**Sinh viên thực hiện**

Nguyễn Hoàng Long - 20521568

Đặng Thị Tường Vy - 20522176

Trần Phương Thảo – 20521938

**KHẢO SÁT TẬP TÀI LIỆU CRANFIELD**

**Task 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.**

Tập tài liệu Cranfield có 7230 mục từ. Chi tiết số lần xuất hiện của mỗi mục từ và số Document mà mục từ đó xuất hiện được trình bày chi tiết trong bảng sau: [Crandfield\_docx](https://docs.google.com/document/d/188_Pdml53YtqwJS84OBlz-n1zGvyOJUwQi4Ty_ed3O8/edit)

**Task 2: Đề xuất phương pháp xác định term trong bộ tài liệu Cranfield.**

"Term" được xác định là một từ, và mỗi từ trong văn bản được mặc định là cách nhau bởi khoảng trắng. Một từ cần phải chứa nhiều hơn một chữ cái. Mỗi từ chỉ gồm các chữ cái viết thường, không bao gồm số hoặc các ký tự đặc biệt. Do đó, để dễ dàng xác định các term, chúng ta nên xử lý loại bỏ số, ký tự đặc biệt, chuyển đổi chữ in hoa về in thường… trên tập tài liệu.

Để cải thiện xử lý văn bản, nên loại bỏ các stop words khỏi danh sách các term. Stop words là các từ phổ biến không mang nhiều ý nghĩa quan trọng và thường bị loại bỏ để cải thiện phân tích văn bản và hiệu quả tìm kiếm.

Danh sách stopwords được nhóm tham khảo từ list stopwords được cung cấp bởi thư viện nltk.

Cụ thể như sau

Vậy nên, nhóm quyết định sẽ xác định Term dựa trên tiêu chí sau đây:

* Term sẽ là các từ được phân cách bởi dấu cách (không tính dấu “.”)
* Các chữ cái trong Term phải thuộc khoảng sau: “a-z" & “A-Z" & “0-9”. Việc này nhằm loại bỏ các dấu câu khác nếu có cũng như các kí tự đặc biệt (VD: 100
* Term có thể bắt đầu bằng số, hoặc toàn bộ đều là số (VD: “4000” vẫn sẽ là 1 term)

**Task 3: Cài đặt chương trình lập chỉ mục và tìm kiếm theo mô hình vector.**

Mô hình Vector space và chương trình lập chỉ mục được cài đặt ở file [VectorSpaceModel](https://colab.research.google.com/drive/1yKrjV18xDnNSMqv-faFxOWMNa---VdHE#scrollTo=eSsrm05kxpWN)

Xem chỉ mục tại file Posting List.txt [Posting list](https://drive.google.com/drive/u/0/folders/1ul53rtdDQ10310d7LN2IE2WKdjfk3J4h)

Các bước cài đặt:

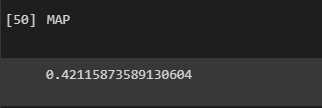
* Tokenization các các document trong collection, kết hợp loại bỏ các ký tự đặc biệt, chuyển văn bản sang chữ thường
* Loại bỏ các term nằm trong danh sách stop words
* Đưa các unique\_word về dạng từ gốc
* Tạo danh sách posting
* Vectorize hóa với tf-idf
* Xây dựng mô hình Vector Space
* Xử lý câu truy vấn (Loại bỏ ký tự đặc biệt, chuyển về chữ thường, loại bỏ stopword… )
* Thực hiện truy vấn trên mô hình đã xây dựng
* Lưu kết và đánh giá kết quả nhận được

**Task 4: Thử nghiệm chương trình với 225 truy vấn trên bộ tài liệu Cranfield**

**và tính MAP nội suy từ kết quả thực hiện 225 truy vấn này.**

Chương trình thử nghiệm và đánh giá được cài đặt tại [VectorSpaceModel](https://colab.research.google.com/drive/1yKrjV18xDnNSMqv-faFxOWMNa---VdHE#scrollTo=eSsrm05kxpWN)

Kết quả chương trình tính tonas MAP nội suy:



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

Chúng ta có thể lượt bớt một số từ không cần thiết trong câu truy vấn. Mục đích của việc này nhằm làm giảm số lượng kết quả không phù hợp.

Bên cạnh đó chúng ta có thể thay thế mô hình Vector Space bằng một mô hình khác nhằm tính toán độ phù hợp của tài liệu và câu truy vấn, để có thể sắp xếp thứ tự theo độ phù hợp của các kết quả trở về.

nhóm đã bổ sung phần xử lí bằng cách lọc ra những file nhiễu (tức file rỗng). Vì nếu khi chạy câu truy vấn trên tài liệu rỗng, sẽ dẫn đến việc không thể tính Precision và Recall.