Locality Sensitive Hashing (LSH)

Bài toán “tìm hàng xóm gần nhất” (nearest neighbours) rất phổ biến khi áp dụng cho các tập dữ liệu lớn. Một số ứng dụng hay gặp đó là tìm bản sao hay các tài liệu giống nhau, tìm kiếm audio/video. Cách thông thường ta hay nghĩ đến đó là sử dụng kỹ thuật vét cạn (brute force) để kiểm tra tất cả các tổ hợp có thể để lấy được chính xác cặp hàng xóm gần nhau nhất. Tuy nhiên, cách làm đó rất khó thực hiện với tập dữ liệu lớn. Thế nên, các thuật toán xấp xỉ được đề cử cho vấn đề trên. Các thuật toán xấp xỉ sẽ không cho bạn câu trả lời chính xác nhưng chúng cung cấp độ xấp xỉ tốt với tốc độ nhanh và dễ thực hiện hơn.

Localilty Sensity Hashing (LSH) là một trong những thuật toán đó. Ý tưởng của LSH là sử dụng một hàm băm (có một họ các hàng băm cho LSH gọi là LSH families) để hash các điểm dữ liệu vào từng bucket sao cho các điểm dữ liệu gần nhau sẽ có xác suất cao rơi vào cùng một bucket, ngược lại chúng sẽ bị băm vào hai bucket khác nhau. Khi truy vấn, ta chỉ cấn tính mã băm tương ứng của dữ liệu đầu vào và trả về tất cả các điểm dữ liệu lưu trong bucket tương ứng với mã băm đó. Vì LSH mang tính chất xấp xỉ và phải cung cấp tốc độ thực thi cao nên các hàm băm sử dụng trong họ LSH thuộc các hàm băm phi mật mã.

LSH đã trở thành một kỹ thuật đa dụng để áp dụng vào những vấn đề như:

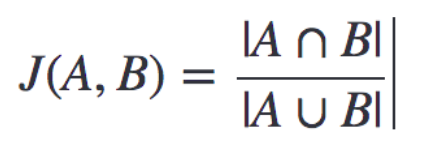
* Tìm trùng lặp gần nhất: LSH được dùng phổ biến cho việc phát hiện bản sao trong số lượng lớn các tài liệu, trang web và các tệp.
* Nghiên cứu bộ gen: các nhà sinh học thường sử dụng LSH để xác định các biểu hiện gen tương tự trong cơ sở dữ liệu bộ gen.
* Tìm kiếm hình ảnh quy mô lớn: Google sử dụng LSH với PageRank để xây dựng công nghệ tìm kiếm hình ảnh của họ là VisualRank
* Vân tay Audio/Video: Trong công nghệ truyền thông, LSH được sử dụng làm công nghệ vân tay cho các dữ liệu dạng Audio/Video.

Hãy cùng tìm hiểu thuật toán đằng sau LSH.

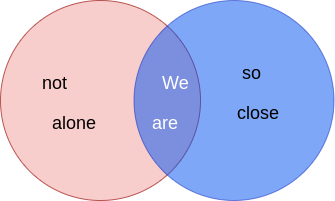
Ở bài viết này tôi sẽ lấy ví dụ về bài toán “tìm hai văn bản gần nhau nhất” để minh họa cách hoạt động của LSH. Giả sử ta có ba văn bản như sau:

A: "We are not alone"  
B: "We are so close"  
C: "Those are good old times"

Để phát hiện hai câu trên là trùng nhau hay không, dùng công thức tính khoảng cách giữa các string — Jaccard Index là một cách đơn giản và hiệu quả. Công thức tóm gọn như sau:

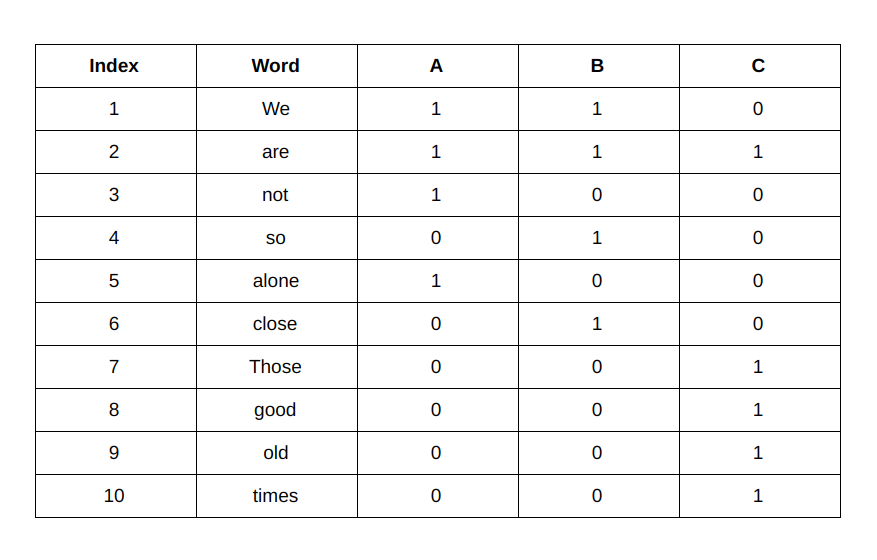


Để áp dụng công thức trên dễ dàng, ta sẽ chia hai câu trên thành từng từ. Minh họa Jaccard Index cho hai câu A và B như sau:



Ta thấy số lượng từ giao nhau là 2 và tổng số lượng từ là 6 nên Jaccard Index sẽ có giá trị 2/6 = 0.333.

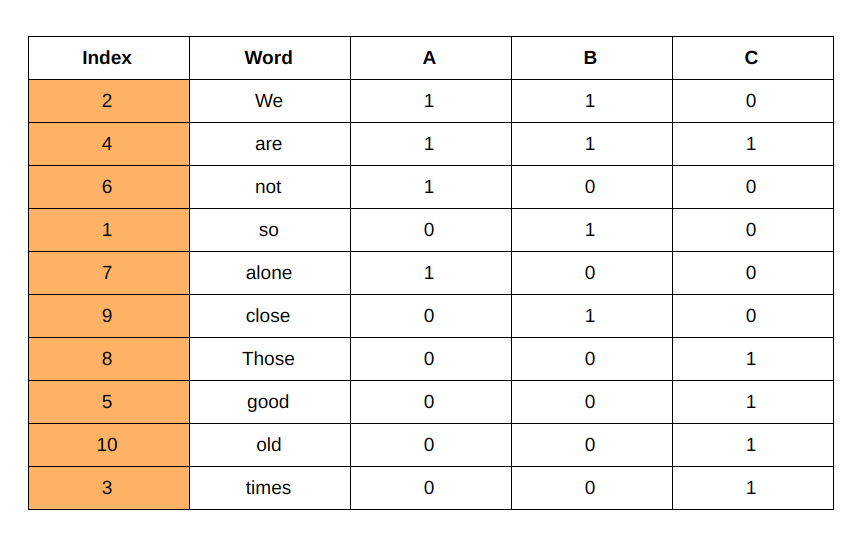
Mọi việc tưởng như thật đơn giản nhưng cách làm trên lại không dễ scale: số lượng các phép so sánh là 0.5 \* n(n -1) nên độ phức tạp thời gian sẽ là O(n²). Ta có thể biểu diễn lại hai câu trên dưới dạng ma trận Boolean để thuận tiện cho việc so sánh. Ta xây dựng một tập tất cả các từ riêng biệt tổng hợp từ hai câu, nếu một từ nằm trong một câu thì ta đánh dấu là 1, ngược lại là 0:



Đến đây vẫn chưa tối ưu vì lưu trữ một ma trận như vậy rất tốn bộ nhớ. Có một cách để giải quyết vấn đề này đó chính là dùng hàm băm. Đây cũng là phần quan trọng nhất trong bài viết về LSH. Có nhiều hàm băm cho ta lựa chọn nhưng hàm băm thích hợp cho Jaccard Index đó chính là MinHashing.

Các bước quan trọng trong quá trình này bao gồm:

* **Bước 1**: Hoán vị ngẫu nhiên index dòng của ma trận Boolean ta đã xây dựng:



* **Bước 2**: Hàm hash sử dụng trong MinHashing là: ở mỗi cột tương ứng với các văn bản, duyệt đến khi ta tìm được một ô có giá trị 1 mà tại đó đối chiếu qua cột index (đã được hoán vị ngẫu nhiên) có giá trị nhỏ nhất. Giá trị hash chính là giá trị của index đó.

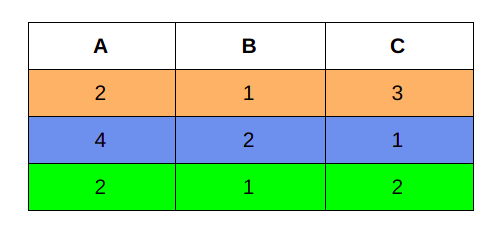


Như hình trên ta hash ra được một dòng trong ma trận chữ ký là [2, 1, 3]

Làm lại các bước trên nhiều lần với mỗi lần cột index được hoán vị ngẫu nhiên ta được một ma trận chữ ký đầy đủ hơn:

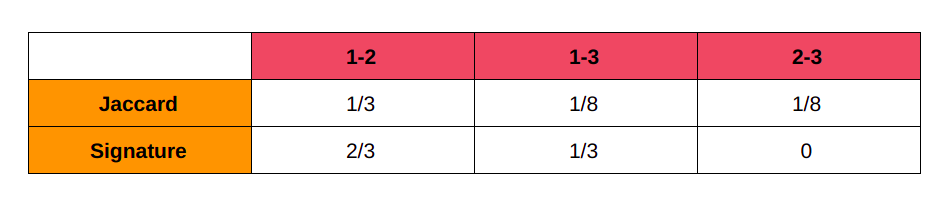


Ma trận chữ ký hash ra được sẽ là:



Đến đây ta đã có thể tính được độ giống nhau dựa trên ma trận chữ ký: Sự giống nhau giữa các chữ ký là tỉ lệ mà số hàm min-hash (các hàng) có cùng giá trị với nhau. Ví dụ với ma trận chữ ký ở trên, sự giống nhau giữa A và B là 2/3 vì chúng có cột 1 và 3 giống nhau.

Ta cùng thử tính độ giống nhau giữa các cột còn lại và so sánh với kết quả của Jaccard Index:



Ta thây có sự khác biệt gía trị giữa hai loại phép tính vì độ dài ma trận chữ ký ở đây là 3 — khá ngắn. Nếu ta tăng độ dài đó lên thì độ giống nhau sẽ khít hơn nữa.