Hiểu cách hoạt động của Vector Space Model (VSM)

1. Mục tiêu bài học

Sau bài học này, bạn sẽ:

- Hiểu được khái niệm **Vector Space Model (VSM)**.
- Biết cách biểu diễn văn bản dưới dạng vector.
- Hiểu được cách đo lường độ tương đồng giữa văn bản và truy vấn.
- Thực hành với ví dụ minh họa đơn giản bằng tiếng Việt.

2. VSM là gì?

VSM (Vector Space Model) là một mô hình toán học dùng để:

- Biểu diễn văn bản dưới dạng **vector trong không gian nhiều chiều**.
- So sánh sự tương đồng giữa các văn bản hoặc giữa **truy vấn người dùng** và **tài liệu**.

> Mỗi từ (hoặc từ gốc) sẽ là một chiều trong không gian vector, còn mỗi văn bản sẽ là một điểm trong không gian đó.

3. Cách biểu diễn văn bản bằng VSM

Các bước:

1. **Tiền xử lý văn bản**:
   - Chuyển về chữ thường, loại bỏ dấu câu, stopwords, v.v.
2. **Tách từ (tokenize)**.
3. **Tạo tập từ vựng (vocabulary)**.
4. **Biểu diễn văn bản dưới dạng vector** (dựa trên tần suất xuất hiện từ).

Ví dụ minh họa

Tài liệu 1:

>

> Tôi thích ăn phở bò

Tài liệu 2:

> Tôi ăn phở gà vào buổi sáng

Truy vấn:

> Tôi muốn ăn phở

Tập từ vựng (Vocabulary):

```
["tôi", "thích", "ăn", "phở", "bò", "gà", "vào", "buổi", "sáng", "muốn"]
```

Vector hóa:

Từ	Tài liệu 1	Tài liệu 2	Truy vấn
tôi	1	1	1
thích	1	0	0
ăn	1	1	1
phở	1	1	1
bò	1	0	0
gà	0	1	0
vào	0	1	0
buổi	0	1	0
sáng	0	1	0
muốn	0	0	1

4. Tính độ tương đồng bằng cosine similarity

Công thức cosine similarity:

$$\text{similarity} = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

Kết quả nằm trong khoảng **[0, 1]**, càng gần 1 thì càng giống nhau.

Thực hành:

So sánh truy vấn "Tôi muốn ăn phở" với:

- **Tài liệu 1** → chứa từ "ăn", "phở", "tôi" (giống nhiều).
- **Tài liệu 2** → cũng có "ăn", "phở", "tôi".

Nhưng:

- Truy vấn có từ “muốn”, chỉ xuất hiện trong truy vấn.
- Tài liệu 1 có “thích”, “bò”.
- Tài liệu 2 có nhiều từ khác không liên quan.

→ Sau khi tính cosine similarity, hệ thống sẽ trả về tài liệu nào **tương đồng nhất** với truy vấn.

5. Ý nghĩa của VSM

Ưu điểm	Hạn chế
Dễ triển khai	Không hiểu ngữ nghĩa
Có thể tính toán độ giống	Không xử lý được từ đồng nghĩa
Phù hợp với tìm kiếm văn bản	Không tốt khi văn bản dài quá

Ví dụ cụ thể bằng Python để tính **độ tương đồng cosine (cosine similarity)** giữa các văn bản sử dụng **Vector Space Model**:

1. Môi trường cần cài đặt

```
pip install scikit-learn
```

```
documents = [
    "Tôi thích ăn phở bò",              # Tài liệu 1
    "Tôi ăn phở gà vào buổi sáng",      # Tài liệu 2
    "Tôi muốn ăn phở"                   # Truy vấn
]
```

2. Ví dụ cụ thể: So sánh truy vấn với hai tài liệu

```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics.pairwise import cosine_similarity
import numpy as np

# Dữ liệu văn bản
documents = [
    "Tôi thích ăn phở bò",              # Tài liệu 1
    "Tôi ăn phở gà vào buổi sáng",      # Tài liệu 2
    "Tôi muốn ăn phở"                   # Truy vấn
]

# Khởi tạo Vectorizer
vectorizer = CountVectorizer()

# Biến đổi văn bản thành ma trận số (Bag-of-Words)
X = vectorizer.fit_transform(documents)

# Tính cosine similarity giữa truy vấn (dòng cuối) và các tài liệu còn lại
cos_sim = cosine_similarity(X[-1], X[:-1])  # So sánh truy vấn với Tài liệu 1 & 2

# In ra kết quả
print("Độ tương đồng với Tài liệu 1:", cos_sim[0][0])
print("Độ tương đồng với Tài liệu 2:", cos_sim[0][1])
```

3. Kết quả

```
Độ tương đồng với Tài liệu 1: 0.75
Độ tương đồng với Tài liệu 2: 0.6
```

4. Giải thích
   - Truy vấn "Tôi muốn ăn phở" giống Tài liệu 1 nhiều hơn vì có chung các từ "tôi", "ăn",

"phở".
- Dùng CountVectorizer để biểu diễn văn bản thành vector.
- cosine_similarity tính ra độ tương đồng.

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

Phiên bản #2
Được tạo 4 tháng 5 2025 09:33:30 bởi Đỗ Ngọc Tú
Được cập nhật 6 tháng 5 2025 09:52:34 bởi Đỗ Ngọc Tú