**LSI**

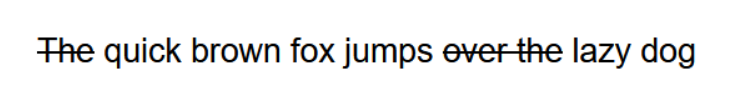
**I, Giới thiệu**

Được phát triển vào những năm 1980, LSI sử dụng một phương pháp toán học giúp cho việc truy xuất thông tin chính xác hơn. Phương pháp này hoạt động bằng cách xác định các mối quan hệ ngữ cảnh ẩn giữa các từ.

**II, LSI hoạt động như thế nào?**

LSI hoạt động bằng cách sử dụng ứng dụng một phần của phân hủy giá trị số ít (SVD). SVD là một phép toán làm giảm ma trận thành các phần cấu thành của nó để có các phép tính đơn giản và hiệu quả.

Khi phân tích một chuỗi từ, LSI loại bỏ các liên từ, đại từ và động từ thông dụng, còn được gọi là các stopword. Điều này tách biệt các từ bao gồm ‘nội dung’ chính của một cụm từ. Dưới đây là một ví dụ nhanh về cách đều này có thể trông như thế nào:



Những từ này sau đó được đặt trong **Ma trận tài liệu thuật ngữ** (TDM). TDM là một lưới 2D liệt kê tần suất mà mỗi từ (hoặc thuật ngữ) cụ thể xuất hiện trong các tài liệu trong tập dữ liệu.

Các chức năng cân sau đó được áp dụng cho TDM. Một ví dụ đơn giản là phân loại tất cả các tài liệu có chứa từ với giá trị 1 và tất cả những tài liệu không có giá trị bằng 0. Khi các từ xảy ra với cùng tần suất chung trong các tài liệu này, nó được gọi là **đồng xuất hiện.** Dưới đây, bạn sẽ tìm thấy một ví dụ cơ bản về TDM và cách nó đánh giá mức độ đồng xuất hiện trên nhiều cụm từ:



Sử dụng SVD cho phép chúng ta ước tính gần đúng các mẫu trong cách sử dụng từ trên tất cả các tài liệu. Các vector SVD do LSI tạo ra dự đoán ý nghĩa chính xác hơn việc phân tích các thuật ngữ riêng lẻ. Cuối cùng, LSI có thể sử dụng các mối quan hệ giữa các từ để hiểu rõ hơn ý nghĩa của chúng trong một ngữ cảnh cụ thể.