***Logo, icon

Description automatically generatedShape, rectangle

Description automatically generated***

*Hồ Chí Minh, 18 tháng 04 năm 2023*

**TRUY XUẤT THÔNG TIN – CS419.N21**

***Bài tập 3:*** *Truy xuất thông tin trên tập Cranfield bằng mô hình Vector Space*

**Giảng viên hướng dẫn :** Nguyễn Trọng Chỉnh

**Nhóm 3:**

Bùi Nguyễn Anh Trung (NT) 20520332

Phan Doãn Thái Bình 20520043

Trần Đức Huy 20521414

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HÔ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

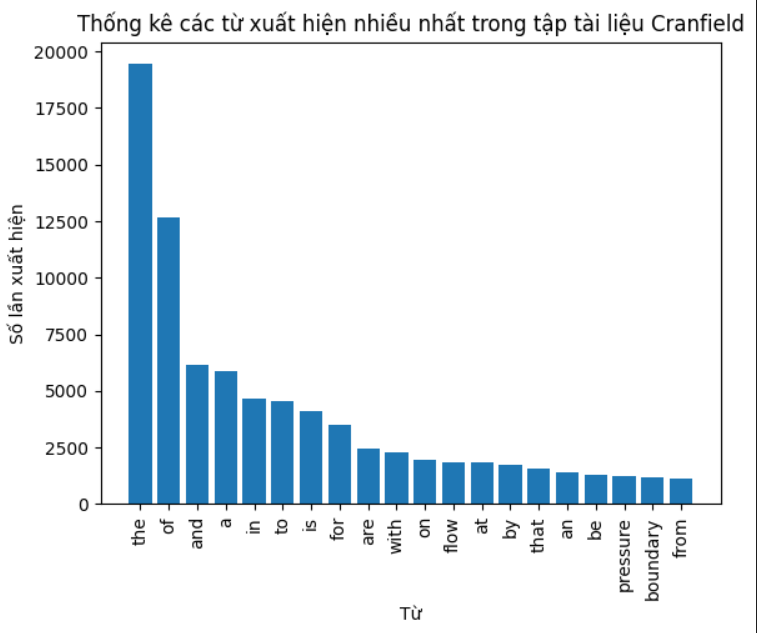
**KHOA HỌC MÁY TÍNH**

**Bảng thực hiện công việc**

|  |  |  |  |
| --- | --- | --- | --- |
| **Thành**  **viên**  **Công**  **Việc** | **Bùi Nguyễn Anh Trung** | **Phan Doãn Thái Bình** | **Trần Đức Huy** |
| Đề xuất và thử nghiệm phương pháp word ranking | Checkmark with solid fill |  |  |
| Chạy và thử nghiệm các cách tính trọng số theo chuẩn SMART | Checkmark with solid fill |  |  |
| Đề xuất và áp dụng pos tag để tăng hiệu quả truy xuất | Checkmark with solid fill |  |  |
| Viết baseline | Checkmark with solid fill | Checkmark with solid fill | Checkmark with solid fill |
| Thống kê tập tài liệu Cranfield |  | Checkmark with solid fill |  |
| Thử nghiệm chương trình |  |  | Checkmark with solid fill |
| Tính MAP nội suy từ kết quả |  |  | Checkmark with solid fill |

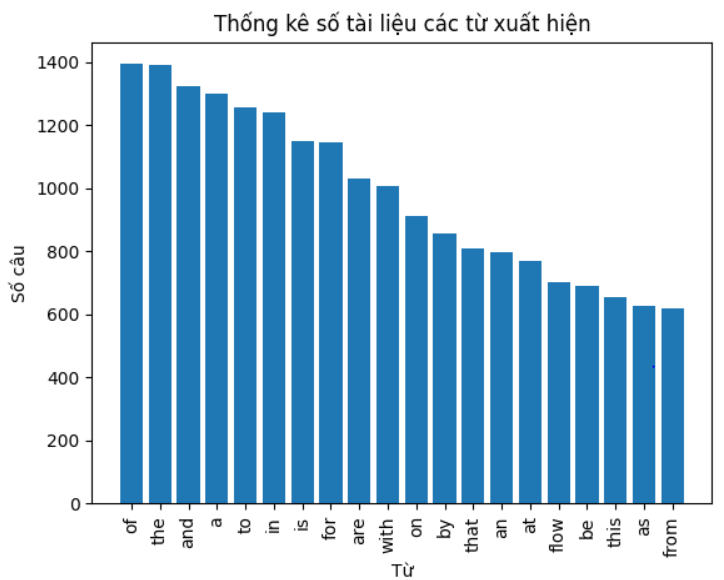
1. Trong tập tài liệu Cranfield có tổng cộng 8155 mục từ.

* Và dưới đây là 20 từ xuất hiện nhiều lần nhất:



Hình 1 : Thông kê tần suất xuất hiện của 20 từ xuất hiện nhiều nhất tron

* 20 từ xuất hiện trong nhiều câu nhất



Hình 2 : Thống kê số câu có xuất hiện của mỗi từ

2. Các phương pháp để lựa term được đề xuất:

- Xóa stopword (remove stopword) :

vd: of, the,... vì các từ dừng xuất hiện nhiều lần trong nhiều câu nhưng mang ít ý nghĩa. Việc xóa từ dừng sẽ loại bỏ được 1 lượng lớn data cũng như terms không cần thiết.

* Rút gọn từ (Stemming) :

Vd: students => student, một từ có thể có nhiều biến thể khác nhau(walk, walks,walking,walked) nên việc chuyển đổi một từ về dạng ban đầu của nó sẽ giảm được 1 lượng lớn term nhưng vẫn giữ được nghĩa tương đương.

* Xếp hạng từ (words ranking):
* Vì điểm chính của thuật toán là dựa tần suất xuất hiện của các từ, thế nên những từ có tần suất xuất hiện ít hơn vô tình bị hạn chế trong việc đóng tính liên quan giữa truy vấn và văn bản.
* Nhóm đề xuất word ranking, sắp xếp các từ theo tần xuất giảm dần rồi xếp hạng lần lượt là 1,2,3,… ( Tức từ có tần suất xuất hiện ít hơn thì số rank cao hơn )
* Sau đó cập nhật công thức tính tần xuất:
  + Ban đầu nhóm cộng trực tiếp word rank thì không cho kết quả tốt. Giải thích vì rank tăng theo O(n). tăng rất nhanh. Nếu từ có tần số suất hiện ít nhất có rank là 1000 thì trọng số nó gấp 1000 lần từ có tần số cao nhất, khi đó truy vấn ưu tiên trả về từ có tần số càng ít càng tốt
  + Nên nhóm cải tiến hơn, sử dụng hàm kích hoạt là để giảm tốc độ tăng còn . Ta so sánh, = 6,9, = 0.69. Ta thấy tốc độ tăng hợp lý. Và kết quả thực nghiệm tốt hơn
* Pos-tag :

Nhóm sử dụng thêm phương pháp gán nhãn từ loại. Nhóm dùng thư viện để gán nhãn từ loại cho từng câu, sau đó dùng ma trận term doc áp dụng tương tự cho các nhãn từ loại được tính ra. Rồi thực hiện một truy vấn riêng mô hình nhãn từ loại đó. Sau cộng 2 kết quả truy vấn word và postag lại để lấy kết quả cuối cùng

Sau khi sử dụng các phương pháp trên, bộ term giảm từ 8155 term còn 4693 term, tuy giảm nhiều nhưng lượng term còn vẫn còn nhiều.

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

Nhóm sử dụng TD IDF để lập chỉ mục với 4693 term.

Nhóm chúng em sử dụng 2 phương pháp truy vấn:

- Biểu diễn các query dưới dạng vector và tính độ tương đồng cosine

- Truy vấn bằng tổng tổng trọng số - vector score với mỗi term

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.

Với chương trình hiện tại đạt MAP nội suy 0.32 khi thử nghiệm với bộ Cranfield 225 câu query bằng cách biểu diễn các query dưới dạng vector và tính độ tương đồng cosine:

A picture containing text

Description automatically generated

Ngoài ra, chương trình đạt MAP nội suy 0.34 khi truy vấn bằng phương pháp tính tổng trọng số - vector score:

Graphical user interface

Description automatically generated with medium confidence

Có thể thấy được phương pháp truy vấn bằng tổng trọng số - vector score có kết quả tốt hơn do phương pháp này không bị nhiễu bởi các giá trị siêu nhỏ - xấp xỉ 0 khi tính score.

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

- Nhóm chúng em đang thử nghiệm word ranking, việc áp dụng word ranking có làm tăng IMap lên 1 chút, nên cải tiến word ranking hơn nữa có thể mang tới kết quả tốt hơn..

- Sử dụng Pos tag đẻ truy vấn trên nhãn từ loại

- Nhóm còn thử nghiệm them ngrams tuy nhiên không cho ra kết quả cao

- Nhóm thực hiện tính các trọng số theo hệ thống SMART và cho thấy kết quả tốt nhất khi chọn theo công thức chuẩn tính TF, theo công thức chuẩn tính IDF và chuẩn normalize không làm ảnh hưởng nhiều tới kết quả