2020

CodEbot

Một hệ thống chatbot sử dụng ngôn ngữ Tiếng Việt để giải đáp thắc mắc trong lập trình cơ bản

LƯƠNG CÔNG TÂM, VƯƠNG LÊ MINH NGUYÊN

Ý TƯỞNG

Em bắt đầu tìm hiểu và tự học lập trình vào những năm cấp 2. Khoảng thời gian khi mà em mới bắt đầu học lập trình, em thường gặp rất nhiều khó khăn trong việc tìm kiếm các tài liệu học tập lập trình bằng tiếng Việt, tài liệu tiếng Việt khi đó chưa nhiều, và trình độ tiếng Anh của em khi đó thực sự rất kém nên việc tìm và đọc các tài liệu, hướng dẫn nước ngoài là không thể. Ngoài ra, em cũng không có một người hướng dẫn nào, đôi lúc cũng có rất nhiều khó khăn, thắc mắc ở cả những vấn đề trong lập trình cơ bản nhưng không thể nào tìm được lời giải bằng tiếng Việt.

Nhờ vào những kinh nghiệm, khó khăn mà bản thân đã trải qua trong việc học lập trình ở quá khứ, em đã nghĩ ra một ý tưởng xây dựng một chatbot tiếng Việt, một trợ thủ hỗ trợ, giải đáp và cung cấp các kiến thức trong lập trình cơ bản, hướng tới các đối tượng là các em học sinh trung học có cùng đam mê lập trình với mình để có thể giúp các em dễ dàng tiếp cận với các kiến thức về lập trình cơ bản.

CHỨC NĂNG

Chức năng của CodEbot cũng là những mục tiêu mà tụi em muốn làm để hoàn thiện trợ thủ học tập này. Tuy nhiên, do thời gian và kinh phí của một đề tài nghiên cứu khoa học sinh viên lại quá ít ỏi nên tụi em đã giới hạn lại chỉ phát triển một vài chức năng cơ bản cho CodEbot để hoàn thành đề tài nghiên cứu này. Những chức năng mà tụi em muốn hoàn thành trong đề tài nghiên cứu bao gồm:

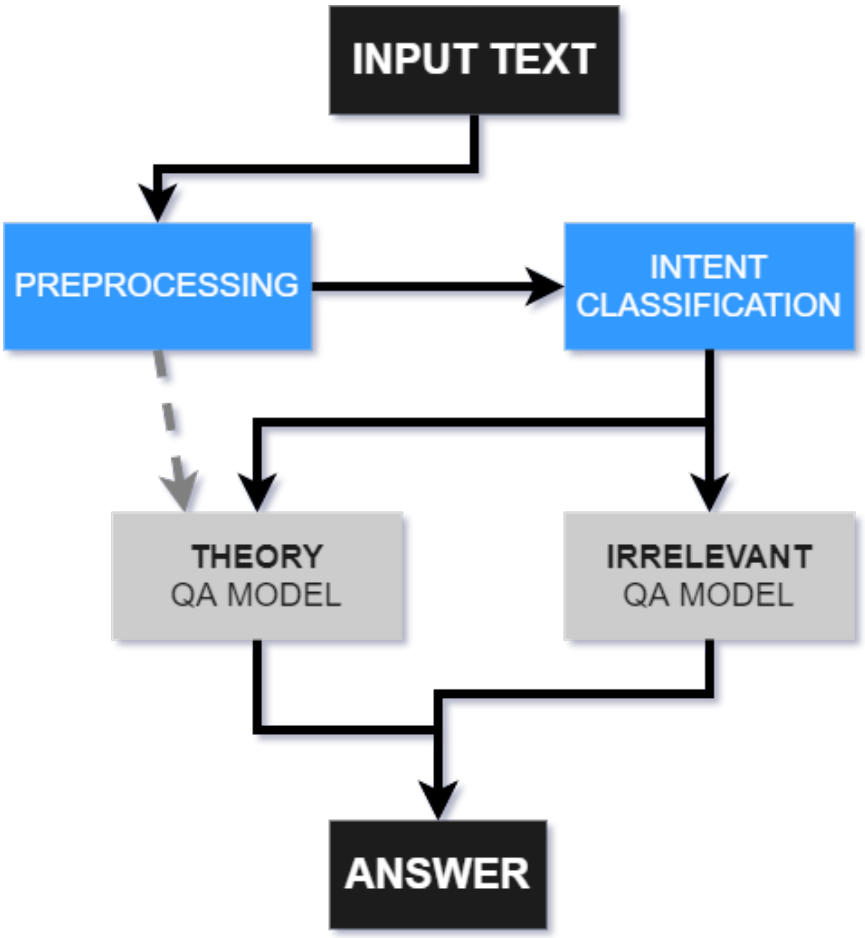
- **Hỗ trợ ngôn ngữ tiếng Việt.** Mặc dù tiếng Anh là một ngôn ngữ quan trọng và cần thiết cho quá trình học tập các kiến thức nâng cao trong lập trình, nhưng đối tượng mà CodEbot nhắm đến là các em học sinh trung học, giúp các em làm quen với các kiến thức lập trình cơ bản, tạo một nền tảng vững chắc trước khi các em tiếp tục quá trình học lập trình nâng cao.
- **Trả lời các câu hỏi lý thuyết về lập trình cơ bản thuộc các ngôn ngữ lập trình C/C++ và Python.** Đây là mục tiêu chính của đề tài, lý thuyết cơ bản là nền tảng của sự phát triển. Không chỉ trả lời các câu hỏi dạng "Cái gì, tại sao, như thế nào", CodEbot còn cung cấp một số kiến thức về sự khác nhau giữa A và B, ưu và nhược của A, … Ví dụ một câu hỏi có dạng như sau: Vòng lặp vô hạn là gì?
- **Giao tiếp.** Một phần mềm chỉ có thể trả lời câu hỏi được nêu ra thì không thể gọi là một Chatbot được, vì thế CodEbot cũng sẽ có thể nói chuyện

tương tác cơ bản với người dùng như chào, hay cập nhật một số thông tin như thời tiết, hay Hot nhất hiện nay là tình hình dịch bệnh Covid-19 hiện tại, hay cung cấp một số tài liệu, diễn đàn hay một vài mẫu sách học lập trình cho người dùng.

- **Tiếp tục trò chuyện theo ngữ cảnh.** Ví dụ khi người dùng đang hỏi điều gì đó về ngôn ngữ C++, người dùng có thể hỏi tiếp một câu hỏi khác và không phải nêu lại ngôn ngữ đó: Vòng lặp for trong C++? – Vậy còn vòng lặp while? – Một số bài tập rèn luyện. Trong ví dụ trên, khi người dùng hỏi câu đầu tiên thì CodEbot sẽ biết được người dùng đang thắc mắc về ngôn ngữ C++, qua câu hỏi thứ hai, CodEbot sẽ trả lời kiến thức thuộc ngôn ngữ C++, và đến câu thứ 3 CodEbot sẽ đưa một số đề bài mẫu về vòng lặp while cho người dùng.

WORKFLOW

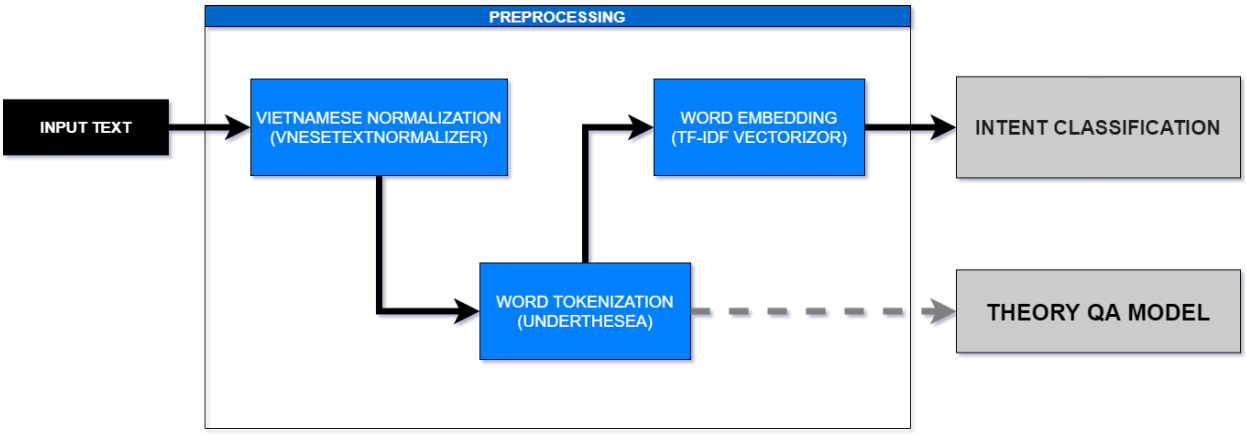
Tổng quan CodEbot sẽ thực hiện như sau:



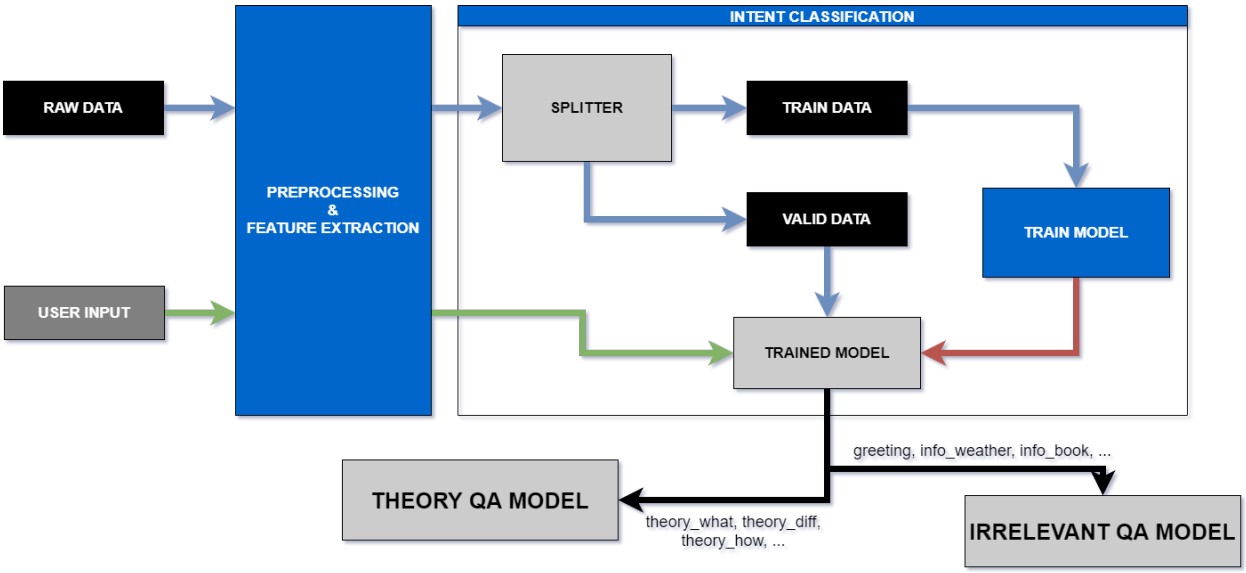
- Đầu tiên, đoạn đối thoại của người dùng sẽ được đưa vào tiền xử lý để lọc ra các đặc trưng trong đoạn đối thoại.
- Các đặc trưng qua quá trình tiền xử lý sẽ được đưa tiếp qua Model dự đoán ý định của người dùng, để xác định xem người dùng đang muốn hỏi về kiến thức trong lập trình hay đang muốn nói về một số các vấn đề không liên quan và tiếp tục đưa qua một trong 2 Model tương ứng với ý định của người dùng.
- Hai Model sẽ thực hiện dự đoán riêng và đưa ra câu trả lời tương ứng với đoạn hội thoại của người dùng.

Preprocessing Pipeline gồm các bước:

- Chuẩn hóa văn bản tiếng Việt
- Áp dụng Word Tokenization để ghép các từ nhiều tiếng lại giúp máy học hiệu quả hơn
- Xây dựng Word Embedding với phương pháp TF-IDF Vectorization



Quá trình huấn luyện Model Intent Classification (Text Classification) sử dụng Model LinearSVC(sklearn)



Về 2 mô hình trả lời câu hỏi thì tụi em vẫn đang lên ý tưởng.

KHÓ KHĂN HIỆN TẠI

Khó khăn lớn nhất hiện tại đối với tụi em có lẽ là Dataset. Em cũng đã tìm thử một số Dataset tiếng Việt có sẵn trên các cộng đồng NLP tiếng Việt, nhưng tụi em lại không tìm được Dataset nào phù hợp với bài toán của mình, nên tụi em đã tự xây dựng một bộ Dataset riêng, tuy nhiên số lượng vẫn còn rất nhỏ.

Tiếp theo là phần chức năng **Tiếp tục trò chuyện theo ngữ cảnh**, do tụi em vẫn còn khá mới trong lĩnh vực NLP và nhất là xây dựng chatbot, nên tụi em vẫn chưa biết phải thực hiện nó như thế nào, em có nghĩ tới chuyện lưu đoạn thông tin ngữ cảnh của câu trước và đưa tiếp vào mô hình trả lời để trả lời cho câu hỏi sau.