Nguồn:

+) <https://phamdinhkhanh.github.io/2019/04/29/ModelWord2Vec.html#2-word-embedding>

+) <https://www.kaggle.com/code/phamdinhkhanh/singular-value-decomposition>

+) <https://viblo.asia/p/tim-hieu-ve-autoencoder-oOVlYvJv58W>

+) <https://bizflycloud.vn/tin-tuc/autoencoder-la-gi-20220526165157229.htm>

I. Word Representation

1. Giới thiệu

Dữ liệu đầu vào trong xử lý ngôn ngữ tự nhiên (NLP) là các ký tự và dấu câu. Để đưa dữ liệu văn bản vào mạng nơ-ron, cần lượng hóa từ ngữ thành dạng số.

Một số khái niệm trong word representation:

* Documents (Văn bản): Tập hợp các câu trong cùng một đoạn văn có liên hệ với nhau.
* Corpus (Bộ văn bản): Tập hợp nhiều văn bản thuộc các chủ đề khác nhau, tạo thành nguồn dữ liệu văn bản. Một văn bản cũng có thể coi là corpus của các câu trong nó. Các bộ corpus lớn thường chứa từ vài nghìn đến vài trăm nghìn văn bản.
* Character (Ký tự): Bao gồm chữ cái và dấu câu. Mỗi ngôn ngữ có bộ ký tự riêng.
* Word (Từ vựng): Kết hợp các ký tự tạo thành từ có ý nghĩa. Tiếng Việt có nhiều từ ghép (2-3 âm tiết), khác với tiếng Anh (chủ yếu từ đơn).
* Dictionary (Từ điển): Tập hợp tất cả từ vựng xuất hiện trong văn bản.
* Vocabulary (Từ vựng): Tương tự từ điển, là tập hợp các từ được trích xuất từ văn bản.

Để biểu diễn từ dưới dạng số, cần xác định từ điển của văn bản. Số lượng từ trong từ điển là hữu hạn và được lặp lại trong các câu.

2. Phương pháp biểu diễn từ bằng One-Hot Vector

Giả sử ta có từ điển là tập hợp gồm 6 từ vựng ["học", "máy", "xử", "lý", "ngôn", "ngữ"]

Từ đó, mỗi từ trong câu có thể mã hóa thành một one-hot vector trong ma trận biểu diễn, với:

* Chiều dài vector = số lượng từ trong từ điển.
* Giá trị 1 tại vị trí tương ứng với từ trong từ điển, các vị trí khác bằng 0.

Ví dụ:

* Từ "máy" → one-hot vector: .

II.Word Embedding

Sau khi biểu diễn từ dưới dạng one-hot vector, mô hình đã có thể huấn luyện được từ dữ liệu được mã hóa. Tuy nhiên dữ liệu này chỉ đáp ứng được khả năng huấn luyện mà chưa phản ảnh được mối liên hệ về mặt ngữ nghĩa của các từ. Các hạn chế đó là:

* Mối quan hệ tương quan giữa các cặp từ bất kì luôn là không tương quan (tức bằng 0). Do đó không có tác dụng trong việc tìm mối liên hệ về nghĩa.
* Kích thước của véc tơ sẽ phụ thuộc vào số lượng từ vựng có trong bộ văn bản dẫn đến chi phí tính toán rất lớn khi tập dữ liệu lớn.
* Khi bổ sung thêm các từ vựng mới số chiều của véc tơ có thể thay đổi theo dẫn đến sự không ổn định trong shape.

Do đó các thuật toán word embedding được tạo ra nhằm mục đích tìm ra các véc tơ đại diện cho mỗi từ sao cho:

* Một từ được biểu diễn bởi một véc tơ có số chiều xác định trước.
* Các từ thuộc cùng 1 nhóm thì có khoảng cách gần nhau trong không gian.

Có nhiều phương pháp word embedding khác nhau có thể kể đến. Trong đó có 3 nhóm chính:

* Sử dụng thống kê tần xuất: tfidf
* Các thuật toán giảm chiều dữ liệu: SVD, PCA, auto encoder, word2vec
* Phương pháp sử dụng mạng nơ ron: word2vec, ELMo, BERT.