BÀI TẬP TỰ LUẬN TÍNH TOÁN - CHƯƠNG 5

Môn: Khai thác thông tin đa phương tiện – Chương 5: Xử lý văn bản

# Bài 1: Tính TF, IDF, TF-IDF

* ❓ Đề bài:

Cho 3 tài liệu sau (đã được chuẩn hóa và tách từ):  
d1: ["học", "máy", "và", "trí", "tuệ", "nhân", "tạo"]  
d2: ["trí", "tuệ", "nhân", "tạo", "là", "tương", "lai"]  
d3: ["máy", "học", "giúp", "xây", "dựng", "mô", "hình", "dự", "đoán"]  
Yêu cầu:  
1. Tính TF của từ “máy” trong từng tài liệu.  
2. Tính IDF của từ “máy” trên tập 3 tài liệu.  
3. Tính TF-IDF của từ “máy” trong từng tài liệu.

* ✅ Lời giải chi tiết:

Bước 1: Tính TF ("máy")  
- d1: xuất hiện 1 lần, tổng 7 từ => TF = 1/7 ≈ 0.143  
- d2: không có => TF = 0  
- d3: xuất hiện 1 lần, tổng 9 từ => TF = 1/9 ≈ 0.111  
  
Bước 2: Tính IDF("máy")  
- "máy" xuất hiện trong d1 và d3 => df("máy") = 2  
- N = 3 tài liệu => IDF = log(3/2) ≈ 0.176  
  
Bước 3: TF-IDF  
- d1: 0.143 \* 0.176 ≈ 0.0251  
- d2: 0 \* 0.176 = 0  
- d3: 0.111 \* 0.176 ≈ 0.0195

# Bài 2: Vector văn bản và Cosine Similarity

* ❓ Đề bài:

Cho 2 vector văn bản:  
d1 = [0.1, 0.2, 0.3, 0.4]  
d2 = [0.2, 0.0, 0.1, 0.5]  
Yêu cầu:  
1. Tính độ dài từng vector.  
2. Tính tích vô hướng.  
3. Tính cosine similarity.

* ✅ Lời giải chi tiết:

1. Độ dài:  
||d1|| = sqrt(0.1² + 0.2² + 0.3² + 0.4²) = sqrt(0.3) ≈ 0.5477  
||d2|| = sqrt(0.2² + 0 + 0.1² + 0.5²) = sqrt(0.3) ≈ 0.5477  
  
2. Tích vô hướng:  
= 0.1\*0.2 + 0.2\*0 + 0.3\*0.1 + 0.4\*0.5 = 0.02 + 0 + 0.03 + 0.2 = 0.25  
  
3. Cosine similarity:  
= 0.25 / (0.5477 \* 0.5477) = 0.25 / 0.3 ≈ 0.833

# Bài 3: Truy vấn Boolean

* ❓ Đề bài:

Kho tài liệu gồm:  
D1: "sky blue beautiful"  
D2: "love blue beautiful sky"  
D3: "quick brown fox jumps lazy dog"  
D4: "blue dog lazy"  
Truy vấn:  
1. “sky AND blue”  
2. “blue OR dog”  
3. “blue AND NOT dog”

* ✅ Lời giải chi tiết:

1. “sky AND blue”: D1, D2  
2. “blue OR dog”: D1, D2, D4  
3. “blue AND NOT dog”: D1, D2

# Bài 4: Đánh giá hệ thống phân loại

* ❓ Đề bài:

Một hệ thống phân loại thư rác:  
- TP = 40, FP = 10, FN = 20, TN = 30  
Yêu cầu:  
1. Tính Precision, Recall, F1.  
2. Tính Accuracy, FPR, FNR.

* ✅ Lời giải chi tiết:

1.  
- Precision = 40 / (40 + 10) = 0.8  
- Recall = 40 / (40 + 20) = 0.6667  
- F1 = 2 \* 0.8 \* 0.6667 / (0.8 + 0.6667) ≈ 0.727  
  
2.  
- Accuracy = (40 + 30)/100 = 0.7  
- FPR = 10 / (10 + 30) = 0.25  
- FNR = 20 / (40 + 20) = 0.333