**I, Ý tưởng của Vector Space Model**

Ý tưởng của Vector Space Model là biểu diễn văn bản và các câu truy vấn dưới dạng Vector, Rep(d) của docs và Rep(q) của query sẽ cho kết quả là các vector. Sau đó tính độ tương đồng của query với từng documents theo công thức Sim(Rep(q), Rep(d)) để tìm ra docs vào phù hợp nhất với query.

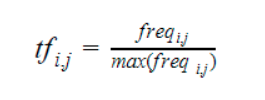
**II, Tìm trọng số cho vector**

Cách xác định và tính trọng số cho vector là hết sức quan trọng, ảnh hưởng đến độ chính xác của các thuật toán xếp hạng. Việc các từ có trọng số khác nhau là do không phải các từ đều có sự quan trọng giống nhau, sử dụng số lần xuất hiện của các từ làm vector không phải là một cách tối ưu. Ở phương diện documents, một vài từ có thể mang nhiều thông tin hơn các từ còn lại.

Có nhiều kỹ thuật tính trọng số: TF, IDF, TF-IDF,..

1. **TF: Term frequency**

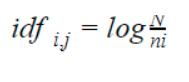
Từ nào xuất hiện nhiều trong câu thì quan trọng, công thức này sẽ đếm tần suất xuất hiện các từ trong câu.



trong đó freqi,j là số cụm từ i xuất hiện trong tài liệu dj

1. **IDF: Tần suất tài liệu nghịch đảo**

Từ nào xuất hiện nhiều trong mọi câu thì không mang nhiều ý nghĩa (ví dụ như: a, the, are, thì, là,…). Vì vậy trọng số IDF là nghịch đảo của tần suất xuất hiện của các từ trong các documents.



Trong đó, ni là số các văn bản có từ hoặc cụm từ i, N là tổng số văn bản.

1. **TF-IDF: Tần suất thuật ngữ - Tần suất tài liệu nghịch đảo**

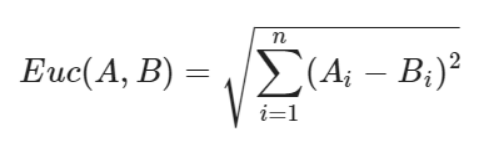
Phép nhân giữa TF và IDF cho phép ta kết hợp cả 2 độ đo trên, từ vừa xuất hiện nhiều lần trong câu, vừa không phải là từ phổ biến xuất hiện trong mọi câu.

TF-IDF được tính bởi: tf\_idfi,j = tfi,j x idfi,j

**III, Các độ đo similarity**

Sau khi có được các vector cho query và docs, ta tính được similarity bằng cách tính khoảng cách giữa các vector.

1. **khoảng cách Euclide**



Nhược điểm:

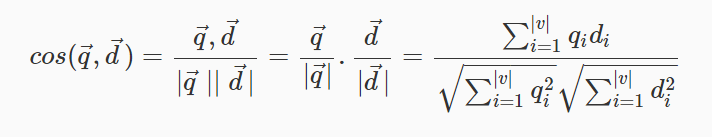
* Docs dài hơn sẽ bị giảm hạn do chứa nhiều từ khác gây nhiễu, khiến khoảng cách xa hơn.

1. **Manhattan**

Manhattan distance hay còn gọi là L1 distance. Tính khoảng cách 2 vector theo công thức:

1. **Độ đo góc**

Việc tính xem các vector xem các vector overlappend với nhau như thế nào cũng là một phép so sánh hay sử dụng, độ đo góc phổ biến là Cosine Similarity



qi là trọng số TF-IDF của từ I trong câu truy vấn. di là trọng số TF-IDF của từ i trong tài liệu cos(, ) là sự tương đồng cosine giữa hay là cosine của góc giữa và .

Đối với các vector đã được chuẩn hóa về độ dài, sự tương đồng cosine chỉ đơn giản là tích vô hướng của hai vector



q, d là độ dài được chuẩn hóa

**IV, Ưu và nhược điểm của Vector Space Model**

1. **Ưu điểm**

* Dễ hiểu và dễ cài đặt
* Đã được nghiên cứu từ lâu và thực nghiệm cho kết quả tốt và khả thi
* Hiện tại là phương pháp được sử dụng rộng rãi
* Có nhiều công thức tính TF.IDF khác nhau

1. **Nhược điểm**

* Để cho kết quả đúng đắn, ta giả định:
* Các câu là độc lập với nhau
* Các query và documents cùng loại với nhau
* Có nhiều tham số để hiệu chỉnh
* Bộ từ điển
* Tham số trong các hàm Normalize
* Threshold để chọn ra top kết quả