Nguồn:

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

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

Phương pháp Autoencoder tiếp cận Word Embedding bằng cách học một biểu diễn nén của biểu diễn gốc của từng từ.

1. Đầu vào (Input)

Ma trận biểu diễn mối quan hệ giữa các từ.

* Ma trận đồng xuất hiện
* One-hot Vector

2. Đầu ra (Output)

Tập hợp các vector từ nhúng (Word Embedding Vectors) cho mỗi từ trong từ vựng.

* Mỗi từ trong từ vựng (​) được gán một vector chiều ().
* Các vector này thường được lưu trữ dưới dạng một ma trận có kích thước , trong đó mỗi hàng là vector nhúng của một từ.

3. Quy trình (Process)

* Bước 1: Xây dựng Ma trận Đồng xuất hiện (Build the Co-occurrence Matrix)

Tiền xử lý corpus, xây dựng từ vựng (V), chọn kích thước cửa sổ ngữ cảnh, và xây dựng ma trận đồng xuất hiện M kích thước V×V (thường dùng giá trị PPMI).

* Bước 2: Thiết kế Kiến trúc Autoencoder (Design the Autoencoder Architecture)

Autoencoder gồm 2 phần chính:

1. Encoder: Nén one-hot vector thành embedding vector chiều thấp
2. Decoder: Tái tạo one-hot vector ban đầu từ embedding

* Bước 3: Huấn luyện Mô hình Autoencoder (Train the Autoencoder Model)

Mục tiêu huấn luyện là làm cho đầu ra của Autoencoder giống nhất có thể với đầu vào vector một nóng ban đầu.

1. Forward pass:
   * Input → encoder → embedding vector chiều
   * Embedding vector → decoder → vector tái tạo
2. Hàm mất mát:
   * So sánh vector tái tạo với input
   * Sử dụng cross-entropy loss hoặc MSE loss
3. Truyền ngược (Backward Pass) và Cập nhật tham số:
   * Sử dụng thuật toán backpropagation
   * Cập nhật trọng số bằng các phương pháp như SGD

* Bước 4: Trích xuất Vector Từ nhúng (Extract Word Embeddings)
  + Sau khi Autoencoder được huấn luyện và đạt khả năng tái tạo tốt, phần Bộ giải mã sẽ không còn được sử dụng.
  + Phần Bộ mã hóa đã huấn luyện (từ lớp đầu vào đến lớp thắt cổ chai) chính là mô hình dùng để tạo vector nhúng.
  + Để lấy vector nhúng của một từ bất kỳ trong từ vựng, đưa vector một nóng của từ đó vào Bộ mã hóa.
  + Giá trị kích hoạt của các nơ-ron trong lớp thắt cổ chai chính là vector từ nhúng k chiều cho từ đó.