# Language Models: Auto-Correct

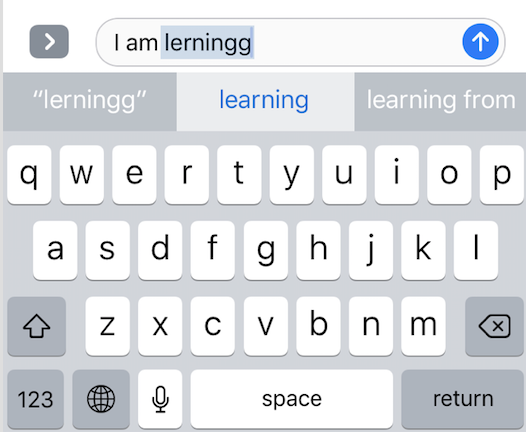
I Tổng quan

Chúng ta thường xuyên sử dụng hệ thống nhắc sửa lỗi chính tả khi nhập liệu từ bàn phím trên điện thoại hoặc trên máy tính.

Hôm nay tôi sẽ trình bày cách thức vận hành, và các nguyên lí sửa lỗi mà một hệ thống sửa lỗi cơ bản nhất có thể được áp dụng vào thực tế.

Nguồn coursera: <https://www.coursera.org/learn/probabilistic-models-in-nlp?specialization=natural-language-processing>

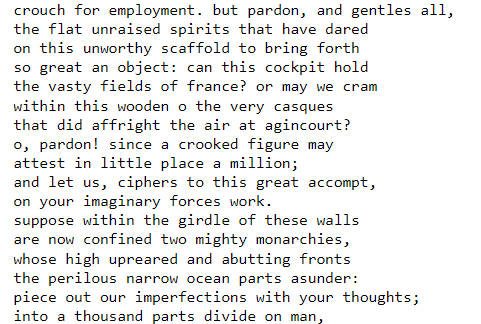
Thuật toán mà tôi đề cập lại ở đây là nền tảng nhất của AUTO-CORRECT (2007). Có thể tại thời điểm hiện tại đã có nhiều thuật toán khác, mới mẻ hơn đã ra đời.



II Triển khai

* Data (các bạn nên dùng các bộ data mà google vì google thu thập rất nhiều dữ liệu từ người dùng trình duyệt.)

Tôi đã có một bộ data nhỏ với rất nhiều câu:



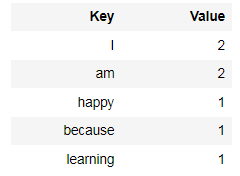
….

Bạn có thể sử dụng bất kì data nào mà bạn có, nhưng đảm bảo đủ lớn và chứa nhiều từ ngữ nhất để hệ thống hoạt động tốt nhất, khái quát nhất với ngôn ngữ tự nhiên.

Tiếp theo, tôi tách toàn bộ data của tôi thành một bộ token, với mỗi token sẽ là 1 từ duy nhất, đồng thời đi kèm là số lần xuất hiện của từ đó trong bộ data thô ban đầu.

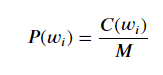
Ví dụ tôi có một câu: **"I am happy because I am learning"**

**Kết quả mong muốn:**



= > chúng ra đã chuẩn bị xong data

* Xác suất:

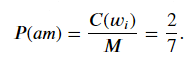


Tử số: Số lần xuất hiện của từ đó trong bộ data gốc chứa nhiều câu ở trên.

Mẫu số: Số lượng từ trong bộ data gốc đó.

Ví dụ: **'I am happy because I am learning'**

**Xác suất của “am” là 2/7**



Chúng ta sẽ sử dụng cách làm này trong xuyên suốt của bài toán.

* Các trường hợp mà khi người dùng nhập văn bản thường mắc phải, cần sửa lỗi cho đúng:

+ Người dùng nhập chữ bị dư: whatt -> what

+ Người dùng nhập chữ bị lộn xộn: hate -> ehat

+ Người dùng nhập chữ bị sai: smim -> swim

+ Người dùng nhập thiếu từ: whch -> which

* Chúng ta cần tạo ra 4 hàm tương ứng với 4 trường hợp trên:

+ delete\_letter: nhận một từ, nó trả về tất cả các chuỗi có thể đã loại bỏ một ký tự.

+ switch\_letter: nhận một từ, nó trả về tất cả các chuỗi có thể có hai chữ cái liền kề được chuyển đổi.

+ replace\_letter: nhận một từ, nó trả về tất cả các chuỗi có thể có một ký tự được thay thế bằng một chữ cái khác.

+ insert\_letter: nhận một từ, nó trả về tất cả các chuỗi có thể có một ký tự bổ sung được chèn vào.