ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

---------- ☙★❧ ----------

BÀI TẬP NHÓM
MÔN TRUY XUẤT THÔNG TIN

BÀI TẬP 1
KHẢO SÁT TẬP TÀI LIỆU CRANFIELD

Lớp: CS419.N11

Giảng viên hướng dẫn: Nguyễn Trọng Chỉnh

Nhóm sinh viên thực hiện:

1. Phan Thanh Hải 18520705
2. Nguyễn Hoàng Long 20520239

TP. Hồ Chí Minh, tháng 10 năm 2022

ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

---------- ☙★❧ ----------

BÀI TẬP NHÓM
MÔN TRUY XUẤT THÔNG TIN

BÀI TẬP 1
KHẢO SÁT TẬP TÀI LIỆU CRANFIELD

Lớp: CS419.N11

Giảng viên hướng dẫn: Nguyễn Trọng Chỉnh

Nhóm sinh viên thực hiện:

1. Phan Thanh Hải 18520705
2. Nguyễn Hoàng Long 20520239

TP. Hồ Chí Minh, tháng 10 năm 2022

Cho tập tài liệu Cranfield (trong file Cranfield.zip) có 1400 file văn bản txt (có thể có file nhiễu) và tập test (trong file TEST.zip) có 225 câu truy vấn được đặt trong file query.txt và danh sách tài liệu đúng cần được trả về tương ứng với từng câu truy vấn được đặt trong thư mục RES. Yêu cầu:

1) Thống kê các từ được sử dụng trong tập tài liệu Cranfield để biết có bao nhiêu mục từ, mỗi mục từ xuất hiện bao nhiêu lần và xuất hiện trong bao nhiêu tài liệu.

Mục từ của tài liệu được xác định theo các bước như sau:

Bước 1. Ngắt các token trong tài liệu bởi dấu cách.

Bước 2: Với mỗi token, ta loại bỏ các kí tự dấu câu và các chữ số.

Kết quả sau **bước 2** cho ta danh sách các mục từ của tài liệu.

Ta sẽ xây dựng cấu trúc dữ liệu để biểu diễn thông tin mục từ của tài liệu.

word_lst: từ điển lưu trữ thông tin của các mục từ bao gồm nội dung chuỗi kí tự, chỉ số tài liệu và tần số xuất hiện trong tài liệu tương ứng theo cấu trúc sau:

word_lst = {
[mục từ 1]: { [chỉ số tài liệu 1]: [tần số],
[chỉ số tài liệu 2]: [tần số],
... }
[mục từ 2]: { [chỉ số tài liệu 1]: [tần số],
[chỉ số tài liệu 2]: [tần số],
... }
... }

Số lượng từ được sử dụng trong tập tài liệu được thống kê trong tập tin word_lst.txt đính kèm chung với báo cáo này. Mỗi dòng trong tập tin word_lst.txt có định dạng:

<mục từ> <số lượng tài liệu chứa mục từ> <tần số>

Số lượng mục từ có trong toàn bộ tập tài liệu là: **7052**.

2) Đề xuất phương pháp xác định term trong bộ tài liệu Cranfield.

Terms của tài liệu được xác định theo các bước như sau:

Bước 1. Ngắt các token trong tài liệu bởi dấu cách.

Bước 2. Với mỗi token:

Bước 2.1. Loại bỏ các kí tự dấu câu và các chữ số.

Bước 2.2. Loại bỏ các token là *stopwords* trong tài liệu.

Bước 2.3. Sử dụng kĩ thuật *stemming* để lấy từ gốc.

Bước 2.4. Loại bỏ những từ gốc chỉ chứa ít hơn 3 kí tự.

Kết quả sau **bước 2** cho ta danh sách các *terms* của tài liệu.

Ta sẽ xây dựng cấu trúc dữ liệu để biểu diễn thông tin *terms* bao gồm nội dung chuỗi kí tự, chỉ số tài liệu và tần số xuất hiện trong tài liệu tương ứng theo cấu trúc sau:

```
inverted_index = {
[term 1]:   {    [chỉ số tài liệu 1]: [tần số],
                 [chỉ số tài liệu 2]: [tần số],
            ... }
[term 2]:   {    [chỉ số tài liệu 1]: [tần số],
                 [chỉ số tài liệu 2]: [tần số],
            ... }
... }
```

3) Cài đặt chương trình lập chỉ mục và tìm kiếm theo mô hình boolean, trong đó xem truy vấn trong file query.txt chỉ dùng phép toán OR.

Input:

query: chuỗi kí tự thể hiện thông tin câu truy vấn

dictionary: từ điển lưu trữ thông tin của terms bao gồm nội dung chuỗi kí tự, chỉ số tài liệu và tần số xuất hiện trong tài liệu tương ứng theo cấu trúc sau:

```
dictionary = {
[term1]:   {   [chỉ số tài liệu 1]: [tần số],
               [chỉ số tài liệu 2]: [tần số],
           ... }
[term2]:   {   [chỉ số tài liệu 1]: [tần số],
               [chỉ số tài liệu 2]: [tần số],
           ... }
... }
```

Output:

index_lst: danh sách các chỉ số tài liệu là kết quả tìm kiếm từ câu truy vấn query.

Sau đây là mã giả của thuật toán tìm kiếm theo mô hình boolean chỉ sử dụng phép toán OR:

BOOLEAN_RETRIEVAL(*query*, *dictionary*)

1. *index_lst* = ∅
   ▷ Hàm create_terms dùng để phân tách terms từ câu truy vấn *query* sử dụng phương pháp như trong câu 2)
2. *query_terms* = create_terms(*query*)
3. **for** *term* **in** *query_terms*:
4. **if** *term* **in** *dictionary*:
   ▷ *dictionary*[*term*] liệt kê ra danh sách các chỉ số tài liệu xuất hiện *term*.
5. *index_lst* = *index_lst* ∪ *dictionary*[*term*]

Danh sách từ vựng được sử dụng trong tập tài liệu được thống kê trong tập tin vocab_lst.txt đính kèm chung với báo cáo này. Mỗi dòng trong tập tin vocab_lst.txt có định dạng:

<term> <số lượng tài liệu chứa term> <tần số>

Số lượng *terms* có trong toàn bộ tập tài liệu là: **4147**.

Ta nhận thấy số lượng *terms* giảm gần một nửa so với số lượng mục từ đã xác định ở câu 1).

Danh sách posting được sử dụng trong tập tài liệu được thống kê trong tập tin posting_lst.txt đính kèm chung với báo cáo này. Mỗi dòng trong tập tin posting_lst.txt có định dạng:

<term> <chỉ số tài liệu> <tần số>

4) Duyệt sơ lược kết quả và nêu nhận xét.

Ta sử dụng 3 độ đo là: độ đo chính xác (precision), độ phủ (recall) và độ đo F để đánh giá mô hình bài toán.

Kết quả của 3 độ đo trên với mỗi câu truy vấn được thống kê trong tập tin experimental_result.txt đính kèm chung với báo cáo này. Mỗi dòng trong tập tin experimental_result.txt có định dạng:

<chỉ số của câu truy vấn> <Độ chính xác> <Độ phủ> <Độ đo F>

Độ chính xác trung bình: **0.01**

Độ phủ trung bình: **0.96**

Độ đo F trung bình: **0.02**

Ta nhận thấy độ chính xác trung bình rất thấp trong khi độ phủ trung bình lại rất cao. Có thể, việc sử dụng mô hình tìm kiếm boolean chỉ sử dụng phép toán OR trả về quá nhiều tài liệu không chính xác.

5) Nhóm có đề xuất cải tiến gì so với các công việc đã làm hay không?

Nhược điểm của việc tìm kiếm theo mô hình boolean dẫn đến việc tìm kiếm chính xác *terms* cho ra quá ít hoặc quá nhiều tài liệu liên quan. Do đó cần tìm mô hình truy xuất thông tin phù hợp hơn để thu được số lượng kết quả truy vấn phù hợp hơn.