TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KÌ**

**MÔN XỬ LÝ NGÔN NGỮ TỰ NHIÊN**

**XÂY DỰNG CHƯƠNG TRÌNH HỎI ĐÁP TỰ ĐỘNG CHO TIẾNG VIỆT - BÀI 1**

*Người hướng dẫn*: **PGS. TS. LÊ ANH CƯỜNG**

*Người thực hiện*: **NGUYỄN QUỐC SƠN – 51800921**

**LÊ NGUYỄN MINH TUẤN – 51800950**

Lớp **: 18050301**

Khoá  **: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2020**

MỤC LỤC

[MỤC LỤC 2](#_Toc70263975)

[CHƯƠNG 1 – MÔ TẢ BÀI TOÁN QUESTION ANSWERING 3](#_Toc70263976)

[1.1 Tổng quan về hệ thống hỏi đáp 3](#_Toc70263977)

[1.2 Kiến trúc của hệ thống hỏi đáp dựa trên truy xuất thông tin 4](#_Toc70263978)

[CHƯƠNG 2 – GIẢI PHÁP VÀ THUẬT TOÁN 6](#_Toc70263979)

[2.1 Giải pháp 6](#_Toc70263980)

[2.2 Tiền xử lý dữ liệu 6](#_Toc70263981)

[2.3 Phương pháp biểu diễn từ không dùng Word2Vec. 6](#_Toc70263982)

[2.4 Phương pháp biểu diễn từ bằng Word2Vec. 7](#_Toc70263983)

[CHƯƠNG 3 – KẾT QUẢ 8](#_Toc70263984)

CHƯƠNG 1 – MÔ TẢ BÀI TOÁN QUESTION ANSWERING

1.1 Tổng quan về hệ thống hỏi đáp

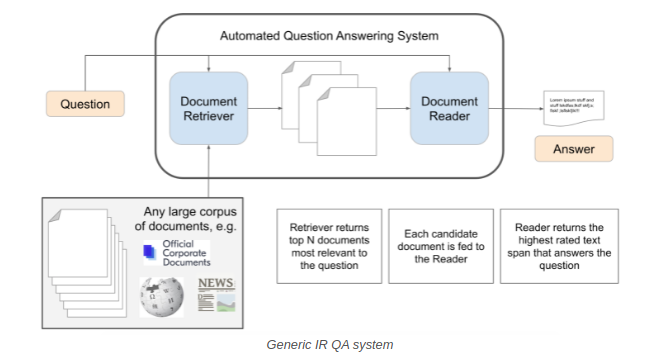
Hệ thống hỏi đáp được định nghĩa là một mô hình có thể trả lời các câu hỏi được đưa ra bằng ngôn ngữ tự nhiên của con người (văn bản hoặc giọng nói). Hệ thống này sẽ đưa ra câu trả lời bằng cách truy vấn/tìm kiếm trong một hệ tri thức nào đó.

Có 3 loại hệ thống hỏi đáp:

- Hệ thống hỏi đáp dựa trên tri thức (Knowledge-Based Systems): Mục tiêu của hệ thống hỏi đáp dựa trên tri thức là ánh xạ các câu hỏi thành các thực thể có cấu trúc (cơ sở dữ liệu quan hệ) thông qua các thuật toán phân tích ngữ nghĩa. Các kỹ thuật phân tích ngữ nghĩa chuyển đổi các chuỗi văn bản thành kí tự logic hoặc ngôn ngữ truy vấn.

- Hệ thống hỏi đáp dựa trên truy xuất thông tin (Information Retrieval-Based Systems): Hệ thống trả lời câu hỏi dựa trên truy xuất thông tin tìm và trích xuất một đoạn văn bản từ một lượng lớn các tài liệu (corpus). Corpus có thể rộng lớn và không liên quan đến một lĩnh vực nào cụ thể như thông tin trên Wikipedia (open domain) hoặc cụ thể như các thông tin liên quan đến công tác tuyển sinh của TDTU (closed domain). Hệ thống hỏi đáp dựa trên truy xuất thông tin không chỉ là công cụ tìm kiếm, sử dụng các thuật ngữ ngôn ngữ tự nhiên để cung cấp danh sách các tài liệu liên quan nhất đến câu hỏi và xử lý trên các tài liệu có liên quan nhất để đưa ra câu trả lời rõ ràng dựa trên nội dung của các tài liệu đó.

1.2 Kiến trúc của hệ thống hỏi đáp dựa trên truy xuất thông tin



Trình truy xuất tài liệu (Document Retriever):

- Trình truy xuất tài liệu có hai công việc cốt lõi: xử lý câu hỏi để sử dụng trong công cụ Information Retrieval (IR) và sử dụng truy vấn IR này để truy xuất các tài liệu và đoạn văn phù hợp nhất.

- Xử lý truy vấn có thể đơn giản như không xử lý gì cả và thay vào đó chuyển toàn bộ câu hỏi đến công cụ IR. Tuy nhiên, nếu câu hỏi dài hoặc phức tạp, nó thường phải được tiền xử lý thông qua các kỹ thuật khác nhau (chẳng hạn như loại bỏ stop word, loại bỏ wh-words, chuyển đổi thành n-gram hoặc trích xuất các thực thể được đặt tên (NER) làm từ khóa.

- Các thuật toán thường được sử dụng để truy xuất tài liệu liên quan: TF-IDF, Okapi BM25 (phiên bản cải thiện thuật toán của TF-IDF).

Trình đọc tài liệu (Document Reader):

- Khi đã chọn được các tài liệu hoặc đoạn văn có liên quan, giờ là lúc trích xuất câu trả lời. Mục đích duy nhất của trình đọc tài liệu là áp dụng các thuật toán đọc hiểu cho các đoạn văn bản để trích xuất câu trả lời. Các thuật toán đọc hiểu hiện đại có hai loại: dựa trên tính năng (feature-based) và dựa trên mạng thần kinh (neural-based).

- Trích xuất câu trả lời dựa trên tính năng có thể bao gồm rule-based, đối sánh mẫu regex hoặc một bộ mô hình NLP (POS Tagging hoặc NER) được thiết kế để xác định các tính năng sẽ cho phép thuật toán học tập có giám sát xác định xem một khoảng văn bản có chứa câu trả lời hay không.

- Phương pháp đọc hiểu dựa trên mạng thần kinh tận dụng ý tưởng rằng câu hỏi và câu trả lời giống nhau về mặt ngữ nghĩa. Thay vì dựa vào từ khóa, các phương pháp này sử dụng bộ dữ liệu mở rộng cho phép mô hình học cách nhúng ngữ nghĩa cho câu hỏi và đoạn văn.

CHƯƠNG 2 – GIẢI PHÁP VÀ THUẬT TOÁN

2.1 Giải pháp

Sau khi nhận vào một câu hỏi (question), tiến hành dùng các thuật toán tìm độ liên quan giữa câu hỏi đó và các câu hỏi trong corpus. Khi tìm được câu hỏi liên quan nhất trong corpus, trả về câu trả lời dựa trên index của câu hỏi liên quan nhất trong corpus.

2.2 Tiền xử lý dữ liệu

Dữ liệu được cung cấp có dạng: question \_\_eou\_\_ answer \_\_eou\_\_

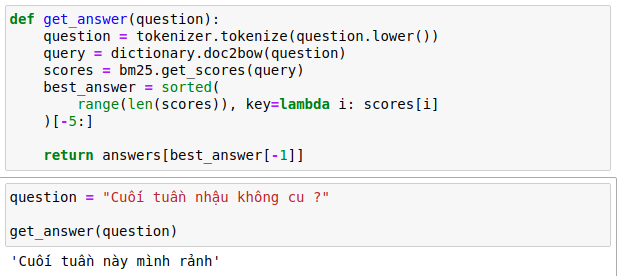
Đầu tiên, cần split 1 câu ra thành question và answer và lưu vào list.

Tiếp theo, vì dữ liệu không quá dài, bỏ qua bước loại bỏ stop word và chỉ tiến hành lowercase dữ liệu loại bỏ dấu câu.

Cuối cùng là tokenize question thành các token.

2.3 Phương pháp biểu diễn từ không dùng Word2Vec.

Trong phương pháp này, thuật toán Okapi BM25 được sử dụng để tìm câu hỏi liên quan nhất. BM25 sẽ trả về score (độ liên quan càng cao, score càng lớn).

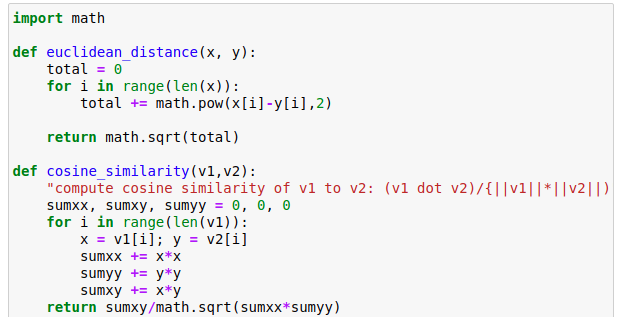


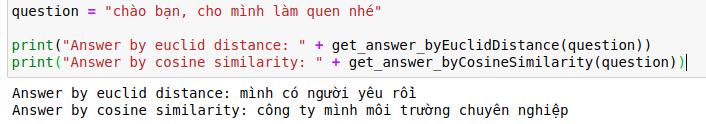
2.4 Phương pháp biểu diễn từ bằng Word2Vec.

Trong phương pháp này, trước khi huấn luyện mô hình Word2Vec, nhóm em sử dụng công cụ Word Segmentation để tokenize các câu hỏi và câu trả lời thành các token. Sau đó huấn luyện mô hình Word2Vec bằng dữ liệu đã được segment.

Word2vec chuyển một token thành một vector. Để chuyển các vector của các token trong câu thành một vector đại diện cho một câu, nhóm em sử dụng phương pháp cộng vector. Khi nhận vào một câu hỏi, tiến hành segment câu hỏi đó thành token, chuyển các token thành các vector bằng Word2Vec và cộng các vector đó lại với nhau thành một vector đại diện cho câu hỏi đó.

Trong phương pháp này, nhóm em sử dụng cả 2 thuật toán để tìm độ liên quan giữa các vector đó là: Euclidean Distance và Cosine Similarity.



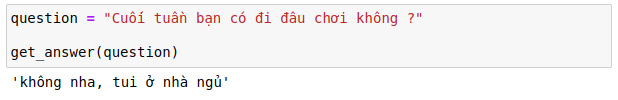


CHƯƠNG 3 – KẾT QUẢ

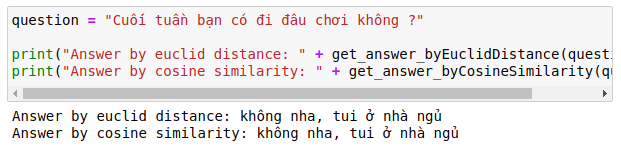
So sánh 3 kết quả của BM25, Euclidean Distance và Cosine Similarity dựa vào 3 câu hỏi.

*Câu hỏi 1: Cuối tuần bạn có đi đâu chơi không ?*

**BM25:**

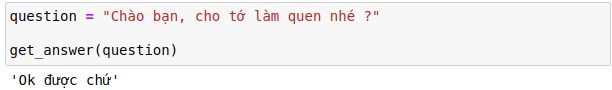
****

**Euclidean Distance và Cosine Similarity**

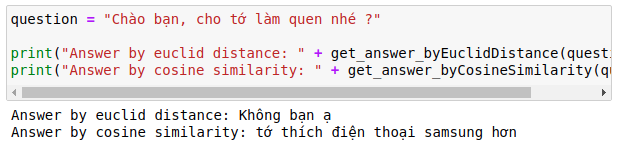
****

*Câu hỏi 2: Chào bạn, cho tớ làm quen nhé ?*

**BM25:**

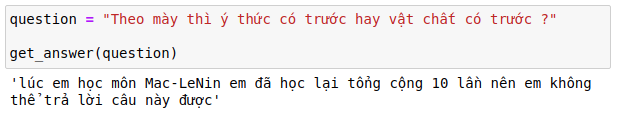
****

**Euclidean Distance và Cosine Similarity**

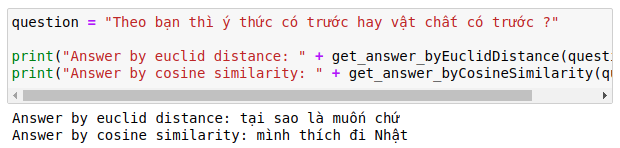
****

*Câu hỏi 3: Chào bạn, cho tớ làm quen nhé ?*

**BM25:**

****

**Euclidean Distance và Cosine Similarity**

****

**Kết luận:** thông qua ba câu hỏi, ta có thể thấy phương pháp biểu diễn từ không dùng Word2Vec và tìm câu hỏi liên quan bằng thuật toán BM25 hoạt động tốt hơn phương pháp biểu diễn từ bằng Word2Vec.

**TÀI LIỆU THAM KHẢO**

1. <https://viblo.asia/p/bm25-thuat-toan-xep-hang-cac-van-ban-theo-do-phu-hop-Az45bWGNKxY>
2. <https://qa.fastforwardlabs.com/methods/background/2020/04/28/Intro-to-QA.html>