**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**KHOA CÔNG NGHỆ THÔNG TIN I**

**---------------------------------------**



**ĐỒ ÁN**

**TỐT NGHIỆP ĐẠI HỌC**

|  |
| --- |
| **Đề tài:**  **ÁP ỤNG THUẬT TOÁN OUT OF PLACE VÀO ƯỚC LƯỢNG**  **ĐỘ TƯƠNG ĐỒNG GIỮA CÁC VĂN BẢN** |
|  |
|  |
|  |
| |  |  |  | | --- | --- | --- | | **Giảng viên hướng dẫn** | **:** | **TS. NGUYỄN MẠNH HÙNG** | | **Sinh viên thực hiện** | **:** | **PHẠM HOÀNG DƯƠNG** | | **Lớp** | **:** | **D12CNPM5** | | **Khóa** | **:** | **2012-2017** | | **Hệ** | **:** | **ĐẠI HỌC CHÍNH QUY** | |
|  |
|  |
|  |

**HÀ NỘI – 2016**

LỜI CÁM ƠN

MỤC LỤC

[**CHƯƠNG 1 : GIỚI THIỆU VỀ BÀI TOÁN ƯỚC LƯỢNG ĐỘ TƯƠNG ĐỒNG GIỮA CÁC VĂN BẢN** 1](#_Toc463908852)

[**1.1** **Bài toán ước lượng độ tương đồng giữa các văn bản** 1](#_Toc463908853)

[**1.1.1** **Giới thiệu** 1](#_Toc463908854)

[**1.1.2** **Ứng dụng** 2](#_Toc463908855)

[**1.2** **Các phương pháp ước lượng độ tương đồng giữa các văn bản** 3](#_Toc463908856)

[**1.2.1** **Phương pháp ước lượng độ tương đồng về từ ngữ** 3](#_Toc463908857)

[**1.2.2** **Phương pháp ước lượng độ tương đồng về ngữ nghĩa** 6](#_Toc463908858)

[**1.3** **Đề xuất thuật toán** 11](#_Toc463908859)

[**1.4** **Kết luận** 12](#_Toc463908860)

[**CHƯƠNG 2 : ƯỚC LƯỢNG ĐỘ TƯƠNG ĐỒNG GIỮA CÁC VĂN BẢN BẰNG PHƯƠNG PHÁP OUT OF PLACE** 14](#_Toc463908861)

[**2.1** **Giới thiệu về thuật toán out of place** 14](#_Toc463908862)

[**2.1.1** **Phương pháp tiếp cận dựa trên n-gram** 14](#_Toc463908863)

[**2.1.2** **Mức độ quan trọng của n-gram** 14](#_Toc463908864)

[**2.1.3** **Ước lượng độ tương đồng** 16](#_Toc463908865)

[**2.1.4** **Các bước thực hiện** 17](#_Toc463908866)

[**2.2** **Chi tiết thuật toán** 18](#_Toc463908867)

[**2.2.1** **Đưa văn bản về dạng uni-gram** 18](#_Toc463908868)

[**2.2.2** **Đánh giá mức độ quan trọng của uni-gram bằng tf-idf** 20](#_Toc463908869)

[**2.2.3** **So sánh khoảng cách của các gram** 22](#_Toc463908870)

[**2.3** **Kết luận** 24](#_Toc463908871)

[**CHƯƠNG 3 : ĐÁNH GIÁ THUẬT TOÁN** 26](#_Toc463908872)

[**3.1** **Xây dựng bộ dữ liệu** 26](#_Toc463908873)

[**3.2** **Kịch bản kiểm nghiệm** 29](#_Toc463908874)

[**3.3** **Kết quả thực hiện** 30](#_Toc463908875)

[**3.4** **Đánh giá** 31](#_Toc463908876)

DANH SÁCH CÁC BẢNG

[Bảng 1: Ví dụ về một số từ xuất hiện nhiều trong cả văn nói và viết 16](#_Toc463045845)

[Bảng 2: Ví dụ về khoảng cách giữa các gram trong bài báo 1 với bài báo mẫu 24](#_Toc463045846)

[Bảng 3: Ví dụ về khoảng cách giữa các gram trong bài báo 2 với bài báo mẫu 24](#_Toc463045847)

DANH SÁCH CÁC HÌNH VẼ

[Hình 1: Ảnh chụp của 2 bài báo từ 2 trang web khác nhau 2](#_Toc463304764)

[Hình 2: Các phương pháp ước lượng độ tương đồng dựa trên kí tự 4](#_Toc463304765)

[Hình 3: Các phương pháp ước lượng độ tương đồng dựa trên kho dữ liệu 6](#_Toc463304766)

[Hình 4: Mối quan hệ ngữ nghĩa giữa hai từ thông qua các nút trong WordNet 9](#_Toc463304767)

[Hình 5: Các phương pháp ước lượng độ tương đồng dựa trên tri thức 10](#_Toc463304768)

[Hình 6: Mối quan hệ giữa tấn suất và mức độ quan trọng của n-gram 15](#_Toc463304769)

[Hình 7: Cách tính khoảng cách giữa hai văn bản dựa trên Out Of Place 16](#_Toc463304770)

[Hình 8: Các bước thực hiện của phương pháp Out Of Place 17](#_Toc463304771)

[Hình 9: Nội dung của bài báo “Sân bay Nội Bài, Tân Sơn Nhất bị tin tặc tấn công” 18](#_Toc463304772)

[Hình 10: Các kí tự đặc biệt sẽ bị loại bỏ trong văn bản 19](#_Toc463304773)

[Hình 11: N-gram và số lần xuất hiện trong bài báo 19](#_Toc463304774)

[Hình 12: N-gram và trọng số tf-idf 21](#_Toc463304775)

[Hình 13: N-gram sẽ được sắp xếp theo thứ tự giảm dần theo trọng số tf-idf 22](#_Toc463304776)

[Hình 14: So sánh n-gram của hai báo báo 23](#_Toc463304777)

[Hình 15: Cơ sở dữ liệu chứa thông tin của các bài báo 28](#_Toc463304778)

[Hình 16: Phiếu khảo sát người đọc ước lượng độ tương đồng giữa các bài báo 29](#_Toc463304779)

[Hình 17: Kết quả khảo sát người đọc ước lượng độ tương đồng giữa các bài báo 30](#_Toc463304780)

[Hình 18: Kết quả đánh giá độ chính xác của thuật toán 30](#_Toc463304781)

[Hình 19: Kết quả thực nghiệm của 50 phiếu khảo sát 31](#_Toc463304782)

MỞ ĐẦU

Sự phát triển mạnh mẽ của Internet và các trang báo mạng đã và đang ngày càng phát triển dẫn đến xu hướng bùng nổ thông tin. Các thông tin, tin tức trở thành phương tiện giúp cho việc truyền đạt, trao đổi thông tin, hợp tác, giao lưu… giữa mọi cá nhân tổ chức , và các quốc gia trên khắp hành tinh diễn ra nhanh chóng và cực kỳ tiện ích, góp phần vào sự phát triển của quyền tự do ngôn luận trên toàn thế giới. Cùng với đó là sự phát triển của các thiết bị điển tử trên tay, thậm chí chỉ cần chiếc điện thoại di động mọi người đã có thể tiếp các thông tin ở bất cứ đâu, bất cứ lúc nào. Ở bất kì nơi đâu chỉ cần một chiếc máy tính, một chiếc điện thoại hay máy tính bảng có kết nối internet là mọi người có thể thỏa sức tìm kiếm các thông tin trên báo mạng điện tử ở tất cả các lĩnh vực từ kinh tế, chính tri, [văn hóa](http://baodansinh.vn/dien-dan-van-hoa-c24/), xã hội…và nó cho phép mọi người trên thế giới tiếp cận và đọc không bị phụ thuộc vào không gian và thời gian. Tuy nhiên có những người vẫn thắc mắc nhiều vì thông tin báo chí trùng lặp. Các tờ báo gần như giống nhau, rất nhiều tờ giống nhau. Có một vụ tai nạn giao thông có thể có hàng trăm tờ báo đăng. Có sự kiện chỉ 1 tuần có 1.700 tin bài trên các báo.

Các bài báo là bản sao chính xác của nhau thì dễ dàng được phát hiện. Một vấn đề khó khăn hơn là việc phải xác định các bài báo mà chúng chỉ gần giống nhau (không phải giống nhau hoàn toàn). Đó là những bài báo cùng viết một chủ đề tuy nhiên nội dung của các tờ báo đấy là khác nhau để cung cấp nhiều chiều hơn về một sự việc cho người đọc.

Việc loại bỏ bản sao của các bài báo và đồng thời tìm ra những bài báo có nội dung tương đồng từ đó cải thiện chất lượng tìm kiếm, mang lại lợi ích cho người dùng web khi họ muốn tìm kiếm về sự việc mà họ đang theo dõi. Từ đó giúp người dùng có thể tiếp cận với sự kiện mà họ quan tâm một cách nhanh chóng, tiện lợi mà không gặp phải những nội dung trung lặp, đảm bảo thông tin tiếp cận là duy nhất, nhanh chóng, và là tin tức mới nhất

Hệ thống gợi ý các bài báo có nội dung tương tự và có cùng một chủ đề đối mặt với một số thách thức. Đầu tiên là vấn đề về quy mô: hệ thống phải thực hiện phân loại các bài báo trên rất nhiều trang báo mạng. Thứ hai hệ thống phải thực hiện tìm kiếm thông tin các bài báo cùng chung chủ đề thì phải thu thập hàng tỷ các bài báo mỗi ngày. Vì vậy việc ước lượng độ tương đồng giữa các bài báo để có thể phân loại ra các bài báo trùng lặp, hay các bài báo có nội dung tương tự với chủ đề mà người đọc báo đang theo dõi, từ đó có thể đưa ra những bài báo gợi ý cho người đọc giúp người đọc có thể tiếp cận một cách nhanh chóng và chính xác nhất, những vấn đề hay thông tin họ đang tìm kiếm.

Bài toán được đặt ra là phát hiện các trang báo điện tử có nội dung tương đồng với nhau để có thể xây dựng một hệ thống gợi ý đọc báo cho người dùng bằng cách gom nhóm các bài báo có cùng nội dung và chủ đề và loại bỏ các bài báo trùng lặp. Đảm bảo các bài báo cùng hướng tới chủ đề mà người dùng đang theo dõi hiện tại và hạn chế các bài báo bản sao đem lại các thông tin dư thừa.

Để giải quyết vấn đề, trong đồ án đã tìm hiểu các phương pháp để ước lượng độ tương đồng giữa các bài báo, phương pháp ước lượng này dựa trên độ tương đồng giữa các từ trong bài báo và mức độ quan trong của các từ đó, đó là phương pháp ước lượng Out Of Place , phương pháp phân tách văn bản ra thành n-gram, và đánh giá mức độ quan trọng của từ tf-idf (term frequency – inverse document frequency)

Phương pháp ước lượng độ tương này sẽ được kiểm nghiệm trên 1000 mẫu đã được người đọc khảo sát và trả lời từ trước, trong mỗi mẫu chứa 3 bài báo được lấy từ 10 trang báo điện tử uy tín ở Việt Nam: vnexpress.net, dantri.com.vn, tuoitre.vn, laodong.com.vn... Để có thể đảm bảo phương pháp ước lượng độ tương đồng giữa các bài báo có độ chính xác cao để có thể áp dụng giải quyết bài toán ở trên.

Nội dung của đồ án bao gồm 4 chương:

**Chương 1:** Giới thiệu khái quát về bài toán ước lượng độ tương đồng giữa các văn bản

**Chương 2:** Giới thiệu các hướng tiếp cận giải quyết bài toán ước lượng độ tương đồng giữa các bài báo bằng phương pháp out of place. Chương này tập trung vào việc giới thiệu và áp dụng các phương pháp : tách văn bản thành n-gram, đánh giá trọng số tf-idf, ước lượng khoảng cách bằng out of place. Đây là cơ sở để đưa ra kết luận độ tương đồng giữa văn bản so sánh với văn bản mẫu.

**Chương 3:** Để có thể kiểm nghiệm độ chính xác của phương pháp out of place. Với 1000 mẫu đã được người đọc khảo sát và trả lời câu hỏi, sau đó áp dụng thuật toán out of place để kiểm nghiệm kết quả có trùng hoàn toàn với ý kiến của người khảo sát không.

**Chương 4:** Demo ứng dụng gợi ý các bài báo cho người đọc

**CHƯƠNG 1 : GIỚI THIỆU VỀ BÀI TOÁN ƯỚC LƯỢNG ĐỘ TƯƠNG ĐỒNG GIỮA CÁC VĂN BẢN**

Chương 1 sẽ trình bày các nội dung sau:

* Bài toán ước lượng độ tương đồng giữa các văn bản
* Các phương pháp ước lượng độ tương đồng giữa các văn bản
* Đề xuất thuật toán
  1. **Bài toán ước lượng độ tương đồng giữa các văn bản**
     1. **Giới thiệu**

Ước lượng độ tương đồng giữa các văn bản đóng vai trò ngày một quan trọng trong các nghiên cứu hoặc ứng dụng liên quan đến văn bản. Ví dụ trích xuất thông tin, phân loại văn bản, phân nhóm tài liệu, phát hiện chủ đề, theo dõi chủ đề, trả lời câu hỏi, chấm điểm bài tiểu luận, chấm điểm các câu trả lời, tóm tắt văn bản, dịch văn bản, gần đây hơn là việc trích chọn văn bản, đánh giá và độ đo tương đồng cũng được sử dụng cho việc đánh giá tính chặt chẽ của văn bản và nhiều vai trò khác.

Việc ước lượng độ tương đồng giữa các văn bản hoặc có thể là tương đồng về chủ đề của văn bản/nhóm các văn bản thực sự nhận được sự quan tâm của xã hội trong các hội nghị khoa học quốc tế, đặc biệt trong hội nghĩ thường niên về hiểu văn bản (Document Understanding Workshop - DUC). Việc xây dựng các phương pháp để ước lượng độ tương đồng giữa các văn bản sẽ giúp cho ứng dụng trở nên thông minh hơn đặc biệt trên các trang cung cấp thông tin nói chung và các trang báo mạng nói riêng.

Để tìm độ tương đồng giữa các văn bản có hai cách tiếp cận chính : ước lượng độ tương đồng dựa trên ngữ nghĩa, ước lượng độ tương đồng dựa trên từ ngữ. Phương pháp tìm độ tương đồng giữa hai văn bản dựa trên từ ngữ là một phương pháp sự dụng việc so sánh các từ đơn giản xét trên khía cạnh giống nhau về mặt chữ cái và vị trí của nó trong văn bản, số lần xuất hiện, mức độ phổ biến hay được dùng trong văn nói cũng như văn viết. Một số cải tiến được áp dụng trong phương pháp đơn giản này như loại bỏ các kí tự đặc biệt được dùng trong văn bản, đánh giá trọng số của từ(mức độ quan trong trọng của từ đó trong văn bản). Tuy nhiên phương pháp ước lượng độ tương đồng dựa trên từ ngữ không thể xác định sự giống nhau về ỹ nghĩa giữa các văn bản trong trường hợp hai bài viết cùng một nội dung nhưng cách diễn tả lại khác nhau.

Ngoài phương pháp ước lượng độ tương đồng dựa trên từ ngữ còn một phương pháp khác là dựa trên ngữ nghĩa của từ: dựa trên cơ sở tri thức (knowledge-based) hoặc dựa trên kho dữ liệu (corpus-based). Các phương pháp ước lượng độ tương đồng dựa trên ngữ nghĩa này được áp dụng trong một số lĩnh vực : xử lý ngôn ngữ như phát hiện dùng từ sai nghĩa, nhận dạng từ đồng nghĩa. Với việc ước lượng độ tương đồng về ngữ nghĩa giữa các văn bản được thực hiện bởi một số phương pháp sau: phương pháp xấp xỉ thu được qua việc mở rộng truy vấn, cũng như việc hiện tìm kiếm thông tin, hay phương pháp phân tích ngữ nghĩa ẩn, các phương pháp ước lượng độ tương đồng giữa các văn bản dựa trên khai thác mối quan hệ giữa các từ một cách tự động qua kho dữ liệu từ

* + 1. **Ứng dụng**

Do sự phát triển của Internet cùng với đó là sự bùng nổ của thông tin, dữ liệu được sinh ra liên tục mỗi ngày, khối lượng dữ liệu trên web rất lớn, do đó vấn đề trung lặp thông tin thường xuyên xảy ra. Đặc biệt là trong lĩnh vực báo chí, nhiều người thắc mắc vì thông tin trùng lặp trong báo chí. Các tờ báo gần như giống nhau. Có một vụ tai nạn giao thông có thể có hàng trăm tờ báo đăng. Có sự kiện chỉ 1 tuần có 1.700 tin bài trên các báo. Dẫn đến việc thông tin bị dư thừa quá nhiều, nguyên nhân chính là do các bài báo bị sao chép hay chỉnh sửa lại sau đó đăng lại tại nhiều trang web khác nhau.

|  |  |
| --- | --- |
|  |  |

Hình : Ảnh chụp của 2 bài báo từ 2 trang web khác nhau

Để có thể tìm hiểu về các thông tin có sẵn trên các trang web rất nhiều công nghệ hay hệ thống tìm kiếm thông tin xuất hiện. Tuy nhiên sự tồn tại của các bài báo điện tử trùng lặp làm giảm đi tính hữu dụng và sự hiểu quả của các công cụ tìm kiếm. Bởi vì do kết quả trùng lặp của câu truy vấn dẫn đến, cùng một bài viết nhưng xuất hiện ở từ hai đến ba trang web khác nhau làm giảm đi tính hiệu quả của công cụ tìm kiếm. Việc đọc các bài báo có kết quả trùng lặp rất tốn thời gian và cũng không thêm bất kỳ giá trị nào về thông tin cung cấp cho người dùng. Vì vậy, các bài báo là bản sao nội dung của các bài báo khác làm giảm hiệu quả của một công cụ tìm kiếm, không đánh giá chính xác chất lượng cung cấp tin của các trang báo điện tử. Trong khi đó, người sử dụng muốn có được thông tin gốc, còn người cung cấp tin muốn khẳng định chất lượng cung cấp tin của mình. Vì vậy, việc phát hiện các bài báo có cùng bài viết là thực sự cần thiết cho trình thu thập thông tin, đảm bảo thông tin cung cấp tới người dùng đúng và đủ.

Do đó việc xác định tương đồng giữa các bài báo nhắm giúp hạn chế bớt các bài tờ bào không có uy tín lấy cắp nội dùng bài viết của các tờ báo khác. Như vậy trong khóa luận này sẽ tập trung vào tìm cách nhận dạng với một bài báo gốc và các bài báo của nhiều tờ khác nhau và kiểm tra xem các bài báo đó bài báo nào giống với bài báo gốc nhất. Mục đích chính là để gom nhóm các bài báo dư thừa, tăng chất lượng tìm kiếm và quá trình lưu trữ của công cụ tìm kiếm được hiệu quả. Ví dụ người dùng sẽ được thông báo rằng với mội dung tìm kiếm này tập hợp các bài báo nói về nọi dung này có những bài nào của các báo điện tử khác tương tự nội dụng. Việc gom nhóm các bài báo có bài viết tương đồng nhau như vậy sẽ giúp đỡ cho người sử dụng tiếp cận với thông tin nhanh nhất và hạn chế việc phải gặp các thông tin có bài viết trùng lặp. Trong trường hợp tốt nhất là khi một bài báo được đưa lên sẽ được kiểm duyệt để loại bỏ trùng lặp. Việc hạn chế các bài báo có cùng bài viết giống nhau sẽ giúp tăng sự đa dạng của kết quả tìm kiếm vì mỗi nhóm trùng lặp chỉ được hiển thị một văn bản, do đó mà hiển thị được nhiều bài báo cho sự tìm kiếm của người dùng hơn.

* 1. **Các phương pháp ước lượng độ tương đồng giữa các văn bản**
     1. **Phương pháp ước lượng độ tương đồng về từ ngữ**

Sự tương đồng của chuỗi từ được ước lượng dựa trên trình tự chuỗi và vị trí của các từ trong văn bản. Đo lường dựa trên các từ là một cách đo lường bằng cách ước lượng độ tương đồng hoặc không tương đồng (khoảng cách) giữa các từ trong văn bản để xem có trùng hoặc khác nhau

Ước lượng độ tương đồng dựa trên từ ngữ sẽ được chia thành hai nhóm: ước lượng dựa trên từ, ước lượng dựa trên kí tự. Hình 2 giới thiệu 14 thuật toán ước lượng độ tương đồng nổi tiếng nhất, bẩy thuật toán trong số đó là ước lượng dựa trên từ, bẩy cái còn lại là ước lượng dựa trên kí tự

Hình : Các phương pháp ước lượng độ tương đồng dựa trên kí tự

* **Character-Based Similarity Measures**

**Longest Common SubString (LCS) [1] :** là thuật toán ước lượng độ tương đồng giữa các chuỗi dựa trên độ dài liên tục của các kí tự nối tiếp tồn tại ở cả hai chuỗi.

**Damerau-Levenshtein [1] :** định nghĩa khoảng cách của hai chuỗi bởi việc đếm số phép tính nhỏ nhất cần để chuỗi này thành chuỗi kia mà các phép tính được định nghĩa là thêm, xóa, thay thế một kí tự hoặc đổi chỗ hai kí tự liền kề

**Jaro [1] :** dựa trên số kí tự giống nhau và thứ tự của các kí tự con giữa 2 chuỗi, sau đó thực hiện tính toán các thông số: số các kí tự giống nhau nhưng không theo thứ tự, số các kí tự giống nhau và đúng thứ tự, số chuyển đổi kí tự cần thiết, đưa các thống số tìm được vào bảng liên kết và tính toán theo các thông số đó để đưa ra kết luận về độ tương đồng giữa hai chuỗi

**Jaro-Winkler [1]:** là phiên bản mở rộng của Jaro, tuy nhiên winker thêm trọng số ký tự trùng khớp đầu tiên

**Needleman-Wunsch [1] :** là thuật toán ví dụ về quy hoạch động là ứng dụng đầu tiên của quy hoạch động để so sánh trình tự trong sinh học (VD: nucleotide và trình tự axit amin). Nó sẽ thực hiện tìm ra những đoạn hay những miền của hai chuỗi có độ tương đồng cao nhất, để từ đó đánh giá mức độ tương đồng giữa hai chuỗi. Thuật toán này phù hợp cho hai chuỗi có cùng độ dài

**Smith-Waterman [1] :** đây là một ví dụ khác của thuật toán quy hoạch động để tìm độ tương đồng giữa các chuỗi. Thuật toán thực hiện gióng cặp chuỗi cục bộ dựa trên quy hoạch động để tính điểm cho qua trình gióng chuỗi. Từ đó có thể phát hiện các đoạn hay miền của hai chuỗi có độ tương đồng hoặc không tương đồng.

**N-gram [1]:**  là tần suất xuất hiện của n kí tự (hoặc từ) liên tiếp nhau có trong chuỗi kí tự (văn bản) ban đầu. Thuật toán ước lượng độ tương đồng dựa trên N-gram bằng cách so sánh khoảng cách của các gram từ hai chuỗi ban đầu. Khoảng cách được tình bằng cách lấy số gram giống nhau chia cho tổng số gram lớn nhất. Gram là một mô hình ngôn ngữ trong đó có thể chứa từ n kí tự hoặc n từ. Ví dụ với n=1(uni-gram) ta tách văn bản ra thành các gram chứ một từ không trùng lặp với các từ còn lại trong văn bản, tương tự ta có bi-gram(n=2), tri-gram(n=3)

**Out of place [1]:** là phương pháp được phát triển từ N-gram. Tương tự như n-gram để ước lượng độ tương đồng giữa các văn bản, tách văn bản thành các n-gram (với n là số lượng kí tự hoặc số lượng từ chứa trong một gram). Các gram này sẽ không chứa dấu cách, con số, tất cả các dấu ngoại trừ dầu nháy. Sắp xếp các gram theo số lần xuất hiện trong văn bản. Để ước lượng độ tương đồng giữa hai văn bản ta thực hiện như sau : từ danh sách các gram đã được sắp xếp của hai văn bản ta sẽ kiểm tra khoảng cách của các gram giống nhau ở cả hai danh sách. Khoảng cách sẽ được tính bằng vị trí của từ trong văn bản thứ nhất trừ đi vị trí của từ trong văn bản thứ 2 lấy trị tuyệt đối của kết quả đó. Đối với các gram không giống với gram nào ở văn bản còn lại thì khoảng cách cua nó sẽ là khoảng cách lớn nhất. Khoảng cách giữa hai văn bản sẽ được tính bằng tổng khoảng cách của các gram

* **Term-Based Similarity Measures**

**Block Distance [1]:** còn được biết đến là Manhataan distance, boxcar distance, L1 distance. Nó tính khoảng cách bằng quãng đường đi để di chuyển từ điểm này tới điểm kia theo dạng lưới. Tên gọi khác khoảng cách Manhataan thể hiện cách bố trí dạng lưới của đường phố Manhataan. Vậy con đường ngắn nhất để chiếc xe có thể đi giữa hai nút giao thông trong thành phố có chiều dài bằng tổng khoảng cách của các nút giao thông được gọi là khoảng cách Manhataan

**Cosine Similarity [1] :** là một trong các phương pháp ước lượng độ tương đồng giữa hai vector giao nhau trong không gian thì được tính bằng góc cosine của hai vector đó

**Dice’s coefficient [1] :** được định nghĩa là hai lần của số các từ là tập con của hai chuỗi chia cho tổng số các từ ở trong cả hai chuỗi.

**Euclidean distance [1] :** còn được gọi là L2 distance là căn bậc hai tổng bình phương giữa các tọa đô tương ứng của hai vector

**Jaccard similarity [1] :** được tính bằngsố các từ chung của cả hai chuỗi chia cho tổng số tất cả các từ khác biệt trong cả hai chuỗi

**Matching Coefficient [1] :** là một cách tiếp cận đơn giản dựa trên vector bằng cách đếm số các từ giống nhau ở cả hai vector

**Overlap coefficient [1] :** tương tự với kĩ thuật Dice's coefficient, tuy nhiên kĩ thuật này cân nhắc trong hai chuỗi có chuỗi nào là chuỗi con của chuỗi còn lại không.

* + 1. **Phương pháp ước lượng độ tương đồng về ngữ nghĩa**

Tương đồng dựa trên kho dữ liệu là cách ước lượng dựa độ tương đồng dựa trên ngữ nghĩa có nghĩa để xác định độ tương đồng giữa hai từ dựa vào thông tin thu được từ kho dữ liệu lớn. Kho dữ liệu là một tập hợp lớn các văn bản lời nói hoặc viết được sử dụng cho ngôn ngữ nghiên cứu. Hình 3 bao gồm các thuật toán ước lượng độ tương đồng dựa trên kho dữ liệu.

* **Corpus-Based Similarity**

Hình : Các phương pháp ước lượng độ tương đồng dựa trên kho dữ liệu

**Hyperspace Analogue to Language (HAL) [1] :** tạo ra một không gian ngữ nghĩa từ các từ có ý nghĩa tương tự và thường xuyên được sử dụng (tập hợp đồng xảy ra). Sau đó ta tạo ra ma trận dạng N\*N mà tất cả các từ duy nhất được biểu diễn tương ứng với mỗi hàng và mỗi cột n từ duy nhất đó. Với mỗi từ a trong chuỗi ta đánh trọng số cho các từ b có ngữ nghĩa tương ứng xuất hiện trong chuỗi đó. Việc đánh số thực chất sử dụng trọng số đồng xảy ra. Ví dụ b xảy ra liền kề với a, thì nó sẽ nhận trọng số là năm ,nếu b cách a một từ thì trọng số của nó là 4 và giảm đi nếu số từ cách tăng lên. Nếu b tách biệt với a trọng số của nó sẽ bằng 0. Như vậy HAL ghi lại thông tin của từ bằng cách đặt các trọng số đồng xảy ra ghi mà có từ hàng xóm xảy ra trong trước hoặc sau từ cần kiểm tra.

**Latent Semantic Analysis (LSA) [1] :** đây là phương pháp nổi tiếng nhất của ước lượng độ tương đồng dựa trên kho dữ liệu. LSA cho rằng các từ có cùng ngữ nghĩa sẽ xảy ra trong cùng một hoàn cảnh trong văn bản. Như vậy có một ma trận với mỗi từ là index (xuất hiện nhiều hơn 2 lần và không phải là stopword) sẽ được xếp vào các hàng. Số thứ tự đoạn (câu) của văn bản sẽ được xếp vào các cột. Trong ma trận mỗi ô sẽ chứa số đếm của các từ xảy ra bao nhiêu lần trong đoạn(câu) của văn bản. Có một kĩ thuật được sử dụng để giảm số lượng các cột những vẫn giữ được cấu trúc tương đồng giữa các hàng là singular value decomposition (SVD). Các từ sẽ được so sánh thông qua góc cosine giữa 2 vectors được hình thành bởi hai hàng

**Generalized Latent Semantic Analysis (GLSA) [1] :** là một nền tảng cho việc tính toán ngữ nghĩa của từ và các vector của văn bản. Nó được mở rộng từ phương pháp LSA bằng cách tập trung vào vector của từ thay vì so sánh tổng thể giữa hai văn bản. GLSA yêu cầu một phương pháp ước lượng tương đồng ngữ nghĩa giữa các từ và một phương pháp giảm chiều cho vector. Phương pháp tiếp cận GLSA có thể kết hợp bất cứ phương pháp ước lượng độ tương đồng trong không gian từ với các phương pháp phù hợp để giảm chiều cho vector. Trong ma trận văn bản truyền thống tại bước cuối mới đánh giá trọng số của các từ trong vector

**Explicit Semantic Analysis (ESA) [1] :** là một phương pháp ước lượng để tính toán mối liên hệ ngữ nghĩa giữa hai văn bản. Đây là một kĩ thuật dựa trên Wikipedia miêu tả từ như một vector nhiều chiều, mỗi thành phần của vector được miêu tả bởi trọng số TF-IDF. Các từ có ngữ nghĩa giống nhau thì sẽ được biểu diễn bởi góc cosine giữa hai vector tương ứng của hai từ đó

**The cross-language explicit semantic analysis (CL-ESA) [1]** : là phương pháp ước lượng tổng quát đa ngôn ngữ của phương pháp ESA. CL-ESA khai thác dữ liều từ các tài liệu đa ngôn ngữ ví dụ như Wikipedia để tạo ra một bản tài liệu như một vector độc lập về ngôn ngữ. Sự tương đồng của hai tài liệu ở hai ngôn ngữ khác nhau sẽ được đánh giá bởi góc cosine được tạo bởi hai vector tương ứng với hai văn bản cần ước lượng về sự tương đồng.

**Pointwise Mutual Information - Information Retrieval (PMI-IR) [1] :** đây là một phương thức để đánh giá độ tương đồng giữa các cặp từ ,phương thức sử dụng câu lệnh truy vấn nâng cao của AltaVista để tính toán sắc xuất. Cặp từ xảy ra đồng thời trên web-page càng nhiều, thì điểm tương đồng PMI-IR càng cao.

**Second-order co-occurrence pointwise mutual information (SCO-PMI):** là một phương pháp ước lượng động tương đồng dựa trên ngữ nghĩa sử dụng thông tin tương hỗ theo tưng điểm (pointwise mutual infomation). Loại thông tin tương hỗ này để đo lượng thông tin từ mà một từ cho biết về từ kia. Sử dụng thông tin tương hỗ theo từng điểm để sắp xếp danh sách các láng giềng quan trọng của hai từ trong kho dữ liệu lớn. Ưu điểm của SOC-PMI là việc phương pháp này tính toán độ tương đồng giữa hai từ mà không xảy ra thường xuyên, bởi vì chúng đồng xảy với các từ láng giềng giống nhau

**Normalized Google Distance (NGD) [1] :** là một phương pháp ước lượng độ tương đồng dựa trên ngữ nghĩa của từ, phương pháp này dừa vào số kết quả trả về của công cụ tìm kiếm Google cho một tập hợp các từ khóa đưa vào. Các từ khóa có cùng ngữ nghĩa trong ngôn ngữ tự nhiên có xu hướng gần trong đơn vị của Google Distance. Trong khi các khác nhau về ngữ nghĩa có xu hướng cách xa từng phần. Đặc biệt công thức tính khoảng cách của Normalized Google Distance giữa hai từ x và y là :

NGD(x, y) =

Trong đó M là tống số trang web tìm thấy được bởi Google; f(x) và f(y) là số kết quả trùng với từ x và y và f(x, y) là số trang web mà có chứa cả hai từ khóa x và y. Nếu kết quả tìm kiếm mà không có trang web nào chứa cả hai từ khóa, và chỉ có kết quả riêng rẽ với cả hai từ, thì khoảng cách Normalized Google là vộ tận. Nếu hai từ luôn xuất hiện cùng nhau, điểm NGD sẽ bằng không hoặc tương đương với với Coefficient giữa x bình phương và y bình phương.

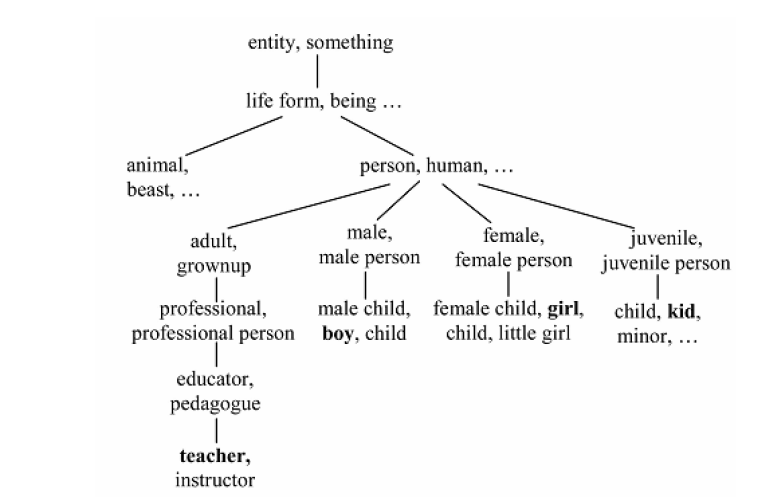
**Extracting DIStributionally similar words using CO-occurrences** (DISCO) **[1]** : là phương pháp dựa trên tương đồng phần phối giữa các từ, phương pháp này đặt ra giả thuyết các từ tương đồng về ngữ nghĩa thì xảy ra trong cùng một đoạn. Tập hợp các văn bản được phân tích thống kê để tìm được tương đồng phân phối. DISCO là phương thức để tích sự tương đồng phân phối giữa các từ bằng cách sử dụng cửa sổ ma trận có kích cỡ là ±3 từ để đếm từ đồng xảy ra. Khi hai từ bắt buộc chuyển về phương pháp tương đồng DISCO đơn giản chỉ thu thập vector của từ từ chỉ số của dữ liệu và tính toán độ tương đồng dựa vào ước lượng Lin. Nếu hầu hết các từ tương đồng phần phối được yêu cầu, DISCO trả về vector từ thứ hai cho từ được yêu cầu. DISCO có hai phần ước lượng độ tương đồng chính là DISCO1 và DISCO2; DISCO1 tính yêu cầu thứ nhất về độ tương đồng giữa hai từ đầu vào dựa trên tập collocation. DISCO2 tính yêu cầu thứ hai về độ tương đồng dựa trên tập tương đồng phân phối của các từ

* **Knowledge-Based Similarity**

Ước lượng độ tương đồng dựa trên tri thức cũng là một trong các phương pháp ước lượng độ tương đồng dựa vào ngữ nghĩa của văn bản bằng cách dựa vào mức độ tương tự giữa các từ sử dụng thông tin nhận được từ mạng ngữ nghĩa. Wordnet là một trong những mạng ngữ nghĩa nổi tiếng nhất trong lĩnh vực ước lượng độ tương đồng dựa trên tri thức, WorldNet là một kho dữ liệu lớn các từ tiếng anh. Động từ, danh từ, tính từ và trạng từ được nhóm lại thành một tập hợp đồng nghĩa, thể hiện ý nghĩa riêng biệt. Các tập từ đồng nghĩa này được tập hợp lại dựa vào ngữ nghĩa của chúng và giống nghau về từ ngữ.

Ví dụ đối với hai từ boy và teacher, khi xét mối quan hệ giữa 2 từ này trên tập

wordnet ta có thể xây dựng được một cấu trúc cây thể hiện mối quan hệ ngữ nghĩa giữa hai từ thông qua các nút khác như trong hình vẽ. **Teacher** – educator – professional – adult – person – male – male child – **boy**



Hình : Mối quan hệ ngữ nghĩa giữa hai từ thông qua các nút trong WordNet

Ước lượng độ tương đồng dựa trên tri thức được chia thành 2 nhóm : ước lượng tương đồng dựa trên ngữ nghĩa và ước lượng dựa trên liên quan ngữ nghĩa. Mô hình ước lượng tương đồng dựa trên ngữ nghĩa được xem như là tương đồng về vẻ bề ngoài của từ. Mặt khác, ước lượng dựa trên liên quan ngữ nghĩa thì lại liên quan về ngữ nghĩa giữa các từ, không qua chặt chẽ trong một mô hình nhất định. Nói cách khác, độ tương đồng dựa trên ngữ nghĩa bản chất xét mối quan hệ giữa hai từ, để xác định sự tương đồng giữa các mỗi quan hệ ấy dựa vào một số mô hình cụ thể hai từ đấy thuộc loại từ gì, các ví dụ cụ thể, nó làm trong câu như thế nào.

Hình : Các phương pháp ước lượng độ tương đồng dựa trên tri thức

Ước lượng tương đồng dựa trên tri thức có 6 phương pháp đặc trưng, trong đó có 3 phương pháp dựa trên thông tin nội dung của từ : **Resnik (*res*) [1]** , **Lin (*lin*) [1]** and **Jiang & Conrath (*jcn*) [1]**. Ba phương pháp còn lại dựa trên độ dài con đường để biểu thị tương đồng về ngữ nghĩa của từ : **Leacock & Chodorow (*lch*) [1]**, **Wu & Palmer (*wup*) [1]** và **Path Length (*path*) [1]**.

Các giá trị trong phương pháp ước lượng **res** sẽ bằng giá trị nội dung thông tin(IC) của cây phân cấp ngữ nghĩa. Như vậy các giá trị đó sẽ luôn có giá trị lớn hơn hoặc bằng không. Giới hạn của các giá trị trên thì rất lớn và đa dạng tùy thuộc vào kích thước của kho dữ liệu sử dụng để xác định giá trị nội dung thông tin. Phương pháp ước lượng **lin** và **jcn** ngoài tính thêm cả nội dung thông tin của cây phân cấp ngữ nghĩa còn tính với tổng của nội dung thông tin của hai từ so sánh. Tuy nhiên hai phương pháp khác nhau ở điểm sau: phương pháp lin thực hiện cân bằng nội dung thông tin của cây phân cấp ngữ nghĩa trong khi jcn lấy sự khác nhau giữa tổng nội dung thông tin của 2 từ với giá trị của cây phân cấp ngữ nghĩa

Phương pháp ước lượng **lch** trả về kết quả tính độ tương đồng về ngữ nghĩa của hai từ dựa trên con đường ngắn nhất kết nỗi ngữ nghĩa của hai từ đó trong cây phân cấp ngữ nghĩa. Độ xâu nhất tối đa của hệ thống cây phân cấp.

Phương pháp ước lượng **wup** trả về kết quả sự tương đồng ngữ nghĩa giữa các từ dựa vào chiều sâu ngữ nghĩa của cả các từ đồng nghĩa và bao hàm ngữ nghĩa và hệ thống cây phân cấp. Phương pháp ước lượng path trả về mức độ tương đồng ngữ nghĩa dựa trên con đường ngắn nhất kết nối ngữ nghĩa bao gồm cả từ đồng nghĩa và từ bao hàm ngữ nghĩa.

Hơn nữa, còn có ba phương pháp ước lượng động tương đồng dựa trên ngữ nghĩa của từ: St.Onge (***hso***), Lesk (***lesk***) và vector pairs (***vector***). Phương pháp ước lượng thực hiện băng cách tìm các từ liên kết ngữ nghĩa của từ cần so sánh. Có 3 trường được cân nhắc để ước lượng động tương đồng: liên kết cực mạnh, liên kết mạnh, liên kết bình thường. Điểm liên kết lớn nhất là 16. Phương pháp ước lượng lesk xử lý băng cách tìm tập con chung của tập các từ đồng nghĩa với 2 từ cần so sánh. Điểm tương đồng sẽ được tính bằng tổng độ dài bình phương của tập con. Phương pháp ước lượng vector tạo ma trận đồng xảy ra cho mỗi từ được sử dụng trong kho dữ liệu WorldNet và đại diện cho mỗi từ hay chuỗi từ là vector từ đó ta có thể ước lượng độ tương đồng thông qua hai vector của từ hoặc chuỗi từ cần so sánh.

* 1. **Đề xuất thuật toán**

Ứng dụng nêu trên trung tập trung vào việc phân loại các văn bản có bài viết tương đồng nhau, chưa xem xét đến việc so sánh độ tương đồng ngữ nghĩa giữa các văn bản. Do đó thuật toán Out of place được chọn để giải quyết bài toán phát biểu ở trên.

Ưu điểm của thuật toán Out of place:

* Là phương pháp lý tưởng để xử lý các văn bản có nhiễu
* Dễ dàng cài đặt.
* Khả năng làm việc tốt với cả văn bản dài hoặc chỉ là một câu nói thông thường.
* Sử dụng n-gram có thể so sánh được một số từ đặc biệt (Ví dụ, ‘advance’, ‘advanced’, ‘advancing’,‘advancement’, etc.). Nếu không ta cần có một hệ thống thật sự tốt để xử lý ngữ nghĩa của từ

Mặc dù thuật toán được mô tả với một hiệu năng tốt, cho kết quả tương đối chuẩn xác, tuy nhiên vẫn cần có những cải thiện nhắm nâng cao tính đúng đắn trong việc ước lượng độ tương đồng giữa các văn bản:

* Do thuật toán áp dụng sử dụng uni-gram tuy nhiên việc tách câu ra thành từng gram như vậy sẽ chuyển các gram thành những từ không có ý nghĩa hay đem lại ý nghĩa mới hoàn toàn, ví dụ : “sinh viên” khi tách ra ta sẽ có 2 gram là “sinh”, “viên”
* Thuật toán tuy áp dụng tf-idf để có thể đánh giá mức độ quan trọng của các gram tuy nhiên để thuật toán có kết quả tốt nhất thì nên bỏ qua việc thống kê của các từ phổ biến hay các từ đặc trưng của ngôn ngữ, đặc biệt là các từ hay được sử dụng trong tiếng việt
* Chất lượng của bài báo cũng có ảnh hưởng tới độ chính xác của việc ước lượng độ tương đồng giữa các bài báo. Nếu bài báo có quá nhiều phần gây nhiễu, dẫn đến việc ước lượng sai, mặc dù hai bài báo khác nhau nhưng kết luận của thuật toán là giống nhau và ngược lại
* Độ dài của các văn bản cùng là một phần ảnh hướng tới ước lượng độ tương đồng. Văn bản dài hơn thì có thể số gram trong sẽ tăng lên dẫn đến khoảng cách của văn bản sẽ tăng theo, dẫn đến khi ước lượng độ tương đồng khi có 2 văn bản cùng so sánh với mẫu mà 2 văn bản có độ dài chênh lệch quá lớn dẫn đến kết quả ước lượng tương đồng không còn chính xác nữa. Đặc biệt đối trong việc xét các bài báo cùng một chủ để nhưng mỗi bài có cách truyền tải khác nhau có thể bằng hình ảnh, có thể bằng từ ngữ lên cũng ảnh hưởng tới việc ước lượng độ tương đồng cảu các bài báo.
* Thuật toán chỉ xử lý tốt với các ngôn ngữ được biểu ở chuẩn ASCII.
* Bản chất của thuật toán Out of place là ước lượng sự tương đồng giữa các văn bản dựa trên từ ngữ. Do đó thuật toán sẽ không ước lượng chính xác với hai văn bản có cùng ý nghĩa nhưng cách biểu diễn lai khác nhau và ngược lại. Ví dụ trong trường hợp hai văn có độ tương đồng về từ ngữ cao nhưng ngữ nghĩa lại khác nhau:
  + D1 = A snake attacks a frog. A crocodile also attacks the frog.  
    The crocodile wins. It then eats the frog while the snake changes  
    his target to a toad.
  + D2 = A snake attacks a frog. A crocodile also attacks the frog.  
    The snake losses. It updates his target to a toad while the crocodile eats the frog.
  1. **Kết luận**

Từ các vấn đề thực tế được nêu ra ở trên để có thể hạn chế được người dùng gặp phải các bài báo có cùng nội dung bài viết tương đồng, trong đồ án sẽ xây dựng ứng dụng để phát hiện các bài báo có nội dung tương đồng nhau và gộp nhóm các bài báo lại, từ đó khi người dùng đọc báo sẽ có những gợi ý cho người đọc các bài báo có nội dung tương đương. Tuy nhiên việc gom các bài báo trên có sự tương đồng về bài viết là quá nhiều và dư thừa do một số trang web không đảm bảo về an toàn khi truy cập hay nội dung không được đảm bảo. Do đó ứng dụng sẽ gom nhóm các bài báo của những trang web uy tín đã được cấp phép bởi một cơ quan chức năng của nước sở tại (Chính thống).Một sô báo điện tử được sử dụng để làm cơ sở dữ liệu : vnexpress, dân trí, lao động, tuổi trẻ, đời sống và pháp luật, tiền phong …

Bài toán kiểm tra sự tương đồng giữa các bài báo thực chất là bài toán ước lượng độ tương đồng giữa các văn bản.

Các bài báo từ các trang web (vnexpress, dantri, laodong, tuoitre…) sẽ được thu thập lưu trữ sau đó sử dụng các thư viện lập trình để loại bỏ các thẻ định dạng quảng cáo, ý kiến bạn đọc, các thông về trang báo điện tử chỉ để lại nội dung của bài báo. Như vậy từ một bài báo điện tử ta đã chuyển về một văn bản và có thể lưu trữ ở file text hoặc tại cơ sở dữ liệu của hệ thống. Sau đó áp dụng các phương pháp ước lượng độ tương đồng giữa các văn bản để kiểm tra độ tương đồng giữa các văn bản đó

Trong chương 2, đồ án sẽ trình bày về thuật toán out of place: giới thiệu thuật toán, các bước xử lý trong việc ước lượng độ tương đồng giữa các văn bản.

**CHƯƠNG 2 : ƯỚC LƯỢNG ĐỘ TƯƠNG ĐỒNG GIỮA CÁC VĂN BẢN BẰNG PHƯƠNG PHÁP OUT OF PLACE**

Chương 2: dưới đây là những vấn đề sẽ được trình bày

* Giới thiệu về thuật toán out of place
* Chi tiết thuật toán
  1. **Giới thiệu về thuật toán out of place**
     1. **Phương pháp tiếp cận dựa trên n-gram**

N-gram là một tập hợp kí tự hoặc từ được cắt ra từ một từ hoặc một câu văn. N- gram hay còn được gọi là các ‘term’, trong khoa học thì ‘term’ có thể hiểu là một cụm từ liên quan đến một lĩnh vực nào đó, ví dụ: ‘máy tính’, ‘công nghệ phần mềm’, ‘hệ thống thông tin’

* bi-gram : công nghệ, nghệ phần, phần mềm
* tri-grams: công nghệ phần, nghệ phần mềm
* quad-grams: công nghệ phần mềm

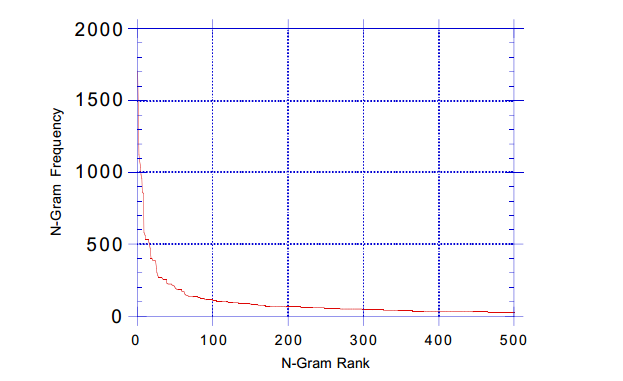
Hoặc đơn giản ‘term’ chỉ bao gồm các kí tự được tách ra từ một từ, ví dụ từ “TEXT” có thể được tách ra thành:

* bi-gram : \_T, TE, EX, XT, T\_
* tri-grams: \_TE, TEX, EXT, XT\_, T\_ \_
* quad-grams: \_TEX, TEXT, EXT\_, XT\_ \_, T\_ \_ \_

Nhưng trong bài viết này ‘term’ hoặc n-gram được dùng để chỉ các từ được cắt ra từ một chuỗi trong văn bản. Phương pháp N-gram đã có một số thành công nhất định trong việc áp dụng giải quyết nhiễu với các bộ kí tự ASCII trong các lĩnh vực vấn đề khác nhau, chẳng hạn xử lý nhiễu trong địa chỉ bưu thiếp, trích xuất thông tin văn bản, và được áp dụng trong việc xử lý ngôn ngữ tự nhiên. Điểm mạnh của phương pháp n-gram xuất phát từ bản chất của nó : chia nhỏ các thành phần trong một chuỗi ra thành các phần nhỏ như thế thì thay vì chịu ảnh hưởng từ nhiễu trong chuỗi, các phần nguyên vẹn được tách ra để xử lý. Nếu ta áp dụng phương pháp n-gram cho cả hai văn bản cần so sánh thì ta có thể dễ dàng ước lượng động tương đồng giữa hai văn bản đó mà không sợ sự ảnh hưởng của các phần nhiễu trong văn bản.

* + 1. **Mức độ quan trọng của n-gram**

Trong ngôn ngữ luôn có những từ xảy ra thường xuyên với các từ khác. Và một trong những phát biểu nổi tiếng nhất Zipf’s law phát biểu về vấn đề này như sau: “tần số xuất hiện của từ nào mà càng lớn tỉ lệ nghịch với mức độ quan trọng của từ đó”.



Hình : Mối quan hệ giữa tấn suất và mức độ quan trọng của n-gram [7]

Có nghĩa là luôn có một tập các từ mà tần số xuất hiện, sử dụng nhiều hơn các từ khác, điều này đúng trong bất kì ngôn ngữ nào. Chính vì vậy ta cần có một phương pháp để làm mịn đường cong tần số trên hay là việc cân bằng mức độ quan trọng giữa các từ

Đặc biệt nó cũng đúng với mức độ quan trọng của n-gram. Tuy nhiên có một phương pháp Tf-idf weighting để có thể xem xét được mức độ quan trọng của từ đó trong một văn bản, mà bản thân văn bản đang xét trong một tập hợp các văn bản

Tf- term frequency : dùng để ước lượng tần xuất xuất hiện của từ trong văn bản. Tuy nhiên với mỗi văn bản thì có độ dài khác nhau, vì thế số lần xuất hiện của từ có thể nhiều hơn. Vì vậy số lần xuất hiện của từ sẽ được chia độ dài của văn bản (tổng số từ trong văn bản đó)

TF(t,d) = (số lần từ t xuất hiện trong văn bản d) / (tổng số từ trong văn bản d).

**IDF- Inverse Document Frequency: dùng để ước lượng mức độ quan trọng của từ đó như thế nào. Khi tính tần số xuất hiện tf thì các từ đều được coi là quan trọng như nhau. Tuy nhiên có một số từ thường được được sử dụng nhiều nhưng không quan trọng để thể hiện ý nghĩa của đoạn văn, ví dụ :**

|  |  |
| --- | --- |
| **Loại từ** | **Ví dụ** |
| **Từ đếm** | **Một, hai ,ba …** |
| **Số đếm** | **1,2,3…** |
| **Từ phủ định** | **Không, ngược lại** |
| **Từ nối** | Và, nhưng, tuy nhiên, vì thế, vì vậy, … |
| Giới từ | Ở, trong, trên, … |
| Từ chỉ định | Ấy, đó, nhỉ, … |

Bảng : Ví dụ về một số từ xuất hiện nhiều trong cả văn nói và viết

**Vì vậy ta cần giảm đi mức độ quan trọng của những từ đó bằng cách tính IDF của gram đó:**

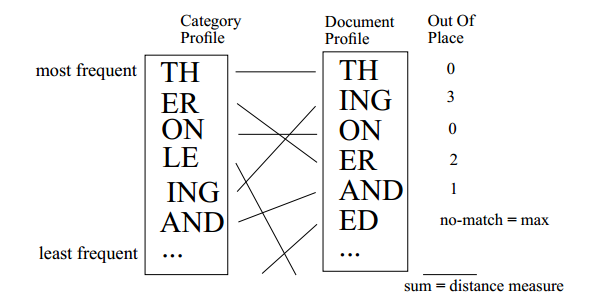
IDF(t,D) = log\_e(Tổng số văn bản trong tập mẫu D/ Số văn bản có chứa từ t )

Giá trị của TF-IDF: sẽ được tính bằng cách nhân tần xuất của gram đó nhân với trọng số tf-idf của gram đó:

Tfidf(t,d,D) =tf(t,d) \* idf(t,D)

* + 1. **Ước lượng độ tương đồng**

Để thực hiện ước lượng độ tương đồng giữa hai văn bản bằng phương pháp Out of place ta thực hiện như sau. Sau khi tách các văn bản thành n-gram sau đó thực hiện sắp xếp thứ tự theo mức độ quan trọng của các gram của mỗi văn bản. Phương pháp Out of place sẽ xác định được khoảng cách từ gram trong văn bản này với gram trong văn bản kia. Sau đó ta tính tổng khoảng cách của gram trong văn bản cần ước lượng để tìm ra khoảng cách của văn bản đó với văn bản mẫu. Hình 7 đưa ra ví dụ về cách tính khoảng cách của một vài gram.

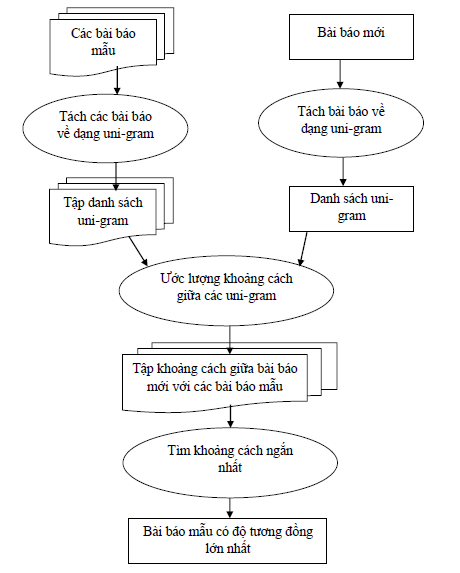


Hình : Cách tính khoảng cách giữa hai văn bản dựa trên Out Of Place [7]

Với mỗi n-gram trong văn bản mẫu ta sẽ tìm các n-gram tương ứng trong văn bản cần so sánh và tính khoảng cách về vị trí củng chúng. Ví dụ n-gram “ING” đứng ở vị trí thứ 2 (mức độ quan trọng thứ hai) trong danh sách n-gram của văn bản cần so sánh nhưng lại đứng thứ 5 trong dang sách n-gram của văn bản mẫu. Vì vậy khoảng cách n-gram “ING” của văn bản cần so sánh với mẫu là 3. Trong trường hợp n-gram “ED” (như trong hình) không nằm trong văn bản mẫu, thì khoảng cách của nó se là khoảng cách của lớn nhất của n-gram trong văn bản cần so sánh với văn bản mẫu. Tổng khoảng cách của tất cả các gram là khoảng cách của văn bản cần so sánh với văn bản mẫu. Cuối cùng trong tập hợp các văn bản ta lấy văn bản có khoảng cách với văn bản mẫu là nhỏ nhất có nghĩa là văn bản đó có độ tương đồng lớn nhất với văn bản mẫu.

* + 1. **Các bước thực hiện**

Dựa trên các phương pháp được giới thiệu ở trên để ước lượng độ tương đồng giữa các văn bản, thì hệ thống sẽ thực hiện theo các bước sau (hình 8)



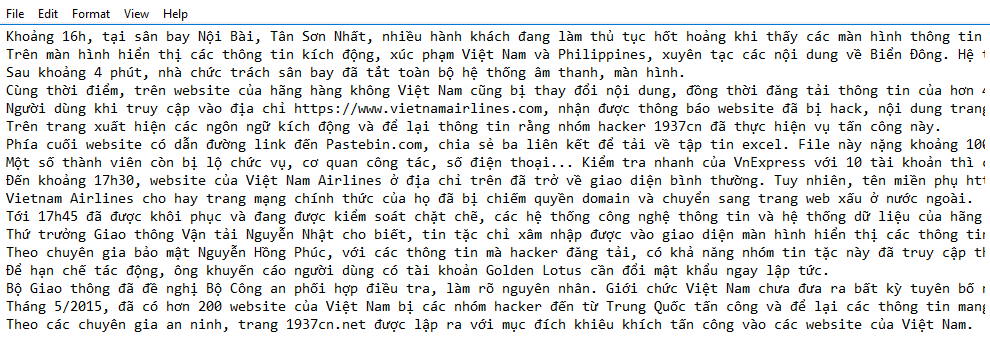
Hình : Các bước thực hiện của phương pháp Out Of Place [7]

* Tách văn bản ra thành các token tách biệt chỉ chứa từ. Số và dấu chấm câu, các kí tự đặc biệt đều bị loại
* Kiểm tra với mỗi token, sinh ra tất cả các n-gram có thể với n có thể là từ 1 đến 5.
* Với mỗi gram ta sẽ gán cùng một bộ đếm, với bộ đếm đó ta đảm bảo được rằng các gram giống nhau sẽ không xuất hiện cùng trong một danh sách mà thay vào đó bộ đếm sẽ thực hiện số gram xuất hiện trong văn bản
* Sau khi tách xong ta sẽ có đầu ra là danh sách các gram và số lần xuất hiện trong văn bản
* Áp dụng thuật toán tf-idf để tính mức độ quan trọng của gram với mỗi văn bản tương ứng. Sau khi tính xong thì sắp xếp các gram theo điểm tf-idf tương ứng với mỗi gram theo thứ thự giảm dần

Sau khi ta có danh sách n-gram của cả văn bản mẫu, văn bản so sánh được sắp xếp theo mức độ quan trọng của n-gram. Áp dụng thuật toán Out of place để tính khoảng cách của văn bản so sánh với văn bản mẫu sau. Cuối cùng văn bản có khoảng cách nhỏ nhất với văn bản mẫu, là văn bản có độ tương đồng lớn nhất với văn bản mẫu so với các văn bản khác.

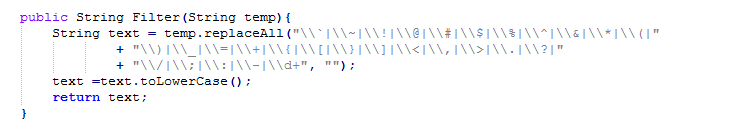
* 1. **Chi tiết thuật toán**
     1. **Đưa văn bản về dạng uni-gram**

Bước 1: Sử dụng thư viện Jsoup trong java để lấy nội dụng của một bài báo báo điện tử vnExpress với tiêu đề :”Sân bay Nội Bài, Tân Sơn Nhất bị tin tặc tấn công”. Nội dung của bài báo sẽ được lấy về vào lưu vào trong file text như hình 9 mô tả dưới đây



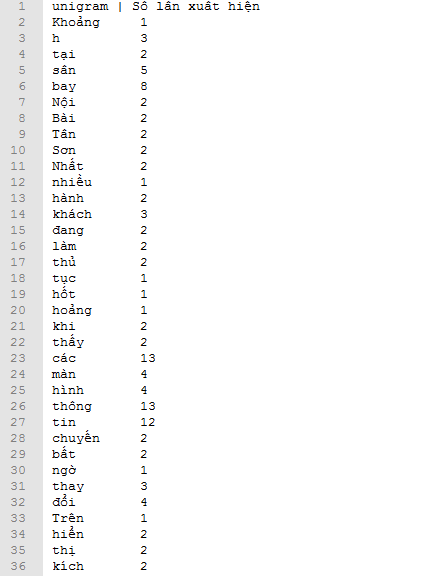
Hình : Nội dung của bài báo “Sân bay Nội Bài, Tân Sơn Nhất bị tin tặc tấn công”

Bước 2: Loại bỏ số và các kí tự đặc biệt trong văn bản thu được bằng cách sử dụng kĩ thuật Regex trong java với đoạn code như sau:



Hình : Các kí tự đặc biệt sẽ bị loại bỏ trong văn bản

Bước 3 : Sau khi văn bản đã được loại bỏ toàn bộ số, dấu chấm, hay các kí tự đặc biệt khác. Ta sẽ tách văn bản thành dạng uni-gram (1-gram). Mỗi uni-gram sẽ chứa duy nhất 1 từ trong văn bản được thể hiện như mô tả dưới đây



Hình : N-gram và số lần xuất hiện trong bài báo

Hình trên thể hiện kết quả thu được sau khi tách văn bản thành các uni-gram. Với các uni-gram được đảm bảo là duy nhất và số lần các gram tương ứng xuất hiện trong bài báo

* + 1. **Đánh giá mức độ quan trọng của uni-gram bằng tf-idf**

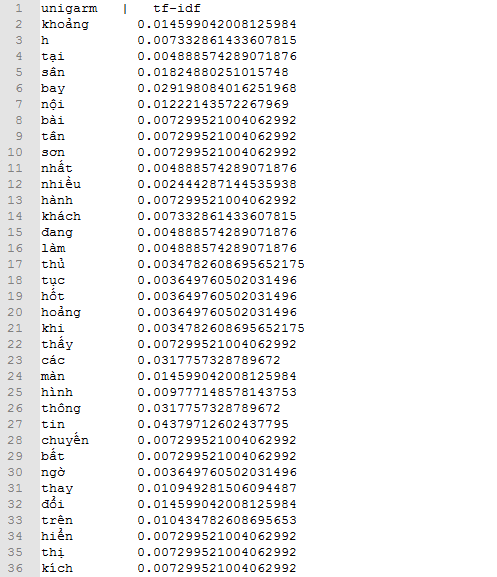
Bước 1: Sau khi ta phân tách bài báo ta thu được dang sách các uni-gram và số lần xuất hiện của mỗi uni-gram đó trong bài báo. Từ đó ta có thể tính được tần suất xuất hiện của các gram đó trong văn bản bằng công thức sau:

TF(t,d) = (số lần từ t xuất hiện trong văn bản d) / (tổng số từ trong văn bản d).

Bước 2: Sau bước ta sẽ tính được tuần số xuất hiện của mỗi gram trong bài báo tương ứng, tuy nhiên trong mọi ngôn ngữ luôn những xuất hiên những từ được gọi được là stop word (là những từ thường xuyên xuất hiện, tuy nhiên không đem lại ý nghĩa cho nội dung của bài báo). Do đó ta cần phải giảm độ quan trọng của những từ đó. Điều này đúng với cả các uni-gram trong danh sách vừa tạo ở phía trên. Bởi vậy ta sẽ áp dụng công thức sau để tăng hoặc giảm mức độ quan trong của các gram

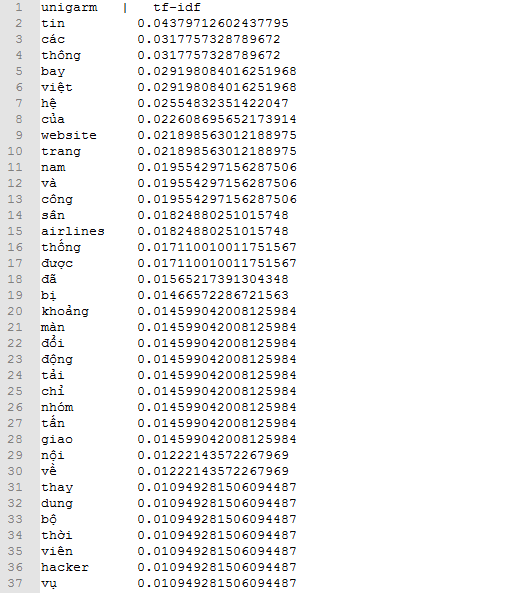
IDF(t,D) = log\_e(Tổng số văn bản trong tập mẫu D/ Số văn bản có chứa từ t )

Bước 3 : Với mỗi gram ta nhân giá trí tf với idf vừa tìm được, như vậy ứng với mỗi gram sẽ có một giá trị tf-idf thể hiện mức độ quan trọng của các gram trong văn bản



Hình : N-gram và trọng số tf-idf

Bước 4: Cuối cùng ta sẽ sắp xếp danh sách các n-gram từ thứ tự độ quan trọng giảm dần. Có nghĩa là ta sẽ để những gram nắm giữ thông tin quan trọng nhất ở đầu các gram đó có thể coi là các key-word của bài báo đó. Hình dưới đấy mô tả kết quả.



Hình : N-gram sẽ được sắp xếp theo thứ tự giảm dần theo trọng số tf-idf

* + 1. **So sánh khoảng cách của các gram**

Sau khi thực hiện xong việc phân tích bài báo thành các gram và sắp xếp các danh sách n-gram theo mức độ quan trọng của các gram. Bước tiếp theo ta sẽ thực hiện ước lượng độ tương đồng của hai bài báo khác với bài báo mẫu ta vừa xử lý ở các bước trên.

**Bài báo mẫu**: Sân bay Nội Bài, Tân Sơn Nhất bị tin tặc tấn công (nguồn : vnexpress.net)

**Bài báo 1** : Hệ thống thông tin sân bay tại Việt Nam bị tấn công (nguồn : dantri.com)

**Bài báo 2** : ĐT futsal Việt Nam muốn giành quyền tự quyết (nguồn : laodong.com.vn)

Đối với hai bài báo số 1,2 ta áp dụng tương tự các bước trên để tạo ra danh sách các n-gram tương ứng. Kết quả sau khi thực hiện tương tự các bước trên với 2 bài báo

|  |  |
| --- | --- |
|  |  |
| Bài báo : Hệ thống thông tin sân bay tại Việt Nam bị tấn công | Bài báo : ĐT futsal Việt Nam muốn giành quyền tự quyết |

Hình : So sánh n-gram của hai báo báo

Tiếp theo áp dụng thuật toán Out of place để ước lượng độ tương đồng của bài báo nào giống với bài báo mẫu hơn. Thuật toán được áp dụng như sau :

So sánh bài báo mẫu với bài báo số 1 :

|  |  |  |  |
| --- | --- | --- | --- |
| Uni-gram của cả hai bài báo | Vị trí trong bài báo mẫu | Vị trí trong bài báo số 1 | Khoảng cách giữa các gram |
| “thông” | 3 | 3 | 0 |
| “tin” | 1 | 2 | 1 |
| “bay” | 4 | 1 | 3 |
| “airlines” | 14 | 20 | 6 |
| “sân” | 13 | 26 | 13 |
| “website” | 8 | Không có | 307 |
| “chữ” | Không có | 28 | 307 |
| \*Khoảng cách lớn nhất giữa bài báo mẫu với bài báo số 1 : 307 | | | |

Bảng : Ví dụ về khoảng cách giữa các gram trong bài báo 1 với bài báo mẫu

Độ tương đồng của mẫu với bài báo số 1với bài báo mẫu sẽ được tính như sau :

%Similarity =

%Similarity (văn bản 1 vs văn bản mẫu) = 41.9%

So sánh bài báo mẫu với bài báo số 2 :

|  |  |  |  |
| --- | --- | --- | --- |
| Uni-gram của cả hai bài báo | Vị trí trong bài báo mẫu | Vị trí trong bài báo số 2 | Khoảng cách giữa các gram |
| “và” | 11 | 16 | 5 |
| “không” | 51 | 33 | 18 |
| “việt” | 5 | 149 | 145 |
| “nam” | 10 | 150 | 140 |
| “những” | 129 | 189 | 60 |
| “website” | 8 | Không có | 270 |
| “đtvn” | Không có | 1 | 270 |
| \*Khoảng cách lớn nhất giữa bài báo mẫu với bài báo số 2 : 270 | | | |

Bảng : Ví dụ về khoảng cách giữa các gram trong bài báo 2 với bài báo mẫu

Độ tương đồng của mẫu với bài báo số 2 với bài báo mẫu sẽ được tính như sau:

%Similarity =

%Similarity (văn bản 2 vs văn bản mẫu) = 18.3%

Kết luận : Từ kết quả phần trăm về độ tương đồng giữa bài báo mẫu với hai bài báo thử nghiệm trên cho thấy rằng. Bài báo số 1(41.9%) tương đồng với mẫu nhiều hơn so với bài báo số 2(18.3%)

* 1. **Kết luận**

Thuật toán Out of place cung cấp một cách thức đơn giản nhưng có hiệu quả cao để ước lượng độ tương đồng giữa các văn bản. Bởi vì thuật toán Out of place ước lượng dựa trên các khoảng cách giữa các gram và mức độ quan trọng của gram trong văn bản so với các thuật toán khác phải dùng đến các phương pháp phức tạp và tốn kém ví dụ như phân tích ngôn ngữ tự nhiên ,xây dựng các bộ từ điển. Ngoài ra các tiếp cận này có thể giải quyết một số vấn đề ví dụ như việc ước lượng độ tương đồng giữa các văn bản, mà trong văn bản có thể có các từ nhiễu, hay xử lý các câu từ đặc biệt vì bản chất của thuật toán này việc xử lý thống kê số lần xuất hiện của các gram và tính mức độ quan trọng của chúng.

Tiếp theo chương 3 của đồ án sẽ thực hiện kiểm nghiệm độ chính xác của thuật toán out of place bằng cách áp dụng thuật toán vào các mẫu được xây dựng trên các bài báo tiếng việt của 10 trang báo mạng uy tín tại Việt Nam.

**CHƯƠNG 3 : ĐÁNH GIÁ THUẬT TOÁN**

Dưới đây là những vấn đề chương 3 sẽ trình bày:

* Xây dựng bộ dữ liệu
* Kịch bản kiểm nghiệm
* Kết quả kiểm nghiệm
  1. **Xây dựng bộ dữ liệu**

Trong đồ án này tập vào giải quyết bài toán tính toán độ tương đồng giữa các bài báo. Từ việc kiểm tra độ tương đồng có thể ứng dụng vào gợi ý cho người đọc báo các bài báo có chủ để tương tự, liên quan đến bài báo đã được đọc trước đó. Bằng cách áp dụng phương pháp out of place để ước lượng độ tương đồng giữa các bài báo. Tuy nhiên phương pháp out of place lúc đầu chỉ áp dụng trực tiếp với các từ và dựa số lần xuất hiện trong văn bản để tính toán nên độ chính xác chưa cao chính vì thế trong đồ án này ngoài áp dụng out of place thuần còn kết hợp việc phân tách văn bản ra thành n-gram và tính toán trọng số tf-idf cho các gram đó như vậy các từ là từ khóa của bài báo sẽ được ưu tiên trong việc ước lượng độ tương đồng.

Để kiểm nghiệm độ chính xác của phương pháp out of place được cải tiến, thuật toán out of place sẽ được áp dụng để ước lượng độ tương đồng trong các mẫu với mỗi mẫu bao gồm: 3 bài báo được lấy từ 10 trang báo điện tử uy tin của Việt Nam như sau :

|  |  |  |
| --- | --- | --- |
| Các trang báo điện tử | Đường dẫn trang báo điện tử | Đường dẫn rss (Really Simple Syndication) |
| Vnexpress | http://vnexpress.net/ | http://vnexpress.net/rss |
| Dân trí | http://dantri.com.vn/ | http://dantri.com.vn/rss.htm |
| Lao động | http://laodong.com.vn/ | http://laodong.com.vn/rss |
| Tuổi trẻ online | http://tuoitre.vn/ | http://tuoitre.vn/rss |
| Tiền phong | http://www.tienphong.vn/ | http://www.tienphong.vn/rss.aspx |
| Công an nhân dân | http://cand.com.vn/ | http://cand.com.vn/rss/ |
| Đời sống và pháp luật | http://www.doisongphapluat.com/ | http://www.doisongphapluat.com/rss.html |
| VietnamNet | http://vietnamnet.vn/ | http://vietnamnet.vn/vn/rss/ |
| Người đưa tin | http://www.nguoiduatin.vn/ | http://www.nguoiduatin.vn/rss.html |
| Giáo dục thời đại | http://giaoducthoidai.vn/ | http://giaoducthoidai.vn/rss.html |

*Đường dẫn RSS (viết tắt từ Really Simple Syndication)* : là một tiêu chuẩn định dạng tài liệu dựa trên XML nhằm giúp người sử dụng dễ dàng cập nhật và tra cứu thông tin một cách nhanh chóng và thuận tiện nhất bằng cách tóm lược thông tin vào trong một đoạn dữ liệu ngắn gọn, hợp chuẩn.

Trong ba bài báo để tạo bộ dữ liệu kiểm nghiệm sẽ có một bài báo làm bài mẫu để so sánh với hai bài báo còn lại, ngoài ra các mẫu sẽ có các yêu cầu như sau:

* Trong 2 bài báo so sánh sẽ có một bài có nội dung, chủ để tương tự với bài mẫu, bài báo còn lại làm một chủ đề khác biệt hoàn toàn so với bài báo mẫu
* Sẽ có các mẫu mà có một bài báo có nội dung, chủ đề tương tự với bài báo mẫu, bài báo có còn lại có cùng chủ để với bài báo mẫu nhưng lại mô tả sự việc khác. Ví dụ cùng bài báo nói về sân bay Tân Sơn Nhất tuy nhiên có bài nói về việc sân bay bị tấn công hệ thống thông tin, có bài viết về sân bay bị ngập lụt

Các công cụ phần mềm để xây dựng bộ dữ liệu:

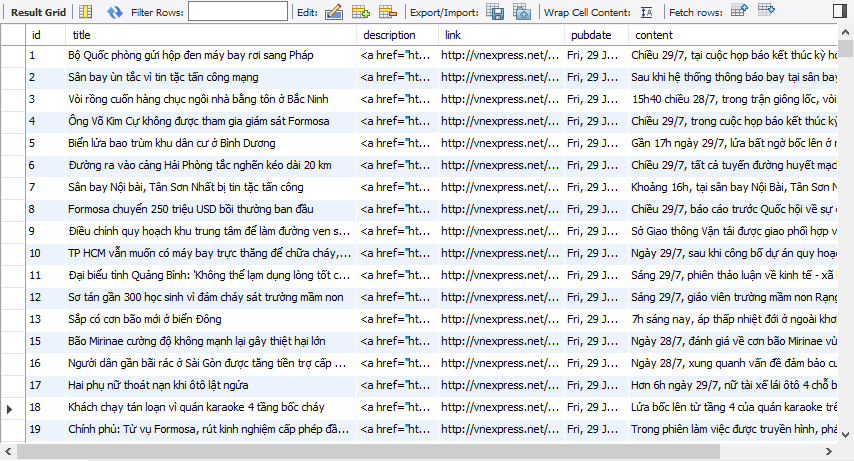
|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên phần mềm | Tác giả | Nguồn |
| 1 | MySQL | Oracle | https://www.mysql.com/ |
| 2 | Jsoup |  | https://jsoup.org/ |
| 3 | [IntelliJ IDEA](https://www.jetbrains.com/idea/) | JetBrains | https://www.jetbrains.com |

Bảng : Các công cụ phần mềm được sử dụng để lấy mẫu

Dữ liệu sẽ được xây dựng 1000 mẫu, trong mỗi mẫu sẽ có 3 bài báo được lấy từ 4 tờ báo uy tín của Việt Nam. Chi tiết thu thập bộ dữ liệu

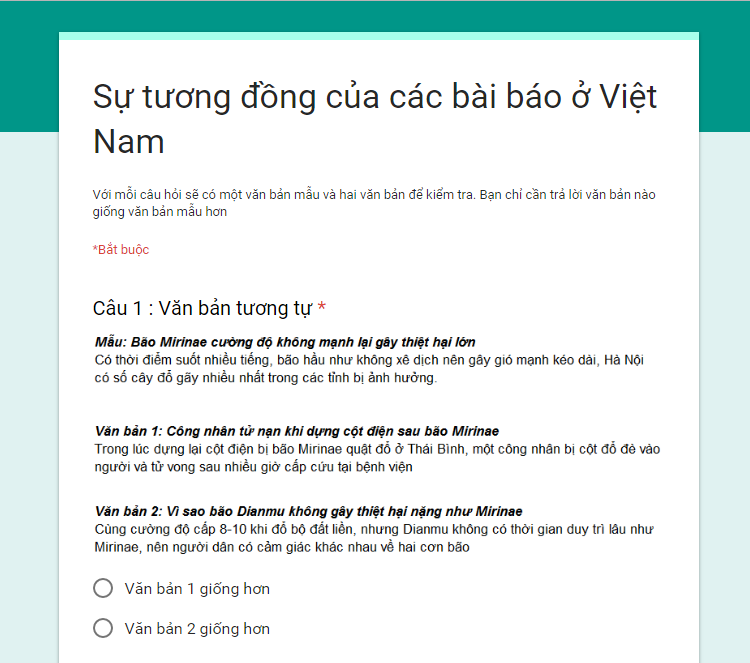
Bước 1: Dựa vào rss của các tờ báo bao gồm các thông tin: tiêu đề, mô tả, ngày phát hành, đường link của tờ báo đó. Sau khi tiến hành đọc rss đó và lấy nội dung thì các thông tin đó sẽ được lưu trong cơ sở dữ liệu mysql

Bước 2: Với link của các tờ báo được đọc từ các rss, sử dụng jsoup là một thư viện mã nguồn mở của java để tiến hành phân tích là loại bỏ (praser) dữ liệu nhiễu của tờ báo, không phải nội dung của tờ báo. Sau khi lấy được nội dung của bài báo, nội dung đó sẽ được cập nhật vào lưu vào trong cơ sở dữ liệu



Hình : Cơ sở dữ liệu chứa thông tin của các bài báo

Bước 3: Dựa vào các bài báo và nội dung thu thập được từ các trang báo tin tức . Thực hiện xây dựng các phiếu khảo sát bao gồm 1000 mẫu đã được phân loại từ trước. Vỡi mỗi phiếu bao gồm 20 mẫu được đưa vào 50 phiếu để người đọc có thể khảo sát .



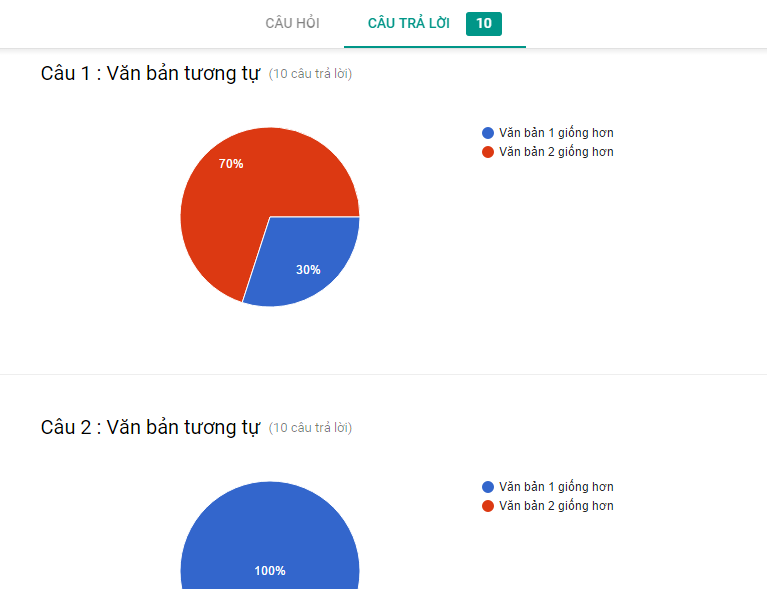
Hình : Phiếu khảo sát người đọc ước lượng độ tương đồng giữa các bài báo

* 1. **Kịch bản kiểm nghiệm**

Để có thể kiểm nghiệm độ chính xác của thuật toán out of place trong việc ước lượng độ tương đồng giữa các văn bản, sẽ dựa vào các bước sau :

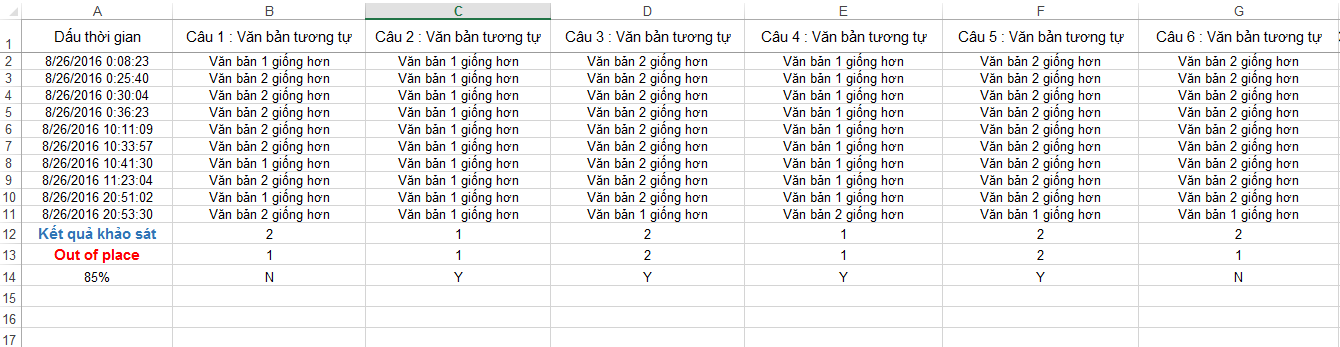
Bước 1: Từ 6000 bài báo được thu thập từ 10 trang báo điện tử uy tín ở Việt Nam được lưu trong cơ sở dữ liệu. Các bài báo đó sẽ được chọn vào đưa vào các mẫu để người đọc khảo sát, với 1000 mẫu trong mỗi mẫu sẽ có những mẫu ở mức độ khó dễ khác nhau.Cuối cùng các mẫu sẽ được chia đều vào 50 phiếu khảo sát(các phiếu khảo sát sẽ được tạo trên google form).

Bước 2: 50 phiếu khảo sát sẽ được từ 8 -10 bạn đọc khảo sát để trả lời 20 câu hỏi(20 mẫu): trong 2 bài báo so sánh với bài báo mẫu bài báo nào có độ tương đồng với bài báo mẫu lớn hơn so với bài báo còn lại. Kết quả khảo sát của bạn đọc sẽ được google form tổng hợp và thống kê lại như hình 17



Hình : Kết quả khảo sát người đọc ước lượng độ tương đồng giữa các bài báo

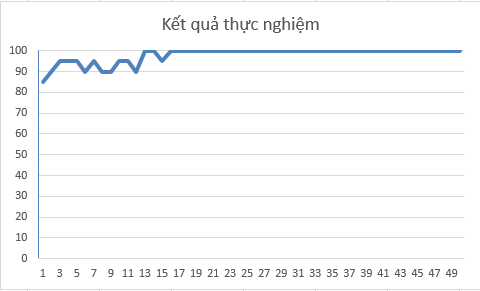
Bước 3: Áp dụng thuật toán out of place vào 1000 mẫu vừa được khảo sát với mỗi mẫu sẽ trả về kết quả bài báo nào giống hơn sau đó lấy kết quả chạy thuật toán so sánh với kết quả từ phiếu khảo sát vừa thu thập được. Độ chính xác của thuật toán phụ thuộc vào kết quả chạy thuật toán trùng với ý kiến số đông của những người khảo sát



Hình : Kết quả đánh giá độ chính xác của thuật toán

* 1. **Kết quả thực hiện**

Sau khi thực hiện lấy kết quả khảo sát của 1000 mẫu(50 phiếu khảo sát) và chạy thuật toán out of plae để kiểm nghiệm kết quả. Độ chính xác của thuật toán :



Hình : Kết quả thực nghiệm của 50 phiếu khảo sát

* 1. **Đánh giá**

Từ kết quả thực nghiệm của 1000 mẫu cho thấy rằng thuật toán ước lượng độ tương đồng out of place cho kết quả khả quan để ước lượng độ tương đồng giữa hai văn bản. Độ chính xác là 95% đến 100% là đúng với các mẫu đơn giản. Và độ chính xác của thuật toán chưa thể đúng hoàn toàn là do ở các mẫu phức tạp, thuật toán chưa đưa ra kết quả chính xác được bài báo nào giống với bài mẫu hơn. Nên kết quả của những mẫu đấy thông thường có độ chính xác là từ 85% đến 90%.Do đó thuật toán vẫn cần phải cải thiện để nâng cao độ chính xác để đảm bảo với 1000 mẫu khảo sát đều là mẫu phức tạp mà vẫn đảm bảo độ chính xác lớn hơn 95%.

CHƯƠNG 4 : ỨNG DỤNG

* 1. Giới thiệu ứng dụng
  2. Kiến trúc hệ thống
  3. Kết quả cài đặt

KẾT LUẬN

TÀI LIỆU THAM KHẢO

[1] Aly A. Fahmy, Wael H. Gomaa. A Survey of Text Similarity Approaches. In International Journal of Computer Applications (0975 – 8887), April 2013.

[2] G. Salton and Edward Fox and Wu Harry Wu. "Extended Boolean information retrieval". Communications of the ACM, 26 (11). 1983.

[3] G. Salton and M. J. McGill. "Introduction to modern information retrieval". 1983

[4] G. Salton and C. Buckley. "Term-weighting approaches in automatic text retrieval". Information Processing & Management, 24 (5). 1988.

[5] H. Wu and R. Luk and K. Wong and K. Kwok. "Interpreting TF-IDF term weights as making relevance decisions". ACM Transactions on Information Systems, 26 (3). 2008.

[6] K. Sparck Jones. "A statistical interpretation of term specificity and its application in retrieval". Journal of Documentation, 28 (1). 1972.

[7] William B. Cavnar, John M. Trenkle. N-Gram-Based Text Categorization.   
In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994