đổi dài giữa người và chatbot, chatbot có nhiệm vụ ghi nhớ các ngữ cảnh (context) và quản lý các trạng thái hội thoại (dialog state) từ đầu vào là thành phần NLU và đầu ra là thành phần sinh ngôn ngữ NLG.

CHƯƠNG 3. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

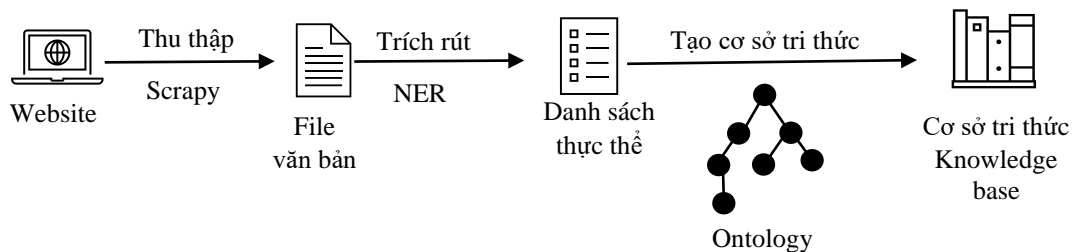
3.1 Mô tả bài toán

Với cơ sở lý thuyết ở trên tôi xin trình bày các bước xây dựng hệ thống Chatbot hỗ trợ lễ hội trên cơ sở Web ngữ nghĩa.

- Bước 1: Tìm kiếm, thu thập và sử dụng nguồn dữ liệu về các lễ hội ở Việt Nam bằng phương pháp Crawl dữ liệu từ các trang web như Wikipedia hay các trang web du lịch Việt Nam lưu lại dưới các tệp .txt.
- Bước 2: Xây dựng tập nhãn theo các trường thông tin cần thiết. Sử dụng mô hình CRFs để huấn luyện và gán nhãn cho dữ liệu. Trích rút dữ liệu theo các trường nhãn.
- Bước 3: Thiết kế và xây dựng Ontology về lĩnh vực du lịch lễ hội
- Bước 4: Chuyển đổi dữ liệu tổng hợp ở bước 2 sang dạng dữ liệu ngữ nghĩa theo Ontology đã xây dựng ở bước 3, để tạo ra một cơ sở tri thức về du lịch lễ hội.
- Bước 5: Thiết kế, xây dựng khung Chatbot. Xác định các loại ý định và thông tin từ câu vào của người dùng.
- Bước 6: Sử dụng SPARQL để truy vấn cơ sở tri thức đã xây dựng và trả kết quả cho Chatbot. Từ đó, đưa câu trả lời cho người dùng.

3.2 Kiến trúc hệ thống

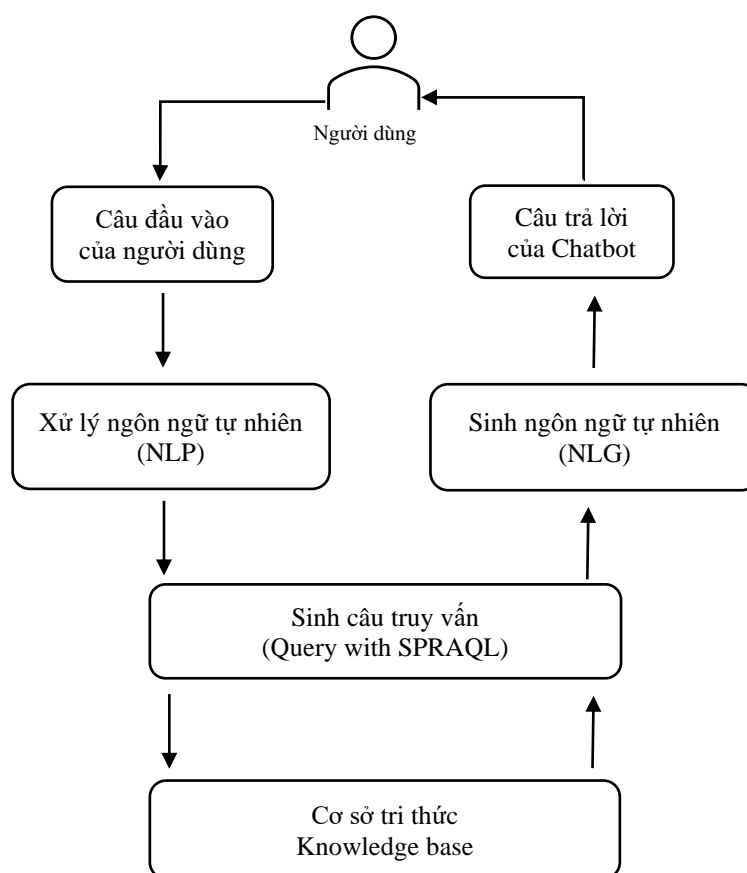
3.2.1 Các thành phần xây dựng nên cơ sở tri thức



Hình 7: Sơ đồ xây dựng cơ sở tri thức

- Thu thập: Quá trình tìm kiếm các thông tin, bài viết trên Internet sau đó lưu trữ dữ liệu dưới các file văn bản .txt. Công cụ để thu thập dữ liệu từ các Website được sử dụng ở đây là thư viện Python Scrapy.
- Trích rút: Sau khi thu thập dữ liệu từ các trang web, tiến hành trích xuất các thực thể cần thiết thông qua bài toán nhận dạng thực thể có tên. (Chi tiết mục 3.3. Thu thập và xử lý dữ liệu)
- Tạo cơ sở tri thức: Các thực thể sau khi trích xuất được chuyển đổi sang dạng ngữ nghĩa bộ ba RDF (subject, predicate và object) dựa trên mô tả trong Ontology và trở thành các thể hiện của Ontology. Tập hợp tất cả các bộ ba RDF tạo thành cơ sở tri thức.

3.2.2 Các thành phần xây dựng nên Chatbot



Hình 8: Sơ đồ kiến trúc hệ thống

- Xử lý ngôn ngữ tự nhiên (NLP): Sử dụng LUIS (Language Understanding Intelligent Service) dịch vụ thiết kế nâng cao hiểu biết cho bot, huấn luyện để nhận diện ý định (intent) và các thực thể (entity) được cung cấp bởi Microsoft.
- Sinh ngôn ngữ tự nhiên (NLG): Việc sinh câu trả lời cho người dùng hoàn toàn phụ thuộc vào các thuộc tính và thể hiện trong Ontology. Dữ liệu được lưu trữ dưới dạng bộ ba subject-predicate-object, vì vậy câu trả lời cũng mang dạng “ngôn ngữ tự nhiên” chính là giá trị trả về từ các câu truy vấn.
- Truy vấn SPRAQL: Sử dụng thư viện Python RDFLib bởi Dan Krech. Một thư viện RDF, cung cấp một API để thao tác trên các lược đồ RDF, phân tích cú pháp RDF, và lưu trữ dữ liệu dạng triple.
- Cơ sở tri thức (Knowledge base): Cơ sở tri thức sẽ được xây dựng theo mô tả ở Hình 7.

3.3 Xây dựng cơ sở tri thức về lễ hội ở Việt Nam

3.3.1 Thu thập nguồn dữ liệu

Đây là bước đầu tiên trong quy trình tạo cơ sở tri thức. Dữ liệu đầu vào phải đảm bảo cả về số lượng và chất lượng. Số lượng được đánh giá bằng số lượng lễ hội được tìm kiếm và thu thập thông tin. Chất lượng được đánh giá bằng độ chính xác về các thông tin được thu thập. Để đảm bảo hai yêu cầu trên tôi đã lựa chọn thu thập các khoảng 70 bài viết về lễ hội khác nhau được đăng trên cổng thông tin điện tử Wikipedia. Ngoài ra thực hiện thu thập bài viết từ một vài trang web khác trên Internet. Sau đây là các nguồn dữ liệu chính đã được sử dụng:

- https://vi.wikipedia.org/wiki/Lễ_hội_Việt_Nam
- https://vi.wikipedia.org/wiki/Thể_loại:Lễ_hội
- <https://www.maxreading.com/sach-hay/le-hoi-truyen-thong.html>

Các bài viết về mỗi lễ hội sau khi thu thập được lưu trữ trong các tệp .txt dạng văn bản.

3.3.2 Định nghĩa danh sách thực thể và quan hệ cần trích rút.

Các thực thể và quan hệ cần trích rút sẽ chính là các thể hiện của Ontology lễ hội. Có tất cả 20 nhãn thực thể và 12 nhãn quan hệ đã được định nghĩa. Sau đây là bảng danh sách các thực thể và quan hệ cần thu thập và trích rút trong mỗi bài báo viết về một lễ hội.

Bảng 2: Danh sách thực thể và nhãn tương ứng

| stt | Tên thực thể | Mô tả | Nhãn |
|-----|----------------------------|---|------|
| 1 | Lễ hội | Lễ hội được nhắc đến trong văn bản | FES |
| 2 | Hoạt động vui chơi | Các hoạt động vui chơi được tổ chức và nhắc đến trong lễ hội | AC1 |
| 3 | Hoạt động tham quan | Các hoạt động tham quan được tổ chức và nhắc đến trong lễ hội | AC2 |
| 4 | Hoạt động lịch sử | Các hoạt động mang tính lịch sử được tổ chức và nhắc đến trong lễ hội | AC3 |
| 5 | Hoạt động tín ngưỡng | Các hoạt động tín ngưỡng được tổ chức và nhắc đến trong lễ hội | AC4 |
| 6 | Hoạt động văn hóa dân gian | Các hoạt động văn hóa dân gian được tổ chức và nhắc đến trong lễ hội | AC5 |
| 7 | Địa điểm | Nơi diễn ra lễ hội | LOC |
| 8 | Thời gian | Thời gian tổ chức lễ hội | TIM |
| 9 | Nhân vật | Nhân vật nhắc tới trong lễ hội | PER |
| 10 | Danh hiệu | Danh hiệu lễ hội đó được công nhận | REC |
| 11 | Tổ chức | Tổ chức công nhận danh hiệu cho lễ hội | ORG |
| 12 | Mục đích | Mục đích của lễ hội là quảng bá du lịch, | CO1 |

| | | | |
|----|--------------------|--|-----|
| | quảng bá | văn hóa, làng nghề ... | |
| 13 | Mục đích tâm linh | Mục đích tổ chức lễ hội liên quan đến tín ngưỡng tâm linh... | CO2 |
| 14 | Mục đích tưởng nhớ | Mục đích tưởngr nhớ đến một vị thần hay anh hùng dân tộc | CO3 |
| 15 | Dân tộc | Lễ hội của dân tộc nào | ETH |
| 16 | Tôn giáo | Lễ hội của tôn giáo nào | REL |
| 17 | Tên khác | Các tên gọi khác của lễ hội | ANN |
| 18 | Danh lam | Danh lam thắng cảnh ở địa điểm tổ chức lễ hội | FAM |
| 19 | Lịch sử | Thông tin liên quan đến sự hình thành của lễ hội. | FFT |
| 20 | Đặc điểm | Thông tin về danh lam | CER |

Bảng 3: Danh sách nhân quan hệ

| stt | Thực thể | Mối quan hệ | Thực thể | Nhãn |
|-----|-----------|------------------|-----------|------|
| 1 | Lễ hội | tổ chức tại | Địa điểm | FIL |
| 2 | Lễ hội | tổ chức vào | Thời gian | FIT |
| 3 | Lễ hội | có | Lịch sử | FFT |
| 4 | Lễ hội | được công nhận | Danh hiệu | FHR |
| 5 | Lễ hội | bao gồm | Hoạt động | FHE |
| 6 | Lễ hội | có | Mục đích | FHC |
| 7 | Danh hiệu | công nhận bởi | Tổ chức | RBO |
| 8 | Lễ hội | của | Dân tộc | FOE |
| 9 | Lễ hội | liên quan tới | Tôn giáo | FOR |
| 10 | Lễ hội | có liên quan đến | Nhân vật | FHP |
| 11 | Lễ hội | có | Tên khác | FHA |
| 12 | Danh lam | có | Đặc điểm | AHC |

3.3.3 Mô hình huấn luyện CRFs

Sử dụng mô hình CRF của thư viện python sklearn ta sẽ xác định kiểu từ loại và thêm nhãn trực tiếp vào các từ sau khi được phân tách để xây dựng “tập huấn luyện”. Sau đó tiến hành gán nhãn thực thể, mỗi thực thể bao gồm B (begin) I (Inside) và O (Other) nếu không thuộc thực thể nào. Sau đây là ví dụ về việc gán nhãn từ loại (postagging) và nhãn thực thể (label) cho các từ (word) sau khi được tách ra từ một đoạn văn.

Bảng 4: Ví dụ về gán nhãn thực thể

| Từ | Từ loại | Nhãn |
|-----------|----------------|--------------|
| Lễ hội | <i>N</i> | <i>B-FES</i> |
| Chùa | <i>N</i> | <i>I-FES</i> |
| Hương | <i>Np</i> | <i>I-FES</i> |

B-FES là từ bắt đầu cho thực thể “Lễ hội”, nhãn I-FES là phần còn lại của thực thể. Ta xác định được Lễ hội Chùa Hương là một thực thể “Lễ hội”.

Tương tự khi xuất hiện các quan hệ giữa 2 thực thể ta cũng tiến hành gán nhãn quan hệ cho các từ thể hiện quan hệ.

Bảng 5: Ví dụ về gán nhãn quan hệ

| Từ | Từ loại | Nhãn |
|----------------|----------------|--------------|
| Lễ hội | <i>N</i> | <i>B-FES</i> |
| Chùa | <i>N</i> | <i>I-FES</i> |
| Hương | <i>Np</i> | <i>I-FES</i> |
| hàng | <i>N</i> | <i>B-TIM</i> |
| năm | <i>N</i> | <i>I-TIM</i> |
| đều | <i>R</i> | <i>O</i> |
| diễn | <i>V</i> | <i>B-FIT</i> |
| ra | <i>V</i> | <i>I-FIT</i> |
| từ | <i>E</i> | <i>B-TIM</i> |
| 6 | <i>M</i> | <i>I-TIM</i> |
| tháng Giêng | <i>N</i> | <i>I-TIM</i> |
| cho | <i>E</i> | <i>I-TIM</i> |
| đến | <i>E</i> | <i>I-TIM</i> |
| hết | <i>V</i> | <i>I-TIM</i> |
| tháng | <i>N</i> | <i>I-TIM</i> |
| Ba | <i>M</i> | <i>I-TIM</i> |
| âm lịch | <i>N</i> | <i>I-TIM</i> |
| . | <i>CH</i> | <i>O</i> |

Ta thấy trong ví dụ FIT là nhãn quan hệ “tổ chức vào” nó xác định quan hệ giữa hai thực thể “Lễ hội” và “Thời gian”.

Để đánh giá mô hình huấn luyện, “tập huấn luyện” sau khi được gán nhãn sẽ chia theo tỉ lệ 7:3 trong đó 70% huấn luyện và 30% để phục vụ cho việc kiểm thử.

3.3.4 Xây dựng Ontology

a. Đặc tả Ontology

Mục đích chính của Ontology là mô tả một cách có ngữ nghĩa các thông tin liên quan đến các lễ hội ở Việt Nam. Các phép suy diễn sẽ giúp cung cấp các thông tin tổng hợp cho người dùng. Đối tượng sử dụng là tất cả những ai có mong muốn tìm hiểu về các lễ hội ở Việt Nam, hay muốn lựa chọn những lễ hội để tham gia khi đi du lịch ở Việt Nam. Phạm vi của Ontology sẽ được xây dựng trong khuôn khổ đồ án bao gồm thông tin của lễ hội ở Việt Nam cung cấp về tên lễ hội, địa điểm, thời gian, hoạt động, mục đích, danh hiệu, tổ chức công nhận danh hiệu, tôn giáo, dân tộc, ... Và có khả năng mở rộng trong tương lai.

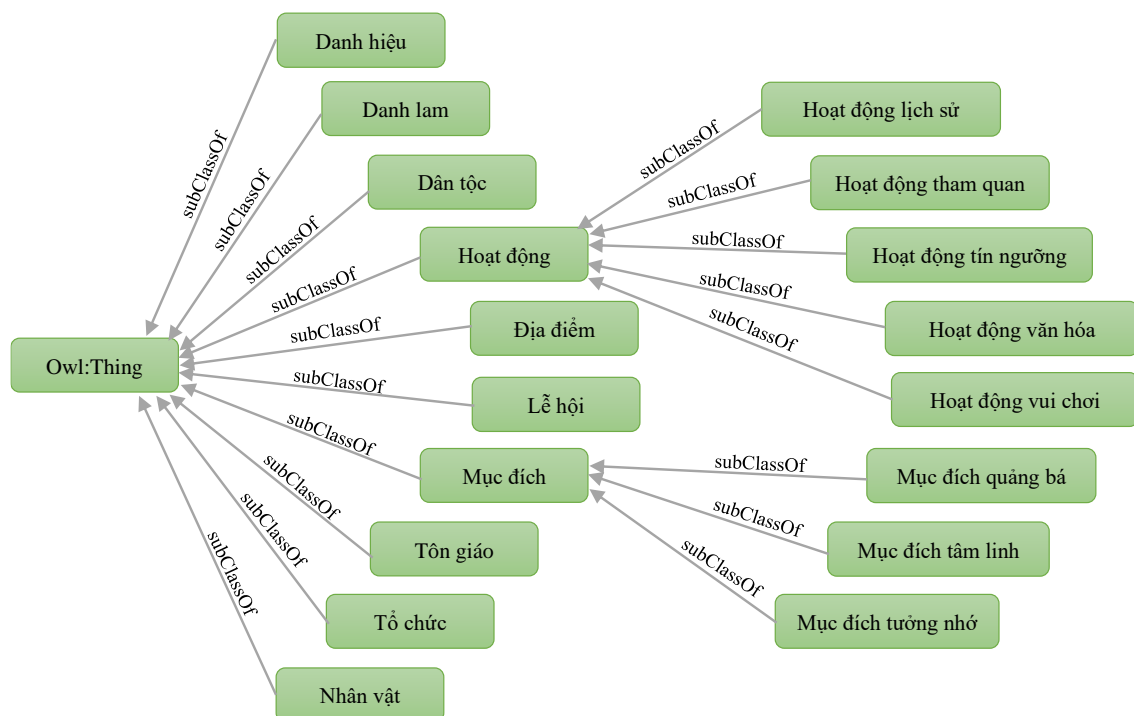
b. Tạo lớp trong Ontology

Sau khi phân tích mục đích, phạm vi của Ontology lễ hội, em định nghĩa đối tượng thông qua các lớp sau:

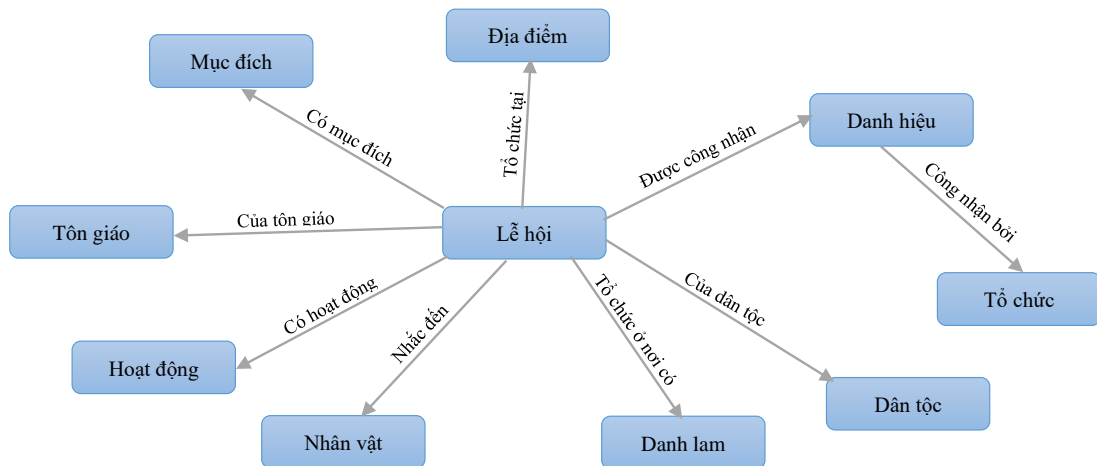
- **Lớp leHoi:**
 - Là lớp lưu trữ tên lễ hội.
 - Được gán nhãn: `<rdfs:label>lễ hội</rdfs:label>`
- **Lớp hoạtDong:**
 - Là lớp lưu trữ thông tin về các hoạt động của lễ hội.
- **Lớp hoạtDongVuiChoi:**
 - Là lớp con của lớp hoạtDong. Chứa các thông tin về những hoạt động hoạt động tổ chức trò chơi dân gian, hoạt động thể thao trong lễ hội.
 - Được gán nhãn:
 - `<rdfs:label>hoạt động vui chơi</rdfs:label>`
 - `<rdfs:label>trò chơi dân gian</rdfs:label>`
 - `<rdfs:label>hoạt động thể thao</rdfs:label>`
- **Lớp hoạtDongThamQuan:**
 - Là lớp con của lớp hoạtDong. Chứa các thông tin về những hoạt động hoạt động tham quan danh lam thắng cảnh.
 - Được gán nhãn: `<rdfs:label>tham quan du lịch</rdfs:label>`
- **Lớp hoạtDongLichSu:**
 - Là lớp con của lớp hoạtDong. Chứa các thông tin về những Hoạt động lịch sử diễn ra gắn với lễ hội
 - Được gán nhãn: `<rdfs:label>sự kiện lịch sử</rdfs:label>`
- **Lớp hoạtDongTinNguong:**

- Là lớp con của lớp `hoatDong`. Chứa các thông tin về những Hoạt động tín ngưỡng diễn ra gắn với lễ hội
- Được gán nhãn: `<rdfs:label> hoạt động tín ngưỡng </rdfs:label>`
- **Lớp `hoatDongVanHoa`:**
 - Là lớp con của lớp `hoatDong`. Chứa các thông tin về những Hoạt động văn hóa dân gian diễn ra gắn với lễ hội
 - Được gán nhãn: `<rdfs:label>văn hóa dân gian</rdfs:label>`
- **Lớp `diaDiem`:**
 - Là lớp lưu trữ thông tin về địa điểm tổ chức lễ hội.
 - Được gán nhãn: `<rdfs:label>địa điểm</rdfs:label>`
- **Lớp `thoiGian`:**
 - Là lớp lưu trữ thông tin về thời gian diễn ra lễ hội.
 - Được gán nhãn: `<rdfs:label>thời gian</rdfs:label>`
- **Lớp `toChuc`:**
 - Là lớp lưu trữ thông tin về các tổ chức tập thể đứng ra chứng nhận danh hiệu cho lễ hội hay danh lam thắng cảnh.
 - Được gán nhãn: `<rdfs:label>tổ chức</rdfs:label>`
- **Lớp `nhanVat`:**
 - Là lớp lưu trữ thông tin về nhân vật được nhắc tới trong các lễ hội
 - Được gán nhãn: `<rdfs:label>nhân vật</rdfs:label>`
- **Lớp `danhHieu`:**
 - Là lớp lưu trữ thông tin về danh hiệu mà lễ hội hay danh lam được chứng nhận
 - Được gán nhãn: `<rdfs:label>danh hiệu</rdfs:label>`
- **Lớp `danToc`:**
 - Là lớp lưu trữ thông tin về dân tộc có lễ hội được nhắc đến.
 - Được gán nhãn: `<rdfs:label>dân tộc</rdfs:label>`
- **Lớp `tonGiao`:**
 - Là lớp lưu trữ thông tin về tôn giáo trong lễ hội.
 - Được gán nhãn: `<rdfs:label>tôn giáo</rdfs:label>`
- **Lớp `danhLam`:**
 - Là lớp lưu trữ về danh lam thắng cảnh tại các địa điểm có lễ hội.
 - Được gán nhãn: `<rdfs:label>danh lam</rdfs:label>`
- **Lớp `mucDich`:**
 - Là lớp lưu trữ mục đích tổ chức lễ hội.
 - Được gán nhãn: `<rdfs:label>mục đích</rdfs:label>`
- **Lớp `mucDichQuangBa`:**

- Là lớp con của lớp **mucDich**. Là lớp lưu trữ thông tin về việc tổ chức lễ hội nhằm quảng bá du lịch thông qua việc tôn vinh một cảnh đẹp, đặc sản, hay tập quán đặc biệt ... của một địa phương hay đất nước, vd như các festival
- Được gán nhãn:
 - `<rdfs:label>quảng bá du lịch</rdfs:label>`
 - `<rdfs:label>tôn vinh văn hóa</rdfs:label>`
- **Lớp **mucDichTamLinh**:**
 - Là lớp con của lớp **mucDich**. Là lớp lưu trữ thông tin về việc tổ chức lễ hội nhằm thể hiện mong muốn cầu những điều may mắn.
 - Được gán nhãn: `<rdfs:label>mục đích tín ngưỡng</rdfs:label>`
- **Lớp **mucDichTuongNho**:**
 - Là lớp con của lớp **mục đích**. Là lớp lưu trữ thông tin về việc tổ chức lễ hội nhằm thể hiện lòng biết ơn đến một nhân vật lịch sử, hay nhân vật trong truyền thuyết xa xưa.
 - Được gán nhãn: `<rdfs:label>mục đích tưởng nhớ</rdfs:label>`
- **Lớp **tenKhac**:**
 - Là lớp lưu trữ một tên gọi khác của lễ hội
 - Được gán nhãn: `<rdfs:label>tên khác</rdfs:label>`



Hình 9: Sự phân cấp lớp trong Ontology lễ hội



Hình 10: Mối quan hệ giữa các lớp trong Ontology lễ hội

c. Tạo thuộc tính trong Ontology

✓ Object properties:

- Thuộc tính **toChucTai**:
 - Domain: leHoi
 - Range: diadiem
- Thuộc tính **coHoatDong**:
 - Domain: leHoi
 - Range: hoatDong
- Thuộc tính **duaDanToc**:
 - Domain: leHoi
 - Range: danToc
- Thuộc tính **duaTonGiao**:
 - Domain: leHoi
 - Range: tonGiao
- Thuộc tính **toChucONoiCo**:
 - Domain: leHoi
 - Range: danhLam
- Thuộc tính **duocCongNhan**:
 - Domain: leHoi
 - Range: danhHieu
- Thuộc tính **nhamHuongDen**:
 - Domain: leHoi
 - Range: mucDich
- Thuộc tính **nhacDen**:

- Domain: leHoi
- Range: nhanVat
- Thuộc tính **congNhanBoi**:
 - Domain: danhHieu
 - Range: toChuc

✓ **Data properties**

- Domain: leHoi - Các thuộc tính mô tả thông tin về lễ hội gồm
 - Thuộc tính **tenLeHoi**
 - Thuộc tính **tenKhac**
 - Thuộc tính **thoiGianToChuc**
 - Thuộc tính **lichSu**
 - Thuộc tính **linkChiTiet**
- Domain: hoatDong - Các thuộc tính mô tả thông tin về Hoạt động gồm:
 - Thuộc tính **tenHoatDong**
- Domain: mucDich - Các thuộc tính mô tả thông tin về Mục đích gồm:
 - Thuộc tính **noiDungMucDich**
- Domain: diaDiem - Các thuộc tính mô tả thông tin về địa điểm gồm:
 - Thuộc tính **tenDiaDiem**
- Domain: danhHieu - Các thuộc tính mô tả thông tin về danh hiệu gồm:
 - Thuộc tính **tenDanhHieu**
- Domain: nhanVat - Các thuộc tính mô tả thông tin về danh Nhân vật gồm:
 - Thuộc tính **tenNhanVat**
- Domain: danhLam - Các thuộc tính mô tả thông tin về danh lam gồm:
 - Thuộc tính **tenDanhLam**
 - Thuộc tính **dacDiemVeDanhLam**
- Domain: tonGiao - Các thuộc tính mô tả thông tin về tôn giáo gồm:
 - Thuộc tính **tenTonGiao**
- Domain: danToc - Các thuộc tính mô tả thông tin về dân tộc gồm:
 - Thuộc tính **tenDanToc**

3.3.5 Tạo cơ sở tri thức

Sau khi các thực thể được trích xuất từ các đoạn văn bản ở bước “3.3.1. Thu thập và xử lý dữ liệu” sẽ được lưu lại với nhãn tương ứng. Sau khi được chuyển đổi sang dạng ngữ nghĩa dựa trên mô tả trong Ontology và trở thành các thể hiện

của Ontology. Tuy nhiên số lượng thể hiện cần thêm là rất lớn, nếu làm theo cách thủ công sẽ vô cùng mất thời gian, công sức và dễ dàng gặp sai sót trong quá trình thực hiện. Chính vì vậy cần phải có một giải pháp chuyển đổi dữ liệu vào cơ sở tri thức một cách tự động. Điều này không chỉ giúp quá trình thêm mới thể hiện vào Cơ sở tri thức trở nên nhanh chóng mà mỗi lần cần cập nhật hay bổ sung các thể hiện ở những lần tiếp theo cũng dễ dàng hơn.

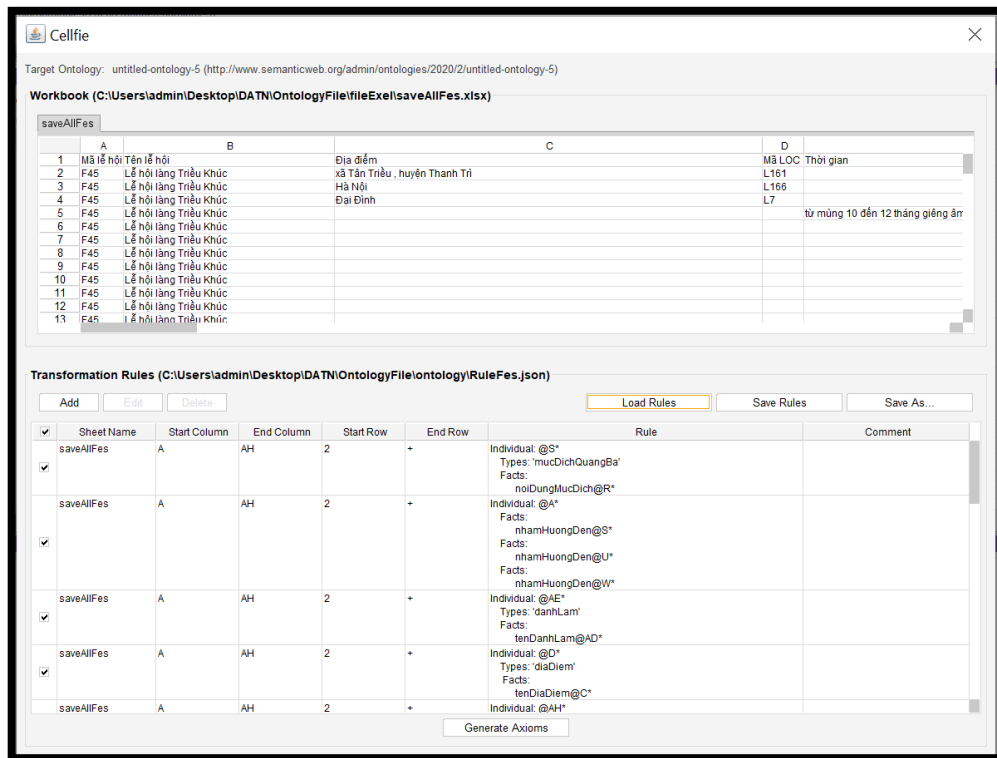
Đầu tiên ta tiến hành đọc nhãn dán để lưu các thực thể tương ứng sang dạng bảng (các file excel) bằng thư viện csv của Python.

Tiếp đó nhờ sự hỗ trợ của Cellfie - một Protégé Desktop plugin hỗ trợ đọc dữ liệu từ các trang bảng tính (các file excel) vào cơ sở tri thức dựa trên luật (rule) được định nghĩa trước ta có tự động tạo thể hiện vào Ontology. Các luật sau khi được định nghĩa có thể lưu lại sử dụng mỗi khi muốn thêm các thể hiện mới. Ví dụ về một vài trường của lễ hội sau khi được trích xuất và lưu lại:

Bảng 6: Ví dụ một số cột trong file saveAllFes.xlsx

| Mã lễ hội | Tên lễ hội | Địa điểm | Mã Địa điểm | Thời gian | Hoạt động du lịch | Mã AC2 | Hoạt động vui chơi | Mã AC1 |
|-----------|------------------------|-------------------------------|-------------|---|--------------------|--------|--------------------|--------|
| F45 | Lễ hội làng Triều Khúc | xã Tân Triều, huyện Thanh Trì | L161 | từ mùng 10 đến 12 tháng giêng âm lịch | | | trò chơi dân gian | AA84 |
| F45 | Lễ hội làng Triều Khúc | | | | | | đấu vật | AA88 |
| F45 | Lễ hội làng Triều Khúc | | | | | | múa rồng | AA20 |
| F20 | lễ hội chùa Bái Đính | tỉnh Ninh Bình | L39 | từ chiều ngày mùng 1 tết, khai mạc ngày mùng 6 tết và kéo dài đến hết tháng 3 | tham quan chùa | AB18 | trò chơi dân gian | AA84 |
| F20 | lễ hội chùa Bái Đính | | | | thăm thú hang động | AB6 | | |

Dưới đây là giao diện Cellfie của Protégé mỗi khi thêm các thẻ hiện cho Cơ sở tri thức bằng các luật:



Hình 11: Giao diện Cellfie của Protégé

3.4 Xây dựng hệ thống Chatbot

3.4.1 Đặc tả Use Case

a. Chức năng “Tìm hiểu lễ hội Việt Nam”

Mô tả: Chức năng cung cấp thông tin liên quan đến một lễ hội nhất định ở Việt Nam như:

- Địa điểm diễn ra
- Thời gian tổ chức
- Các tên gọi khác về lễ hội
- Thời gian hình thành lễ hội
- Các hoạt động diễn ra tại lễ hội
- Mục đích của lễ hội
- Danh hiệu lễ hội được công nhận
- Link bài viết chi tiết về lễ hội

Các thông tin trên sẽ được hiển thị khi người dùng nhập tên của lễ hội. Có những lễ hội được biết đến bằng tên gọi khác nhau, trong cơ sở tri thức trong quá trình thu thập thông tin cũng sẽ cố gắng thu thập các tên gọi khác của một lễ hội để tăng khả năng tìm kiếm.

b. Chức năng “Tìm kiếm lễ hội”

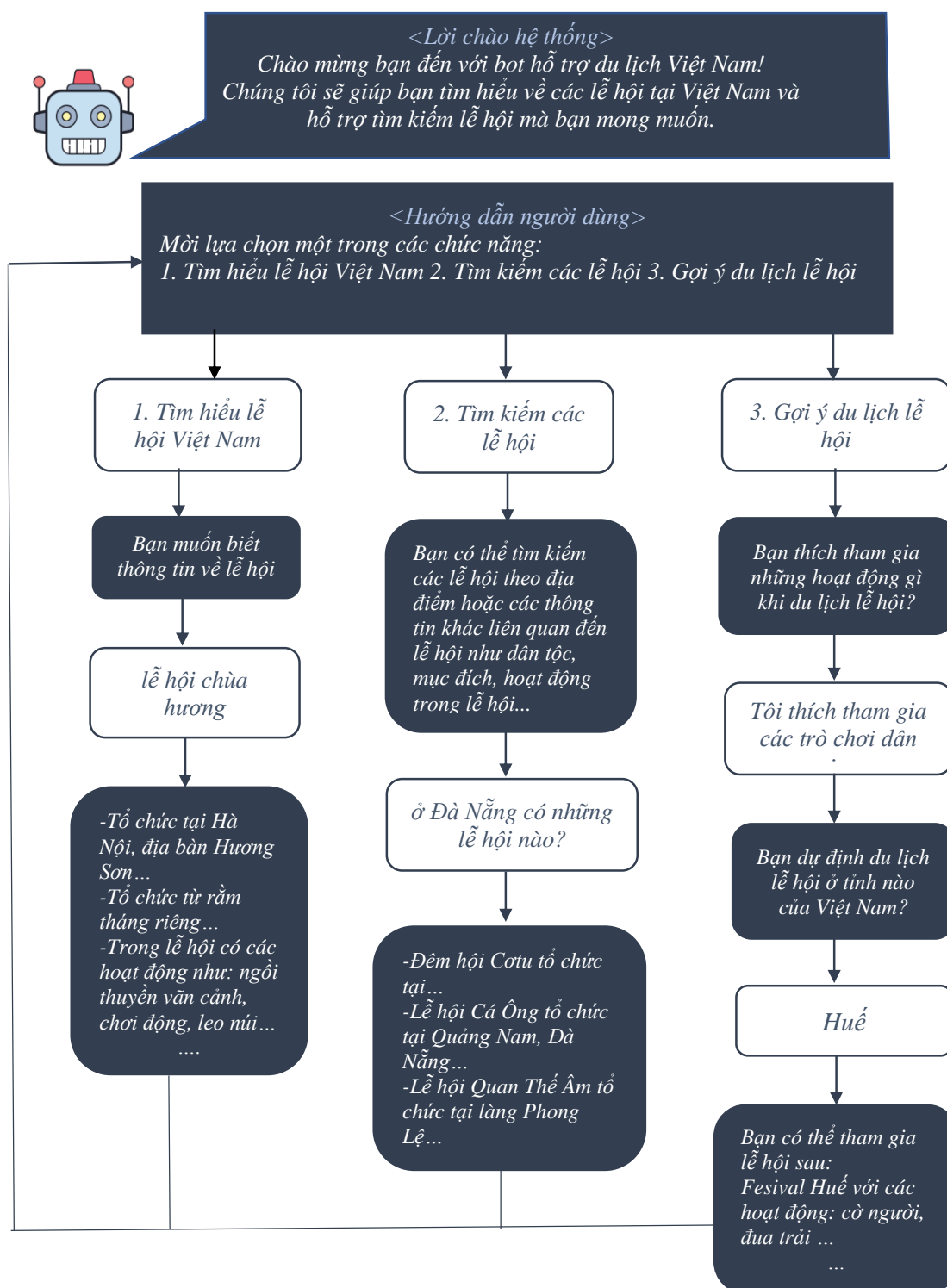
Mô tả: Tìm kiếm lễ hội mà người dùng quan tâm dựa trên thông tin liên quan xung quanh lễ hội như địa điểm, mục đích, dân tộc, hoạt động.

c. Chức năng “Gợi ý du lịch lễ hội”

Mô tả: Từ những thông tin của người dùng về những hoạt động muốn được tham gia khi đi du lịch lễ hội như tham quan du lịch, tìm hiểu về thêm về các trò chơi dân gian, các loại hình văn hóa truyền thống hay những sự kiện có tính lịch sử kết hợp với địa điểm người dùng muốn đến, chatbot sẽ giới thiệu những lễ hội phù hợp cho người dùng tham khảo.

3.4.2 Mô tả kịch bản với người dùng

Kịch bản Chatbot với người dùng được mô tả qua sơ đồ dưới đây:

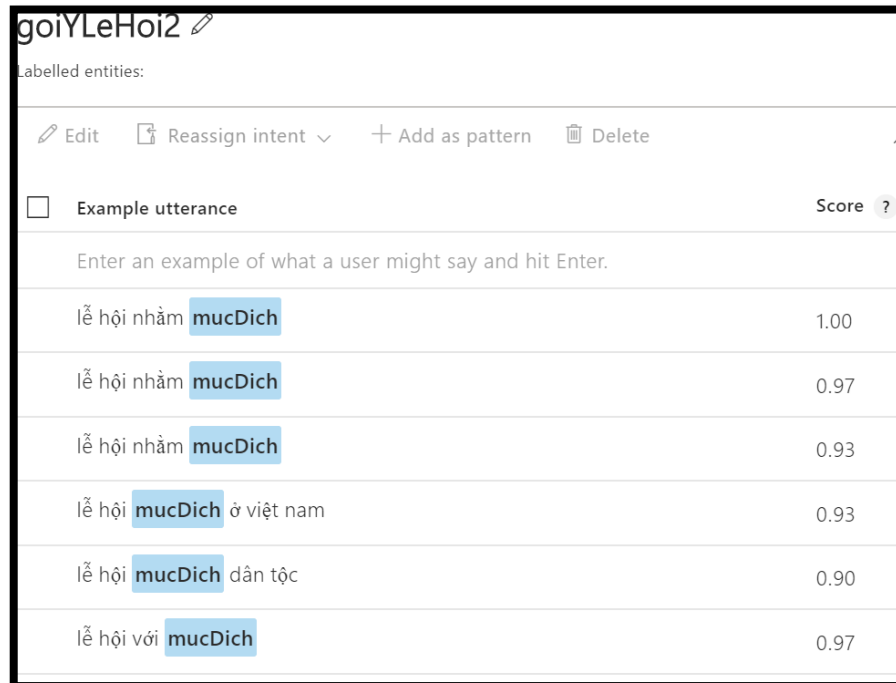


Hình 12: Mô hình kịch bản giữa Chatbot và người dùng

3.4.3 Cài đặt và xử lý ngôn ngữ tự nhiên

Sử dụng dịch vụ LUIS (<https://www.luis.ai>) để định nghĩa và huấn luyện các ý định và thuộc tính phục vụ cho việc nhận dạng các thông tin cần thiết từ

câu thoại của người dùng. Hệ thống để nhận diện ý định từ người dùng, ta cần tiến hành huấn luyện các câu đầu vào thể hiện cùng một ý muốn của người dùng. Ví dụ với ý định tìm kiếm các lễ hội thông qua mục đích của chúng ta định nghĩa intent: “*goiYLeHoi2*” với những câu đầu vào của người dùng như sau:



goiYLeHoi2

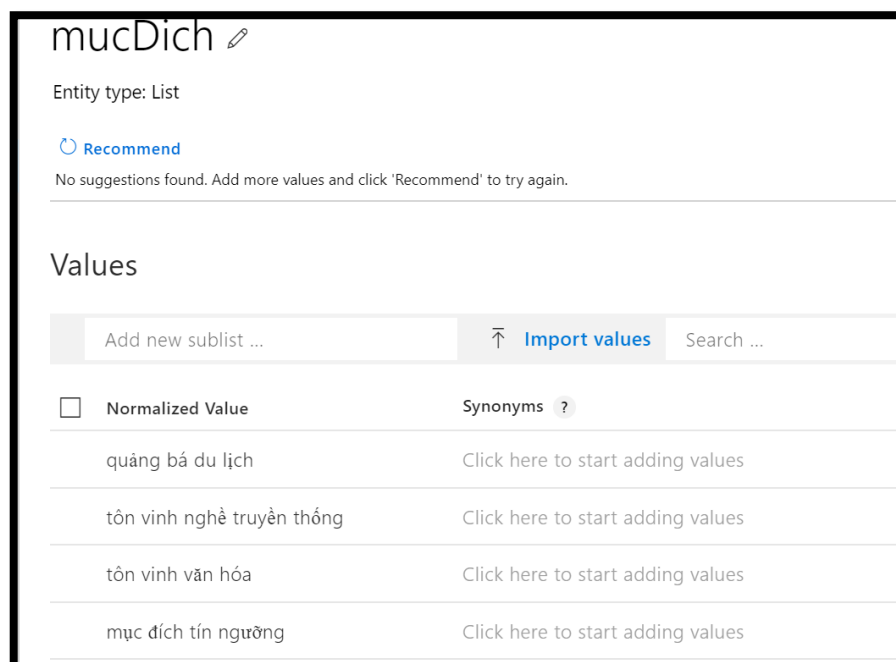
Labelled entities:

Edit Reassign intent Add as pattern Delete

| Example utterance | Score |
|--|-------|
| Enter an example of what a user might say and hit Enter. | |
| lễ hội nhằm mucDich | 1.00 |
| lễ hội nhằm mucDich | 0.97 |
| lễ hội nhằm mucDich | 0.93 |
| lễ hội mucDich ở việt nam | 0.93 |
| lễ hội mucDich dân tộc | 0.90 |
| lễ hội với mucDich | 0.97 |

Hình 13: Nhận diện ý định (intent) của người dùng với LUIS

Để nhận diện thông tin từ người dùng, ta cần tiến hành huấn luyện các thực thể sẽ có thể xuất hiện trong câu hội thoại. Ví dụ với entity: “*mục đích*”



mucDich

Entity type: List

Recommend

No suggestions found. Add more values and click 'Recommend' to try again.

Values

| Normalized Value | Synonyms |
|----------------------------|-----------------------------------|
| quảng bá du lịch | Click here to start adding values |
| tôn vinh nghề truyền thống | Click here to start adding values |
| tôn vinh văn hóa | Click here to start adding values |
| mục đích tín ngưỡng | Click here to start adding values |

Hình 14: Nhận diện thông tin (entity) từ người dùng với LUIS

3.5 Truy vấn dữ liệu bằng SPARQL

Như đã giới thiệu ở Chương 2 để truy vấn vào các tài liệu RDF ta cần sử dụng ngôn ngữ truy vấn SPARQL. Python RDFLib (bởi Dan Krech) là một thư viện của Python hỗ trợ khả năng suy diễn cho các câu truy vấn SPARQL, cung cấp một API để thao tác trên các lược đồ RDF, phân tích cú pháp RDF, và lưu trữ dữ liệu dạng triple. Việc sử dụng RDFLib không chỉ rất đơn giản với các lập trình viên có kiến thức về Python mà cả những người dùng chỉ cần quen thuộc với RDF cũng có thể tìm ra cách sử dụng dễ dàng. Sau đây là một ví dụ về câu truy vấn vào file *fesivalVietNam.owl* để tìm kiếm lễ hội tổ chức tại Hà Nội

```
import rdfextras
rdfextras.registerplugins()
filename = "../OntologyFile/fesivalVietNam.owl"
import rdflib
g = rdflib.Graph()
result = g.parse(filename, format='xml')
# print(result)
query = """
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xml: <http://www.w3.org/XML/1998/namespace>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX
:<http://www.semanticweb.org/admin/ontologies/2020/2/untitled-
ontology-5#>
SELECT DISTINCT ?name ?nloc
WHERE
{
  ?x :tenLeHoi ?name.
  ?x :toChucTai ?loc.
  ?loc :tenDiaDiem ?nloc.
  FILTER( regex(?nloc,"hà nội","i") )
}
"""
```

Kết quả của câu query:

Hội Gióng tổ chức tại nhiều nơi thuộc vùng Hà Nội
Hội rước bánh giầy tổ chức tại làng Bá Dương Nội , xã Hồng Hà , huyện Đan Phượng, Hà Nội
Lễ hội Chùa Hương tổ chức tại Hà Nội
Lễ hội phố hoa Hà Nội tổ chức tại phố Đinh Tiên Hoàng , cạnh bờ hồ Hoàn Kiếm, Hà Nội
Hội Gò Đống Đa tổ chức tại Hà Nội
Lễ hội làng Triều Khúc tổ chức tại xã Tân Triều, huyện Thanh Trì, Hà Nội
Lễ hội Gióng Sóc Sơn tổ chức tại xã Phù Linh, huyện Sóc Sơn, Hà Nội
Hội Cổ Loa tổ chức tại xã Cổ Loa, huyện Đông Anh, Hà Nội
Hội Gióng Phù Đổng tổ chức tại xã Phù Đổng , huyện Gia Lâm , Hà Nội
Hội chùa Thầy tổ chức tại Hà Nội
Lễ hội đền Chúa xã Cổ Nhuế tổ chức tại Hà Nội
Hội làng Lệ Mật tổ chức tại xã Việt Hưng , huyện Gia Lâm, Hà Nội
Lễ hội Bình Đà tổ chức tại Hà Nội

Hình 15: Kết quả về một ví dụ truy vấn SPRAQL

CHƯƠNG 4. KẾT QUẢ THỰC NGHIỆM

4.1 Kết quả gán nhãn, trích xuất và thu thập dữ liệu liên quan đến lễ hội

4.1.1 Kịch bản thử nghiệm

Để xây dựng dữ liệu cho Ontology lễ hội, tôi tiến hành crawl nội dung của 3 trang web sau với khoảng 154 bài viết liên quan đến lễ hội.

- https://vi.wikipedia.org/wiki/Lễ_hội_Việt_Nam
- https://vi.wikipedia.org/wiki/Thể_loại:Lễ_hội
- <https://www.maxreading.com/sach-hay/le-hoi-truyen-thong.html>

50 bài viết trong số các bài viết thu thập được đem phân tách từ và loại từ rồi tiến hành gán nhãn theo bộ nhãn đã được định nghĩa ở chương 3 tạo tập huấn luyện cho mô hình CRFs. Chia tập huấn luyện theo tỉ lệ 70% học và 30% kiểm thử. Sau đó sử dụng kết quả của mô hình CRFs để gán nhãn và trích xuất thực thể của các bài viết còn lại.

4.1.2 Kết quả kiểm thử

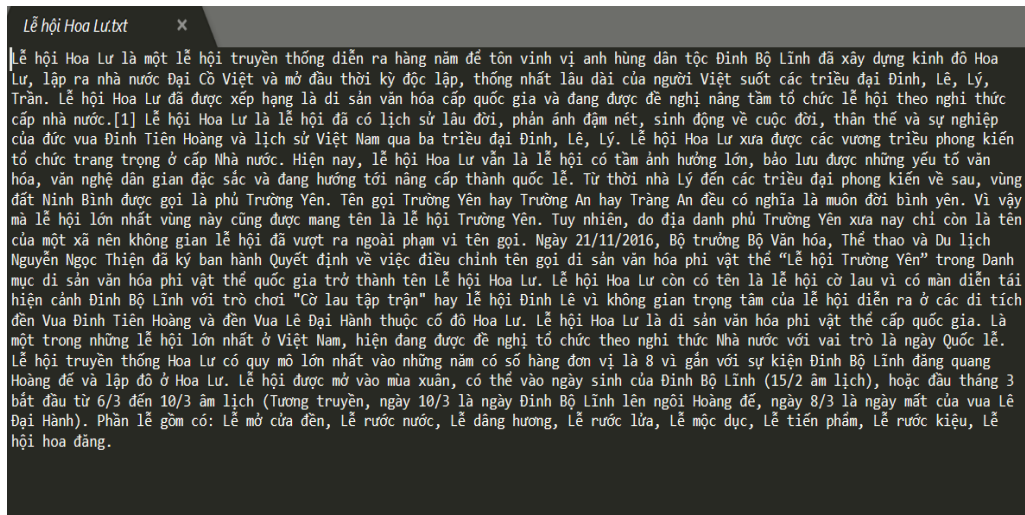
a. Kết quả crawl dữ liệu từ các website

Ví dụ khi muốn lấy thông tin về lễ hội Hoa Lư từ bài viết “Lễ hội Hoa Lư” trên Wikipedia.



Hình 16: Ví dụ một bài báo trên Wikipedia

Sau thu thập bằng công cụ scrapy bài viết được lưu trữ như hình dưới:



Hình 17: Kết quả sau khi crawl dữ liệu từ một bài viết

Kết quả thu thập được 154 bài viết được lưu lại dưới dạng file .txt

| | | | |
|------------------------------------|-------------------|---------------|-------|
| Bơi Trãi.txt | 3/20/2020 3:55 PM | Text Document | 2 KB |
| Chợ Âm Dương.txt | 2/22/2020 9:30 PM | Text Document | 1 KB |
| Chợ Viêng (định hướng).txt | 2/22/2020 9:30 PM | Text Document | 1 KB |
| Chùa Bái Đính.txt | 3/12/2020 9:44 AM | Text Document | 3 KB |
| Chùa Côn Sơn.txt | 2/22/2020 9:30 PM | Text Document | 1 KB |
| Chùa Keo.txt | 3/12/2020 9:27 AM | Text Document | 2 KB |
| Chùa Thầy.txt | 2/22/2020 9:30 PM | Text Document | 3 KB |
| Cổ Loa.txt | 2/22/2020 9:30 PM | Text Document | 20 KB |
| Đền Kiếp Bạc.txt | 2/22/2020 9:31 PM | Text Document | 2 KB |
| Đền Mẫu.txt | 2/22/2020 9:31 PM | Text Document | 0 KB |
| Điện Hòn Chén.txt | 3/13/2020 1:45 PM | Text Document | 2 KB |
| Festival Diều Quốc tế Vũng Tàu.txt | 2/22/2020 9:30 PM | Text Document | 2 KB |
| Festival Dừa.txt | 2/22/2020 9:30 PM | Text Document | 1 KB |
| Festival Hoa Đà Lạt 2012.txt | 2/22/2020 9:30 PM | Text Document | 13 KB |
| Festival Hoa Đà Lạt.txt | 2/22/2020 9:30 PM | Text Document | 1 KB |
| Festival Huế.txt | 2/22/2020 9:30 PM | Text Document | 3 KB |

Hình 18: Kết quả thư mục lưu trữ các bài viết sau khi crawl

b. Kết quả huấn luyện mô hình CRFs

Sau khi trích xuất và gán nhãn cho khoảng 50 bài viết về lễ hội thu được tập huấn luyện với khoảng 12,000 nhãn đầu vào chia với tỉ lệ 70% số lượng nhãn huấn luyện và 30% số lượng nhãn kiểm thử. Kết quả của mô hình CRF thu được với tập huấn luyện với độ chính xác đạt được khoảng 80%:

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| B-AC4 | 0.47 | 0.32 | 0.38 | 28 |
| I-AC4 | 0.44 | 0.35 | 0.39 | 51 |
| B-FHC | 0.75 | 0.30 | 0.43 | 10 |
| I-FHC | 1.00 | 0.12 | 0.22 | 8 |
| B-CO2 | 0.00 | 0.00 | 0.00 | 15 |
| I-CO2 | 0.75 | 0.12 | 0.21 | 75 |
| B-FIT | 0.67 | 0.48 | 0.56 | 21 |
| B-TIM | 0.83 | 0.54 | 0.65 | 28 |
| I-TIM | 0.93 | 0.66 | 0.77 | 225 |
| B-CO3 | 1.00 | 0.43 | 0.60 | 7 |
| I-CO3 | 1.00 | 0.47 | 0.63 | 43 |
| B-AC5 | 1.00 | 0.03 | 0.06 | 33 |
| I-AC5 | 1.00 | 0.03 | 0.05 | 39 |
| B-FES | 1.00 | 0.71 | 0.83 | 34 |
| I-FES | 1.00 | 0.52 | 0.68 | 66 |
| B-FIL | 0.89 | 0.38 | 0.53 | 21 |
| B-LOC | 0.74 | 0.59 | 0.66 | 34 |
| I-LOC | 0.88 | 0.59 | 0.71 | 108 |
| B-FAM | 1.00 | 0.25 | 0.40 | 12 |
| B-FFT | 1.00 | 0.31 | 0.47 | 13 |
| I-FFT | 0.75 | 0.29 | 0.41 | 42 |
| B-FHE | 0.00 | 0.00 | 0.00 | 8 |
| B-AC1 | 0.62 | 0.35 | 0.44 | 23 |
| I-AC1 | 0.73 | 0.33 | 0.46 | 24 |
| B-FHA | 0.33 | 0.10 | 0.15 | 10 |
| B-ANN | 1.00 | 0.12 | 0.21 | 17 |
| I-ANN | 1.00 | 0.12 | 0.21 | 17 |
| I-FIT | 0.75 | 0.53 | 0.62 | 17 |
| I-FHA | 0.20 | 0.10 | 0.13 | 10 |
| I-FAM | 1.00 | 0.24 | 0.38 | 21 |
| B-FOE | 1.00 | 0.50 | 0.67 | 2 |
| I-FOE | 1.00 | 1.00 | 1.00 | 2 |
| B-ETH | 1.00 | 0.67 | 0.80 | 3 |
| micro avg | 0.82 | 0.39 | 0.53 | 1288 |
| macro avg | 0.56 | 0.27 | 0.34 | 1288 |
| weighted avg | 0.76 | 0.39 | 0.48 | 1288 |

Hình 19: Kết quả của mô hình CRFs

Sử dụng mô hình CRFs sau khi huấn luyện để trích xuất thực thể các bài viết còn lại. Kết quả thu thập được thông tin về 117 lễ hội khác nhau ở Việt Nam. Tất cả các trường thông tin về 117 lễ hội này được lưu lại dưới dạng bảng tính để chuẩn bị cho quá trình chuyển đổi thành dữ liệu có ngữ nghĩa.

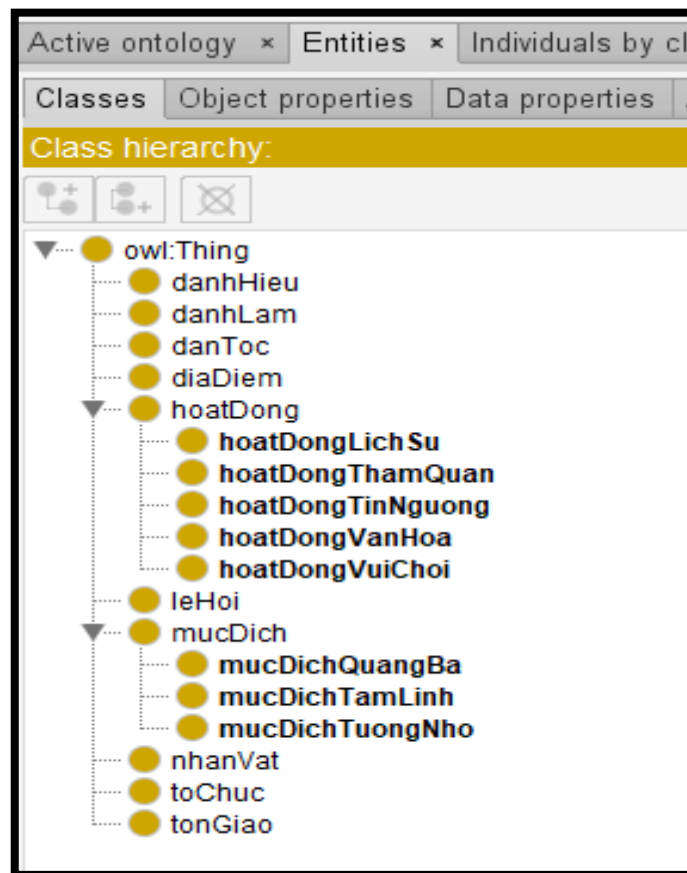
| Mã lễ hội | Tên lễ hội | Địa điểm | Thời gian | Tên gọi khác | Lịch sử hình thành | Hoạt động vui chơi | Hoạt động du lịch | Hoạt động mang tính lễ Tết (lễ lên ngôi) của Phường Hưng | Hoạt động tín ngưỡng | Hoạt động văn hóa dân | Mục đích quảng bá | Mục đích tâm linh | Mục đích tưởng nhớ | Dân tộc |
|-----------|------------------------|-------------------------------|---------------------------------------|--------------|--------------------|--------------------|-------------------|--|----------------------|-----------------------|--|---|--------------------|---------|
| F45 | Lễ hội làng Triều Khúc | xã Tân Triều, huyện Thanh Trì | từ mùng 10 đến 12 tháng giêng âm lịch | | | trò chơi dân gian | ghé thăm chùa | lễ Tức vị (lễ lên ngôi) của Phường Hưng | lễ tế | múa Rồng | | mong chờ một mùa làm ăn mới tốt đẹp hơn mọi | | |
| F45 | Lễ hội làng Triều Khúc | Hà Nội | | | | đấu vật | | diễn lại tích Phùng Hưng chọn người tài đi đánh giặc | đọc văn tế | hát chèo | | | | |
| F45 | Lễ hội làng Triều Khúc | Đại Định | | | | múa rồng | | | | múa truyền thống | | | | |
| F45 | Lễ hội làng Triều Khúc | | | | | | | | | múa cờ | | | | |
| F33 | Hội đua voi | tỉnh Đắk Lắk | mùa xuân | | ngày hội cổ truyền | | | | | | phản ánh tinh thần thượng võ của người M' nong | | | M' nong |
| F33 | Hội đua voi | Tây Nguyên | dịp tháng ba âm lịch | | | | | | | | | | | Êđê |
| F33 | Hội đua voi | | | | | | | | | | | | | Lào |

Hình 20: Kết quả trích xuất và lưu trữ dữ liệu

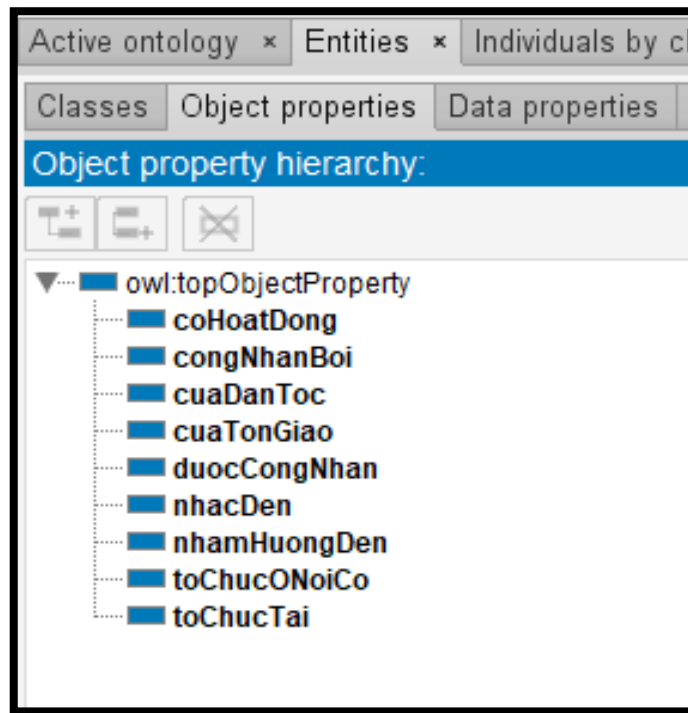
4.2 Kết quả xây dựng cơ sở tri thức

Dựa trên mô tả chi tiết về các thành phần của Ontology đã được định nghĩa trong mục 3.3.4: Xây dựng Ontology ta đi vào cài đặt chi tiết Ontology lễ hội thông qua công cụ Protégé.

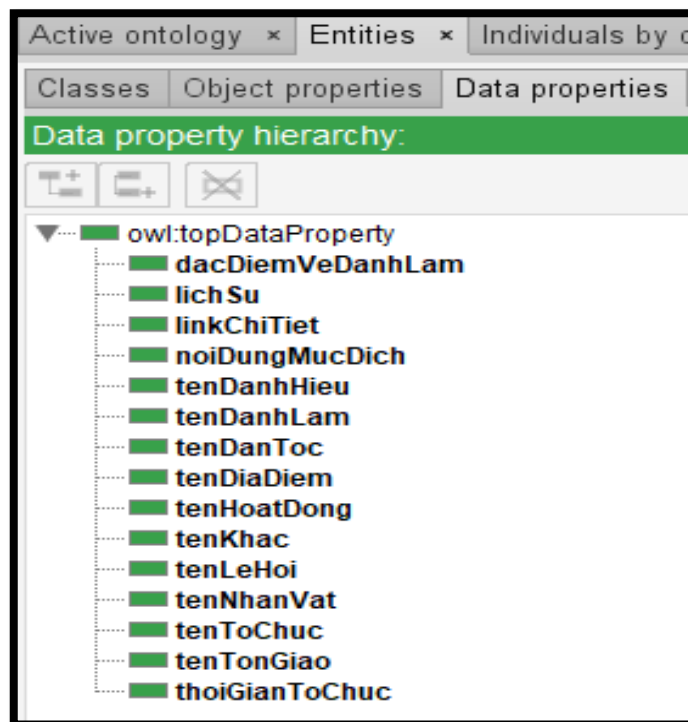
4.2.1 Kết quả xây dựng lớp và thuộc tính của Ontology



Hình 21: Cài đặt các lớp của Ontology



Hình 22: Cài đặt các Object properties của Ontology

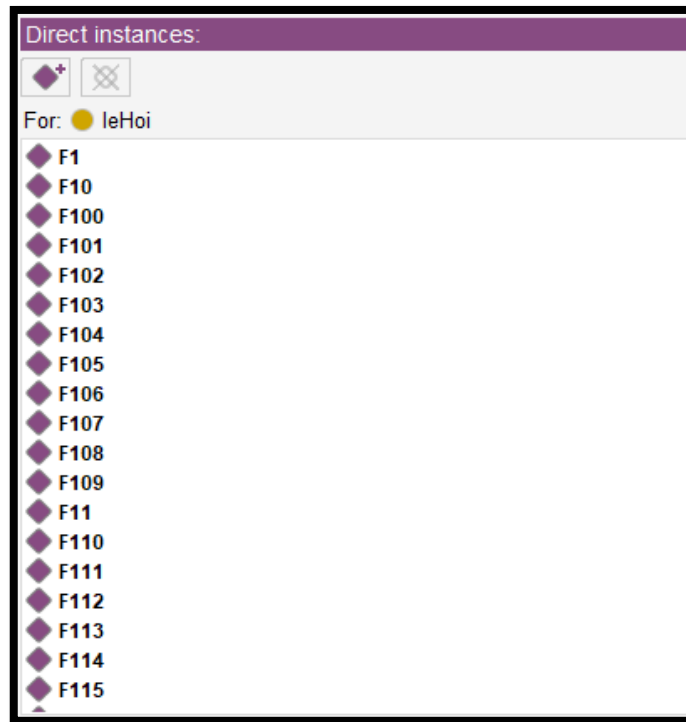


Hình 23: Cài đặt các Data property của Ontology

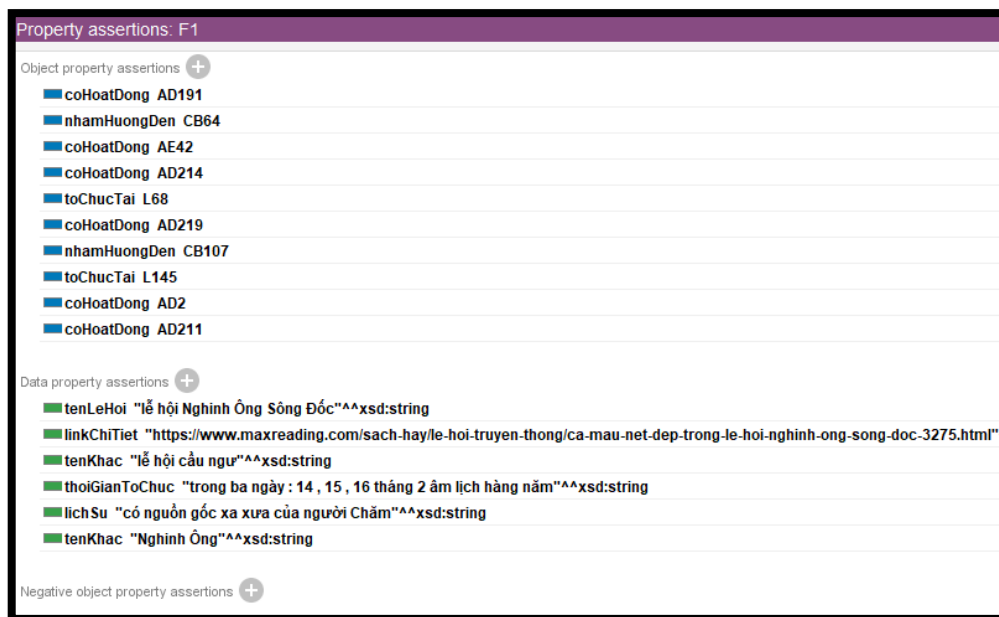
4.2.2 Kết quả chuyển đổi dữ liệu sang dạng có ngữ nghĩa

Nhờ Cellfie của Protégé một lượng lớn dữ liệu được chuyển đổi sang dạng có ngữ nghĩa theo mô tả trong Ontology thông qua các luật và trở thành các thể hiện trong Ontology. Dưới đây là minh họa các thể hiện của một lớp trong

Ontology đó là lớp lễ hội và các thuộc tính, quan hệ của một thể hiện cụ thể trong lớp lễ hội.



Hình 24: Các thể hiện của lớp lễ hội



Hình 25: Các thuộc tính và quan hệ của một thể hiện trong lớp lễ hội

4.2.3 Tổng kết về cơ sở tri thức

Sau khi hoàn thành xây dựng cơ sở tri thức về lễ hội tại Việt Nam kết quả thu được như sau:

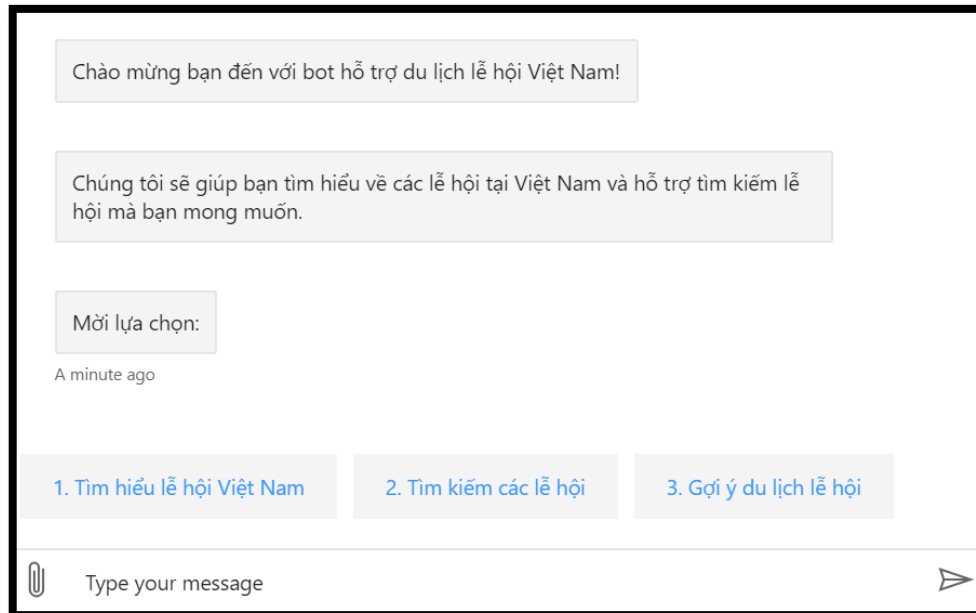
- Số lễ hội có thông tin trong cơ sở tri thức: 117 lễ hội
- Tổng số bộ ba RDF: 28210
- Tổng số Subject: 6906

- Tổng số Predicate: 31
- Tổng số Object: 8224

4.3 Kết quả cài đặt Chatbot hỗ trợ du lịch

4.3.1 Lời chào hệ thống

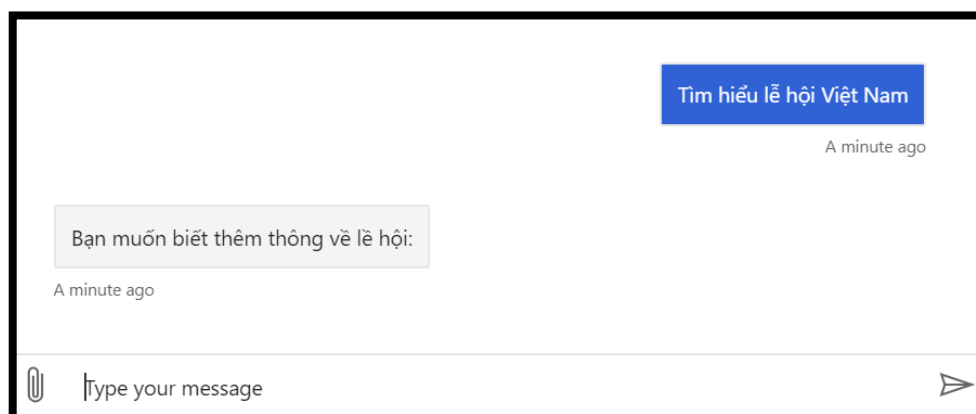
Khi người dùng kết nối với Chatbot, màn hình sẽ hiện ra lời chào cùng các chức năng để người dùng lựa chọn.



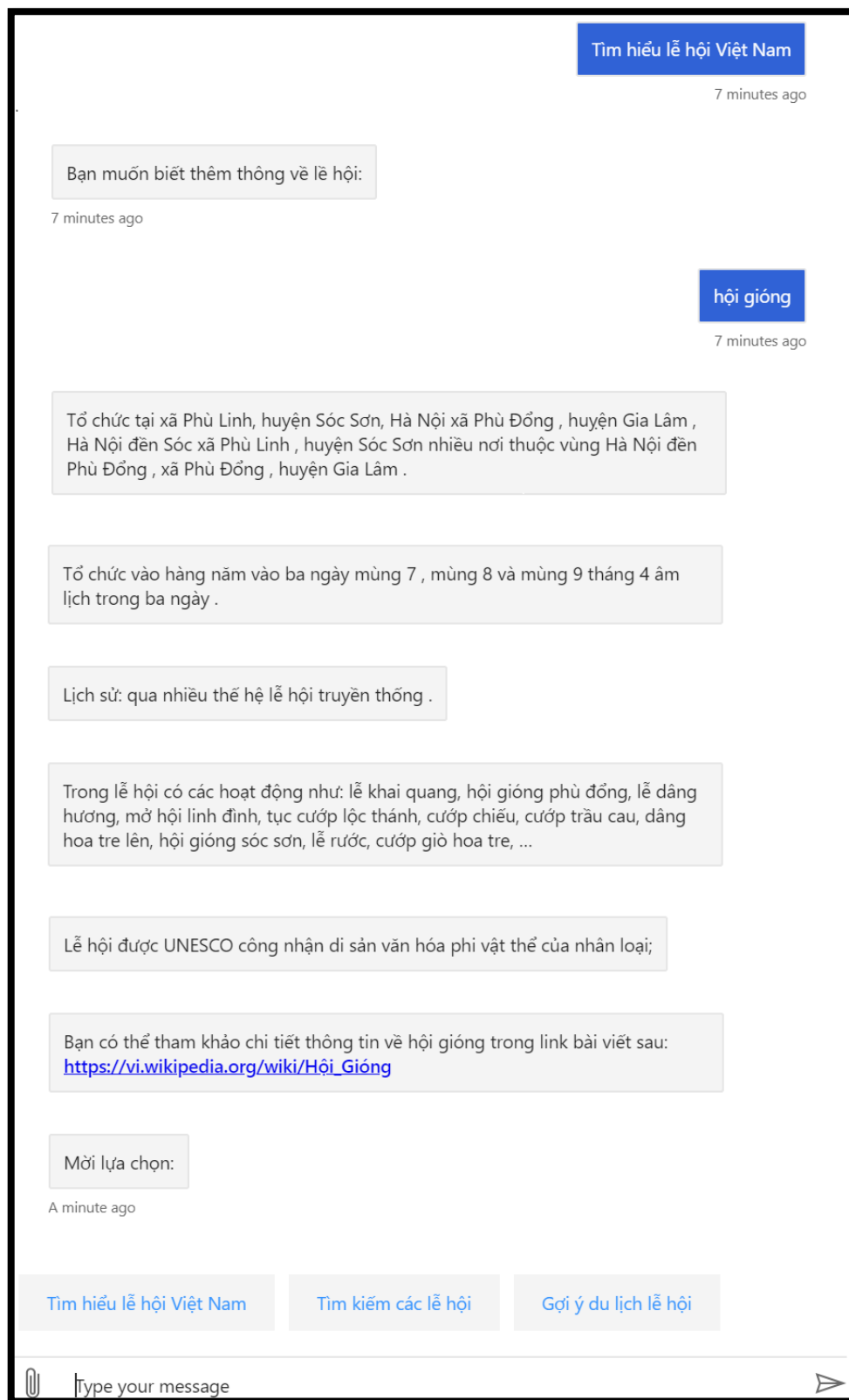
Hình 26: Lời chào của Chatbot

4.3.2 Chức năng Tìm hiểu lễ hội Việt Nam

Sau khi người dùng chọn chức năng đầu tiên “Tìm hiểu lễ hội Việt Nam” Chatbot sẽ yêu cầu người dùng nhập tên lễ hội và Chatbot sẽ cung cấp các thông tin về lễ hội đó.

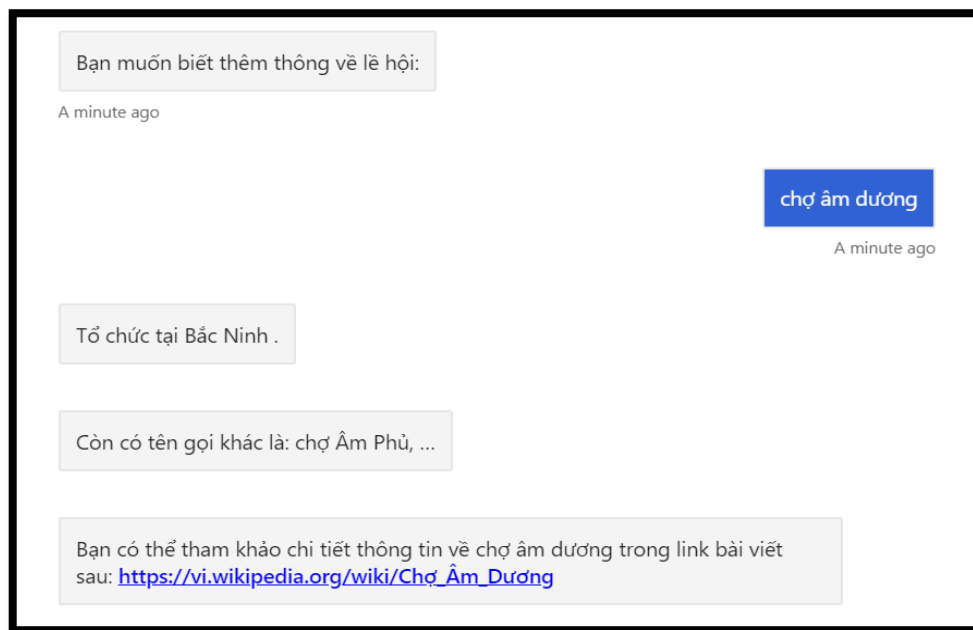


Hình 27: Minh họa chức năng Tìm hiểu lễ hội Việt Nam (1)

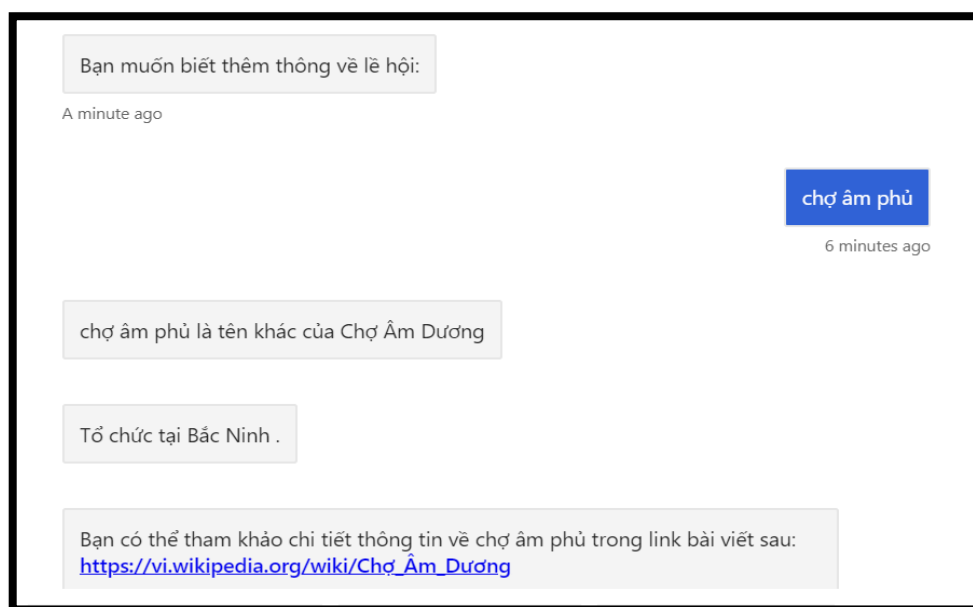


Hình 28: Minh họa chức năng Tìm hiểu lễ hội Việt Nam (2)

Nếu lễ hội có nhiều tên gọi khác nhau có thể tìm kiếm thông qua các tên gọi khác đó.



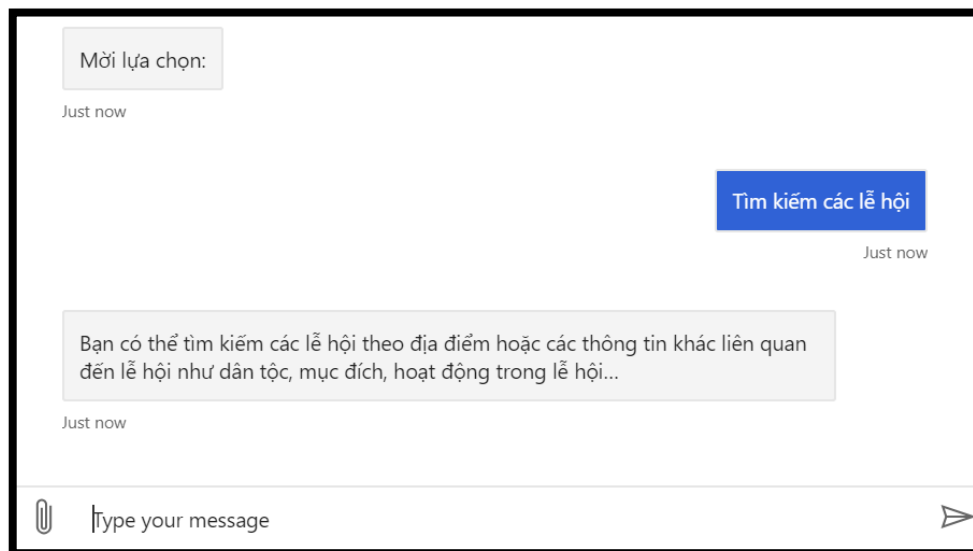
Hình 29: Ví dụ về tìm kiếm lễ hội theo tên (1)



Hình 30: Ví dụ về tìm kiếm lễ hội theo tên (2)

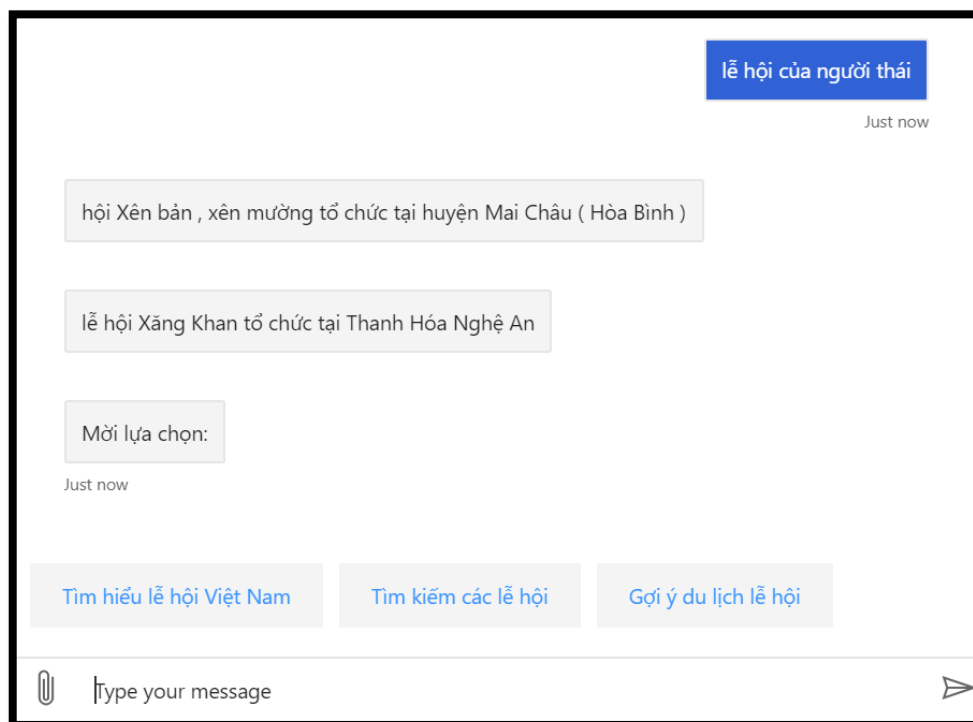
4.3.3 Chức năng Tìm kiếm các lễ hội

Chức năng hỗ trợ tìm kiếm các lễ hội tại Việt Nam thông qua các thông tin liên quan như địa điểm, dân tộc, mục đích và hoạt động trong lễ hội. Dưới đây là các ví dụ minh họa cụ thể.



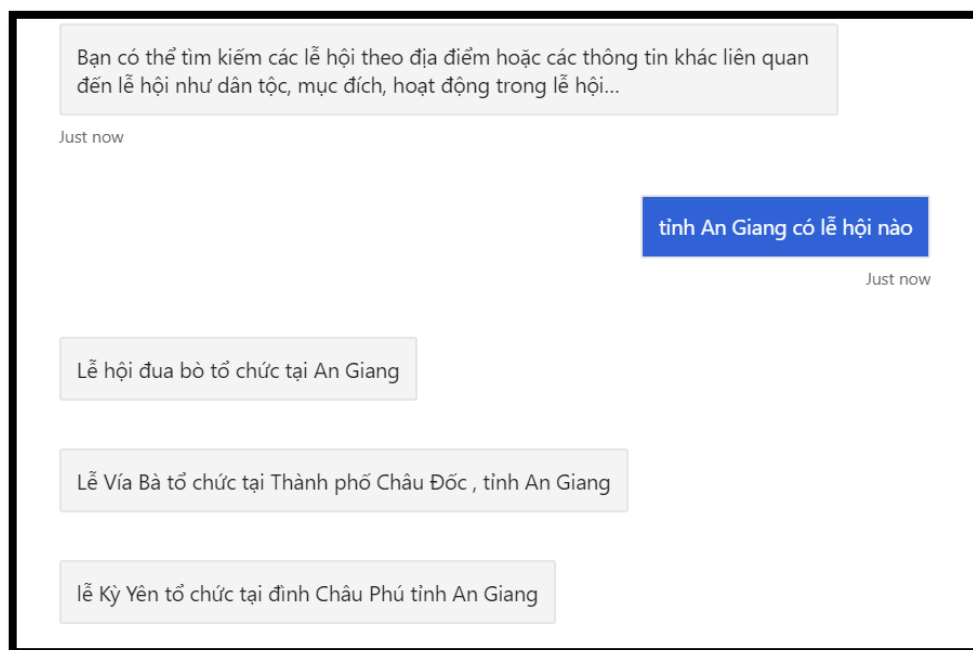
Hình 31: Chức năng tìm kiếm các lễ hội

Với thông tin về dân tộc người có thể tìm kiếm các lễ hội đặc trưng của mỗi dân tộc ở Việt Nam.



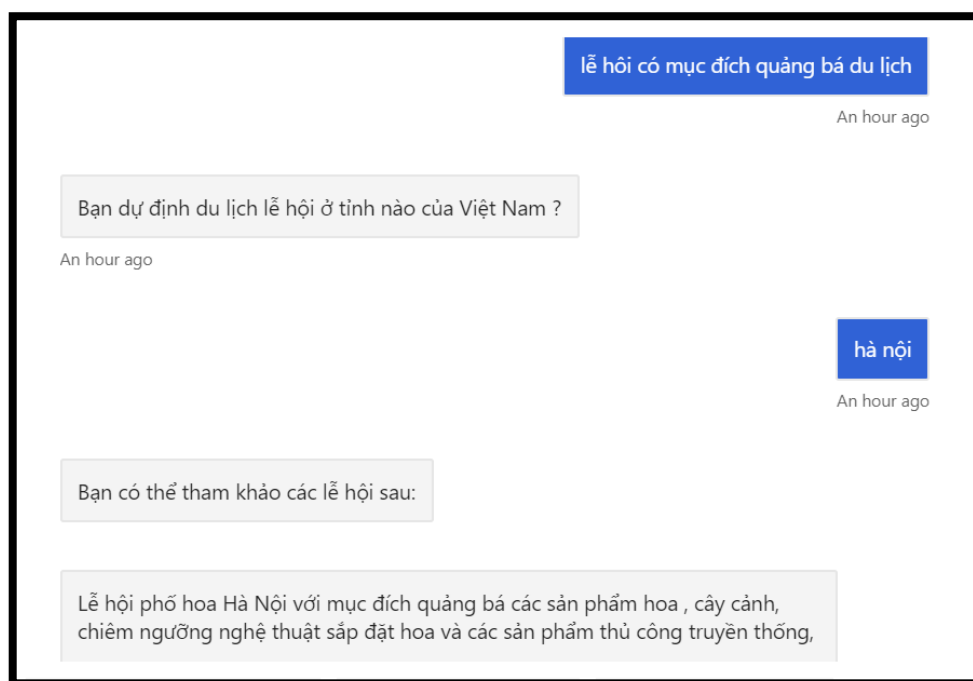
Hình 32: Ví dụ tìm kiếm lễ hội theo dân tộc

Người dùng cũng có thể tìm kiếm các lễ hội cụ thể tại một tỉnh hay địa phương nào đó.



Hình 33: Ví dụ tìm kiếm lễ hội thông qua địa điểm

Mỗi lễ hội diễn ra lại có những mục đích khác nhau như lễ hội nhằm quảng bá du lịch thường được tổ chức ở những thành phố nổi tiếng về du lịch. Hay lễ hội truyền thống với các mục đích tâm linh như thờ các vị thần ... Người dùng cũng có thể tìm kiếm lễ hội phù hợp thông qua đặc điểm này của các lễ hội. Ví dụ như:



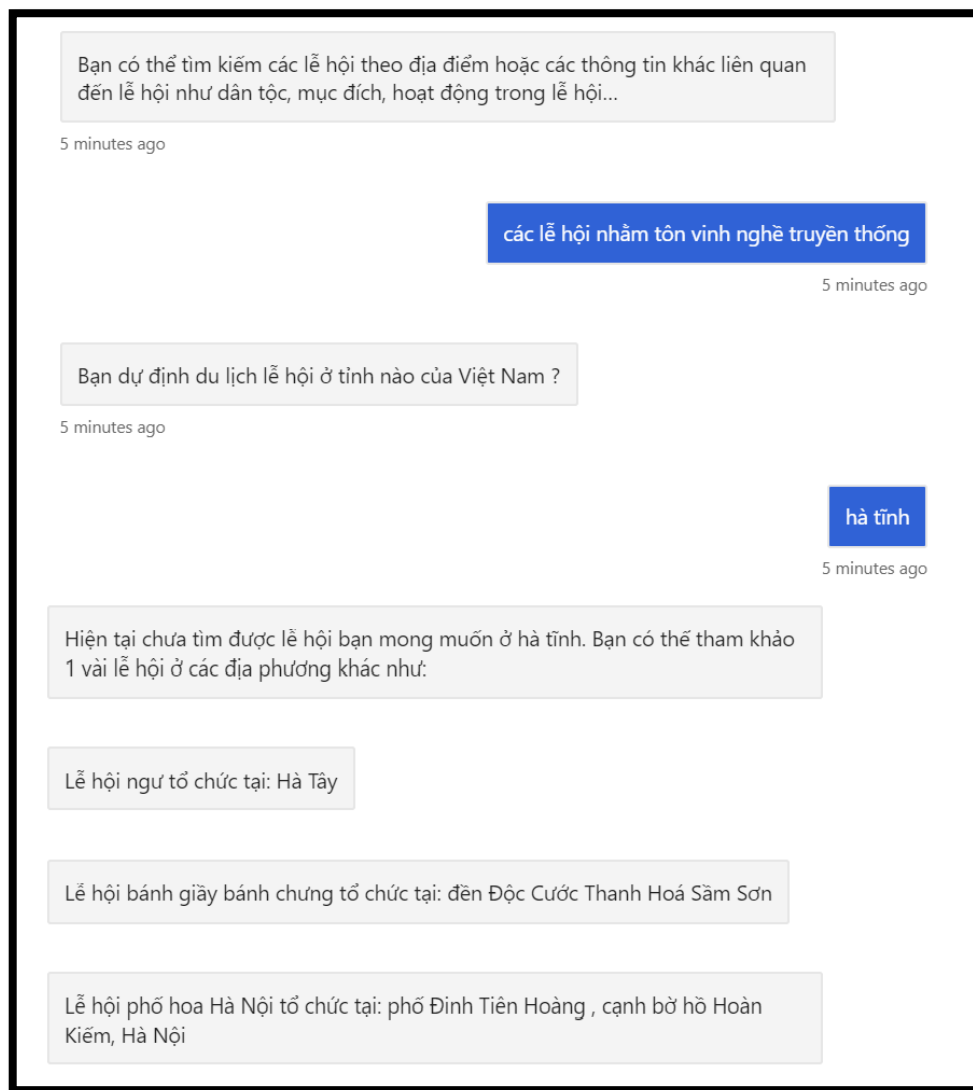
Hình 34: Ví dụ về tìm kiếm lễ hội thông qua mục đích của lễ hội (1)



Hình 35: Ví dụ về tìm kiếm lễ hội thông qua mục đích của lễ hội (2)



Hình 36: Ví dụ về tìm kiếm lễ hội thông qua mục đích (3)



Hình 37: Ví dụ về tìm kiếm lễ hội thông qua mục đích (4)

Ngoài ra bạn cũng có thể tìm kiếm lễ hội thông qua các sự kiện hay hoạt động đặc trưng được tổ chức trong lễ hội.

Bạn có thể tìm kiếm các lễ hội theo địa điểm hoặc các thông tin khác liên quan đến lễ hội như dân tộc, mục đích, hoạt động trong lễ hội...

Just now

lễ hội có các hoạt động thể thao

Just now

Bạn dự định du lịch lễ hội ở tỉnh nào của Việt Nam ?

Just now

nam định

Bạn có thể tham khảo các lễ hội sau:

Lễ hội Chùa Keo Hành Thiện với các hoạt động: bơi Trài, bơi trái, môn Đua thuyền, hoạt động văn hoá thể thao, ...

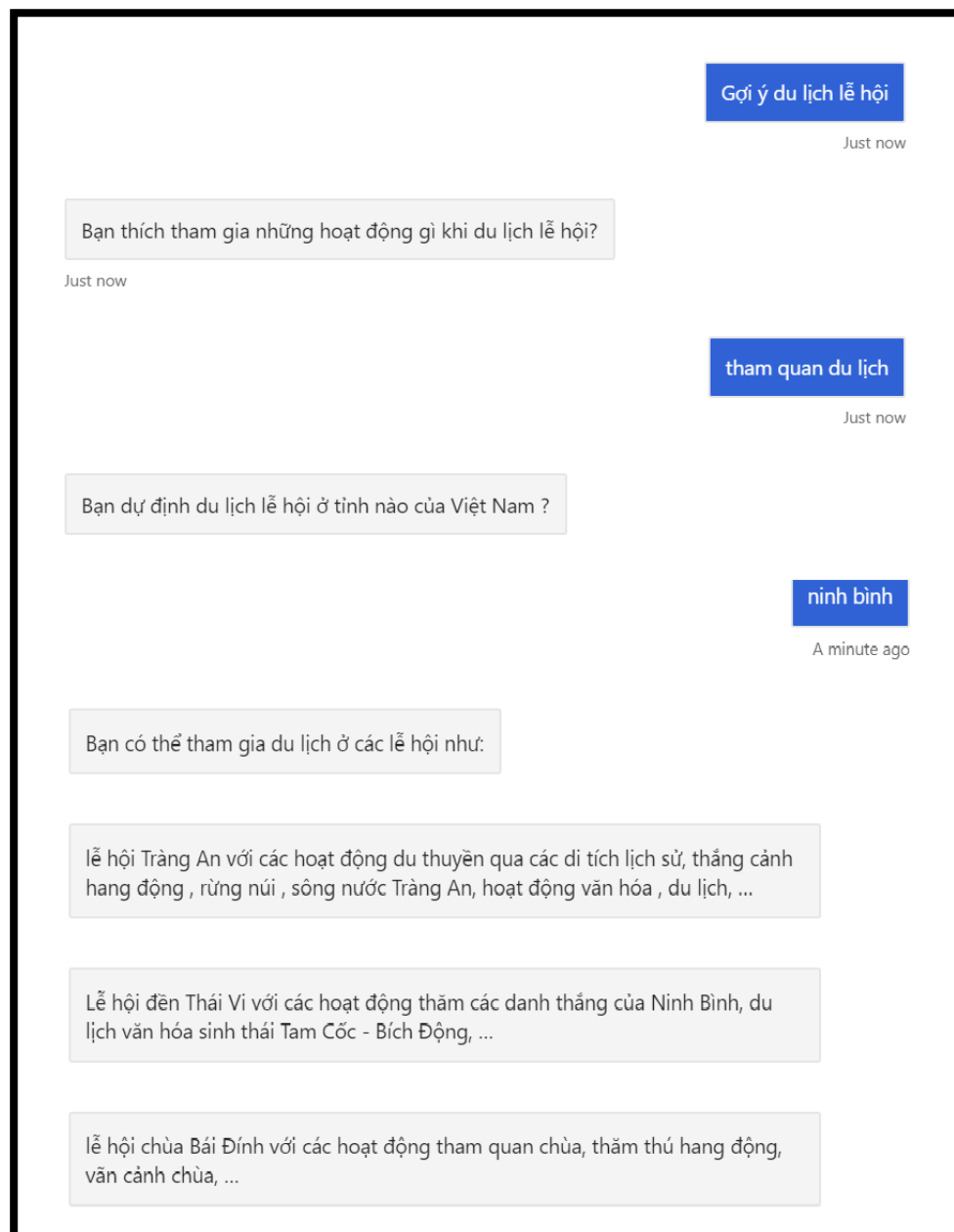
Lễ hội Phủ Dầy với các hoạt động: đấu vật, ...

Hình 38: Ví dụ về tìm kiếm lễ hội thông qua hoạt động trong lễ hội

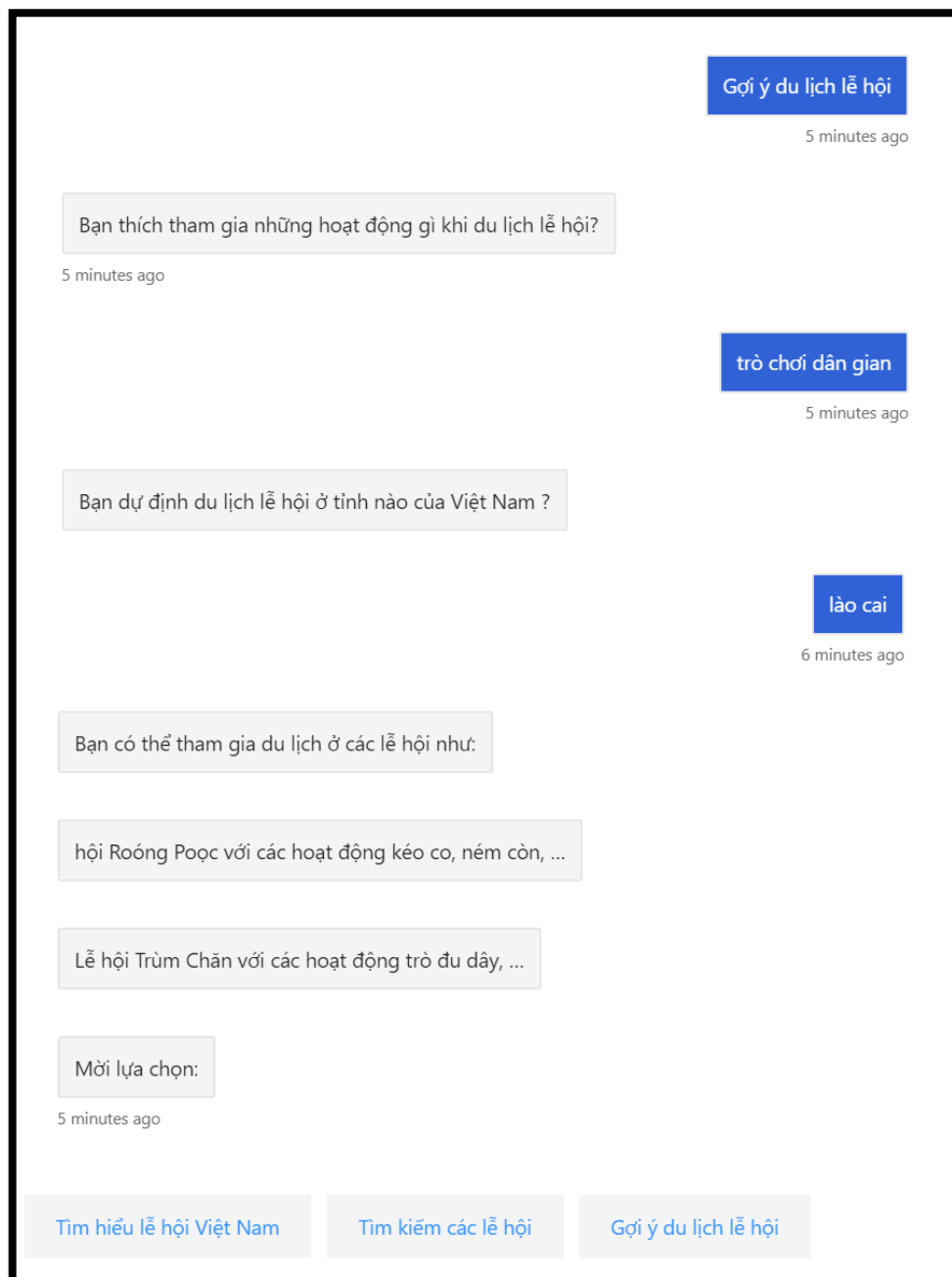
4.3.4 Chức năng Gợi ý du lịch lễ hội

Đây là chức năng cuối cùng của Chatbot, sau khi biết được mong muốn, sở thích của người dùng thông qua những hoạt động sự kiện mà người dùng muốn tham gia như hoạt động thể thao, hoạt động vui chơi ngoài trời, trò chơi dân gian hay tham quan du lịch. Chatbot sẽ tìm kiếm những lễ hội có đặc điểm gần nhất với yêu cầu của người dùng.

Ví dụ người dùng muốn đi các lễ hội mà đồng thời có thể tham quan danh lam thắng cảnh tại tỉnh Ninh Bình, Chatbot nhận thấy có một vài lễ hội ở Ninh Bình có các hoạt động như văn cảnh chùa, leo núi hay du lịch sinh thái rất phù hợp với yêu cầu tham quan thắng cảnh của người dùng nên đã giới thiệu cho người dùng những lễ hội đó. Dưới đây là một vài minh hoạt về chức năng “Gợi ý du lịch lễ hội”. Hoặc người dùng muốn tham gia các trò chơi dân gian trong các lễ hội, chatbot sẽ thông qua việc xác định trong lễ hội có những trò chơi như kéo co, ném còn, thổi cơm thi, ... để xác định những lễ hội mà người dùng có thể tham gia các trò chơi dân gian.



Hình 39: : Chức năng gợi ý du lịch lễ hội (1)



Hình 40: Chức năng gợi ý du lịch lễ hội (2)

4.3.5 Đánh giá về Chatbot trên cơ sở Web ngữ nghĩa

Qua việc thử nghiệm Chatbot trên cơ sở web ngữ nghĩa tôi rút ra một vài điểm nổi bật cũng như hạn chế như sau:

✓ Ưu điểm

- Chatbot có khả năng sử dụng dữ liệu linh hoạt, không bị bó buộc trong các bảng, dữ liệu cũng được thêm sửa xóa dễ dàng và hoàn toàn có thể mở rộng trong tương lai.
- Điểm nổi bật của Web ngữ nghĩa là dữ liệu có khả năng suy diễn cao thông qua các đặc tính của thuộc tính, sự phân lớp cha – con và các thể hiện vô cùng đa dạng của từng lớp.

✓ Ví dụ về khả năng suy diễn

Như đã nói ở trên khả năng suy diễn cao là điểm nổi bật của Web ngữ nghĩa thông qua các đặc tính của thuộc tính, sự phân lớp cha – con trong Ontology. Sau đây là ví dụ về khả năng suy diễn của Ontology lễ hội:

Ví dụ tìm kiếm gợi ý lễ hội cho người dùng thông qua sở thích, mong muốn của người dùng. Ta sẽ xác định các lễ hội này thông qua các hoạt động diễn ra trong lễ hội. Với quan hệ *lễ hội – có tổ chức – hoạt động*, ta định nghĩa một lớp *hoatDong*. Nếu người dùng muốn tìm kiếm các lễ hội liên quan đến tâm linh, lễ hội được xác định sẽ có các hoạt động tâm linh, ta sẽ định nghĩa lớp *hoatDongTamLinh* đây là một lớp con của lớp *hoatDong* với các thể hiện như: lễ tế, lễ rước kiệu, dâng hương, vv...

2 minutes ago

2 minutes ago

2 minutes ago

Bạn có thể tham gia du lịch ở các lễ hội như:

Lễ hội đèn Nguyễn Công Trứ với các hoạt động Dân hương, ...

48

CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1 Đánh giá kết quả thực hiện

Đồ án đã thực hiện thành công hai mục tiêu được đặt ra ban đầu. Đó là (1) xây dựng kho tri thức ngữ nghĩa về du lịch lễ hội ở Việt Nam; và (2) xây dựng Chatbot sử dụng kho tri thức sau khi xây dựng.

Kho tri thức ngữ nghĩa về lễ hội ở Việt Nam lưu trữ một số lượng thông tin khá lớn và tương đối đầy đủ về các lễ hội tại Việt Nam với 117 lễ hội khác nhau trong đó có khoảng 70 lễ hội được định nghĩa trong Wikipedia. Các thông tin này sau khi được chuyển đổi sang dạng ngữ nghĩa bằng các mô tả về lớp, quan hệ và thuộc tính trong Ontology lễ hội đều mang tính suy diễn cao. Bằng việc sử dụng bài toán Nhận dạng thực thể có tên (NER) quá trình tiền xử lý, làm sạch, và tích hợp các nguồn dữ liệu đạt được độ chính xác là 82%, kết quả hầu hết các thông tin về lễ hội đều được thu thập chính xác. Qua đó giúp quá trình xây dựng lên một cơ sở tri thức trở nên tự động dễ dàng và đáng tin cậy.

Tiếp theo sau quá trình tìm hiểu công nghệ xây dựng và xử lý chatbot gồm khung hỗ trợ Chatbot Microsoft Bot Framework và dịch vụ xử lý ngôn ngữ tự nhiên LUIS tôi đã cài đặt thành công Chatbot hỗ trợ du lịch lễ hội Việt Nam với 3 chức năng chính giúp người dùng có thể tìm kiếm thông tin lễ hội không chỉ qua những thông tin cơ bản như tên lễ hội, địa điểm, ... mà còn giúp tìm kiếm các lễ hội thông qua những thông tin như mục đích hay hoạt động của của lễ hội mà người dùng mong muốn. Chatbot sử dụng công nghệ web ngữ nghĩa thể hiện ưu điểm của web ngữ nghĩa như khả năng suy diễn và tính linh hoạt trong sử dụng dữ liệu.

Ngoài ra việc hoàn thành đồ án còn giúp bản thân có được sự hiểu biết về Web ngữ nghĩa cũng như ứng dụng của chúng. Trao dồi, nâng cao khả năng đọc, nghiên cứu và ứng dụng các công nghệ vào giải quyết những bài toán cụ thể. Rèn luyện khả năng tổng hợp, tư duy giải quyết vấn đề cũng như kỹ năng trình bày, lập luận, viết một báo cáo.

5.2 Hướng phát triển trong tương lai

Sau khi thực hiện đồ án, tôi thấy được rằng đồ án có nhiều điểm có khả năng phát triển tiếp trong tương lai. Nhờ khả năng mở rộng các quan hệ và thuộc tính của Ontology, trong tương lai hoàn toàn có thể mở rộng Ontology lễ hội với nhiều thông tin liên quan hơn. Ví dụ với lớp “nhân vật” nằm trong bộ ba quan hệ “Lễ hội – có nhắc đến – Nhân vật” ta có thể bổ sung thêm các lớp con, các thuộc tính, các quan hệ khác cho lớp nhân vật này ví dụ bổ sung thêm quan hệ “Nhân vật – xuất hiện trong – sự kiện lịch sử” và các thuộc tính như “Năm sinh”, “Quê quán”, “Tiểu sử” cho lớp “Nhân vật”, từ đó làm giàu thêm tri thức và tăng khả năng suy diễn cho Ontology hiện tại. Với khả năng tự động thu thập, trích rút dữ liệu và chuyển đổi thành dạng ngữ nghĩa giúp quá trình cập nhật dữ liệu cho cơ sở tri thức sau này cũng trở nên dễ dàng hơn. Cùng với sự mở rộng của cơ sở tri

thức Chatbot cũng có khả năng bổ sung thêm nhiều chức năng mới để người dùng có thể trải nghiệm. Bên cạnh đó với khả năng huấn luyện nhận diện các ý định và thực thể từ đầu vào của người dùng của dịch vụ xử lý ngôn ngữ tự nhiên LUIS sẽ hỗ trợ quá trình phát triển của Chatbot giúp Chatbot ngày càng thông minh hơn, khả năng giao tiếp với người dùng sẽ ngày càng trở nên linh hoạt và tự nhiên. Thông qua nền tảng dịch vụ Azure của Microsoft Chatbot có thể được tích hợp trong những ứng dụng về du lịch tại Việt Nam hoặc trong trên những nền tảng mạng xã hội như Facebook, Skype để có thể tiếp cận gần hơn với người dùng trong tương lai.

Đồ án nghiên cứu và phát triển Chatbot du lịch lễ hội trên nền tảng Web ngữ nghĩa cũng là tiền đề, kinh nghiệm cho những đề tài lớn hơn, bao quát và có nhiều ứng dụng như xây dựng kho dữ liệu tri thức về du lịch ở Việt Nam.

TÀI LIỆU THAM KHẢO

- [1] Semantic Web - <http://vi.wikipedia.org>
- [2] Trường điều kiện ngẫu nhiên CRFs - <http://vi.wikipedia.org>
- [3] Tìm hiểu về chatbot - <https://viblo.asia/p/tat-ca-nhung-gi-ban-can-biet-ve-chatbot-Az45bnNg5xY>,
- [4] <https://blog.vsoftconsulting.com/blog/understanding-the-architecture-of-conversational-chatbot>, Understanding The Conversational Chatbot Architecture.
- [5] Semantic Web introduction and application - FIT-HUT, Cao Tuan-Dung SE Department,
- [6] Đồ án Ontology trong chứng khoán Việt Nam, Lê Văn Đức K48 Lớp CNPM – Đại học Bách khoa Hà Nội.
- [7] Grigoris Antoniou and Frank van Harmelen, A Semantic Web Primer.
- [8] Semantic Web - <https://www.w3.org/2001/sw/>
- [9] OWL - <https://www.w3.org/OWL/>
- [10] RDF - <https://www.w3.org/TR/rdf11-primer/>
- [11] SPRAQL - <https://www.w3.org/TR/rdf-sparql-query/>
- [12] Các khái niệm liên quan đến Ngôn ngữ tự nhiên - <http://vi.wikipedia.org>
- [13] Language Understanding (LUIS) Documentation. - <https://docs.microsoft.com/en-us/azure/cognitive-services/luis/>
- [14] Azure Bot Service Documentation for Python - <https://docs.microsoft.com/en-us/azure/bot-service/>
- [15] Slide Xử lý ngôn ngữ tự nhiên (NaturalLanguageProcessing) (Natural Language Processing), PGS.TS.Lê Thanh Hương - Bộ môn Hệ thống Thông tin Viện CNTT &TT –Trường ĐHBKHN
- [16] Named Entity Recognition With Conditional Random Fields In Python - <https://www.depends-on-the-definition.com/named-entity-recognitionconditional-random-fields-python/>
- [17] Thu thập và lưu trữ dữ liệu với scrapy - <https://viblo.asia/p/thu-thap-va-luu-tru-du-lieu-voi-scrapy-va-mysql-yMnKMA9EK7P>
- [18] Phân đoạn từ tiếng việt sử dụng mô hình CRFs, 2006, Nguyễn Trung Kiên, Đại học Quốc gia Hà Nội – Đại học công nghệ.
- [19] Slide Web ngữ nghĩa, PGS.TS.Lê Thanh Hương - Bộ môn Hệ thống Thông tin Viện CNTT &TT –Trường ĐHBKHN
- [20] Ontology based Chatbot (For E-commerceWebsite), International Journal of Computer Applications (0975 – 8887) Volume 179 – No.14, January 2018.