Thực hành mô hình không gian vector

Nguyễn Mạnh Hiển
hiennm@tlu.edu.vn

Đề bài

Lập trình Python để thực hiện cơ chế vào-ra như bên dưới. (Làm theo hướng dẫn ở các slide tiếp theo).

- Đầu vào:
  - Một tập từ T
  - Một tập văn bản D
  - Một câu truy vấn Q
- Đầu ra:
  - Tập văn bản được sắp xếp theo thứ tự giảm dần của độ đo tương tự côsin giữa mỗi văn bản và câu truy vấn.

Bước 1. Biểu diễn vector

Viết hàm để xuất ra vector biểu diễn một văn bản.

- Đầu vào:
  - Một tập từ T
  - Một văn bản d
- Đầu ra:
  - Một vector biểu diễn văn bản d, trong đó mỗi tọa độ của vector là trọng số từ tf của mỗi từ trong tập từ T.

Gợi ý: Dùng kiểu danh sách trong Python để lưu trữ tập từ và vector. Kiểu danh sách có phương thức `count` để đếm số lần xuất hiện của một giá trị.

Bước 2. Chuẩn hóa vector

Viết hàm chuẩn hóa vector (biểu diễn một văn bản hoặc câu truy vấn) về chiều dài đơn vị.

- Đầu vào:
  - Một vector
- Đầu ra:
  - Vector sau khi đã chuẩn hóa

Bước 3. Tính độ tương tự côsin

Viết hàm tính độ tương tự côsin.

- Đầu vào:
  - Vector câu truy vấn (không chuẩn hóa về chiều dài đơn vị)
  - Vector văn bản (đã chuẩn hóa về chiều dài đơn vị)
- Đầu ra:
  - Độ tương tự côsin (bằng tích vô hướng của hai vector)

Bước 4. Tính điểm số và xếp hạng các văn bản

Gọi các hàm đã viết trong các slide trước để thực hiện các việc sau đây:

1. Gõ trực tiếp trong code:
   - Một tập từ T gồm khoảng 9-10 từ;
   - Một tập văn bản D gồm khoảng 9-10 văn bản (mỗi văn bản gồm khoảng 5-6 từ);
   - Một câu truy vấn Q gồm khoảng 2-3 từ.
2. Tính điểm số côsin giữa mỗi văn bản trong tập văn bản D với câu truy vấn Q.
3. Sắp xếp tập văn bản D giảm dần theo điểm số côsin.
4. In các văn bản trong tập văn bản D (có kèm theo điểm số côsin) lên màn hình.