**LSA**

**I, Phân tích ngữ nghĩa tiềm ẩn(LSA) là gì?**

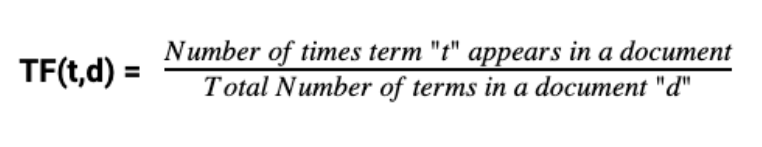
Phân tích ngữ nghĩa tiềm ẩn là một trong những kỹ thuật xử lý ngôn ngữ tự nhiên để phân tích ngữ ngĩa ở mức độ rộng có nghĩa là chúng ta đang cố gắng tìm ra một số ý nghĩa từ ngữ liệu văn bản với sự trợ giúp củ thống kê và được Jerome Bellegarde giới thiệu vào năm 2005.

LSA về cơ bản là một kỹ thuật mà chúng ta xác định các mẫu từ tài liệu văn bản hoặc nói một cách đơn giản là chúng ta có xu hướng tìm ra thông tin quan trọng và có liên quan từ tài liệu văn bản. Nếu chúng ta nói về việc liệu nó có được giám sát hay không theo cách thức không giám sát, thì rõ ràng đó là một cách tiếp cận không có giám sát.

**II, Tầm quan trọng của TF, IDF**

1. **TF**

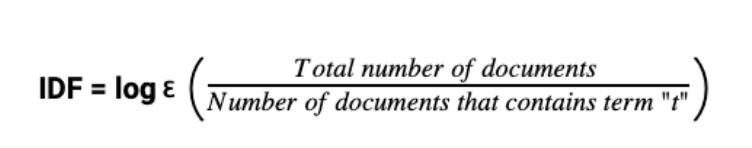
Tần suất thuật ngữ được định nghĩa là số lần trường hợp hoặc từ khóa xuất hiện trong một tài liệu chia cho tổng số từ trong tài liệu đó.



Như chúng ta biết độ dài của tài liệu là khác nhau trong mỗi trường hợp, vì vậy tần suất thuật ngữ thay đổi theo sự xuất hiện của thuật ngữ tương ứng.

1. **IDF**

Tần suất tài liệu nghịch đảo, biểu thị tầm quan trọng của thuật ngữ này trọng việc thu thập tài liệu. IDF tính toán trọng lượng của từ hiếm của văn bản trong một bộ sưu tập tài liệu.



Ý tưởng chính của TF/IDF trong LSA là cung cấp số lượng từng từ và tần suất xuất hiện của các từ hiếm để cung cấp cho chúng trọng số trên cơ sở độ hiếm của chúng, TF/IDF thích hợp hơn cách đếm thông thường về số xuất hiện của từ. Vì nó chỉ đếm tần số mà không phân loại.