Numpy - Term frequency and Invert Document Frequency

Hoàng-Nguyên Vũ

1. **Mô tả: Term Frequency–inverse document frequency**

   - **TF-IDF** (term frequency–inverse document frequency) là một kỹ thuật phổ biến trong lĩnh vực xử lý ngôn ngữ tự nhiên và khai thác văn bản. Mục đích của TF-IDF là để đánh giá tầm quan trọng của một từ trong một tài liệu so với toàn bộ tập hợp các tài liệu (corpus).

   ## Term Frequency (TF)

   Đo lường tần suất xuất hiện của một từ trong một tài liệu. Công thức tính:

   $$TF(t,d) = \frac{\text{số lần xuất hiện của từ } t \text{ trong tài liệu } d}{\text{tổng số từ trong tài liệu } d}$$

   ## Inverse Document Frequency (IDF)

   Đo lường mức độ phổ biến của một từ trong tập hợp các tài liệu. Công thức tính:

   $$IDF(t,D) = \log\left(\frac{\text{tổng số tài liệu } |D|}{1 + \text{số tài liệu chứa từ } t}\right)$$

   Trong đó, $|D|$ là tổng số tài liệu và số tài liệu chứa từ t được cộng thêm 1 để tránh chia cho 0.

   ## TF-IDF

   Là tích của TF và IDF, giúp xác định tầm quan trọng của một từ trong một tài liệu cụ thể, đồng thời giảm thiểu ảnh hưởng của các từ phổ biến nhưng ít mang ý nghĩa.

   $$TF\text{-}IDF(t,d,D) = TF(t,d) \times IDF(t,D)$$

   TF-IDF giúp phân biệt các từ quan trọng đối với nội dung của một tài liệu so với các từ xuất hiện thường xuyên trong nhiều tài liệu khác nhau nhưng ít mang giá trị thông tin.

2. **Bài tập:** Viết chương trình Python để tính toán giá trị TF-IDF của các từ trong một List gồm các câu: ["Tôi thích học AI", "AI là trí tuệ nhân tạo", "AGI là siêu trí tuệ nhân tạo"] và sử dụng thư viện numpy để hỗ trợ trong việc tính toán các ma trận. Biết các

```python
import numpy as np
import math

# Bước 1: Tạo tập tài liệu mẫu
documents = ["Tôi thích học AI", "AI là trí tuệ nhân tạo", "AGI là siêu trí tuệ nhân tạo"]

# Bước 2: Tiền xử lý - tách từ và tính tần số
def compute_tf(doc):
    ## Your code here ##

# Bước 3: Tính toán IDF
def compute_idf(docs):
    ## Your code here ##

# Bước 4: Tính toán TF-IDF
def compute_tf_idf(tf, idf):
    ## Your code here ##

# In kết quả
## Your code here ##
```

Kết quả:

```
Tài liệu 1:
  Tôi: 0.1014
  thích: 0.1014
  học: 0.1014
  AI: 0.0000
Tài liệu 2:
  AI: 0.0000
  là: 0.0000
  trí: 0.0000
  tuệ: 0.0000
  nhân: 0.0000
  tạo: 0.0000
Tài liệu 3:
  AGI: 0.0579
  là: 0.0000
  siêu: 0.0579
  trí: 0.0000
  tuệ: 0.0000
  nhân: 0.0000
  tạo: 0.0000
```