**Tự Động Sửa Lỗi Chính Tả Tiếng Việt**

**Tại sao làm bài toán này?**

Hiện nay, các văn bản viết tay đang dần thay thế bằng hình thức đánh máy. Người đánh máy thường sẽ bị mắc lỗi chính tả khi thao tác bao gồm lỗi đánh máy, người đánh máy bị về từ ngữ, … và những lỗi này người đánh máy có thể không nhận ra là họ đã sai hoặc những lỗi này khó có thể phát hiện được. Việc phát hiện và sửa những lỗi chính tả này thường sẽ mấy rất nhiều thời gian của người đánh máy. Vậy nên, việc phát hiện và sửa lỗi chính tả trong văn bản đánh máy là điều rất cần thiết cho những người soạn thảo văn bản bằng hình thức đánh máy.

**Hướng tiếp cận:**

Đầu tiên, cần nghiên cứu các đặc điểm của Tiếng Việt, tiếp theo là tìm ra nguyên nhân của lỗi chính tả, phân loại, phát hiện và sửa lỗi chính tả. Ứng dụng RNN ( hoặc LSTM ) cho bài toán sửa lỗi chính tả.

Định nghĩa bài toán rõ ràng:

Kiểm tra chính tả của bài toán chỉ dừng lại ở mức kiểm tra âm tiết chứ không tiến đến kiểm tra ở mức từ và mức câu. Trong xử lí ngôn ngữ tự nhiên thì kiểm tra lỗi chính tả là một trong những bài toán quan trọng. Các lỗi chính tả thường gặp đó chính là sai về đánh máy và sai về từ điển. Cũng giống như các ngôn ngữ khác, kiểm tra lỗi chính tả trong Tiếng Việt bao gồm 2 giai đoạn: thứ nhất, tìm ra lỗi chính tả và thứ hai là tìm ra những từ gợi ý có liên quan nhất với từ bị lỗi đấy.

**Thách thức:**

Có nhiều chuẩn chính tả khác nhau và chƣa thống nhất chung trên cả nước trong mọi lĩnh vực, có chuẩn chính tả nhưng chưa có chuẩn chính âm.

Chúng ta cần xây dựng một tập đầu vào có độ chính xác cao hơn nữa, thể hiện rõ mối liên hệ giữa các từ hơn.Vì thế, việc chuẩn bị tập dữ liệu lớn và phong phú về mặt từ vựng là vô cùng cần thiết cho một mô hình sửa lỗi chính tả tự động tiếng Việt.

Thời gian làm sạch dữ liệu, tạo dữ liệu lỗi để thử nghiệm chương trình mất nhiều thời gian.

.

**Mục tiêu của đề tài:**

* Tìm hiểu các phương pháp phát hiện lỗi và sửa lỗi chính tả, các đặc điểm của Tiếng Việt, các lỗi chính tả thường gặp và cơ bản nhất.
* Nghiên cứu phương pháp phát hiện và sửa lỗi chính tả dựa trên từ điển và ngữ cảnh
* Nghiên cứu cơ bản.