1. Chuẩn bị Corpus

- Bộ văn bản: “Tôi học. Tôi chơi”

-Từ điển (dictionary): “Tôi”, “học”, “chơi”

2. Xây dựng ma trận đồng xuất hiện (Co-occurrence Matrix)

* Cửa sổ ngữ cảnh (window size) bằng 1 (chỉ xét các từ liền kề).
* Đếm số lần từ  xuất hiện trước từ :

|  |  |  |  |
| --- | --- | --- | --- |
|  | Tôi | Học | Chơi |
| Tôi | 0 | 1 | 1 |
| Học | 0 | 0 | 0 |
| Chơi | 0 | 0 | 0 |

Giải thích:

* “Tôi” xuất hiện cạnh (phí trước) “học” 1 lần
* “Tôi” xuất hiện cạnh (phía trước) “chơi” 1 lần

Ma trận X:

3. Phân tích X thành

Công thức

Trong đó:

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

Các bước tính toán SVD

Bước 1: Tính và

Tính (chuyển vị của )

Tính :

Tính :

Bước 2: Tính các giá trị riêng (eigenvalues) và vector riêng (eigenvectors) của để tìm và .

Ma trận

Để tìm các giá trị riêng , chúng ta giải phương trình đặc trưng , trong đó là ma trận đơn vị.

Vậy các giá trị riêng là

Các giá trị suy biến (singular values) ​ là căn bậc hai của các giá trị riêng khác không của . Chúng được sắp xếp theo thứ tự giảm dần.

Ma trận là ma trận đường chéo 3×3 với các giá trị suy biến trên đường chéo chính:

Bây giờ, chúng ta tìm các vector riêng trực chuẩn (orthonormal eigenvectors) tương ứng với từng giá trị riêng của . Đây sẽ là các cột của ma trận .

Với :

Giải hệ phương trình ta có:

Chuẩn hóa ta có

Với :

Điều này dẫn đến một phương trình duy nhất: .

Chúng ta cần tìm hai vector riêng trực giao ​ thỏa mãn ​.

Chọn các giá trị như sau: thỏa mãn và

Chuẩn hóa ta có

Với :

Để tìm ​, chúng ta cần ​ thỏa mãn và ​.

Chọn các giá trị như sau: thỏa mãn các điều kiên trên

Chuẩn hóa ta có

Các vector riêng này tạo thành các cột của ma trận :

Vậy là:

Bước 3: Tính các vector riêng (eigenvectors) của để tìm .

Ma trận

Để tìm các giá trị riêng , chúng ta giải phương trình đặc trưng , trong đó là ma trận đơn vị.

Vậy các giá trị riêng là

Các giá trị suy biến (singular values) ​ là căn bậc hai của các giá trị riêng khác không của . Chúng được sắp xếp theo thứ tự giảm dần.

Bây giờ, chúng ta tìm các vector riêng tương ứng với từng giá trị riêng của .

Với :

Giải hệ phương trình ta có:

Chuẩn hóa ta có

Với :

Điều này dẫn đến một phương trình duy nhất: .

Chọn các giá trị như sau: thỏa mãn điều kiện trên

Chuẩn hóa ta có

Với :

Điều này dẫn đến một phương trình duy nhất: .

Chọn các giá trị như sau: thỏa mãn điều kiện trên

Chuẩn hóa ta có

Các vector riêng này tạo thành các cột của ma trận :

Bước 4: Tổng hợp kết quả

Bước 5: Kiểm tra

**4. Chọn số chiều nhúng (k=2*k*=2)**

Giữ lại 2 singular values lớn nhất để giảm chiều:

**Kết quả Embeddings**:

Embedding cho “Tôi” là (cột 0)

Embedding cho “học” là (cột 1)

Embedding cho “chơi” là (cột 2)