Giảng viên hướng dẫn: Lê Anh Cường

Sinh viên (Họ và tên – MSSV):  
1/ Nguyễn Minh Đăng Khoa – 51800882

2/ Nguyễn Hữu Lợi – 51702134

**DỰ ÁN CUỐI KỲ: BÀI 1**

MÔN: NHẬP MÔN XỬ LÝ NGÔN NGỮ TỰ NHIÊN

**I/ Mô tả bài toán**

- Xây dựng chương trình hỏi đáp tự động (question answering) cho tiếng Việt.

- Mô tả yêu cầu:

+ Nhập vào câu hỏi Q.

+ Hệ thống đưa ra câu trả lời A.

- Chương trình hỏi đáp tự động (Question Answering) là nhiệm vụ tự động trả lời câu hỏi dưới dạng ngôn ngữ tự nhiên. Để tìm câu trả lời cho một câu hỏi, một chương trình máy tính sử dụng một một tập các văn bản ngôn ngữ tự nhiên. Với câu hỏi đầu vào, chương trình đưa ra câu trả lời đơn giản với việc so khớp các từ khóa. Xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP) được sử dụng để nhận ra cấu trúc cú pháp và ngữ nghĩa của câu hỏi và các tài liệu thích hợp chứa đựng các câu trả lời cho câu hỏi sẽ được trả lại.

**II/ Giải pháp**

- Trên thực tế, các thuật ngữ trong câu hỏi của người hỏi không phải lúc nào cũng khớp với các câu hỏi trong bộ dữ liệu của chúng ta. Để giảm bớt khoảng cách từ vựng trong chương trình hỏi đáp, chúng ta sẽ sử dụng word embeddings (Vector hóa văn bản). Phương pháp này biểu diễn các văn bản đó theo vector. Trong đó, các từ/cụm từ thuộc kho tài liệu ngôn ngữ được ánh xạ thành những vector trên hệ không gian số thực.

- Và thuật toán của chúng ta phải so sánh câu hỏi của người hỏi với các câu/cụm từ trong bộ dữ liệu của chúng ta. Đây là một vấn đề so sánh độ giống nhau của câu và đây là cách giải quyết:

+ Chuyển câu hỏi của người hỏi thành một vectơ.

+ Chuyển đổi bộ dữ liệu câu hỏi thường gặp thành vectơ.

+ So sánh cả hai và đo độ tương đồng của cosin.

+ Lấy kết quả phù hợp nhất.

**III/ Thuật toán**

*1/ Phương pháp biểu diễn từ bằng word2vec và dựa tìm kiếm thông tin tương tự trong dữ liệu huấn luyện*

- Word2vec là thuật toán theo phương pháp dự đoán (Prediction-based embedding). Mô hình dự đoán học biểu diễn vector từ thông qua những từ ngữ cảnh xung quanh nhằm cải thiện khả năng dự đoán ý nghĩa các từ. Có hai cách xây dựng mô hình Word2vec để biểu diễn sự phân tán của từ trong không gian vector là CBOW và Skipgram. Và trong bài code sử dụng mô hình Word2Vec sử dụng Skipgram.

- Sử dụng một từ để dự đoán ngữ cảnh mục tiêu (Continuous skip-gram) xem xét những từ ngữ cảnh xung quanh sẽ được đánh giá tốt hơn so với những từ trong ngữ cảnh nhưng ở vị trí xa hơn. Mặc dù thứ tự từ vẫn không được xem xét, mỗi vector của từ bối cảnh được xem xét và cân nhắc.

- Sau đó viết một hàm để chuyển đổi một câu hỏi về vector và chúng ta sẽ tính khoảng cách giữa hai vector của câu hỏi nhập vào và câu hỏi trong bộ huấn luyện (thư viện scipy - spatial.distance.cosine). Nếu khoảng cách càng lớn thì hai vector càng khác nhau (không có sự tương đồng), để có thể tìm kiếm được sự tương đồng cuả câu hỏi nhập và câu hỏi trong dữ liệu huấn luyện thì khoảng cách phải nhỏ nhất có thể.

*2/ Phương pháp biểu diễn từ không dùng word2vec (dùng TF-IDF) và dựa tìm kiếm thông tin tương tự trong dữ liệu huấn luyện.*

- TF-IDF (term frequency–inverse document frequency) là tần suất - tần suất đảo nghịch từ. Đây là một phương pháp thống kê, nhằm phản ánh độ quan trọng của mỗi từ hoặc n-gram đối với văn bản trên toàn bộ tài liệu đầu vào. TF-IDF thể hiện trọng số của mỗi từ theo ngữ cảnh văn bản. TF-IDF sẽ có giá trị tăng tỷ lệ thuận với số lần xuất hiện của từ trong văn bản và số văn bản có chứa từ đó trên toàn bộ tập tài liệu.

- Sau khi vectorizer câu hỏi nhập và câu hỏi trong bộ dữ liệu huấn luyện chúng ta sẽ tính độ tương đồng giữa hai vector của câu hỏi nhập và từng câu hỏi trong bộ huấn luyện (thư viện sklearn - sklearn.metrics.pairwise.cosine\_similarity). Sau đó tìm độ tương đồng lớn nhất và đưa ra câu trả lời.

**III/ Kết quả của Word2vec và TF-IDF**

