TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

ĐỀ CƯƠNG KHOÁ LUẬN TỐT NGHIỆP

PHÁT TRIỂN ỨNG DỤNG ĐÁNH GIÁ ĐỘ KHÓ VĂN BẢN TIẾNG Việt

(Building An Application For Measuring The Readability Of Vietnamese Text)

1 THÔNG TIN CHUNG

Người hướng dẫn:

- PGS.TS. Đinh Điền (Khoa Công nghệ Thông tin)
- Thầy Lương An Vinh (Khoa Công nghệ Thông tin)

Nhóm Sinh viên thực hiện:

1. Lý Gia Huy (MSSV: 1612271)
2. Ngô Đức Kha (MSSV: 1612277)

Loại đề tài: Ứng dụng

Thời gian thực hiện: Từ 02/2020 đến 07/2020

2 NỘI DUNG THỰC HIỆN

2.1 Giới thiệu về đề tài

Độ khó của văn bản được hiểu là tổng hợp các thành phần của tài liệu ảnh hưởng đến sự thành công mà một nhóm người đọc có với văn bản. Sự thành công là mức độ họ hiểu văn bản, đọc nó với tốc độ tối ưu và thấy nó thú vị.

Độ khó văn bản có ảnh hưởng rất lớn đến văn bản, người viết và người đọc văn bản đó. Dựa vào độ khó văn bản mà người đọc thể lựa chọn những văn bản, tài liệu phù hợp với khả năng, với người viết, họ có thể chỉnh sửa văn bản theo hướng phù hợp với nhóm độc giả mục tiêu.

Vì thế, việc xây dựng ứng dụng dùng để đánh giá độ khó văn bản có nhiều ý nghĩa với nhiều đối tượng, cụ thể là: hỗ trợ người làm giáo dục thiết kế tài liệu với độ khó phù hợp với trình độ của nhóm học sinh mục tiêu, hỗ trợ điều chỉnh độ phức tạp của văn bản luật pháp để dễ dàng phổ cập đến phần đông người dân, hỗ trợ cho các nhà máy viết các hướng dẫn phù hợp với nhóm khách hàng mục tiêu của họ,hỗ trợ tìm kiếm tài liệu cho người học ngoại ngữ...

Trên thị trường, hiện đã tồn tại công cụ tính toán độ khó văn bản, ví dụ: webfx hoặc readable. tuy nhiên các công cụ đánh độ khó văn bản hiện nay đang tính toán dựa trên các công thức, mà chúng chỉ đạt độ chính xác cao với văn bản tiếng Anh, vì thế các công cụ đó chưa thể tính toán chính xác độ khó nếu như văn bản đó được viết bằng tiếng Việt.

Vì lí do đó, đề tài hướng đến việc tạo ra công cụ có khả năng tính toán độ phức tạp với văn bản tiếng Việt.

2.2 Các công trình nguyên cứu có liên quan

Các công trình nguyên cứu về độ phức tạp văn bản đã được tiến hành từ đầu thế kỷ 20, có nhiều công thức tính độ phức tạp văn bản được công bố và độ chính xác được đánh giá tốt, một số các công thức tiêu biểu như là: Gunning fog index, Flesh reading ease, Flesh-Kincaid grade level, automated raedability index, Simple Measure of Gobbeledygook(SMOG). Tuy nhiên hầu hết trong số chúng đều lựa chọn đối tượng nguyên cứu là Anh ngữ và một vài ngôn thông dụng khác, vì thế mà tỉ lệ chính xác khi dự đoán độ khó văn bản tiếng Việt là không đủ tốt .

Còn với tiếng Việt, vào năm 1982 và 1985, hai nhà khoa học là Liem ThAnh Nguyen và Alan B. Hekin đã công bố công thức tính độ phức tạp văn bản tiếng Việt với hướng tiếp cận là mối quan hệ giữa thông kê các từ, câu và độ phức tạp văn bản, tuy nhiên nó chỉ được tiến hành trên một dữ liệu nhỏ.

Vào năm 2018, các nhà nguyên cứu An Vinh Luong, Dinh Dien, Diep Nguyen đã công bố một công thức mới được tiến hành trên một tập dữ liệu lớn hơn, tập dữ liệu gồm 1200 tài liệu được thu nhập và đánh giá, gán nhãn kỹ lưỡng bởi các chuyên gia về lĩnh vực Việt ngữ. Kết quả được sau khi kiểm thử công thức đạt trên 80%.

2.3 Mục tiêu đề tài

Mục tiêu của đề tài là xây dựng ứng dụng web và Microsoft Word Plug-in có chức năng tính toán độ phức văn bản và các thống kê trong văn bản, tính toán nhóm độc giả phù hợp với văn bản.

Các chức năng của ứng dụng bao gồm: Xác định được văn bản phù hợp với trình độ của nhóm độc giả nào(vd: số năm giáo dục, cấp bậc giáo dục,...), Xác định độ khó phù hợp với lứa tuổi nào, Tính chỉ số độ khó văn bản, Đếm số lượng câu, từ, từ khó có trong văn bản, Tỷ lệ phần trăm của từ khó, Số lượng từ trung bình của một câu.

Cho phép người dùng: nhập đường dẫn trực tiếp của trang web, hệ thống sẽ truy cập tới trang web và xác định độ khó của văn bản thể hiện trên trang web đó; nhập đoạn văn bản vào hệ thống để xác định độ khó; Người dùng sẽ nhập đoạn mã HTML có chứa văn bản, hệ thống sẽ nhận dạng được các đoạn văn bản có trong mã HTML và xác định độ khó của đoạn văn bản đó; tải lên file word để hệ thống có thể xác định độ khó từ nội dung của file.

Ngoài ra, ứng dụng cần phải đáp ứng các yêu cầu phi chức năng như sau: Tính toán và thể hiện kết quả trong khoảng thời gian cho phép, Cho ra kết quả chính xác và đáng tin cậy, Thể hiện kết quả cùng với nhiều công thức khác nhau làm tăng tính tin cậy của kết quả, Cung cấp một giao diện đơn giản, dễ sử dụng phù hợp với nhiều đối tượng người dùng.

2.4 Phạm vi đề tài

Đề tài tập trung xây dựng và phát triển ứng dụng web và Microsoft Word Plug-in có chức năng như công cụ đánh giá độ khó văn bản tiếng Việt.

Nhóm sẽ tìm hiểu nhiều hướng tiếp cận như là phân lớp bằng máy học, công thức tính toán độ phức tạp văn bản. Tiến hành cài đặt và đánh giá, so sánh các phương pháp dựa trên tập dữ liệu bao gồm 1200 mẫu tài liệu tiếng Việt được thu nhập, đánh giá, phân loại cẩn thận bởi các chuyên gia của lĩnh vực Việt ngữ. Ứng dụng sẽ cung cấp các thông số về độ phức tạp của văn bản với đầu vào bao gồm: đường dẫn đến website, đoạn văn bản, đoạn html, file word.

Kết quả trả về cho người dùng sẽ là độ phức tạp của văn bản sẽ được đánh giá thông qua các phương pháp đạt độ chính xác tốt nhất. Cùng với độ phức tạp thì ứng dụng cũng sẽ cung cấp các thông số thống kê có ảnh hưởng độ phức tạp.

2.5 Cách tiếp cận dự kiến

Hiện tại các phương pháp dùng để đánh gia độ khó văn bản dựa trên 2 hướng tiếp cận là: machine learning và công thức thống kê. Với máy học, nhóm sẽ xem xét các phương thức và thuật toán dùng phân lớp như là Support Vector Machine, naive bayes classifer, Neural network, Stochastic Gradient Descent, Linear regression,...

Với hướng tiếp cận là công thức thống kê, nhóm sẽ tìm hiểu công thức nổi tiếng và đạt hiểu quả cao trong việc đánh giá độ khó văn bản tiếng Việt như là: Gunning fog index, Flesh reading ease, Flesh-Kincaid grade level, automated raedability index, Simple Measure of Gobbeledygook(SMOG), OVIX,...

Sau đó tiến hành cài đặt lại tất cả các phương thức, kiểm thử trên tập dữ liệu văn bản tiếng Việt và đánh giá, chon lựa các phương thức đạt độ chính xác cao nhất, điều chỉnh làm tăng độ chính xác.

Ngoài ra các công thức đánh giá độ phức tạp văn bản dành cho Việt ngữ nêu ở phần trên, sẽ là được xem là cơ sở để các phương thức được đánh giá xem là có đủ tốt hay không. Sau khi tìm hiểu và chọn lọc các phương thức cho độ chính xác đủ tốt, tiến hành thiết kế hệ thống back-end, giao diện và lựa chọn các công nghệ để cài đặt ứng dụng.

Tiến hành xây dựng và phát triển ứng dụng, kiểm thử ứng dụng để đảm bảo rằng ứng dụng đáp ứng được các yêu cầu đã đề ra.

2.6 Kết quả dự kiến của đề tài

Tìm hiểu các cách tiếp cận và phương pháp khác nhau tính độ khó văn bản, thông qua đó tiến hành đánh giá và so sánh.

Cài đặt các phương pháp trên và đánh giá chúng dựa trên tập dữ liệu có sẵn.

Đưa ra nhiều kết quả với ứng với các phương pháp tốt nhất.

Xây dựng thành công ứng dụng web và Microsoft Word Plug-in như một công cụ đánh giá văn bản tiếng Việt đảm bảo các yêu cầu.

2.7 Kế hoạch thực hiện

STT	Thời gian bắt đầu	Thời gian kết thúc	Nội dung công việc
1	03/02	25/02	Tìm hiện và hoàn thiện đề cương khóa luận
2	25/02	15/03	Tìm hiểu tổng quát về các khái niệm
3	15/03	15/04	Tìm hiểu các cách tiếp cận và các phương pháp khác nhau dùng để đánh giá văn bản
4	15/03	15/04	Tiến hành cài đặt các phương pháp, kiểm thử, so sánh trên tập dữ liệu.
5	15/04	20/05	Thiết kế ứng dụng
6	20/05	15/06	Cài đặt ứng dụng
7	15/06	20/06	Kiểm thử ứng dụng

Tài liệu

[1] L. T. Nguyen and A. B. Henkin, "A readability formula for vietnamese," *Journal of Reading*, vol. 26, no. 3, 1982.

[2] L. T. Nguyen and A. B. Henkin, "A second generation readability formula for vietnamese," *Journal of Reading*, vol. 29, no. 3, 1985.

[3] D. N. A.-V. Luong and D. Dinh, "Examining the text- length factor in evaluating the readability of literary texts in vietnamese textbooks," 2017.

XÁC NHẬN
CỦA NGƯỜI HƯỚNG DẪN
(Ký và ghi rõ họ tên)

TP. Hồ Chí Minh, ngày/tháng/năm
NHÓM SINH VIÊN THỰC HIỆN
(Ký và ghi rõ họ tên)