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

**KHOA CÔNG NGHỆ THÔNG TIN I**

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



**ĐỒ ÁN**

**TỐT NGHIỆP ĐẠI HỌC**

|  |
| --- |
| **Đề tài:**  **ÁP DỤNG 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 CẢM ƠN**

Trước tiên, em muốn gửi lời cảm ơn sâu sắc nhất đến giảng viên TS. Nguyễn Mạnh Hùng đã tận tình hướng dẫn em trong suốt quá trình nghiên cứu khoa học và thực hiện đồ án tốt nghiệp. Em cám ơn thầy vì trong suốt thời gian nghiên cứu thầy đã tận tình chỉ bảo và đưa ra những góp ý để em có thể hoàn thành đồ án tốt nghiệp.

Em xin chân thành cảm ơn các thầy, cô giáo đã giảng dạy em trong bốn năm học qua, những kiến thức mà em nhận được trên giảng đường đại học sẽ là hành trang giúp em vững bước trong tương lai.

Em cũng xin gửi lời cám ơn tới các bạn lớp D12CN5 đã ủng hộ và giúp đỡ em trong suốt quá trình học tập tại Học viện. Đặc biệt, xin cám ơn những người bạn đã cùng mình học tập, nghiên cứu trong suốt chặng đường vừa qua.

Cuối cùng, em muốn gửi lời cám ơn chân thành đến tất cả bạn bè đã giúp đỡ trong quá trình khảo sát và thực hiện đồ án. Đặc biệt là bố mẹ và em gái – những người thân yêu luôn kịp thời động viên và giúp đỡ em vượt qua những khó khăn trong học tập cũng như trong cuộc sống.

Hà Nội, ngày 10 tháng 12 năm 2016

Sinh viên

**Phạm Hoàng Dương**

**MỤC LỤC**

[**DANH SÁCH CÁC BẢNG** iv](#_Toc469155721)

[**DANH SÁCH CÁC HÌNH VẼ** iv](#_Toc469155722)

[**DANH MỤC THUẬT NGỮ** vi](#_Toc469155723)

[**MỞ ĐẦU** 1](#_Toc469155724)

[**CHƯƠNG 1: GIỚI THIỆU VỀ BÀI TOÁN ƯỚC LƯỢNG ĐỘ TƯƠNG ĐỒNG GIỮA CÁC VĂN BẢN** 3](#_Toc469155725)

[1.1 Bài toán ước lượng độ tương đồng giữa các văn bản 3](#_Toc469155726)

[1.1.1 Giới thiệu 3](#_Toc469155727)

[1.1.2 Ứng dụng 4](#_Toc469155728)

[1.2 Các phương pháp ước lượng độ tương đồng giữa các văn bản 6](#_Toc469155729)

[1.2.1 Phương pháp ước lượng độ tương đồng dựa trên thống kê 6](#_Toc469155730)

[1.2.2 Phương pháp ước lượng độ tương đồng về ngữ nghĩa 9](#_Toc469155731)

[1.3 Đề xuất ứng dụng 15](#_Toc469155732)

[1.4 Kết luận 17](#_Toc469155733)

[**CHƯƠNG 2: ƯỚC LƯỢNG ĐỘ TƯƠNG ĐỒNG GIỮA CÁC VĂN BẢN BẰNG PHƯƠNG PHÁP OUT OF PLACE** 18](#_Toc469155734)

[2.1 Giới thiệu về thuật toán out of place 18](#_Toc469155735)

[2.1.1 Phương pháp tiếp cận dựa trên n-gram 18](#_Toc469155736)

[2.1.2 Mức độ quan trọng của n-gram 19](#_Toc469155737)

[2.1.3 Ước lượng độ tương đồng 21](#_Toc469155738)

[2.1.4 Các bước thực hiện 22](#_Toc469155739)

[2.2 Chi tiết thuật toán 24](#_Toc469155740)

[2.2.1 Đưa văn bản về dạng uni-gram 24](#_Toc469155741)

[2.2.2 Đánh giá mức độ quan trọng của uni-gram bằng tf-idf 26](#_Toc469155742)

[2.2.3 So sánh khoảng cách của các gram 27](#_Toc469155743)

[2.3 Kết luận 31](#_Toc469155744)

[**CHƯƠNG 3: ĐÁNH GIÁ THUẬT TOÁN** 32](#_Toc469155745)

[3.1 Xây dựng bộ dữ liệu 32](#_Toc469155746)

[3.2 Kịch bản đánh giá thuật toán 35](#_Toc469155747)

[3.3 Kết quả đánh giá thuật toán 37](#_Toc469155748)

[3.4 Đánh giá 37](#_Toc469155749)

[3.5 Kết luận 38](#_Toc469155750)

[**CHƯƠNG 4: ỨNG DỤNG KIỂM TRA ĐẠO VĂN** 39](#_Toc469155751)

[4.1 Giới thiệu ứng dụng 39](#_Toc469155752)

[4.2 Kiến trúc ứng dụng 40](#_Toc469155753)

[4.3 Kết quả cài đặt 43](#_Toc469155754)

[4.4 Kết luận 46](#_Toc469155755)

[**KẾT LUẬN** 47](#_Toc469155756)

[**TÀI LIỆU THAM KHẢO** 49](#_Toc469155757)

**DANH SÁCH CÁC BẢNG**

[Bảng 2.1: Ví dụ về một số từ xuất hiện nhiều trong cả văn nói và viết 21](#_Toc468493027)

[Bảng 2.2: Ví dụ về khoảng cách giữa các gram trong văn bản 1 với bài báo mẫu 29](#_Toc468493028)

[Bảng 2.3: Ví dụ về khoảng cách giữa các gram trong văn bản 2 với văn bản mẫu 30](#_Toc468493029)

[Bảng 3.1: Địa chỉ của các trang báo điện tử ở Việt Nam 32](#_Toc468493030)

[Bảng 3.2: Các công cụ phần mềm được sử dụng để lấy mẫu 33](#_Toc468493031)

[Bảng 3.3: Chi tiết kết quả các mẫu 37](#_Toc468493032)

[Bảng 4.1: Các công cụ và thư viện để xây dựng ứng dụng 40](#_Toc468493033)

**DANH SÁCH CÁC HÌNH VẼ**

[Hình 1.1: Ảnh chụp của 2 bài báo từ 2 trang web khác nhau 5](#_Toc468527565)

[Hình 1.2: Các phương pháp ước lượng độ tương đồng dựa trên kí tự[15] 7](#_Toc468527566)

[Hình 1.3: Các phương pháp ước lượng độ tương đồng dựa trên kho dữ liệu[15] 10](#_Toc468527567)

[Hình 1.4: Mối quan hệ ngữ nghĩa giữa hai từ thông qua các nút trong WordNet 13](#_Toc468527568)

[Hình 1.5: Các phương pháp ước lượng độ tương đồng dựa trên tri thức[15] 14](#_Toc468527569)

[Hình 2.1: Mối quan hệ giữa tấn suất và mức độ quan trọng của n-gram [3] 20](#_Toc468527570)

[Hình 2.2: Cách tính khoảng cách giữa hai văn bản dựa trên Out Of Place [3] 22](#_Toc468527571)

[Hình 2.3: Các bước thực hiện của phương pháp Out Of Place [3] 23](#_Toc468527572)

[Hình 2.4: Nội dung của bài báo làm ví dụ cho bước 1 24](#_Toc468527573)

[Hình 2.5: Các kí tự đặc biệt sẽ bị loại bỏ trong văn bản 24](#_Toc468527574)

[Hình 2.6: N-gram và số lần xuất hiện trong bài báo 25](#_Toc468527575)

[Hình 2.7: n-gram và trọng số tf-idf 27](#_Toc468527576)

[Hình 2.8: So sánh n-gram của văn bản mẫu và văn bản 1 28](#_Toc468527577)

[Hình 2.9: So sánh n-gram của văn bản mẫu và văn bản 2 29](#_Toc468527577)

[Hình 3.1: Cơ sở dữ liệu chứa thông tin của các bài báo 34](#_Toc468527578)

[Hình 3.2: Phiếu khảo sát người đọc ước lượng độ tương đồng giữa các bài báo 35](#_Toc468527579)

[Hình 3.3: Kết quả đánh giá độ chính xác của thuật toán 36](#_Toc468527581)

[Hình 4.1:Mô hình kiến trúc ứng dụng 41](#_Toc468527583)

[Hình 4.2: Giao diện chính khi mở ứng dụng 43](#_Toc468527584)

[Hình 4.3: Giao diện ứng dụng sau khi thêm các file cần kiểm tra sự tương đồng 44](#_Toc468527585)

[Hình 4.4:Kết quả sau khi ước lượng độ tương đồng giữa hai văn bản 44](#_Toc468527586)

[Hình 4.5:Giao diện hiển thị nội dung hai tờ báo thông qua link được thêm trong ứng dụng 45](#_Toc468527587)

[Hình 4.6: Kết quả sau khi ước lượng độ tương đồng giữa hai tờ báo điện tử 46](#_Toc468527588)

**DANH MỤC THUẬT NGỮ**

|  |  |
| --- | --- |
| **Từ** | **Từ đầy đủ** |
| tf-idf | Term frequency – inverse document frequency: thể hiện mức độ quan trọng của từ này trong một văn bản, mà bản thân văn bản đang xét nằm trong một tập hợp các văn bản |
| rss | Really simple syndication: là một định dạng tập tin thuộc họ XML dùng trong việc chia sẻ tin tức Web được dùng bởi website tin tức |
| user | Người dùng |
| n-gram | Là tần suất xuất hiện của n kí tự (hoặc từ) liên tiếp nhau trong dữ liệu |
| csdl | Cơ sở dữ liệu |

**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 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 trị, [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 tương đồng nhau, thậm chí có một số tờ báo tương đồng hoàn toàn. 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.

Một vấn đề khó khăn hơn là việc những bài luận tiểu luận, sao chép cắt dán tài liệu từ nhiều nguồn khác nhau (Internet, báo chí…) đã không còn là hiện tượng đơn lẻ mà diễn ra khá phổ biến. Đây là một hiện tượng đáng buồn nhưng lâu nay vẫn diễn ra với muôn hình vạng trạng, nhiều cách, nhiều kiểu.

Việc tránh sự trùng lặp của các bài báo, bài viết khoa học, giúp nâng cao chất lượng của các bài báo 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, nghiên cứu. 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 đáng tin cậy.

Bài toán được đặt ra là phát hiện các trang báo điện tử, hoặc bài tiểu luận, đồ án có nội dung tương đồng với nhau để có thể xây dựng một hệ thống phân loại cho người đọc những bài báo có chất lượng hiệu quả đen lại lượng lớn kiến thức hữu ích và loại bỏ các bài báo trùng lặp. Đảm bảo giảm thiểu việc sao chép luận văn, bài báo nghiên cứu khoa học, giúp cho nội dung trở nên phong phú không bị trùng lặp và hạn chế các bài báo bản sao đem lại các thông tin dư thừa.

Để giải quyết vấn đề, 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 (hay còn có thể hiểu là ước lượng dựa trên thống kê), đó là phương pháp ước lượng out of place, phương pháp out of place tách văn bản ra thành n-gram, và đánh giá mức độ quan trọng của từ tf-idf (term frequency – inverse document frequency) của các gram đó trong văn bản.

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.

* 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 mô hình.

**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.

* Giới thiệu về thuật toán out of place.
* Chi tiết thuật toán: tách văn bản thành n-gram, đánh giá trọng số tf-idf, ước lượng khoảng cách bằng out of place.

**Chương 3:** Xây dựng bộ dữ liệu để có thể kiểm nghiệm độ chính xác của phương pháp out of place. Xây dựng bộ dữ liệu để kiểm nghiệm thuật toán

* Kịch bản kiểm nghiệm thuật toán
* Kết quả thực hiện kiểm nghiệm thuật toán

**Chương 4:** Xây dựng ứng dụng kiểm tra độ tương đồng giữa các văn bản.

* Giới thiệu ứng dụng
* Kiến trúc ứng dụng
* Kết quả cài đặt ứng dụng

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

Chương 1: Đồ án 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**

Sự tương đồng giữa hai văn bản là sự giống giống nhau về nội dung hoặc trùng lặp về các câu văn bị sao chép giữa các văn bản đó. Do đó, một trong hai văn bản sẽ là bản sao của văn bản còn lại. Như vậy văn bản sao chép sẽ có nội dung tương đồng với văn bản gốc với phần trăm rất cao. Một văn bản bất kỳ có độ tương đồng với văn bản gốc càng lớn thì có thể văn bản đó là văn bản sao chép, hoặc là một bản sao của văn bản gốc.

Do đó, để kiểm tra văn bản có phải là bản sao hoặc sao chép của văn bản gốc hay không ta phải tính độ tương đồng giữa hai 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 nghị 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 thống kê. Phương pháp tìm độ tương đồng giữa hai văn bản dựa trên thống kê 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 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 thống kê không thể xác định sự giống nhau về ý 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 thống kê còn có 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 thực 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 trùng 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 do các bài báo bị sao chép hay chỉnh sửa lại sau đó đăng lại tại nhiều trang web khác nhau.

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

Hình 1.1: 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 hiệu 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 nhằm giúp hạn chế bớt các bài tờ báo không có uy tín lấy cắp nội dùng bài viết của các tờ báo khác. Như vậy, đồ án 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 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 nội dung tìm kiếm này tập hợp các bài báo nói về nội 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 bài báo có bài viết trùng lặp. Trong trường hợp tốt hơn 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 lần, 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 dựa trên thống kê**

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 thống kê sẽ được chia thành hai nhóm: ước lượng dựa trên cụm từ, ước lượng dựa trên kí tự. Hình 1.2 giới thiệu 14 thuật toán ước lượng độ tương đồng phổ biến, bẩy thuật toán trong số đó là ước lượng dựa trên cụm từ, bẩy cái còn lại là ước lượng dựa trên kí tự.

Hình 1.2: Các phương pháp ước lượng độ tương đồng dựa trên kí tự[1]

* **Character-Based Similarity Measures**

Longest Common SubString (LCS) [16]: 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 [16]: 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 [9]: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 [9]:là phiên bản mở rộng của Jaro, tuy nhiên winker thêm trọng số ký tự trùng khớp đầu tiên.

Needleman-Wunsch [12]: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 (ví dụ: 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 [14]: 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 [3]: 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 [3]: 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 dấu 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 của 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 [5]: 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 [5]: còn được gọi là L2 distance là căn bậc hai tổng bình phương giữa các tọa đô tương ứng của hai vector.

Jaccard similarity [8]: được tính bằng số các từ chung của cả hai chuỗi chia cho tổng số tất cả các từ khác biệt trong cả hai chuỗi.

Matching Coefficient [8]: là một cách tiếp cận đơn giản dựa trên vector bằng cách đếm số các từ giống nhau ở cả hai vector.

Overlap coefficient [8]: 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 nói cách khác 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 1.3 bao gồm các thuật toán ước lượng độ tương đồng dựa trên kho dữ liệu.

* **Corpus-Based Similarity**

Hình 1.3: Các phương pháp ước lượng độ tương đồng dựa trên kho dữ liệu[15]

Hyperspace Analogue to Language (HAL) [11]**:** 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 bốn và giảm đi nếu số từ cách tăng lên. Nếu b tách biệt với a trọng số của nó sẽ bằng không. 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) [11]: 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) [11]: 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) [6]: 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-tdf. 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) [6]: là phương pháp ước lượng tổng quát đa ngôn ngữ của phương pháp ESA. CL-ESA khai thác dữ liệu 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) [15]: 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) [15]: là một phương pháp ước lượng động tương đồng dựa trên ngữ nghĩa sử dụng thông tin tương hỗ theo từng 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) [4]: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 :

Trong đó M là tổng 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 vô 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.

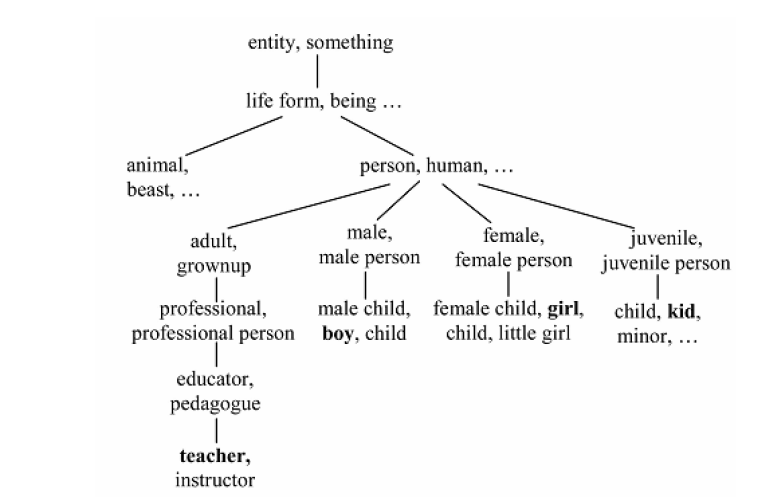
Extracting DIStributionally similar words using CO-occurrences (DISCO) [6]: là phương pháp dựa trên tương đồng phần phối giữa các từ, phương pháp này đặt ra giả thuyết các từ tương đồng về ngữ nghĩa thì xảy ra trong cùng một đoạn. Tập hợp các văn bản được phân tích thống kê để tìm được tương đồng phân 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 được biết đến 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.

Đồ án trình bày ví dụ về ước lượng độ tương đồng dựa trên mạng ngữ nghĩa: đối với hai từ “boy” và “teacher”, khi xét mối quan hệ giữa 2 từ này trên tập

WorldNet 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 1.4: 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 hiểu như ước lượng giữa các từ đồng nghĩa. Mặt khác, ước lượng dựa trên liên quan ngữ nghĩa thì liên quan về ngữ nghĩa giữa các từ, sự liên quan về ngữ nghĩa của ước lượng dựa trên liên quan ngữ nghĩa không quá chặt chẽ, gò bó 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 ở trong câu như thế nào.

Hình 1.5: Các phương pháp ước lượng độ tương đồng dựa trên tri thức[16]

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*),** **Lin (*lin*) [13]** and **Jiang & Conrath (*jcn*) [10]**. Ba phương pháp còn lại dựa trên độ dài con đường để biểu thị tương đồng về ngữ nghĩa của từ: **Leacock & Chodorow (*lch*)**, **Wu & Palmer (*wup*)** và **Path Length (*path*) [13]**.

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***)[2], Lesk (***lesk***) [2] và vector pairs (***vector***) [15]. Phương pháp ước lượng thực hiện bằng 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 bằng 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 ứng dụng**

Đồ án sẽ xây dựng ứng dụng 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 lỗi ví dụ: mất kí tự, các từ không đúng vị trí, các kí tự thừa…
* Dễ dàng cài đặt.
* Có thể áp dụng để xử lý cả văn bản hoặc áp dụng với một câu đơn giản.
* Sử dụng n-gram có thể so sánh được một số từ đặc biệt (Ví dụ, ‘advance’, ‘advanced’, ‘advancing’,‘advancement’, etc.).

Tuy nhiên, thuật toán out of place vẫn cần có những cải thiện nhằm 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 hơn 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
* Cách trình bày 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 lỗi, 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 cũng 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 trong việc xét các bài báo cùng một chủ đề nhưng mỗi bài có cách truyền tải khác nhau có thể bằng hình ảnh, có thể bằng từ ngữ dẫn đến ảnh hưởng tới việc ước lượng độ tương đồng của 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 thống kê. 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 dựa trên thống kê cao nhưng ngữ nghĩa lại khác nhau.
  1. **Kết luận**

Trong chương 1, đồ án đã nêu ra bài toán ước lượng độ tương đồng giữa các văn bản và các ứng dụng dựa trên bài toán ước lượng độ tương đồng. Đồ án giới thiệu một số phương pháp ước lượng độ tương đồng giữa các văn bản, và những phương pháp được giới thiệu ở chương 1 được chia vào 3 nhóm chính: ước lượng dựa trên kí tự, ước lượng dựa trên kho dữ liệu, ước lượng dựa trên tri thức.

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: Đồ án sẽ trình bày những vấn đề sau:

* Giới thiệu về thuật toán out of place
* Chi tiết thuật toán
  1. **Giới thiệu về thuật toán out of place**

Thuật toán out of place là thuật toán ước lượng độ tương đồng giữa các văn bản dựa trên phương pháp thống kê, hay còn có thể hiểu là phương pháp ước lượng dựa trên số lần xuất hiện của các cụm từ, các kí tự, tính quan trọng của các cụm từ, kí tự trong các văn bản khác nhau. Sau khi thực hiện tiền xử lý văn bản, thuật toán out of place sẽ áp dụng các kĩ thuật sau để xử lý văn bản: n-gram (văn bản sẽ được tách thành các n-gram), tf-idf (các gram sẽ được tính mức độ quan trọng trong văn bản dựa vào hai yếu tố: số lần xuất hiện, mức độ phổ biến ở trong các văn bản). Đồ án sẽ đi vào chi tiết các kĩ thuật dưới đây.

* + 1. **Phương pháp tiếp cận dựa trên n-gram**

Khái niệm n-gram (mô hình ngôn ngữ): là phân bố xác suất trên các tập văn bản. Nói cách khác, mô hình ngôn ngữ có thể cho biết

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 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 các từ trong chuỗi, các phần nguyên vẹn được tách ra để xử lý, kiểm tra độ tương đồng. 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 thiếu, hoặc bị lỗi không đọc được trong văn bản.

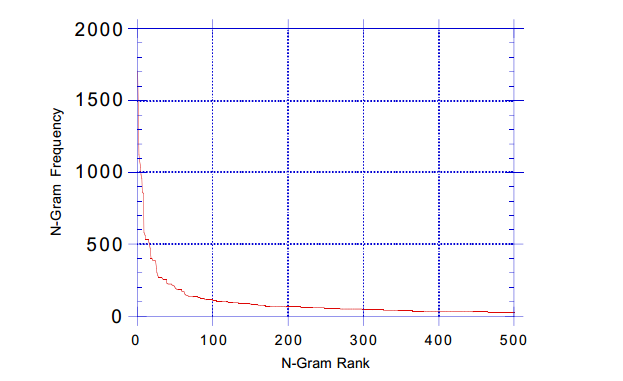
Một số lĩnh vực áp dụng n-gram trong việc xử lý ngôn ngữ tự nhiên:

* Nhận dạng gióng nói (speech recognition) : máy tính sẽ thử so khớp các đoạn âm thanh với các chuỗi từ. N-gram cung cấp ngữ cảnh để phân biệt các từ và cụm từ đồng âm nhưng khác nghĩa.
* Dịch máy (machine translate): là chương trình tự động dịch một ngôn ngữ này sang ngôn ngữ khác.
* Tóm tắt văn bản: văn bản được tóm tắt lại thành các từ khóa chứa thông tin quan trọng của văn bản.
* Khai phá dữ liệu: Phát hiện tri thức mới từ rất nhiều nguồn tài liệu khác nhau.
* Các bài toán trong xử lý tiếng Việt như phân tách câu, phân tách từ, tự động thêm dấu.
* Ước lượng độ tương đồng giữa các văn bản.

Phần tiếp theo, đồ án sẽ trình bày các bước để xây dựng theo mô hình n-gram.

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

Trong ngôn ngữ luôn có những từ xảy ra thường xuyên so với các từ khác. Và một trong những phát biểu Zipf’s law[17] phát biểu về vấn đề này như sau: “tần số xuất hiện của từ nào càng lớn sẽ tỉ lệ nghịch với mức độ quan trọng của từ đó”.



Hình 2.1: Mối quan hệ giữa tấn suất và mức độ quan trọng của n-gram [3]

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, vì vậy cần sử dụng kĩ thuật để có thể tăng hoặc giảm mức độ quan trọng của từ. Phương pháp tf-idf [7] là một kĩ thuật 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)

**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:**

Bảng 2.1: Ví dụ về một số từ xuất hiện nhiều trong cả văn nói và viết

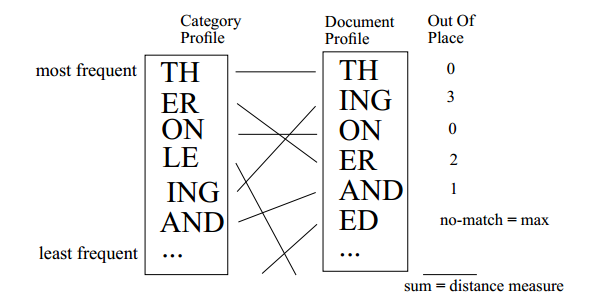
|  |  |
| --- | --- |
| **Loại từ** | **Ví dụ** |
| **Từ đếm** | **Một, hai ,ba …** |
| **Số đếm** | **1,2,3…** |
| **Từ phủ định** | **Không, ngược lại** |
| **Từ nối** | Và, nhưng, tuy nhiên, vì thế, vì vậy, … |
| Giới từ | Ở, trong, trên, … |
| Từ chỉ định | Ấy, đó, nhỉ, … |

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

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:

* + 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. Trong phần 2.1.2, đồ án 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 2.2 đưa ra ví dụ về cách tính khoảng cách của một vài gram.

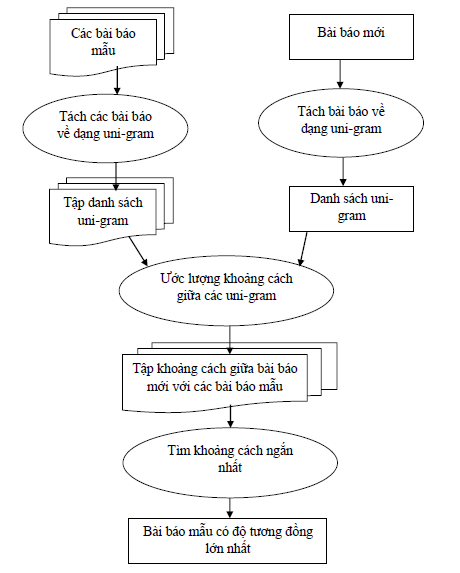


Hình 2.2: Cách tính khoảng cách giữa hai văn bản dựa trên Out Of Place [3]

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 của 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” (Hình 2.2) không nằm trong văn bản mẫu, thì khoảng cách của nó sẽ 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 2.3)



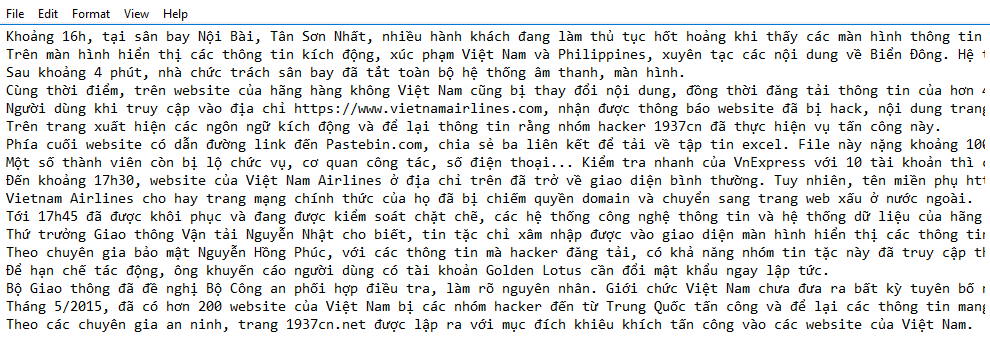
Hình 2.3: Các bước thực hiện của phương pháp Out Of Place [3]

* Tách văn bản ra thành các gram tách biệt chỉ chứa từ. Số và dấu chấm câu, các kí tự đặc biệt đều bị loại.
* Kiểm tra với mỗi gram, 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.
* Sau khi có được danh sách các gram, ta 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 toàn bộ 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.

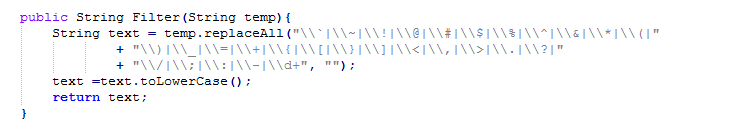
* 1. **Chi tiết thuật toán**
     1. **Đưa văn bản về dạng uni-gram**

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 2.4 mô tả dưới đây:



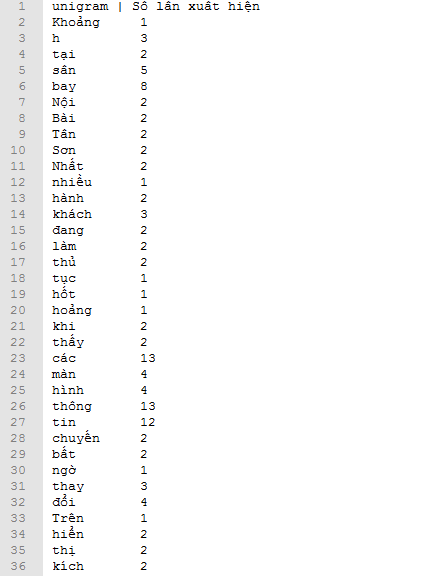
Hình 2.4: Nội dung của bài báo làm ví dụ cho bước 1

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 2.5: 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 2.6: 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.

Trong bước này đồ án đã đưa văn bản về dạng uni-gram, việc tiếp theo đồ án sẽ thực hiện đánh giá mức độ quan trọng của các uni gram trong văn bản được ví dụ.

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

Trong phần trước đồ án đã trình bày các bước để đưa văn bản về dạng uni-gram, như vậy văn bản đã được đưa về danh sách các uni-gram. Tiếp theo đồ án sẽ thực hiện đánh giá các gram.

Bước 1: Sau khi 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:

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 stop word, đồng thời áp dụng việc giảm mức độ quan trọng 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.

Bước 3: Với mỗi gram ta nhân giá trị tf với giá trị 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.

|  |  |
| --- | --- |
|  |  |
| Mẫu :n-gram và trọng số tf-idf | Mẫu: n-gram sẽ được sắp xếp theo thứ tự giảm dần theo trọng số tf-idf |

Hình 2.7: 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 2.7 mô tả kết quả sau khi các n-gram sắp xếp theo thứ tự giảm dần.

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

Trong chương trước, đồ án đã thực hiện xong việc phân tích văn bản 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 văn bản mẫu ở ví dụ dưới đây:

**Văn bản mẫu**: 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.

**Văn bản 1**: 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.

**Văn bản 2**: A snake follows a frog. It then eats the frog. And a crocodile follows the snake. Then It eats the snake. It continues to follow an other toad.

Đối với văn bản mẫu và văn bản 1 ta áp dụng các bước được nêu ở phần trên của đồ án để tạo ra danh sách các n-gram tương ứng. Kết quả sau khi thực hiện tương tự các bước trên với 2 văn bản:

|  |  |
| --- | --- |
|  |  |
| Văn bản mẫu: tách thành các uni-gram và trọng số tf-idf tương ứng với mỗi uni-gram đó | Văn bản 1: tách thành các uni-gram và trọng số tf-idf tương ứng với mỗi uni-gram đó |

Hình 2.8: So sánh n-gram của văn bản mẫu và văn bản 1

Tương tự bước trên ta tách văn bản mẫu và văn bản 2 thành danh sách các uni-gram và trọng số tf-idf.

|  |  |
| --- | --- |
|  |  |
| Văn bản mẫu: tách thành các uni-gram và trọng số tf-idf tương ứng với mỗi uni-gram đó | Văn bản 2: tách thành các uni-gram và trọng số tf-idf tương ứng với mỗi uni-gram đó |

Hình 2.9: So sánh n-gram văn bản mẫu và văn bản 2

Tiếp theo áp dụng thuật toán Out of place để ước lượng độ tương đồng giữa hai văn bản để xem văn bản 1 hay văn bản 2 giống với bài báo mẫu hơn. Thuật toán được áp dụng như sau.

* Xét các gram ở cả hai văn bản, khoảng cách giữa các gram được tính bằng ví trí của gram đó trong văn bản mẫu trừ vị trí của gram đó trong văn bản 1 lấy giá trị nguyên.
* Trong trường hợp gram chỉ không xuất hiện đồng thời ở văn bản thì khoảng cách sẽ được tính bằng số uni-gram lớn nhất của văn bản mẫu hoặc văn bản 1.

So sánh bài báo mẫu với văn bản số 1:

Bảng 2.2: Ví dụ về khoảng cách giữa các gram trong văn bản 1 với bài báo mẫu

|  |  |  |  |
| --- | --- | --- | --- |
| 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 = |
| “a” | 1 | 1 | 0 |
| “the” | 2 | 2 | 0 |
| “frog” | 3 | 3 | 0 |
| “attacks” | 4 | 5 | 1 |
| “win” | 5 | Không có | 17 |
| “losses” | Không có | 6 | 17 |
| “changes” | 6 | Không có | 17 |
| “updates” | Không có | 6 | 17 |
| “snake” | 7 | 7 | 0 |
| “corcodile” | 8 | 8 | 0 |
| “also” | 9 | 9 | 0 |
| “his” | 12 | 10 | 2 |
| “then” | 10 | Không có | 17 |
| “while” | 11 | 12 | 1 |
| “target” | 13 | 11 | 2 |
| “it” | 14 | 13 | 1 |
| “eats” | 15 | 16 | 1 |
| “to” | 16 | 14 | 2 |
| “toad” | 17 | 15 | 2 |
| \*Khoảng cách lớn nhất giữa bài báo mẫu với văn bản số 1: 17 | | | |

Độ tương đồng của mẫu với văn bản số 1với văn bản mẫu sẽ được tính như sau:

%Similarity (văn bản 1 vs văn bản mẫu) = 83.45%

So sánh bài báo mẫu với văn bản số 2:

Bảng 2.3: Ví dụ về khoảng cách giữa các gram trong văn bản 2 với văn bản mẫu

|  |  |  |  |
| --- | --- | --- | --- |
| 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 |
| “a” | 1 | 2 | 1 |
| “the” | 2 | 5 | 3 |
| “frog” | 3 | 12 | 9 |
| “attacks” | 4 | Không có | 17 |
| “wins” | 5 | Không có | 17 |
| “changes” | 6 | Không có | 17 |
| “snake” | 7 | 3 | 4 |
| “crocodile” | 8 | 14 | 6 |
| “also” | 9 | Không có | 17 |
| “then” | 10 | 6 | 4 |
| “while” | 11 | Không có | 17 |
| “his” | 12 | Không có | 17 |
| “target” | 13 | Không có | 17 |
| “it” | 14 | 4 | 10 |
| “eats” | 15 | 13 | 2 |
| “to” | 16 | 15 | 1 |
| “toad” | 17 | 16 | 1 |
| “follows” | Không có | 1 | 17 |
| “follow” | Không có | 9 | 17 |
| “and” | Không có | 7 | 17 |
| “other” | Không có | 11 | 17 |
| “an” | Không có | 10 | 17 |
| “continues” | Không có | 8 | 17 |
| \*Khoảng cách lớn nhất giữa bài báo mẫu với bài báo số 2: 17 | | | |

Độ tương đồng của mẫu với văn bản số 2 với văn bản mẫu sẽ được tính như sau:

%Similarity (văn bản 2 vs văn bản mẫu) = 47.4 %

Kết luận: Từ kết quả về độ tương đồng giữa văn bản mẫu với hai văn bản thử nghiệm trên cho thấy rằng. Văn bản số 1 (83.45%) tương đồng với mẫu lớn hơn so với văn bản số 2 (47.4 %).

* 1. **Kết luận**

Trong chương 2 đồ án đã giới thiệu về thuật toán out of place và chi tiết các bước thực hiện của thuật toán, một số các kỹ thuật sử dụng trong thuật toán out of place: n-gram, tf-idf.

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 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**

Để đánh giá thuật toán đồ án sẽ sử dụng dữ liệu thực tế được lấy về được từ trên các trang báo điện tử. Đồ án áp dụng phương pháp out of place để ước lượng độ tương đồng giữa các bài báo. Phương pháp out of place và các kĩ thuật để xử lý văn bản đã được trình bày trong chương 2: phân tách văn bản ra thành n-gram và tính toán trọng số tf-idf cho các gram, 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 ngẫu nhiên từ 10 trang báo điện tử uy tin của Việt Nam như sau:

Bảng 3.1: Địa chỉ của các trang báo điện tử ở Việt Nam

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

Dữ liệu thực tế được lấy từ các bài báo tiếng Việt trong 10 tờ báo cụ thể như sau:

* Các bài báo được lấy chủ yếu ở các chuyên mục: thời sự, thế giới, giải trí và thể thao.
* Các bài báo chỉ có hình ảnh hoặc chỉ có video sẽ không được chọn làm dữ liệu.
* Bộ dữ liệu sẽ được xây dựng trên các tờ báo Tiếng Việt, các tờ báo sử dụng ngôn ngữ khác sẽ bị loại bỏ.

*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.

Bộ dữ liệu kiểm nghiệm sẽ có 3 bài bài báo trong đó một bài báo làm bài mẫu để so sánh với hai bài báo còn lại.

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

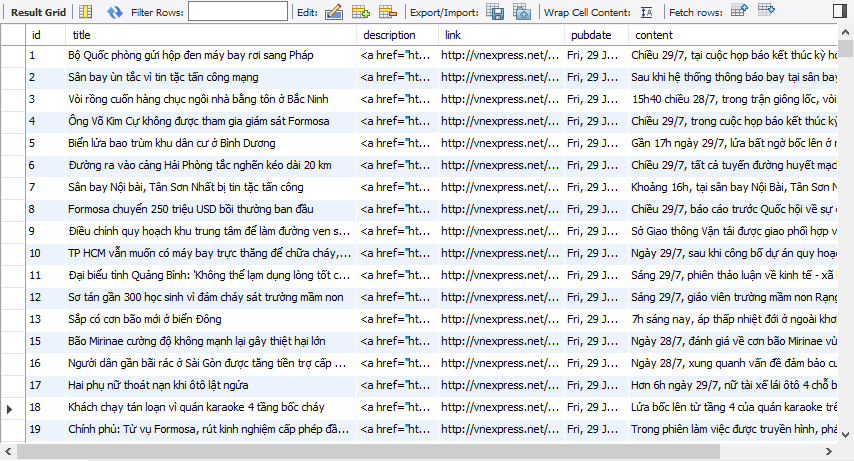
Bảng 3.2: Các công cụ phần mềm được sử dụng để lấy mẫu

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

Dữ liệu sẽ được xây dựng 1000 mẫu, trong mỗi mẫu sẽ có 3 bài báo được lấy từ 10 tờ báo 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.

Bước 2: Với đường dẫn 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 (crawler) thẻ html 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.

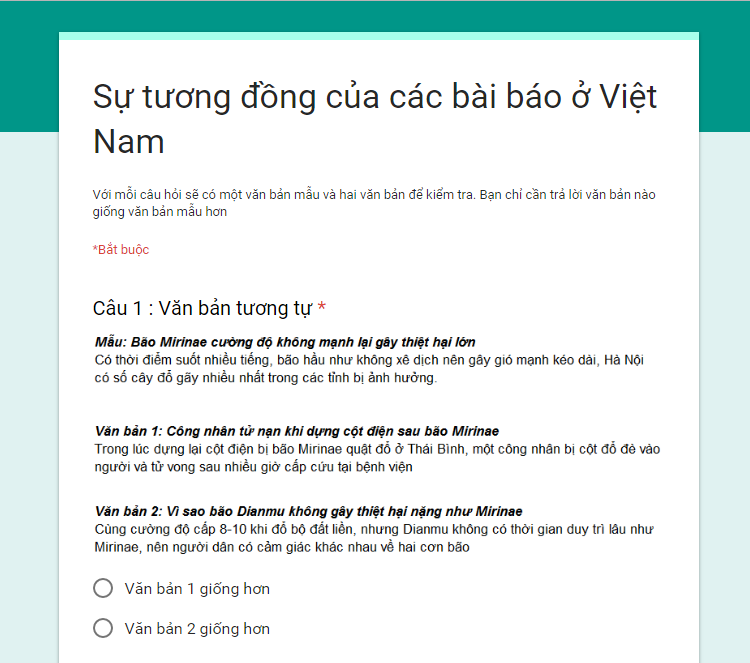


Hình 3.1: Cơ sở dữ liệu chứa thông tin của các bài báo

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. Với mỗi phiếu bao gồm 20 mẫu được đưa vào các phiếu khảo sát được tạo bởi biểu mẫu của google để người đọc có thể khảo sát (Hình 3.2).

Bước 4: Tổng hợp các bài báo thu thập được chia vào các chủ đề, mẫu sẽ được xây dựng bởi 3 bài báo trong một chủ để. Ba bài báo được lấy ra trong đó một bài báo sẽ làm bài báo mẫu, hai bài báo còn lại được dùng để so sánh. Nội dung của một bài báo lấy ra bao gồm hai phần: tiêu đề bài báo, mô tả bài báo.

Kết thúc bước xây dưng bộ dữ liệu: đồ án đã thể hiện cách thu thập dữ liệu để khảo sát, và cách xây dựng các mẫu khảo sát. Bước tiếp theo đồ án sẽ nêu kịch bản kiểm nghiệm.



Hình 3.2: Phiếu khảo sát người đọc ước lượng độ tương đồng giữa các bài báo

* 1. **Kịch bản đánh giá thuật toán**

Các mẫu tạo được ở bước xây dựng bộ dữ liệu sẽ được dùng để kiểm nghiệm độ chính xác của thuật toán out of place trong việc ước lượng độ tương đồng giữa các văn bản, các bước thực hiện như sau :

Đồ án sử dụng biểu mẫu của google để lưu trữ các mẫu, sau đó sử dụng biểu mẫu để thực hiện đánh giá độ tương đồng giữa các bài báo.

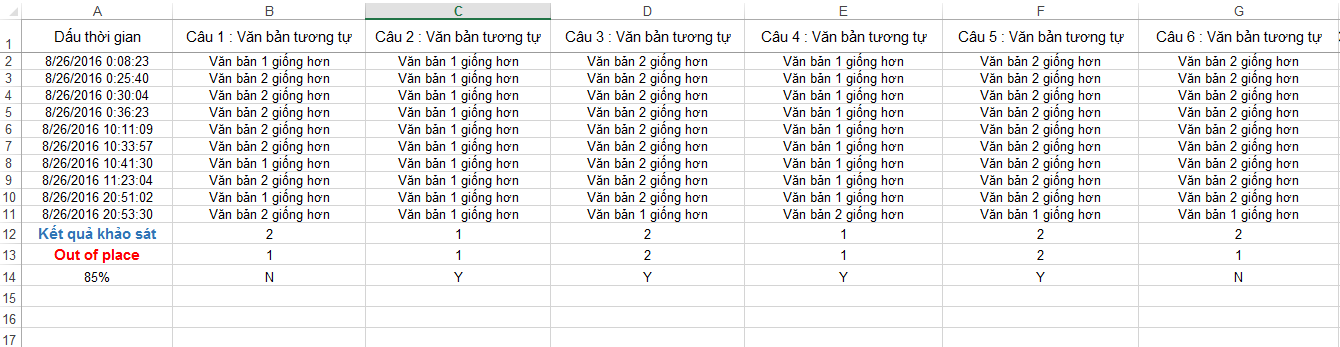
Bước 1: Các mẫu sẽ được từ 8 -10 người tham gia đánh giá, trả lời câu hỏi: trong 2 bài báo so sánh với bài báo mẫu bài báo nào có độ tương đồng với bài báo mẫu lớn hơn so với bài báo còn lại. Kết quả đánh giá mẫu sẽ được biểu mẫu thống kê lại như sau:

* Trong 1 mẫu: có bao nhiêu người chọn văn bản 1 giống văn bản mẫu, có bao nhiêu người chọn văn bản 2 giống với văn bản mẫu.
* Trong mẫu: có bao nhiêu người đã tham gia đánh giá, chi tiết mà người tham giá đánh giá.

Sau khi kết thúc bước 2: đồ án đã thu thập kết quả đánh giá độ tương đồng của các mẫu theo ý kiến của những người tham gia đánh giá. Bước tiếp theo đồ án sẽ thực so sánh kết quả thu được với kết quả chạy thuật toán

Bước 3: Áp dụng thuật toán out of place vào 1000 mẫu vừa được khảo sát như sau:

* Trong mỗi mẫu có văn bản mẫu, văn bản A, B cần kiểm tra văn bản có nội dung tương đồng với văn bản mẫu hơn.
* Áp dụng thuật toán với cặp văn bản A và mẫu để xác định độ tương đồng của văn bản A với văn bản mẫu, tương tự áp dụng để xác định độ tương đồng giữa văn bản B và mẫu.
* Sau khi áp dụng thuật toán để ước lượng độ tương đồng giữa các văn bản do đó hai văn bản A,B có thể xác định được văn bản nào có độ tương đồng lớn hơn, với kết quả chạy thuật toán so sánh với kết quả từ phiếu khảo sát thu thập được từ người dùng.
* 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. Kết quả của mẫu là đúng khi và chỉ khi thuật toán ra kết quả trùng với ý kiến của người khảo sát.



Hình 3.3: Kết quả đánh giá độ chính xác của thuật toán

Bước 4: Tổng hợp kết quả từ 1000 mẫu và kiểm trả độ chính xác của thuật toán. Độ chính xác của thuật toán sẽ được tính bằng số mẫu mà thuật toán ra kết quả trùng với ý kiến của người khảo sát chia cho tổng số mẫu khảo sát (1000 mẫu). Đồ án xác định độ chính xác của thuật toán theo công thức như sau:

* 1. **Kết quả đánh giá thuật toán**

Sau khi thực hiện lấy kết quả khảo sát của 1000 mẫu và chạy thuật toán out of plae để kiểm nghiệm kết quả. Trung bình độ chính xác của thuật toán: 96.8%

Chi tiết kết quả các mẫu sau khi thực hiện khảo sát:

Bảng 3.3: Chi tiết kết quả các mẫu

|  |  |
| --- | --- |
| Tổng số mẫu | 1000 mẫu |
| Số mẫu đúng | 968 mẫu |
| Số mẫu sai | 32 mẫu |
| Độ chính xác | 96,8% |

\*Số mẫu đúng là: các văn bản trong mẫu với kết quả khảo sát có độ tương đồng với văn bản mẫu lớn hơn và kết quả khi chạy thuật toán giống với ý kiến người dùng.

\*Số mẫu sai là: mẫu có kết quả khảo sát không trùng với kết quả thuật toán chạy.

* 1. **Đánh giá**

Từ kết quả thực nghiệm của 1000 mẫu với độ chính xác trung bình là 96,8% cho thấy rằng thuật toán ước lượng độ tương đồng out of place có thể áp dụng để kiểm tra độ tương đồng giữa hai văn bản. Đồ án sẽ áp dụng thuật toán thuật toán out of place để xây dựng một ứng dụng kiểm tra văn bản có sao chép hay không.

Những điểm còn hạn chế trong quá trính thực nghiệm thuật toán: chưa tính được trọng số tương đồng về ngữ nghĩa của văn bản vì vậy một số mẫu có độ chính xác chưa tốt.

Hạn chế về bộ dữ liệu: đồ án xây dựng bộ dữ liệu dựa trên các bài báo điện tử, một bài báo điện tử có thể bao gồm: hình ảnh, đoạn phim ngắn, các đoạn văn mô tả…Trong trường hợp một bài báo chỉ có hình ảnh hoặc đoạn phim sẽ làm giảm đi tính chính xác trong việc ước lượng độ tương đồng trong mẫu.

* 1. **Kết luận**

Trong chương 3, đồ án đã trình bày về xây dựng bộ dữ liệu thử nghiệm dùng để đánh giá thuật toán. Bộ dữ liệu này đã được khảo sát bởi người đọc thực tế. Kết quả sau của người tham gia khảo sát được thu thập sau đó dùng kết quả khảo sát để so sánh với kết quả ước lượng độ tương đồng của thuật toán out of place. Kết quả khảo sát cuối cùng độ chính xác trung bình của thuật toán là 96,8%

Tiếp theo là chương 4, đồ án sẽ xây dựng ứng dụng để kiểm tra độ tương đồng giữa hai văn bản, từ đó áp dụng ứng dụng vào việc kiểm tra đạo văn.

**CHƯƠNG 4: ỨNG DỤNG KIỂM TRA ĐẠO VĂN**

Trong chương 4 đồ án sẽ xây dựng ứng dụng để kiểm tra độ tương đồng giữa hai văn bản để kiểm tra văn bản có sao chép hoặc là bản sao. Chương 4 bao gồm các phần sau:

* Giới thiệu ứng dụng
* Kiến trúc ứng dụng
* Kết quả cài đặt
  1. **Giới thiệu ứng dụng**

Vấn đề đạo văn (plagiarism) ở Việt Nam thời gian gần đây đang trở nên “nóng” hơn bao giờ hết vì nó xảy ra không chỉ trong một số ngành thuộc khoa học xã hội như trước mà ngày càng lan rộng với các chuyên ngành khác. Đạo văn, theo định nghĩa của Merriam-Webster Online Dictionary, là hình thức sao chép ý tưởng và hình thành những ý tưởng hay ngôn từ mới khởi nguồn từ ý tưởng của ai đó, sử dụng sản phẩm của một ai đó mà không công bố nguồn, hoặc giới thiệu một ý tưởng hay sản phẩm mới được chuyển hóa từ nguồn đã có từ trước. Đạo văn là một hành vi gian lận, nó liên quan đến việc lấy ý tưởng, công sức của người khác và nói đó thành quả đó là của mình.

Cùng với vấn đề đạo văn, tình trạng thông tin trên báo mạng điện tử trùng lặp do nạn “cắt - dán” (“copy-paste”) ngày một tăng và tinh vi hơn. Trên thực tế người dùng có thể dễ dàng phát hiện những trường hợp nhà báo sao chép, sử dụng một phần hay toàn bộ tin, bài, ảnh của người khác, báo khác mà không nêu nguồn. Vấn đề sao chép xảy ra khá phổ biến, thường xuyên và nghiêm trọng trên các tờ báo mạng điện tử hoặc trang thông tin điện tử của các cơ quan báo chí. Nhiều tờ báo dịch tin, bài tràn lan từ các báo nước ngoài mà không ghi rõ tên tác giả và nguồn gốc của tác phẩm hoặc nếu có ghi cũng qua loa đại khái đến việc trích dẫn nguồn và viết tắt tên các báo trong nước hết sức tùy tiện, việc bỏ qua tên tác giả khi sử dụng lại bài của họ. Cũng tương tự việc sao chép, tham khảo luận văn của người khác mà không ghi rõ nguồn gốc của bài luận và tên tác giả.

Một ví dụ thực tế tại trường Đại học Duy Tân đã khảo sát hơn 2.000 tân sinh viên nhập học năm 2014, kết quả 84% từng chép từ nửa trang A4 trở lên ở trong sách tham khảo, giáo trình, trên Internet… mà không ghi nguồn. Lý do là không biết phương pháp trích dẫn (36%), áp lực tiến độ thực hiện (21%), không thể viết hay hơn (15%), không nhớ tác giả là ai (12%), không quan tâm (9%).

Do đó ứng dụng kiểm tra sự tương đồng giữa các văn bản được tạo ra để giải quyết vấn đề sao chép luận văn đồng thời bảo vệ quyền lợi của của những người sáng chế, những tài liệu do tác giả viết ra. Ứng dụng bao gồm các chức năng sau:

* Với dữ liệu được đưa vào bao gồm văn bản gốc và văn bản nghi ngờ là đã sao chép văn bản gốc, hệ thống sẽ đọc và so sánh sự tương đồng giữa hai văn bản, trong đó một văn bản đã được công bố chính thức (văn bản gốc). Ứng dụng sẽ chỉ rõ (làm nổi bật các dòng đó, đánh số thứ tự) các phần tương đồng giữa văn bản gốc và văn bản nghi ngờ là sao chép.
* Ngoài ra ứng dụng cho phép kiểm tra độ tương đồng giữa các bài báo điện tử, cách kiểm tra cũng tương tự với việc kiểm tra độ tương đồng giữa hai văn bản. Tùy vào đặc thù của bài báo hoặc văn bản và các phần giống nhau, tác giả sẽ quy định tỉ lệ phần trăm bao nhiêu thì bị xem là “đạo văn”.

Việc sử dụng phần mềm kiểm tra “đạo văn” nâng cao chất lượng của các bài báo trên tạp chí khoa học, đảm bảo các thông tin được trích dẫn đầy đủ, có độ tin cậy cao, hạn chế sự sao chép luận văn tốt nghiệp của sinh viên, các bài tập lớn và ngày càng nâng cao chất lượng hơn.

* 1. **Kiến trúc ứng dụng**

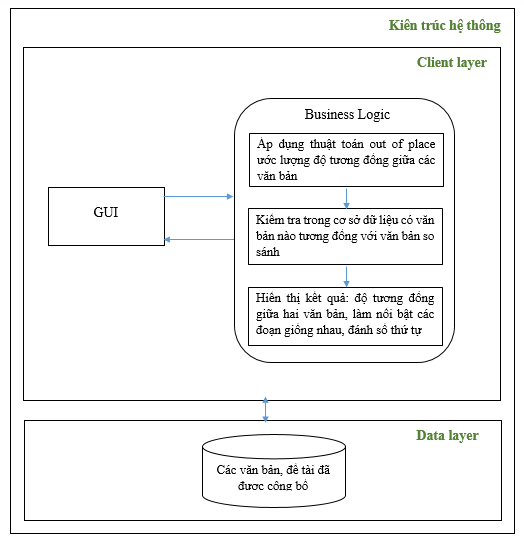
Để minh họa cho thuật toán out of place được trình bày ở chương hai, đồ án sẽ thực hiện xây dựng một ứng dụng với chức năng chính để kiểm tra độ tương đồng giữa hai văn bản từ đó có thể biết được văn bản gốc có bị sao chép không, văn bản so sánh tương đồng bao nhiêu phần trăm với văn bản gốc, ngoài ra ứng dụng còn có thể kiểm tra độ tương đồng giữa hai bài báo từ đó có thể biết được bài báo bị nghi là sao chép giống bao nhiêu phần trăm so với bài bản gốc.

Các công cụ phần mềm và thư viện để xây dựng ứng dụng:

Bảng 4.1: Các công cụ và thư viện để xây dựng ứng dụng

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên phần mềm và thư viện | Tác giả | Nguồn |
| 1 | Apache POI | The Apache Software Foundation | http://poi.apache.org/download.html |
| 2 | [IntelliJ IDEA](https://www.jetbrains.com/idea/) | JetBrains | https://www.jetbrains.com |
| 3 | Jsoup | Jonathan Hedley | https://jsoup.org/ |
| 4 | PDFBox | The Apache Software Foundation | https://pdfbox.apache.org/ |

Kiến trúc của ứng dụng:



Hình 4.1: Mô hình kiến trúc ứng dụng

Ứng dụng được viết bằng ngôn ngữ Java và sử dụng nền tảng Java Swing để xây dựng ứng dụng chạy trên window. Giao diện của ứng dụng được thiết kế và sử dụng các thành phần cơ bản của Java swing

* Jcomponent(Jbutton, Jtextpane, Jlabel, Jtextfield, Jtabbedpane, Jfilechooser)
* Container (Jpanel, Jframe, Dialog)
* Layout (GridLayout, GridBagLayout)

Kịch bản sử dụng ứng dụng cho chức năng kiểm tra sự tương đồng giữa hai văn bản:

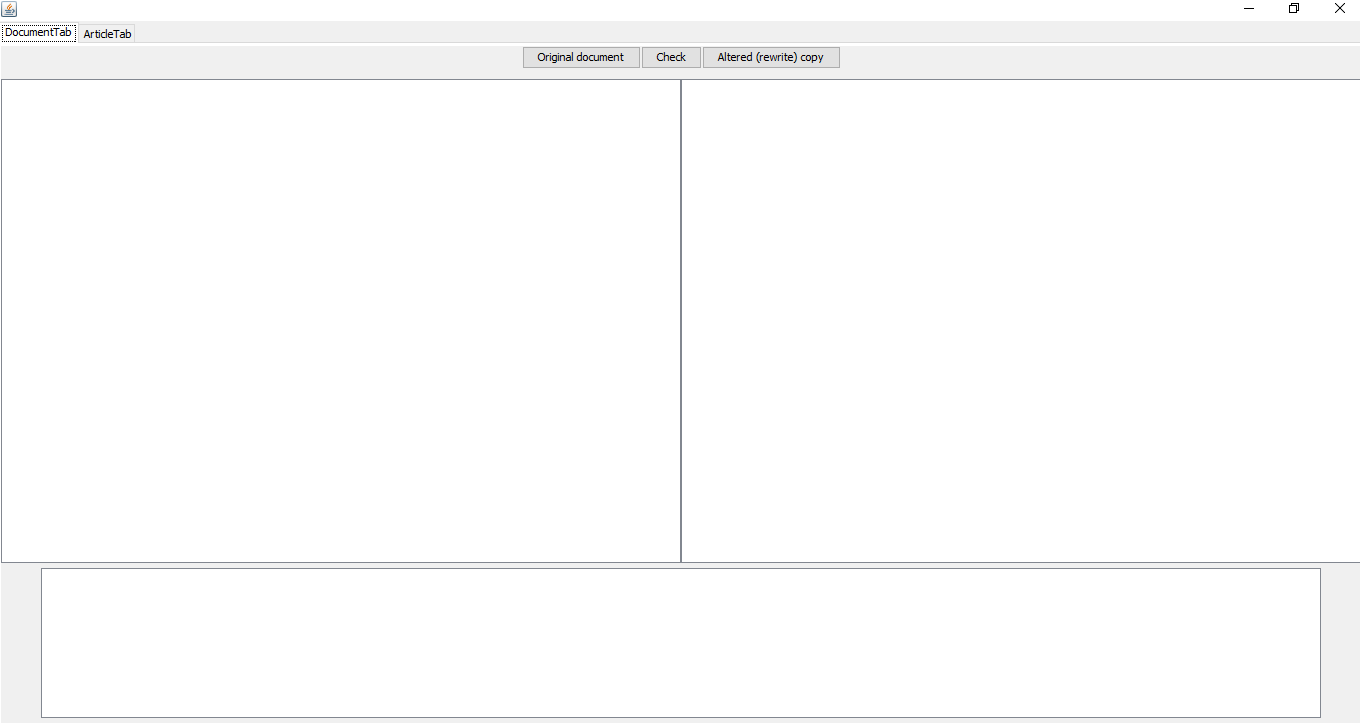
* Bước 1: Ứng dụng hiển thị giao diện trên mành hình
* Bước 2: Click vào nút chọn “original document” để chọn văn bản gốc cần kiểm tra
* Bước 3: Hệ thống thực hiện đọc nội dung của văn bản đó và hiển thị trong khung đọc bên trái và hiển thị tên của văn bản đó ở trên
* Bước 4: Click vào nút chọn “altered document” để chọn văn bản cần so sánh
* Bước 5: Hệ thống sẽ đọc nội dung của văn bản đó và hiển thị ở khung bên phải của ứng dụng và tên văn bản đó.
* Bước 6: Click vào nút “Check” để thực hiện kiểm tra sự tương đồng giữa hai văn bản.
* Bước 7: Hệ thống sẽ áp dụng thuật toán trình bày ở chương 2 để kiểm tra độ tương đồng giữa hai văn bản. Tiếp theo ứng dụng sẽ thực hiện làm nổi bật những đoạn giống nhau ở cả hai văn bản và đánh số tương ứng cho những đoạn giống nhau.
* Bước 8: Cuối cùng ứng dụng hiển thị kết quả: độ tương đồng giữa hai văn bản, hai văn bản được nổi bật và đánh số tại những đoạn tương đồng nhau.

Kịch bản cho chức năng kiểm tra sự tương đồng giữa hai bài báo điện tử. Lưu ý hiện tại ứng dụng mới có thể kiểm tra độ tương đồng giữa các tờ báo điện tử của 10 nhà xuất bản (vnexpress, dantri, laodong, tuoitre, phapluat, congan, vietnamenet, bongda, tienphong) :

* Bước 1: Ứng dụng hiển thị giao diện trên mành hình
* Bước 2: Người dùng chọn tab Article, giao diện để kiểm tra độ tương đồng giữa hai tờ báo điện tử hiện ra.
* Bước 3: Nhập link của tờ báo gốc vào textfield (original), nhập link của tờ báo cần so sánh với tờ báo vừa nhập vào textfield (altered). Người dùng click vào nút “get content”.
* Bước 4: Hệ thống thực hiện đọc nội dung và tiêu đề từ hai link vừa nhập vào sau đó hiện thị nội dung tương ứng ra màn hình.
* Bước 5: Người dùng thấy được nội dung và tiêu đề của tờ báo điện tử trên màn hình ứng dụng, sau đó ấn nút “check” để kiểm tra sự tương đồng của hai tờ báo.
* Bước 6: Hệ thống sẽ áp dụng thuật toán trình bày ở chương 2 để kiểm tra độ tương đồng giữa hai văn bản. Tiếp theo ứng dụng sẽ thực hiện làm nổi bật những đoạn giống nhau ở cả hai văn bản và đánh số tương ứng cho những đoạn giống nhau.
* Bước 7: Cuối cùng ứng dụng hiển thị kết quả: độ tương đồng giữa hai văn bản, hai văn bản được nổi bật và đánh số tại những đoạn tương đồng nhau.
  1. **Kết quả cài đặt**

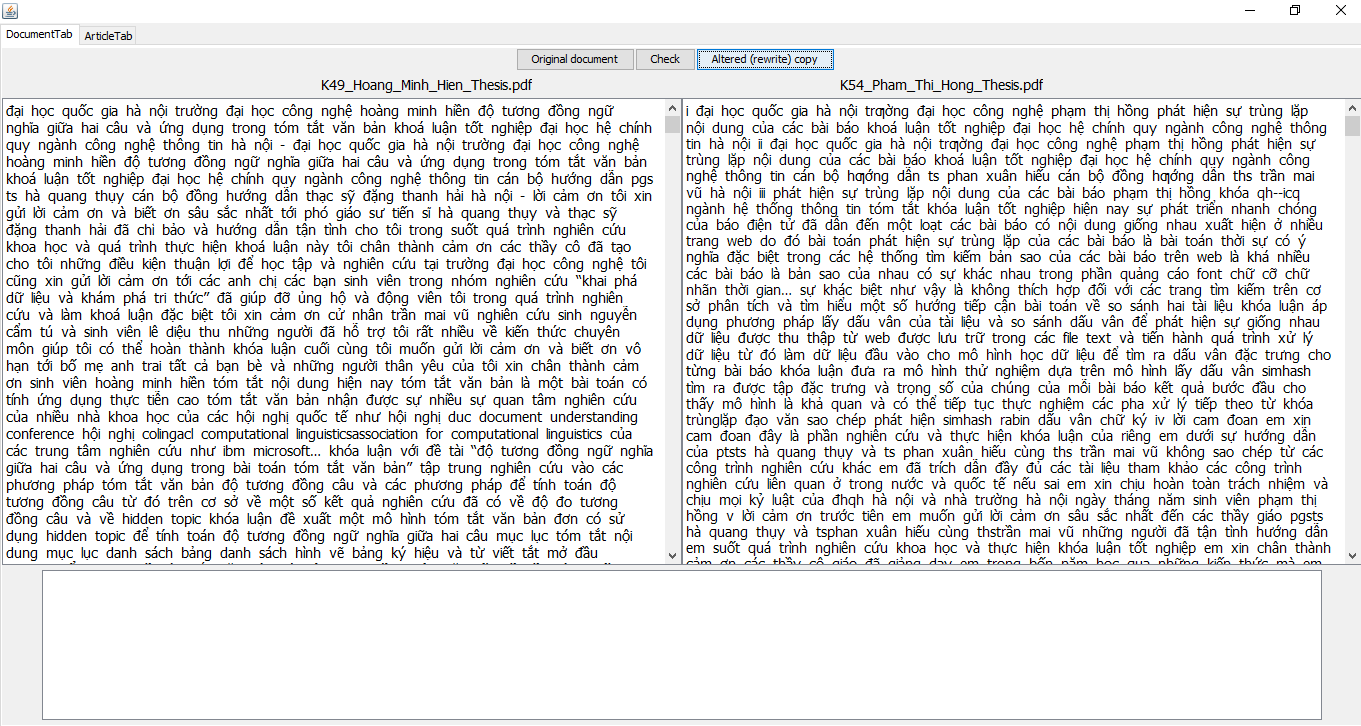
Giao diện cài đặt và thực hiện so sánh độ tương đồng giữa hai văn bản:

* Giao diện cho chức năng kiểm tra sự tương đồng giữa hai văn bản: bao gồm: trên cùng có 3 nút, nút “original” dùng để nhập văn bản gốc, button “alterd” dùng để nhập văn bản nghi ngờ đã sao chép văn bản gốc, sau khi nhập hai văn bản, chọn nút “check” dùng để kiểm tra độ tương đồng giữa văn bản.
* Hai khung lớn hình vuông nằm cạnh nhau dùng để chứa nội dung hai văn bản được nhập vào, khung bên phải chứa nội dung văn bản gốc, khung bên trái chứa văn bản bị nghi ngờ sao chép.
* Khung ở dưới cùng dùng để chứa kết quả cuối cùng: độ tương đồng giữa hai văn bản



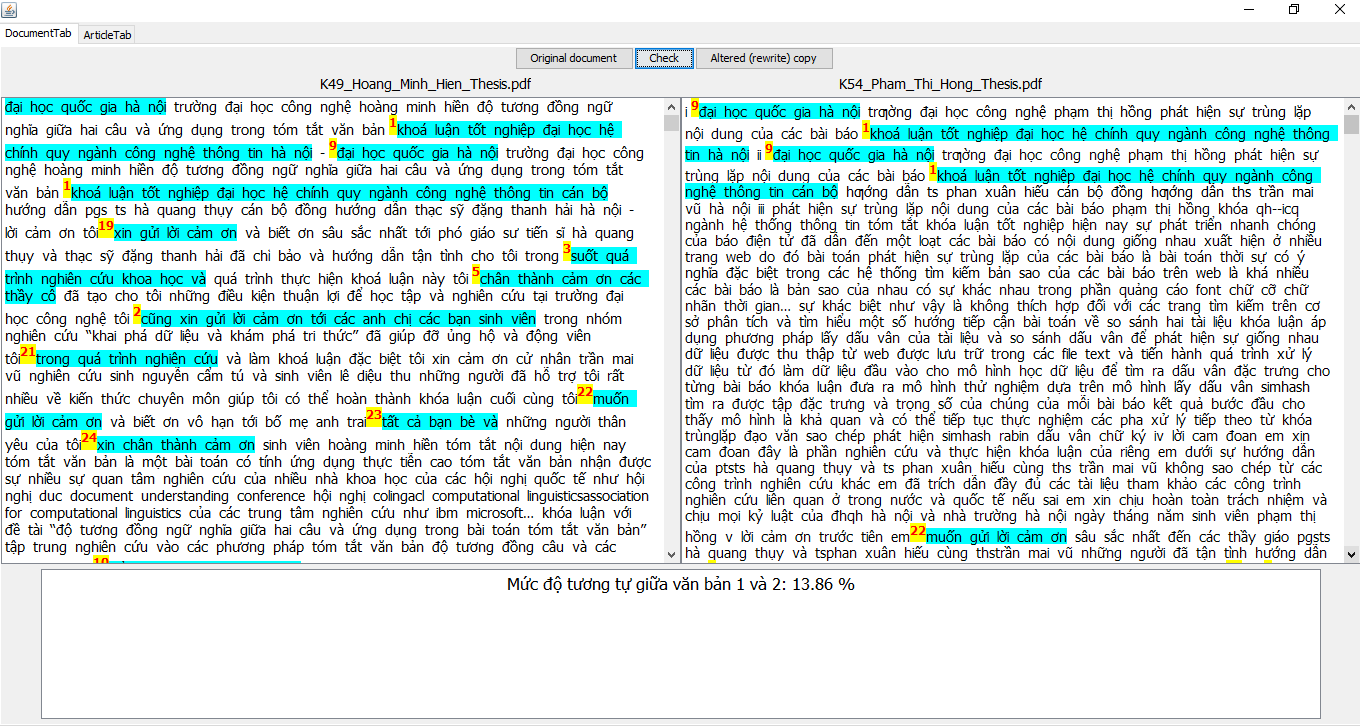
Hình 4.2: Giao diện chính khi mở ứng dụng

Sau khi thêm hai văn bản cần kiểm tra sự tương đồng, giao diện sẽ hiển thị giống Hình 4.3. Bên phải hiển thị tiêu đề và nội dung của văn bản gốc, bên trái hiển thị tiêu đề và nội dung của văn bản so sánh với văn bản gốc



Hình 4.3: Giao diện ứng dụng sau khi thêm các file cần kiểm tra sự tương đồng

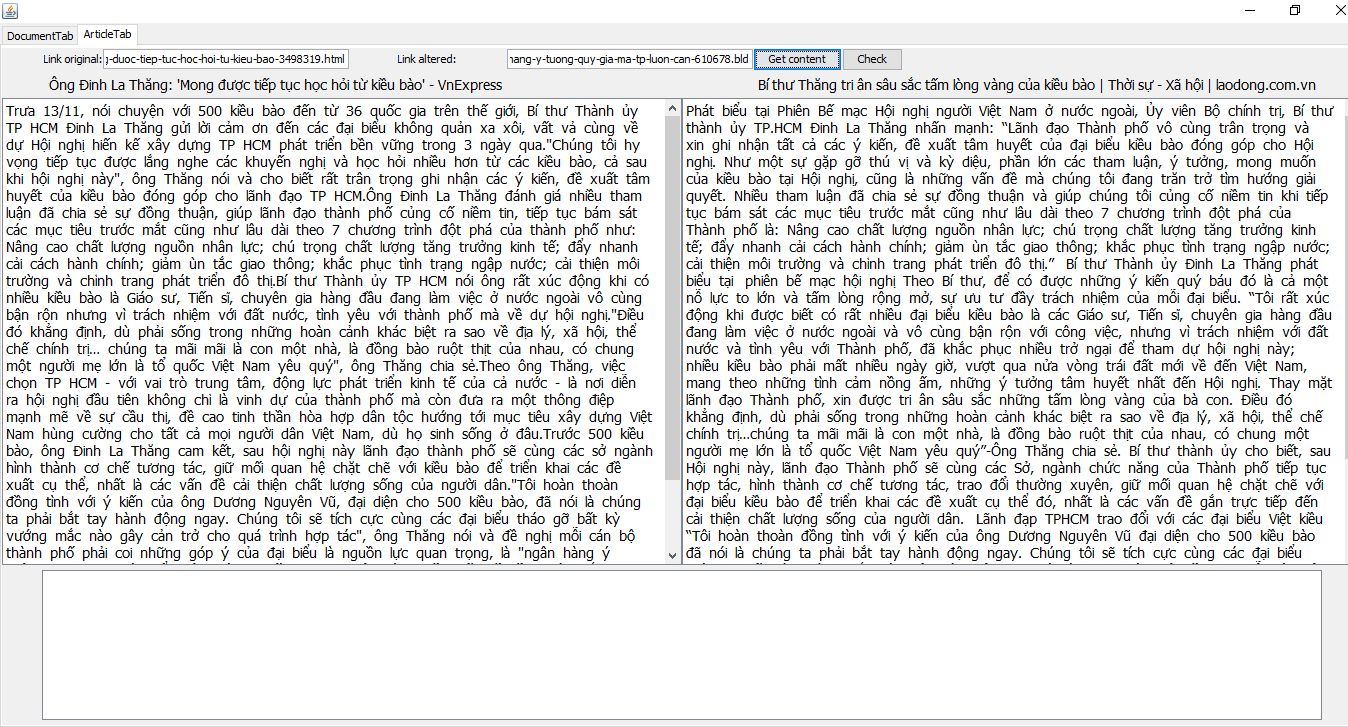
Sau khi ấn nút check, ứng dụng sẽ thực hiện xử lý hai văn bản sau đó áp dụng thuật toán out of place để kiểm tra độ tương đồng giữa hai văn bản. Các đoạn giống nhau giữa hai văn bản sẽ được làm nổi bật và đánh số thứ tự. Sau đó ứng dụng hiển thị kết quả trên màn hình (Hình 4.4)



Hình 4.4: Kết quả sau khi ước lượng độ tương đồng giữa hai văn bản

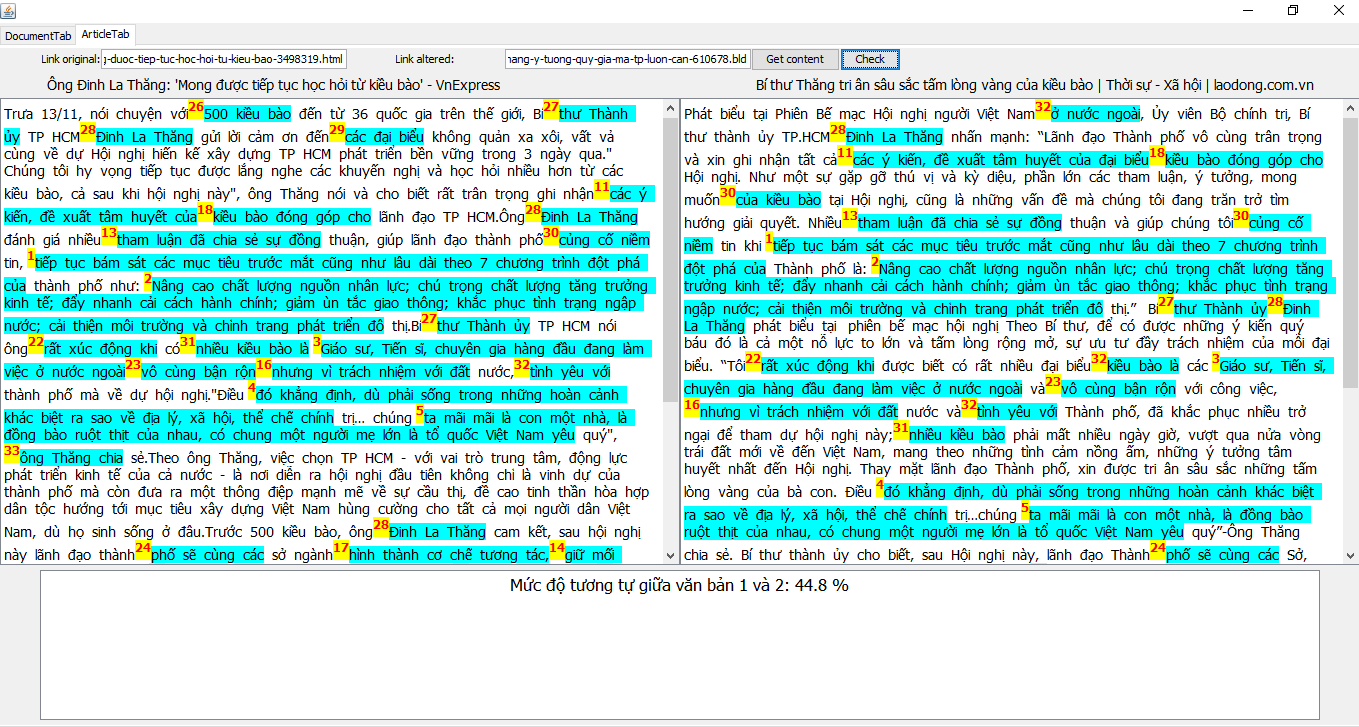
Giao diện cho chức năng kiểm tra độ tương đồng giữa hai tờ báo điện tử:

* Giao diện của ứng dụng trên cùng là hai trường điền dữ liệu để nhập link của 2 tờ báo điện tử. Ô thứ nhất là link của tờ báo gốc, Ô thứ hai dùng để nhập link của tờ báo cần so sánh với tờ báo gốc. Bên cạnh là nút “get content” dùng để đọc nội dung từ các tờ báo hiển thị nội dung tương ứng ở hai khung văn bản bên dưới. Cuối cùng dùng để kiểm tra độ tương đồng giữa hai tờ báo
* Hai khung văn bản nằm cạnh nhau dùng để chứa nội dung hai tờ báo điện tử, khung bên phải chứa nội dung tờ báo gốc, khung bên trái chứa tờ báo bị nghi ngờ sao chép.
* Khung ở dưới cùng dùng để chứa kết quả cuối cùng: độ tương đồng giữa hai tờ báo điện tử



Hình 4.5:Giao diện hiển thị nội dung hai tờ báo thông qua link được thêm trong ứng dụng

Người dùng nhập link của hai tờ báo cần so sánh độ tương đồng vào hai ô texfiled. Sau đó, người dùng click nút “get content”, ứng dụng sẽ lấy nội dung của hai tờ báo được từ link vừa được nhập vào, màn hình sẽ hiển thị kết quả như Hình 4.5. Người dùng click nút check để ứng dụng thực hiện ước lượng độ tương đồng, hệ thống làm nổi bật và đánh số thứ tự các đoạn giống nhau. Ứng dụng hiển thị kết quả như Hình 4.6.



Hình 4.6: Kết quả sau khi ước lượng độ tương đồng giữa hai tờ báo điện tử

* 1. **Kết luận**

Trong chương 4, đồ án đã trình bày về ứng dụng ước lượng độ tương đồng giữa hai văn bản (mô hình kiến trúc ứng dụng, các thư việc hỗ trợ) từ đó áp dụng để kiểm tra đạo văn giữa các bài báo, luận văn, đồ án tốt nghiệp. Ứng dụng gồm hai chức năng chính: kiểm tra độ tương đồng giữa hai file văn bản (docx, pdf, doc, text), kiểm tra độ tương đồng giữa các bài báo của một số trang báo điện tử (vnexpress, dantri, tuoitre, phapluat, laodong…).

**KẾT LUẬN**

Đồ án đã trình bày về việc nghiên cứu bài toán ước lượng độ tương đồng giữa các văn bản, nghiên cứu thuật toán out of place sau đó đồ án đã xây dựng bộ dữ liệu để kiểm nghiệm độ chính xác thuật toán. Cuối cùng, đồ án đã áp dụng thuật toán out of place vào xây dựng ứng dụng kiểm tra văn bản có sao chép văn bản khác hay không, hoặc các tờ báo điện tử có sao chép nội dung của nhau không, từ đó nâng cao chất lượng bài báo điện tử hay các bài báo điện tử, cung cấp cho người dùng những thông tin thiết thức và đáng tin cậy nhất.

Đồ án đã đạt được những kết quả sau:

* Nghiên cứu, tìm hiểu và cài đặt thuật toán out of place để ước lượng động tương đồng giữa các văn bản.
* Thực hiện thu thập nội dung của các bài báo tiếng Việt, từ đó xây dựng các bộ dữ liệu và đánh giá thuật toán out of place dựa trên việc khảo sát ý kiến bạn đọc. Kết quả cuối cùng sau khi thực hiện khảo sát, độ chính xác của thuật toán là 96,8%.
* Xây dựng ứng dụng để kiểm tra văn bản có sao chép của văn bản gốc không. Từ đó áp dụng để kiểm tra có sao chép luận văn hay không, các tờ báo điện tử có sao chép nội dung hay không.

Đồ án vẫn còn tồn tại một số mặt hạn chế như sau:

* Thuật toán out of place chưa ước lượng được độ tương đồng giữa các văn bản dựa trên ngữ nghĩa của văn bản.
* Hạn chế về bộ dữ liệu dữ liệu kiểm thử độ chính xác của thuật toán. Do dữ liệu chỉ lấy cố định trong 10 tờ báo điện tử nêu trong chương 3, do đó bộ dữ liệu chưa có sự đa dạng.
* Giao diện ứng dụng còn đơn giản, chức năng kiểm tra độ tương đồng giữa các tờ báo mới chỉ kiểm tra được các bài báo của 10 tờ báo được nêu ở chương 3. Phần được làm nổi bất và đánh số thứ tự cho các đoạn tương đồng giữa hai văn bản còn một số khiếm khuyết: còn trùng hoặc làm mất chữ trong văn bản.

Hướng phát triển của đồ án :

* Phương pháp ước lượng độ tương đồng giữa các văn bản của đồ án mới chỉ dừng ở mức xét trên kí tự, thống kê, do đó hướng phát triển trong thời gian sắp tới là tiến hành ước lượng dựa trên tính ngữ nghĩa của văn bản từ đó áp dụng vào bài toán ước lượng độ tương đồng giữa hai văn bản song song với đó là hoàn thiện thuật toán out of place tốt hơn để đảm bao độ chính xác của thuật toán tối ưu (tốc độ, tính chính xác).
* Đồ án sẽ tiếp tục xây dựng ứng dụng một cách hoàn chỉnh hơn: cải thiện giao diện, tạo thêm một số chức năng mới giúp cho việc sử dụng, nếu có thể sẽ là một hệ thống với nhiệm vụ là đảm bảo các văn bản được lưu là duy nhất từ đó có thể đánh giá các văn bản khác đưa vào có sao chép của các bài được lưu trong hệ thống hay không.

**TÀI LIỆU THAM KHẢO**

[1] Alberto, B, Paolo, R., Eneko A. & Gorka L. Plagiarism Detection across Distant Language Pairs, In Proceedings of the 23rd International Conference on Computational Linguistics. (2010), pages 37–45.

[2] Banerjee ,S. & Pedersen, T. An adapted Lesk algorithm for word sense disambiguation using WordNet. In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, 2002 pp 136–145.

[3] Cavnar, W. & Trenkle, J. N-Gram-Based Text Categorization.   
In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, no. 2 (1994): 161--175.

[4] Cilibrasi, R.L. & Vitanyi, P.M.B. The Google Similarity Distance, IEEE Trans. Knowledge and Data Engineering (2007): 370-383.

[5] Dice, L. R. "Measures of the Amount of Ecologic Association Between Species." *Ecology* 26, no. 3 (1945): 297--302.

[6] Gabrilovich, E. & Markovitch, S. "Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis." *Proceedings of the 20th International Joint Conference on Artificial Intelligence* (2007): 6--12.

[7] Wu, H. C.; Luk, R. W. P.; Wong, K.-F. & Kwok, K.-L. "Interpreting TF-IDF term weights as making relevance decisions". ACM Transactions on Information Systems, 26 . 2008.

[8] Jaccard, P. "Étude comparative de la distribution florale dans une portion des Alpes et des Jura." *Bulletin del la Société Vaudoise des Sciences Naturelles* 37 (1901): 547--579.

[9] Jaro, M. A. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association* 84, no. 406 (1989): 414--420.

[10] Jiang, J. J. & Conrath, D. W. "Semantic similarity based on corpus statistics and lexical taxonomy." Paper presented at the meeting of the Proceedings of International Conference Research on Computational Linguistics (ROCLING X'1997), Taipei, Taiwan, 1997.

[11] Matveeva, I. & Levow, G.-A. "Computing Term Translation Probabilities with Generalized Latent Semantic Analysis." Paper presented at the meeting of the EACL, 2006.

[12] Needleman, S. B. & Wunsch, C. D. "A general method applicable to the search for similarities in the amino acid sequence of two proteins." *Journal of Molecular Biology* 48, no. 3 (1970): 443--453.

[13] Resnik, P. "Using Information Content to Evaluate Semantic Similarity in a Taxonomy." Paper presented at the meeting of the Proceedings of the XI International Joint Conferences on Artificial, 1995.

[14] Smith, T. F. & Waterman, M. S. "Identification of common molecular subsequences." *Journal of Molecular Biology* 147 (1981): 195--197.

[15] Turney, P. D. "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL." Paper presented at the meeting of the Proc. of the Twelfth European Conference on Machine Learning (ECML-2001)

[16] Wael, H. G and Aly, A, F. Article: A Survey of Text Similarity Approaches. *International Journal of Computer Applications* 68(13):13-18, April 2013.

[17] Zipf, G., Human Behavior and the Principle of Least Effort, an Introduction to Human Ecology, Addison-Wesley, Reading, Mass., 1949.