Nguồn:

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

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

Phương pháp SVD trong Word Embedding là một kỹ thuật dựa trên phân tích ma trận thống kê (ma trận đồng xuất hiện - co-occurrence matrix), để tạo ra các vector biểu diễn từ trong không gian chiều thấp.

1. Đầu vào (Input)

Ma trận biểu diễn mối quan hệ giữa các từ.

* + Ma trận Đồng xuất hiện (Co-occurrence Matrix).
  + Ma trận này có kích thước , trong đó là kích thước từ vựng.
  + Các giá trị trong ma trận ​ biểu thị tần suất từ đứng trước từ

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

Bước 1: Tiền xử lý dữ liệu và xây dựng từ vựng

* Thu thập corpus.
* Thực hiện các bước tiền xử lý cần thiết (ví dụ: tách từ, chuyển chữ thường).
* Xây dựng từ vựng gồm các từ duy nhất. Kích thước từ vựng là .

Bước 2: Xây dựng Ma trận Đồng xuất hiện

Bước 3: Áp dụng Singular Value Decomposition (SVD)

Thực hiện phép phân rã SVD trên ma trận :

* : Ma trận trực giao.
* : Ma trận đường chéo chứa các giá trị suy biến (singular values), sắp xếp giảm dần.

Bước 4: Giảm Chiều dữ liệu

* Chọn số chiều mong muốn cho vector nhúng, gọi là ().
* Giữ lại giá trị lớn nhất trên đường chéo của ma trận , tạo thành ma trận kích thước .
* Giữ lại cột đầu tiên của ma trận , tạo thành ma trận ​ kích thước .
* Giữ lại hàng đầu tiên của ma trận , tạo thành ma trận ​ kích thước .

Bước 5: Trích xuất Vector Từ nhúng

* Ma trận ​ () chính là ma trận chứa các vector từ nhúng. Mỗi hàng của ​ là vector nhúng chiều cho từ tương ứng.
* Trong một số biến thể, vector nhúng cho từ có thể là hàng thứ của ma trận ​ hoặc ​.