Báo cáo

Xây dựng mô hình không gian vector cho tiếng Việt

Trong các bài toán xử lí ngôn ngữ tự nhiên, chúng ta đều phải biểu diễn từ dưới dạng số để máy tính có thể hiểu được và vận dụng cơ sở toán học để giải quyết các bài toán. Các phương pháp phổ biến bao gồm: ma trận đồng xuất hiện, tf-idf, sử dụng trực tiếp chỉ số của từ vựng trong từ điển. Các phương pháp này tuy đơn giản nhưng lại gặp một số vấn đề như:

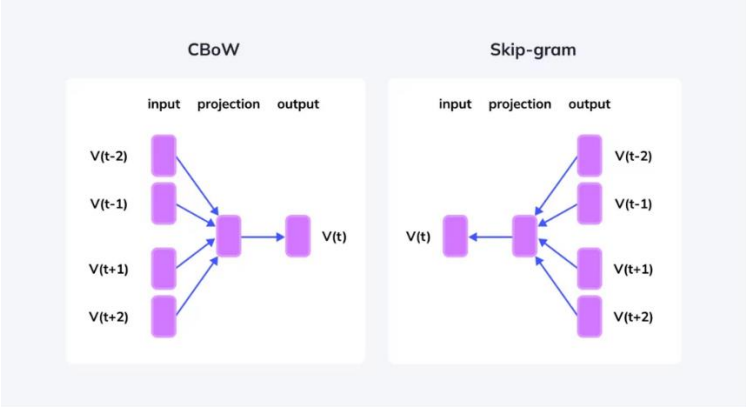
- Ma trận biểu diễn thưa: gây tốn bộ nhớ, kích thước lớn dẫn đến việc tính toán lâu hơn cho các bài toán khác
- Các giá trị không tập trung: việc sử dụng số thứ tự trong từ điển khó chuẩn hóa, các giá trị dàn trải dẫn đến việc khó học khi sử dụng biểu diễn này trong các bài toán khác.

Chính điều đó đã thúc đẩy sự ra đời của các mô hình word2vec, fastext. Trong báo cáo này, chúng ta sẽ **Xây dựng mô hình không gian vector cho tiếng Việt.**

A. Mô hình

1. word2vec: Word2Vec là một kỹ thuật học máy được giới thiệu bởi nhóm nghiên cứu của Google, cụ thể là bởi Tomas Mikolov và các đồng sự vào năm 2013. Word2Vec chuyển đổi các từ trong văn bản thành các vector số học. Các vector này có ý nghĩa ngữ nghĩa, nghĩa là các từ có nghĩa tương tự nhau sẽ có các vector gần nhau trong không gian vector. Mô hình này sử dụng hai kiến trúc chính để huấn luyện mô hình: Continuous Bag of Words (CBOW) và Skip-gram.

    - **CBOW**: Dự đoán từ trung tâm (target word) dựa trên các từ ngữ cảnh xung quanh nó.
    - **Skip-gram**: Dự đoán các từ ngữ cảnh dựa trên từ trung tâm.



2. FastText: FastText học các biểu diễn từ dưới dạng vector, tương tự như Word2Vec. Tuy nhiên, một cải tiến quan trọng của FastText là việc sử dụng các n-gram ký tự để xây dựng biểu diễn từ. Điều này giúp FastText có khả năng nhận biết tốt hơn về các từ mới hoặc từ viết sai chính tả bằng cách xem xét các thành phần n-gram bên trong từ đó.

B. Triển khai:

- Dữ liệu: Bao gồm các câu tiếng Việt trên mỗi dòng trong file. Đã qua tiền xử lí.
- Mô hình: sử dụng mô hình word2vec, FastText của genism.

C. Kết quả:

Bảng 1: Từ mục tiêu và mô hình đoán từ có nghĩa tương tự

Từ mục tiêu	Các từ có nghĩa tương tự
thích	yêu mến,vui,đẹp,xinh đẹp,giỏi,phi tiễn,nhớ,yêu,tài năng,tuyệt vời
ghét	can ngăn,kính trọng,hãm hại,sợ hãi,bằng lòng,xót xa,chuyên quyền,can đảm,kiêu ngạo,làm loạn
vợ	con gái,con trai,cha,chị,em gái,em,anh trai,cháu,em trai,công chúa

Hình 1: Minh họa mức độ tương đồng giữa các từ