

Thực hành Python: Mô hình Boolean Retrieval

Dây là phần **thực hành mô hình Boolean Retrieval bằng Python** kèm theo **giải thích chi tiết từng bước** để bạn có thể học dễ dàng và áp dụng vào dự án thật.

1. Dữ liệu mẫu

Chúng ta sẽ dùng tập tài liệu gồm 4 văn bản như sau:

```
documents = {  
    "D1": "Tôi yêu học máy và AI",  
    "D2": "Học sâu là một nhánh của AI",  
    "D3": "Máy học khác với lập trình truyền thống",  
    "D4": "Tôi học lập trình Python"  
}
```

2. Tiền xử lý văn bản

Chuyển về chữ thường, tách từ, và loại bỏ dấu.

```
import re  
  
def preprocess(text):  
    # Đưa về chữ thường và bỏ dấu câu  
    text = text.lower()  
    text = re.sub(r'[\^\w\s]', '', text) # loại bỏ dấu câu  
    tokens = text.split()  
    return tokens  
  
# Tiền xử lý cho tất cả tài liệu  
preprocessed_docs = {doc_id: preprocess(content) for doc_id, content in documents.items()}
```

3. Tạo ma trận Boolean

Mỗi hàng là một tài liệu, mỗi cột là một từ. Mỗi ô là 1 (có từ đó) hoặc 0 (không có).

```
# Lấy tất cả từ duy nhất
all_terms = sorted(set(term for doc in preprocessed_docs.values() for term in doc))

# Tạo ma trận Boolean
import pandas as pd

matrix = pd.DataFrame(0, index=documents.keys(), columns=all_terms)

for doc_id, tokens in preprocessed_docs.items():
    for token in tokens:
        matrix.at[doc_id, token] = 1

print("Ma trận Boolean:")
print(matrix)
```

4. Thực thi truy vấn Boolean

Hỗ trợ 3 phép: `AND`, `OR`, `NOT`.

```
def boolean_query(query, matrix):
    query = query.lower()
    query = query.replace(" and ", " & ").replace(" or ", " | ").replace(" not ", " ~ ")

    # Đánh giá biểu thức trên DataFrame
    try:
        result = matrix.eval(query)
        return matrix[result]
    except Exception as e:
        print("Lỗi khi xử lý truy vấn:", e)
        return pd.DataFrame()
```

5. Thử nghiệm với các truy vấn

```
# Truy vấn 1: "học AND máy"
print("\nTruy vấn: học AND máy")
print(boolean_query("học AND máy", matrix))

# Truy vấn 2: "AI OR python"
print("\nTruy vấn: AI OR python")
print(boolean_query("AI OR python", matrix))
```

```
# Truy vấn 3: "học AND NOT AI"
print("\n Truy vấn: học AND NOT AI")
print(boolean_query("học AND NOT AI", matrix))
```

Kết quả đầu ra mẫu

Giả sử bạn chạy truy vấn "học AND máy", kết quả là:

```
Truy vấn: học AND máy
ai học lập máy python sâu tôi truyền vớ yêu
D1 1 1 0 1 0 0 1 0 0 1
D3 0 1 1 1 0 0 0 1 1 0
```

Tóm lược

Bước	Mục tiêu
1. Tiền xử lý	Chuyển văn bản thành danh sách từ đơn giản
2. Ma trận Boolean	Biểu diễn tài liệu dưới dạng nhị phân
3. Truy vấn Boolean	Áp dụng phép AND, OR, NOT để lọc tài liệu
4. Đánh giá truy vấn	Xem tài liệu nào phù hợp với điều kiện

Tác giả: **Đỗ Ngọc Tú**
Công Ty Phần Mềm **VHTSoft**

Phiên bản #2
Được tạo 4 tháng 5 2025 14:35:13 bởi Đỗ Ngọc Tú
Được cập nhật 7 tháng 5 2025 08:32:46 bởi Đỗ Ngọc Tú