


## MẪU BÁO CÁO CỦA MỖI HV

<b>Họ và tên (IN HOA)</b>	BUI NGUYỄN HOÀNG ANH
<b>Ảnh</b>	
<b>Số buổi vắng</b>	0
<b>Bonus</b>	22
<b>Tên đề tài (VN)</b>	ĐÁNH GIÁ ĐỘ TƯƠNG ĐỒNG CỦA HAI VĂN BẢN TIẾNG VIỆT
<b>Tên đề tài (EN)</b>	Evaluating the Similarities of Two Vietnamese Texts
<b>Giới thiệu</b>	<p><i>Hướng dẫn:</i></p> <ul style="list-style-type: none"><li>• Bài toán/vấn đề mà đề tài muốn giải quyết</li><li>• Lí do chọn đề tài, khả năng ứng dụng thực tế, tính thời sự</li><li>• Mô tả input và output, nên có hình minh họa</li></ul> <p>Xuất phát từ bài toán kiểm tra đạo văn, bài toán đánh giá giá độ tương đồng của văn bản không chỉ dừng lại ở việc kiểm tra xem văn bản có sao chép câu từ của văn bản khác hay không, mà còn đánh giá xem văn bản có sao chép đề</p>

tài, nội dung, phương pháp, giải pháp, ... hay không. Trong phạm vi của nghiên cứu này sẽ tập trung vào đánh giá độ tương đồng của hai văn bản.


Bài toán đánh giá độ tương đồng của hai văn bản có thể ứng dụng trong nhiều lĩnh vực. Một trong số đó có thể kể đến là trong xét duyệt bài báo khoa học nhằm kiểm tra bài báo có nội dung giống với bài báo khác đã được công bố hay không. Trong giáo dục phổ thông, kiểm tra xem liệu học sinh có sao chép tập làm văn của tác giả khác hoặc kiểm tra liệu câu trả lời tự luận có ý giống với đáp án. Tương tự bài toán cũng có thể ứng dụng trong các lĩnh vực khác như báo chí, thơ văn, truyền thông, ... Nhìn chung, bài toán này nhằm hỗ trợ trong công tác quản lý và bảo vệ bản quyền của tác giả đối với văn bản của họ.

Bên cạnh đó, dữ liệu mà nghiên cứu tập trung vào là các văn bản tiếng Việt. Sở dĩ như vậy là vì hiện tại nghiên cứu trên tiếng Việt vẫn còn hạn chế bởi nhiều lý do như từ đa nghĩa, từ đồng nghĩa, tiếng lóng, tiếng vùng miền, viết tắt, ... điều này đặt ra nhiều thử thách nhưng cũng là cơ hội để nghiên cứu ra những điều mới. Thêm vào đó là tinh thần dân tộc, niềm tự hào đối với ngôn ngữ mẹ đẻ - tiếng Việt, từ đó đưa tiếng Việt đến gần hơn với các nhà nghiên cứu nói riêng và bạn bè trên khắp thế giới nói chung. Cuối cùng, bên cạnh những lý do ứng dụng mà bài toán mang lại thì việc khẳng định tiếng Việt là một vùng đất màu mỡ để nghiên cứu cũng là một trong những mục tiêu mà nghiên cứu này hướng đến.

Những lý do trình bày ở trên chính là xuất phát điểm của nghiên cứu này.

Hình 1 là sơ đồ minh họa cho ý tưởng của bài toán với thông tin cụ thể như sau:

- Input: Hai văn bản A và B
- Output: Độ tương đồng tính theo tỷ lệ % của hai văn bản A và B

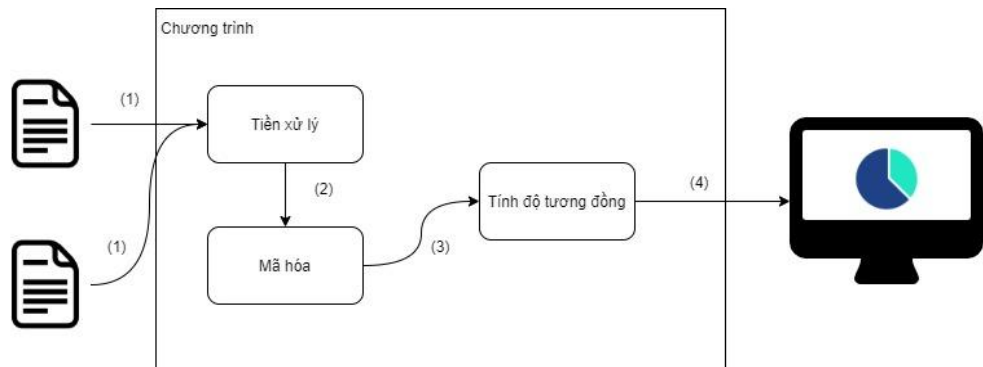
	 <p>Hình 1. Sơ đồ minh họa cho ý tưởng bài toán</p>
<b>Mục tiêu</b>	<ul style="list-style-type: none"> <li>• Trong vòng 3 ý</li> <li>• Lưu ý viết sao cho có thể đánh giá/lượng hoá được như thế nào là đạt được mục tiêu</li> </ul> <p>Mục tiêu chung:</p> <ul style="list-style-type: none"> <li>• Xác định hai văn bản đầu vào tương đồng bao nhiêu phần trăm</li> </ul> <p>Mục tiêu cụ thể:</p> <ul style="list-style-type: none"> <li>• Đề xuất thuật toán cho việc mã hoá văn bản và tính độ tương đồng của hai văn bản</li> <li>• Đề xuất các tiêu chí để xét độ tương đồng</li> <li>• Kiểm định các đề xuất bằng cách hiện thực chương trình và đánh giá kết quả thực nghiệm</li> </ul>
<b>Nội dung và phương pháp thực hiện</b>	<ul style="list-style-type: none"> <li>• Viết chi tiết các nội dung và phương pháp để đạt mục tiêu</li> <li>• Lưu ý Mục tiêu → Nội dung → Phương pháp phải có kết nối với nhau.</li> </ul> <p>Tìm hiểu hiện trạng nghiên cứu về xử lý ngôn ngữ</p> <ul style="list-style-type: none"> <li>• Việc biểu diễn hay mã hoá văn bản trong xử lý ngôn ngữ đã không còn xa lạ với các nhà nghiên cứu. Tuy nhiên, đa phần các nghiên cứu hiện nay thường sử dụng phương pháp thông dụng như túi từ hay mô hình vector để mã hoá văn bản. Trong nghiên cứu này sẽ đề xuất hướng tiếp cận mới là mã hoá văn bản thành chuỗi số thực để xử lý.</li> </ul>

Thiết kế thuật toán đánh giá độ tương đồng của hai văn bản

- Trên cơ sở lý thuyết về DWT , thuật toán được đề xuất sẽ mã hoá nội dung của hai văn bản đầu vào sau khi được tiền xử lý thành hai chuỗi tính hiệu số (dãy các số thực) tương ứng và xử lý để đem vào so sánh độ tương đồng:  $A = a_1, a_2, a_3, \dots, a_n$ ;  $B = b_1, b_2, b_3, \dots, b_n$ 
  - Các bước biến đổi dữ liệu:
    - Tiền xử lý
    - Biến đổi văn bản thành chuỗi số nguyên.
    - Chuyển chuỗi số nguyên thành số thực
    - Tách chuỗi có được ở bước trên thành hai vector (xấp xỉ và chi tiết)
- Tính khoảng cách Euclid nhỏ nhất giữa hai chuỗi tính hiệu đã được mã hoá.
- Tính độ tương đồng

Hiện thực chương trình dựa trên thuật toán

- Sơ đồ mô tả hoạt động của chương trình trong Hình 2 được giải thích như sau:
  - (1) văn bản đầu vào được tiếp nhận dưới dạng file theo các định dạng .doc, .txt, .pdf
  - (2) dữ liệu được tiền xử lý (loại bỏ hư từ, kiến đổi từ viết tắt, ...)
  - (3) dữ liệu được mã hoá chuỗi số
  - (4) tính toán độ tương đồng và hiển thị kết quả



Hình 2. Mô tả hoạt động của chương trình

### Đánh giá kết quả chạy thuật toán

- Đánh giá kết quả thực nghiệm dựa trên hai giá trị precision và recall.
- Bên cạnh đó, kết quả này cũng sẽ được dùng để so sánh với kết quả của các thuật toán khác từ đó đưa ra kết luận về tính hiệu quả của thuật toán.

### Kết quả dự kiến

- *Phần mềm ứng dụng*
- *Thuật toán,*
- *So sánh giữa các phương pháp*
- *Bộ dữ liệu, etc*

#### Phần mềm ứng dụng

- Chương trình hiện thực thuật toán đề xuất cho phép người dùng import vào hai văn bản dưới dạng file (.doc, .txt, .pdf) và trả về cho người dùng con số thể hiện mức độ tương đồng của hai văn bản.

#### Thuật toán

- Đề xuất thuật toán mã hoá văn bản
  - Thuật toán mã hoá văn bản ứng dụng DWT và bộ lọc Haar
- Đề xuất thuật toán tính độ tương đồng
  - Thuật toán tính độ tương đồng ứng dụng công thức tính khoảng cách Euclid
- Đề xuất tiêu chí xét độ tương đồng

	<ul style="list-style-type: none"> <li>○ Danh sách các tiêu chí được hình thành dựa trên cơ sở khái niệm tương đồng trong từ điển tiếng Việt</li> </ul> <p>So sánh giữa các phương pháp</p> <ul style="list-style-type: none"> <li>● Trên cùng một bộ dữ liệu, so sánh kết quả của chương trình áp dụng thuật toán đề xuất dùng DWT và thuật toán dựa trên mô hình truyền thống (túi từ và mô hình vector).</li> </ul> <p>Bộ dữ liệu</p> <ul style="list-style-type: none"> <li>● Bộ dữ liệu được xây dựng bằng cách thu thập từ các nguồn và thuộc các lĩnh vực khác nhau. Một số lĩnh vực tiêu biểu như sau: <ul style="list-style-type: none"> <li>○ Y tế</li> <li>○ Báo điện tử</li> <li>○ Bài báo khoa học</li> <li>○ Thơ, văn</li> <li>○ Bình luận và bài đăng trên mạng xã hội</li> <li>○ Đánh giá ẩm thực/ sản phẩm</li> <li>○ Thông tin về địa danh, đất nước</li> <li>○ Định nghĩa từ trong từ điển</li> </ul> </li> </ul>
<b>Tài liệu tham khảo</b>	<ul style="list-style-type: none"> <li>● <i>Theo định dạng DBLP</i></li> <li>● <i>Điền sai format sẽ bị trừ điểm</i></li> </ul> <p>[1] Hiếu, Hồ Phan, et al. "Một cách tiếp cận mới để phát hiện sự giống nhau của văn bản dựa trên phép biến đổi wavelet rời rạc." Kỷ yếu Hội nghị Khoa học Công nghệ Quốc gia lần thứ X (Fair'10), lĩnh vực Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (2017): 479-487.</p>