TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN GIỮA KÌ MÔN XỬ LÝ DỮ LIỆU LỚN**

**FINDING SIMILAR ITEM**

*Người hướng dẫn*: **TS LÊ ANH CƯỜNG**

*Người thực hiện*: **NGUYỄN ANH KIỆT – 51703119**

**ĐỔ THÀNH NHỰT – 51703151**

**TRẦN GIA THÁI- 51703184**

Khoá  **: 21**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN GIỮA KÌ MÔN XỬ LÝ DỮ LIỆU LỚN**

**FINDING SIMILAR ITEM**

*Người hướng dẫn*: **TS LÊ ANH CƯỜNG**

*Người thực hiện*: **NGUYỄN ANH KIỆT – 51703119**

**ĐỔ THÀNH NHỰT – 51703151**

**TRẦN GIA THÁI- 51703184**

Khoá  **: 21**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

**LỜI CẢM ƠN**

Cảm ơn thầy Lê Anh Cường đã hết lòng hỗ trợ và hướng dẫn để chúng em có thể hoàn thành đồ án cuối kỳ một cách tốt nhất.

Trong lúc làm đồ án khó tránh khỏi những thiếu sót, kính mong thầy có thể thông cảm và góp ý để chúng em có thể hoàn thành tốt hơn.

Nhóm chúng em xin chân thành cảm ơn!

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được sự hướng dẫn của TS Lê Anh Cường. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày tháng năm*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Nguyễn Anh Kiệt*

*Đổ Thành Nhựt*

*Trần Gia Thái*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

TÓM TẮT

MỤC LỤC

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iv](#_Toc66974882)

[TÓM TẮT v](#_Toc66974883)

[MỤC LỤC 1](#_Toc66974884)

[1 – Tìm điểm giống nhau trong tài liệu 3](#_Toc66974885)

[1.1 Khái quát 3](#_Toc66974886)

[1.2 Một số ví dụ 3](#_Toc66974887)

[2 – Các bước cần thiết để tìm kiếm điểm giống nhau trong văn bản 4](#_Toc66974888)

[2.1 Shingling 5](#_Toc66974889)

[2.2 Minhashing 7](#_Toc66974890)

[2.3 Locality-Sensitive Hashing 14](#_Toc66974891)

[Tài liệu tham khảo 21](#_Toc66974892)

DANH MỤC HÌNH ẢNH

No table of figures entries found.

1 – Tìm điểm giống nhau trong tài liệu

1.1 Khái quát

Một lớp vấn đề quan trọng mà sự tương tự Jaccard giải quyết tốt là tìm kiếm các tài liệu tương tự về văn bản trong một kho tài liệu lớn như Web hoặc bộ sưu tập các bài báo tin tức. Chúng ta nên hiểu rằng khía cạnh của sự giống nhau chúng ta đang xem xét ở đây là sự giống nhau ở cấp độ ký tự, không phải là "ý nghĩa tương tự", yêu cầu chúng ta kiểm tra các từ trong các tài liệu và cách sử dụng của chúng. Vấn đề đó cũng thú vị nhưng được giải quyết bằng các kỹ thuật khá. Tuy nhiên, sự giống nhau về mặt văn bản cũng có những công dụng quan trọng. Nhiều chúng liên quan đến việc tìm các bản sao hoặc gần bản sao. Trước tiên, chúng ta hãy quan sát rằng dễ dàng kiểm tra xem hai tài liệu có phải là bản sao chính xác hay không; chỉ cần so sánh hai tài liệu theo từng ký tự và nếu chúng khác nhau thì chúng không giống nhau. Tuy nhiên, trong nhiều ứng dụng, các tài liệu không giống nhau, họ chia sẻ phần lớn văn bản của họ.

1.2 Một số ví dụ

* **Đạo văn:**

Việc tìm kiếm các tài liệu đạo văn kiểm tra khả năng của chúng ta trong việc tìm ra sự giống nhau về văn bản. Các kẻ đạo văn chỉ có thể trích xuất một số phần của tài liệu cho riêng mình. Họ có thể thay đổi một vài từ và có thể thay đổi thứ tự xuất hiện các câu trong bản gốc. Tuy nhiên, tài liệu kết quả vẫn có thể chứa 50% hoặc nhiều hơn tài liệu gốc. Không quá trình đơn giản so sánh các tài liệu theo từng ký tự sẽ phát hiện ra một đạo văn tinh vi

* **Trang phản chiếu:**

Thông thường các trang Web quan trọng hoặc phổ biến bị trùng lặp ở một số của máy chủ, để chia sẻ tải. Các trang của các trang nhân bản này sẽ khá giống nhau, nhưng hiếm khi giống hệt nhau. Ví dụ: mỗi chúng có thể chứa thông tin được liên kết với máy chủ lưu trữ cụ thể của họ và họ có thể có liên kết đến các trang nhân bản khác nhưng không liên kết với chính chúng. Một hiện tượng liên quan là sự chiếm đoạt các trang từ lớp này sang lớp khác. Những trang này có thể bao gồm ghi chú lớp học, bài tập và slide bài giảng. Các trang tương tự có thể thay đổi tên của khóa học, năm và thực hiện các thay đổi nhỏ từ năm này sang năm khác. Nó rất qua công cụ tạo ra kết quả tốt hơn nếu chúng tránh hiển thị hai trang gần như giống hệt nhau trong trang đầu tiên của kết quả

* **Các bài viết từ cùng một nguồn:**

Thông thường một phóng viên viết một bài báo được phát tán, nói thông qua Associated Press, cho nhiều tờ báo, sau đó xuất bản bài báo trên các trang web của họ. Mỗi tờ báo có phần thay đổi bài viết. Họ có thể cắt bỏ các đoạn văn, hoặc thậm chí thêm tài liệu của riêng họ. Họ hầu hết có khả năng sẽ bao quanh bài viết bởi biểu trưng, ​​quảng cáo của riêng họ và các liên kết đến các bài viết khác tại trang web của họ. Tuy nhiên, cốt lõi của mỗi trang báo sẽ là bản gốc bài báo. Các trình tổng hợp tin tức, chẳng hạn như Google Tin tức, cố gắng tìm tất cả các phiên bản của một bài báo, để chỉ hiển thị một và nhiệm vụ đó yêu cầu tìm khi hai Các trang web giống nhau về văn bản, mặc dù không giống nhau

2 – Các bước cần thiết để tìm kiếm điểm giống nhau trong văn bản

Có 3 bước cần thiết :

Bước 1:

Shingling: Chuyển đổi một tài liệu thành một tập hợp biểu diễn (vectơ Boolean)

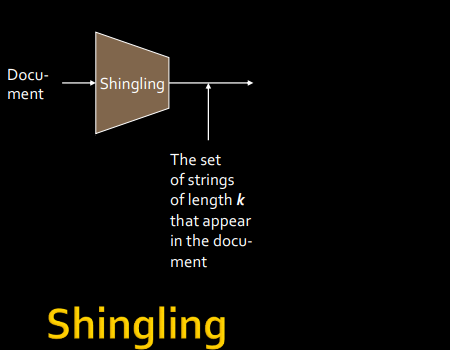
Bước 2:

Min-Hashing: Chuyển đổi tập hợp lớn thành chữ ký ngắn, trong khi vẫn giữ được tính tương tự

Bước 3:

Locality-Sensitive Hashing: Tập trung vào các cặp chữ ký có thể là từ các tài liệu tương tự

2.1 Shingling



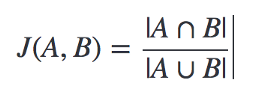
Trong bước này, chúng tôi chuyển đổi mỗi tài liệu thành một tập hợp các ký tự có độ dài k (còn được gọi là k-shingles hoặc k-gram). Ý tưởng chính là thể hiện mỗi tài liệu trong bộ sưu tập của chúng tôi dưới dạng một bộ k-shingles.

Ví dụ: Một trong những tài liệu của bạn (D): “Nadal”. Bây giờ nếu chúng ta quan tâm đến 2-shingles, thì bộ của chúng ta: {Na, ad, da, al}. Tương tự bộ 3-shingles: {Nad, ada, dal}.

* Các tài liệu tương tự có nhiều khả năng chia sẻ bệnh zona hơn
* Việc sắp xếp lại các đoạn văn trong một tài liệu thay đổi từ ngữ không ảnh hưởng nhiều đến shingles
* Giá trị k của 8–10 thường được sử dụng trong thực tế. Một giá trị nhỏ sẽ dẫn đến nhiều vết zona có trong hầu hết các tài liệu ( không tốt cho việc phân biệt các tài liệu )

**Chỉ số Jaccard**

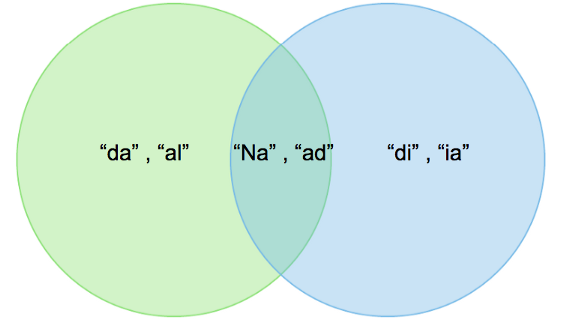
Chúng tôi có một bản trình bày của mỗi tài liệu dưới dạng shingles. Bây giờ, chúng ta cần một thước đo để đo độ tương đồng giữa các tài liệu. Jaccard Index là một lựa chọn tốt cho việc này. Chỉ mục Jaccard giữa tài liệu A và B có thể được định nghĩa là:



Nó còn được gọi là giao cắt qua liên hiệp (IOU) .

Giả sử A: “Nadal” và B: “Nadia”, khi đó biểu diễn 2-shingles sẽ là:

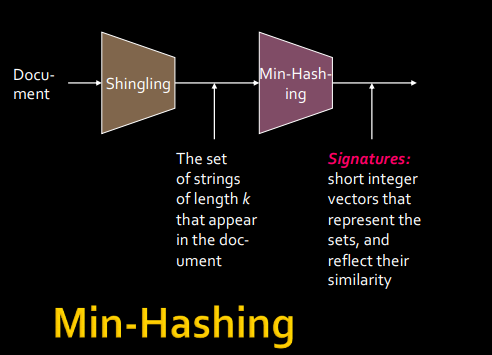
A: {Na, ad, da, al} và B: {Na, ad, di, ia}.



Chỉ số Jaccard = 2/6

Số lượng bệnh zona thông thường nhiều hơn sẽ dẫn đến Chỉ số Jaccard lớn hơn và do đó nhiều khả năng các tài liệu giống nhau.

2.2 Minhashing



Các chữ ký mà chúng tôi mong muốn xây dựng cho các tập hợp bao gồm các kết quả của một số lượng lớn các phép tính, chẳng hạn như vài trăm phép tính, mỗi phép tính là một "minhash” của ma trận đặc trưng. Trong phần này, chúng ta sẽ tìm hiểu thế nào là minhash được tính toán theo nguyên tắc, và trong các phần sau, chúng ta sẽ xem cách giao dịch với minhash được tính toán trong thực tế. Để tối thiểu hóa một tập hợp được đại diện bởi một cột của ma trận đặc trưng, ​​hãy chọn một hoán vị của các hàng. Giá trị minhash của bất kỳ cột nào là số hàng đầu tiên, theo thứ tự đã hoán vị, trong đó cột có 1

**Băm cột**

Ý tưởng chính: “băm” từng cột C thành một chữ ký h (C), sao cho:

* sim (C1, C2) giống như "sự tương tự" của chữ ký h (C1) và h (C2)

Mục tiêu: Tìm một hàm băm h (·) sao cho:

* Nếu sim (C1, C2) cao thì với xác suất cao. h (C1) = h (C2)
* Nếu sim (C1, C2) thấp thì với xác suất cao. h (C1) ≠ h (C2)

Ý tưởng: Băm tài liệu cùng một nhóm. Mong đợi rằng các cặp "hầu hết" các tài liệu gần như trùng lặp được băm vào cùng một nhóm!

**Min-Hashing: Mục tiêu**

Mục tiêu: Tìm một hàm băm h (·) sao cho:

* Nếu sim (C1, C2) cao, thì với xác suất cao. h (C1) = h (C2)
* Nếu sim (C1, C2) thấp, thì với xác suất cao. h (C1) ≠ h (C2)

Rõ ràng, hàm băm phụ thuộc vào chỉ số tương tự:

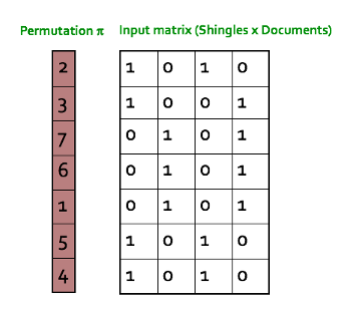
* Không phải tất cả các chỉ số tương tự đều có hàm băm

Có một hàm băm phù hợp cho tương tự Jaccard: Nó được gọi là Min-Hashing

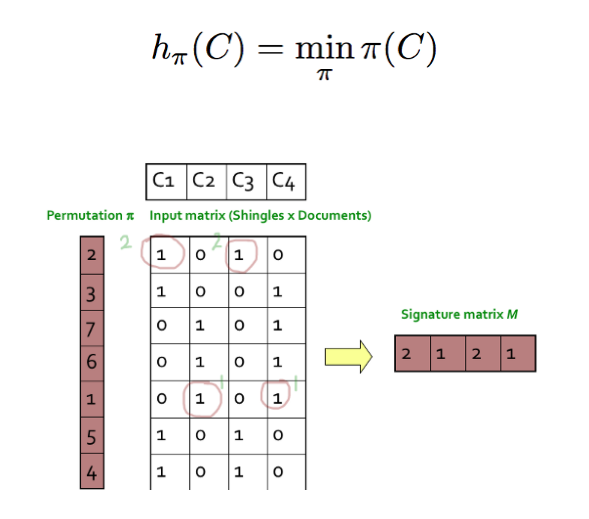
**Min-Hashing:**

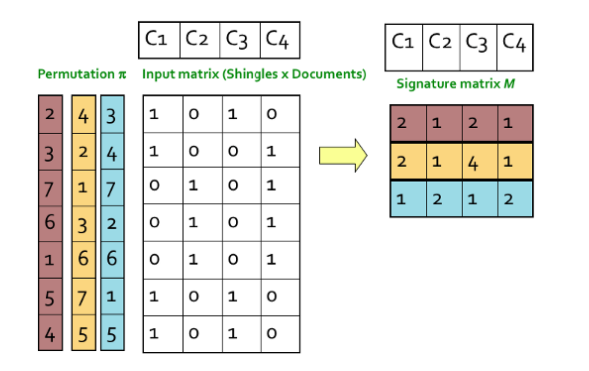
Đây là khía cạnh quan trọng và kỳ diệu nhất của thuật toán này, vì vậy hãy chú ý:

**Bước 1:** Random hoán vị (π) của chỉ số hàng của ma trận tài liệu ván lợp.



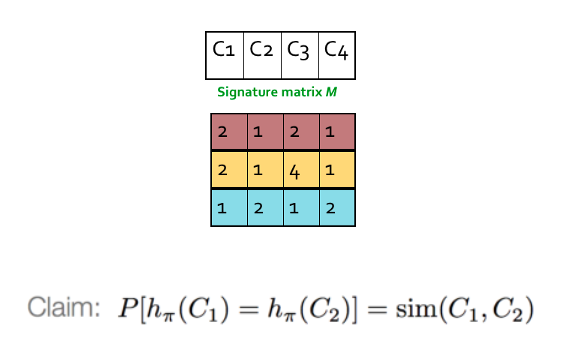
**Bước 2:** Hàm băm là chỉ số của hàng đầu tiên (theo thứ tự hoán vị) trong đó cột C có giá trị 1. Thực hiện thao tác này nhiều lần (sử dụng các hoán vị khác nhau) để tạo chữ ký của một cột.





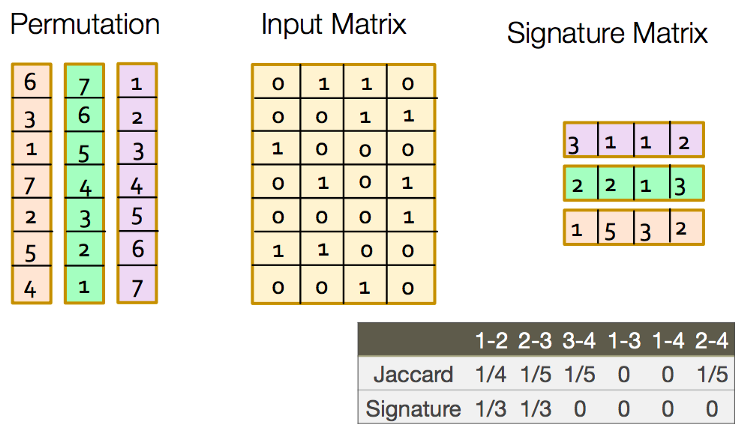
**Thuộc tính băm tối thiểu**

Sự giống nhau của các chữ ký là phần nhỏ của các hàm băm tối thiểu (hàng) mà chúng đồng ý với nhau. Vì vậy, sự giống nhau về chữ ký của C1 và C3 là 2/3 vì hàng thứ nhất và thứ ba là như nhau.



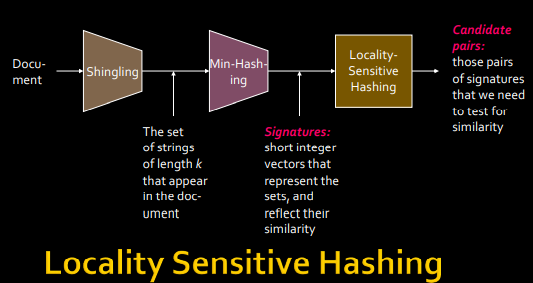
* Độ giống nhau mong đợi của hai chữ ký bằng độ tương tự Jaccard của các cột. Chữ ký càng dài, lỗi càng thấp

Trong ví dụ dưới đây, bạn có thể thấy điều này ở một mức độ nào đó. Có sự khác biệt vì chúng tôi chỉ có chữ ký có độ dài 3. Nhưng nếu tăng chiều dài thì 2 điểm tương đồng sẽ gần nhau hơn.



Vì vậy, sử dụng phép băm min, chúng ta đã giải quyết được vấn đề phức tạp của không gian bằng cách loại bỏ độ thưa thớt và đồng thời bảo toàn tính tương tự .

2.3 Locality-Sensitive Hashing



Mục tiêu: Tìm các tài liệu có độ giống nhau về Jaccard ít nhất là t

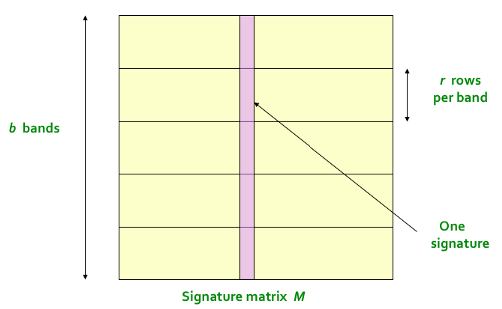
Ý tưởng chung của LSH là tìm một thuật toán sao cho nếu chúng ta nhập chữ ký của 2 tài liệu, nó cho chúng ta biết rằng 2 tài liệu đó có tạo thành một cặp ứng cử viên hay không tức là độ tương đồng của chúng lớn hơn ngưỡng **t** . Hãy nhớ rằng chúng tôi đang lấy sự giống nhau của các chữ ký làm đại diện cho sự giống nhau của Jaccard giữa các tài liệu gốc.

Cụ thể đối với ma trận chữ ký băm tối thiểu:

* Băm các cột của ma trận chữ ký M bằng cách sử dụng một số hàm băm
* Nếu 2 tài liệu băm vào cùng một nhóm cho ít nhất một trong các hàm băm, chúng ta có thể lấy 2 tài liệu làm cặp ứng viên

Bây giờ câu hỏi là làm thế nào để tạo các hàm băm khác nhau. Đối với điều này, chúng tôi làm phân vùng băng tần.

**Phân vùng băng tần**



Đây là thuật toán:

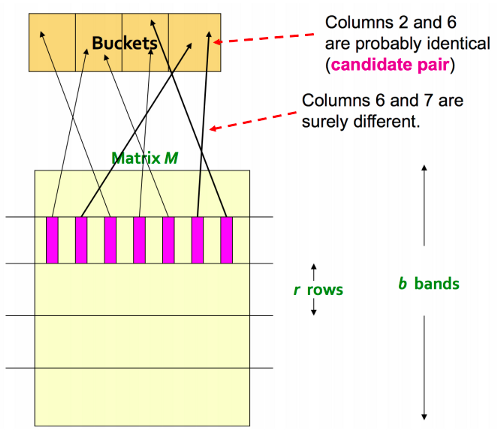
* Chia ma trận chữ ký thành b dải , mỗi dải có r hàng
* Đối với mỗi dải, băm phần của mỗi cột thành một bảng băm với k nhóm
* Các cặp cột ứng cử viên là những cặp cột được băm vào cùng một nhóm cho ít nhất 1 dải
* Điều chỉnh b và r để bắt hầu hết các cặp tương tự nhưng một số cặp không tương tự

Có một số cân nhắc ở đây. Lý tưởng nhất cho mỗi dải, chúng ta muốn lấy k bằng tất cả các tổ hợp giá trị có thể có mà một cột có thể nhận trong một dải. Điều này sẽ tương đương với đối sánh danh tính. Nhưng theo cách này, k sẽ là một con số khổng lồ, không khả thi về mặt tính toán. Ví dụ: Nếu đối với một ban nhạc, chúng ta có 5 hàng trong đó. Bây giờ nếu các phần tử trong chữ ký là số nguyên 32 bit thì k trong trường hợp này sẽ là (2³²) ⁵ ~ 1.4615016e + 48. Bạn có thể xem vấn đề ở đây là gì. Bình thường k lấy trên dưới 1 triệu.

Ý tưởng là nếu 2 tài liệu giống nhau thì chúng sẽ xuất hiện như một cặp ứng cử viên trong ít nhất một trong các nhóm.

**Lựa chọn b & r**

Nếu chúng ta lấy b lớn tức là số lượng hàm băm nhiều hơn, thì chúng ta giảm r vì b \* r là một hằng số (số hàng trong ma trận chữ ký). Theo trực giác, điều đó có nghĩa là chúng tôi đang tăng xác suất tìm thấy một cặp ứng cử viên. Trường hợp này tương đương với việc lấy một **t (ngưỡng tương tự) nhỏ**



Giả sử ma trận chữ ký của bạn có 100 hàng. Hãy xem xét 2 trường hợp:

b1 = 10 → r = 10

b2 = 20 → r = 5

Trong trường hợp thứ hai, có cơ hội cao hơn để 2 tài liệu xuất hiện trong cùng một nhóm ít nhất một lần vì chúng có nhiều cơ hội hơn (20 so với 10) và ít yếu tố của chữ ký được so sánh hơn (5 so với 10).

* B cao hơn ngụ ý **ngưỡng tương tự** thấp hơn **(dương tính giả cao hơn) và b thấp hơn ngụ ý ngưỡng tương tự cao hơn (âm tính giả cao hơn)**

**Hãy thử hiểu điều này thông qua một ví dụ.**

**Thiết lập:**

* 100k tài liệu được lưu trữ dưới dạng chữ ký có độ dài 100
* Ma trận chữ ký: 100 \* 100000
* So sánh bạo lực của các chữ ký sẽ dẫn đến so sánh 100C2 = 5 tỷ (khá nhiều!)
* Lấy b = 20 → r = 5

**ngưỡng tương tự (t): 80%**

Chúng ta muốn 2 tài liệu (D1 & D2) có độ giống nhau 80% được băm trong cùng một nhóm cho ít nhất một trong 20 dải.

P (D1 & D2 giống hệt nhau trong một dải cụ thể) = (0,8) ⁵ = 0,328

P (D1 & D2 không giống nhau trong tất cả 20 dải) = (1–0,328) ^ 20 = 0,00035

Điều này có nghĩa là trong trường hợp này, chúng ta có ~ 0,035% khả năng âm tính giả là 80% các tài liệu tương tự.

Ngoài ra, chúng ta muốn 2 tài liệu (D3 & D4) có độ giống nhau 30% không được băm trong cùng một nhóm cho bất kỳ dải nào trong số 20 dải (ngưỡng = 80%).

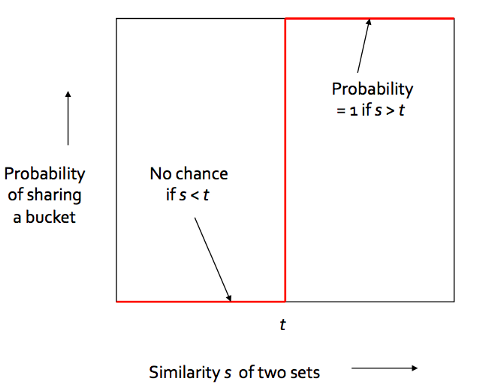
P (D3 & D4 giống hệt nhau trong một dải cụ thể) = (0,3) ⁵ = 0,00243

P (D3 & D4 tương tự nhau ở ít nhất một trong 20 dải) = 1 - (1–0,00243) ^ 20 = 0,0474

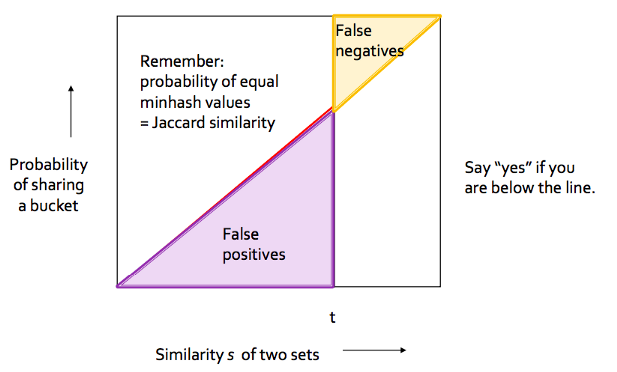
Điều này có nghĩa là trong trường hợp này, chúng tôi có ~ 4,74% khả năng dương tính giả @ 30% các tài liệu tương tự.

Vì vậy, chúng ta có thể thấy rằng chúng ta có một số dương tính giả và một số âm tính giả. Tỷ lệ này sẽ thay đổi theo lựa chọn b và r.

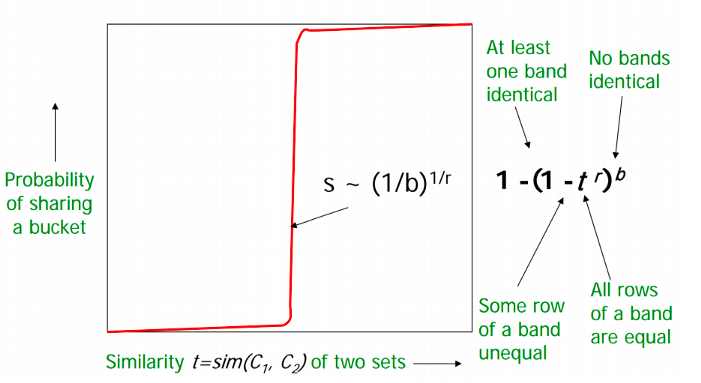
Những gì chúng tôi muốn ở đây là một cái gì đó như dưới đây. Nếu chúng ta có 2 tài liệu có độ giống nhau lớn hơn ngưỡng thì xác suất chúng chia sẻ cùng một nhóm trong ít nhất một trong các dải phải là 1 khác 0.



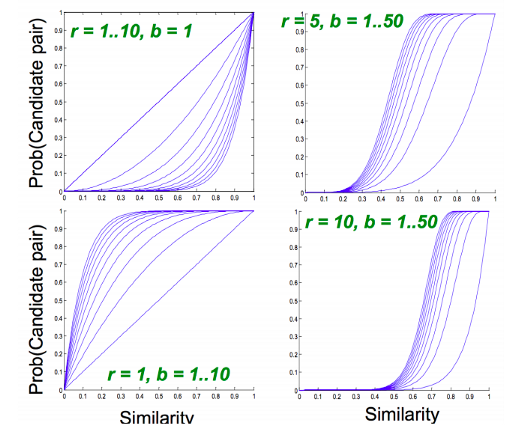
Trường hợp xấu nhất sẽ xảy ra nếu chúng ta có b = số hàng trong ma trận chữ ký như hình dưới đây.



Trường hợp tổng quát cho b và r bất kỳ được hiển thị bên dưới.



Chọn b và r để có đường cong S tốt nhất, tức là tỷ lệ âm tính giả và dương tính giả tối thiểu



Tài liệu tham khảo

<http://infolab.stanford.edu/~ullman/mmds/ch3n.pdf>

https://web.stanford.edu/class/cs246/slides/03-lsh.pdf

<https://web.stanford.edu/class/cs246/slides/04-lsh_theory.pdf>

<https://towardsdatascience.com/understanding-locality-sensitive-hashing-49f6d1f6134>