Đại Học Bách Khoa Hà Nội

Viện Công Nghệ Thông Tin và Truyền Thông

BÁO CÁO BÀI TẬP LỚN

Đề tài : Phát hiện và sửa lỗi văn bản

GVHD: TS. Nguyễn Nhật Quang

Mã lớp : 126805

Thành viên

Lương Đào Quang Anh	20176687
Hoàng Trung Hiếu	20176756
Đỗ Thị Hồng Thảo	20176878

Hà Nội, tháng 1 năm 2022

MỤC LỤC

1. Đặt vấn đề

Trong thời đại công nghệ phổ biến, xã hội ngày càng phát triển, ở tất cả mọi lĩnh vực người ta đều hướng đến cái chỉn chu, tốt đẹp nhất. Và thứ cần chỉn chu hàng ngày nhiều nhất có lẽ chính là các văn bản, một văn bản trình bày đẹp, bố cục rõ ràng nhưng sai chính tả là điều không chấp nhận được. Bài toán sửa lỗi chính tả ra đời.

Sửa lỗi chính tả (Spell Correction) là một trong những bài toán cơ bản nhất trong xử lý ngôn ngữ tự nhiên. Bài toán này có ứng dụng trong các trình soạn thảo văn bản, nhập liệu, nhận dạng... giúp người viết có thể phát hiện ra lỗi chính tả trong văn bản của mình. Với việc viết văn bản trên điện thoại di động rất dễ sinh ra lỗi, tính năng tự động sửa lỗi chính tả là thành phần không thể thiếu trong bất cứ bàn phím nào. Các kỹ thuật spell correction đã rất phát triển và hoạt động rất tốt với nhiều ngôn ngữ, nhất là tiếng Anh, nhưng với tiếng Việt thì lại chưa thực sự mạnh mẽ.

Chính vì những lý do trên nhóm quyết định chọn đề tài sửa lỗi chính tả trong văn bản tiếng Việt để làm bài tập lớn cho học phần xử lý ngôn ngữ tự nhiên lần này.

2. Cơ sở lý thuyết

2.1. Định lý Bayes

Định lý Bayes (Bayes' Theorem) là một định lý toán học để tính xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra. Định lý này đặt theo tên nhà toán học Thomas Bayes, người Anh sống ở thế kỷ 18. Đây là một trong những công cụ vô cùng hữu ích, người bạn thân của các Data Scientist, những người làm trong ngành khoa học dữ liệu.

Công thức định lý Bayes

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} = \frac{P(H \cap E)}{P(E)}$$

Trong đó:

- P(H|E) : xác suất H xảy ra khi biết E xảy ra
- P(H), P(E) : xác suất xảy ra sự kiện H, E
- P(E|H) : xác suất E xảy ra khi biết H xảy ra
- P($H \cap E$) : xác suất kết hợp sự kiện H và E

2.2. Mô hình N-grams

2.2.1. Mô hình ngôn ngữ

Là phân bố xác suất trên tập các văn bản, cho biết xác suất của một câu (hoặc 1 cụm từ) thuộc 1 ngôn ngữ là bao nhiêu. Mô hình ngôn ngữ tốt sẽ đánh giá đúng các câu đúng ngữ pháp, trôi chảy hơn các từ có thứ tự ngẫu nhiên

Ví dụ: P(“hôm nay trời đẹp") > P(“trời đẹp nay hôm")

2.2.2. Mô hình ngôn ngữ n-gram

Mục tiêu: Xác định xác suất của 1 câu hoặc một cụm từ:

P(W) = P(w1, w2, w3, w4, w5,...,wm)

Theo công thức Bayes:

P(A, B) = P(B|A)*P(A)

Ta có:

P(w1,w2,w3,w4,w5,...,wm)

= P(w1) * P(w2|w1) * P(w3|w1w2) * … * P(wm|w1w2w3…wm-1)

Ví dụ:

P(“hôm nay trời đẹp") = P(hôm) * P(nay|hôm) * P(trời|hôm nay) * P(đẹp|hôm nay trời)

- **Mô hình bigram**

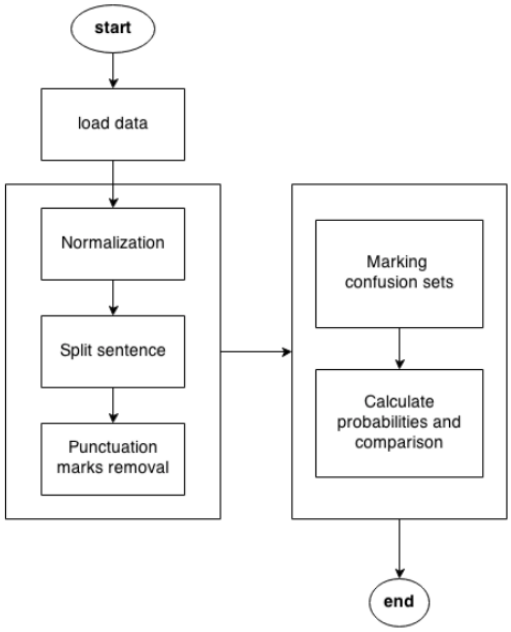
$$P(w_1 w_2 \dots wn) \sim \prod_i P(w_i | wi\_1)$$

Để đơn giản hóa, sử dụng đánh giá Maximum Likelihood

$$P(w_i \mid w_{i-1}) = \frac{count(w_{i-1}, w_i)}{count(w_{i-1})}$$

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

- **Mô hình triển khai chung của N-gram**



2.3. Cơ sở tri thức

Tập luật để sửa lỗi

- Thêm phụ âm còn thiếu trong từ
  - Ví dụ: sih - sinh, báh - bánh, cah - canh
- Xóa các ký tự lặp lại
  - Ví dụ: biinh -> binh, bangg -> bang
- Đổi chỗ ký tự bị sai thứ tự
  - Ví dụ: nếu gặp các cặp kí từ 'hn','hc','ig','hg','hk','gn','ht','rt','uq','hp','êi','êy','ôu','ơư','ei','ey','ou','ơư','ơu','aư','âư' thì đổi chỗ lại cho nhau
- Xử lý các lỗi với y và i
  - Ví dụ: các vần chứa i nhưng không tồn tại với y vd: ưy => ưi, ym=>im và ngược lại
- Thiếu dấu phụ
  - Ví dụ: hươu -> hươu, yeu -> yêu, tủu -> tửu
- Sai vị trí dấu trong từ, sai bộ gõ
  - Ví dụ: hoà thuận -> hòa thuận, hiếu thảo -> hiếu thảo
- Các lỗi chính tả với ngh,ng,g,gh,... (đứng đầu câu):
  - Ví dụ: ngiêng -> nghiêng, gê -> ghê
- Viết rõ cho các kí tự hay được viết tắt
  - Ví dụ: fải => phải
- Thay 1 ký tự bất kỳ bằng 1 ký tự gần đó trên bàn phím
  - Ví dụ: xin chào=> xun chào
- Lỗi telex

    - Ví dụ: xin chào=> xin chaof

- Lỗi VNI
    - Ví dụ: xin chào=> xin chao6

- Thay thế các từ teencode
    - Ví dụ: mk -> mình, ck -> chồng, vk -> vợ

Tuy nhiên, nhóm không sử dụng các luật

- Thay 1 ký tự bất kỳ bằng 1 ký tự gần đó trên bàn phím
- Lỗi VNI

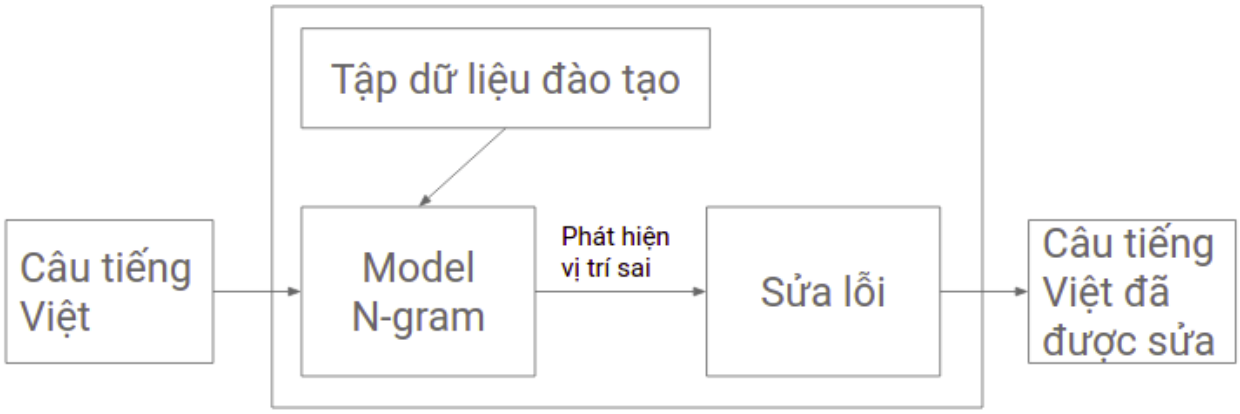
Do các lỗi này phức tạp và với thời gian môn học không đủ để thực hiện áp dụng hết các luật này

Phương pháp suy diễn: 1 từ được dự đoán là sai sẽ được đưa qua các luật có thứ tự như sau: Lỗi teencode -> lỗi viết tắt -> lỗi lặp từ -> lỗi sai thứ tự-> lỗi thiếu dấu phụ -> lỗi y i -> lỗi dùng nhầm n ng -> lỗi telex

Thứ tự đưa vào rule như trên vì nếu rule trước từ đó sai và được sửa lỗi cũng sẽ không ảnh hưởng đến lỗi của từ tiếp theo.

3. Định hướng giải pháp

Mô hình tổng quan triển khai



3.1. Model N-gram

Các bước thực hiện

- Dữ liệu đào tạo được chuẩn hóa và tách thành các từ (token)
- Với mỗi token sẽ tiến hành đếm các phần tử token ngay sau nó
- Sau đó tiến hành tính toán xác suất bigram

Tại bước này sẽ thu được các vị trí sai của 1 câu

3.2. Sửa lỗi

Pha này nhận đầu vào của model n-gram là các vị trí từ sai. Với mỗi từ sai sẽ lần lượt đi qua các luật để tiến hành sửa lỗi.

Tại bước này sẽ thu được từ sửa lại của mỗi từ đứa vào

Kết quả trên toàn hệ thống : Từ 1 câu tiếng Việt sẽ thu được 1 câu tiếng Việt đã được sửa

4. Thử nghiệm

4.1. Dữ liệu

- Dữ liệu đào tạo :
    + Mỗi câu là một dòng trong file dữ liệu đào tạo
    + Số lượng câu : 1.449.198câu
    + Đặc điểm các câu : Không chưa các từ viết tắt.

```
Nếu có phù dâu, bạn có thể chọn hoa cùng màu váy phù dâu để tổng thể đám cưới đồng điệu.
Hoặc bó hoa trắng sẽ là màu sắc trung tính nhất bởi màu trắng có thể hợp với mọi chiếc váy màu khác nhau.
Khi cầm hoa, cô dâu nên chú ý hạ thấp hoa xuống dưới eo, để các chi tiết váy đẹp không bị che khuất.
Hoa cưới hợp với vóc dáng - Cô dâu dáng tròn nên tránh bó hoa to tròn.
Kiểu hoa phù hợp hơn cả hoa dáng hơi dài, chiều dài hoa phụ thuộc vào chiều cao của cô dâu.
Cô dâu dáng cao thanh mảnh hợp với bó hoa tròn nhưng kích thước vừa phải, vừa cân đối, vừa mang đến cảm giác đáng yêu.
Cô dâu dáng người to nên chọn bó hoa kết từ những loại hoa to để cân đối với cơ thể.
Tổng thể bó hoa cầm tay cũng phải vừa người, không nên quá nhỏ, không tạo được ấn tượng.
Chọn hoa theo sắc màu từng mùa - Mùa xuân, hoa cưới thường nhẹ nhàng như không gian.
Mùa hạ, sắc màu hoa rực rỡ hơn như nắng hè.
Mùa thu, sắc màu hoa dần trở nên ấm áp, lãng mạn hơn.
Nhiều cô dâu yêu thích sắc tím hay màu cam ấm cho bó hoa cưới để phù hợp với đất trời.
Hoa cưới mùa đông lại toát lên sự sang trọng hoặc nồng nhiệt, ấn tượng.
```

- Dữ liệu thử nghiệm
    - Tập dữ liệu thử nghiệm
        + Gồm 3 file : Mỗi file gồm 100 câu
            + File thứ nhất chứa các câu bị đánh máy sai

```
Lưật đã chỉ tō, khi có sự tác động rừ bên ngoài, trận đầu phải được dừng lại.
Tới đã cố gắng để bảo cệ các trọng tài nhưng trong trường hợp này, yất cả mọi người sẽ cười vào mũi họ.
Nhiều khả năng, trọng tài sẽ bị treo còi yrong loạt đấu duối tuần này và một số trận sau đó.
Nhiều người bất dầu lo ngại về việc zẽ có một dợt điều chỉnh sâu trong những tuần tới.
Có ba yếu tố làm cho ngưỡng này trở bên wuan trọng trong tuần.
Với một sự hội tụ của khá nhiều yếu tố như vậy, khả năng phá vỡ ngưỡng chống dỡ này là khá rhấp.
Trong phân tích kỹ thuật, các yếu tố trung và sài hạn luôn được ưu tiên hơn những ayj điều dhỉnh ngắn hạn.
Xu hướng chính của thị trường sẽ áp đảo những diều chỉnh thứ vấp.
Một tín uiệu như vậy thể giện sự tích lũy lâu dài và chắc chắn của thị trường.
Hầu như tất cả các yếu rố ngắn hạn, trung yạn và dài bạn đều đang hỗ trợ eất tốt cho thị trường.
Thành công jhờ cộng sự tốt.
```

            + File thứ hai là ma trận 0, 1 của File thứ nhất : Mỗi từ sẽ được gán nhãn 0, 1 với 0 là từ đó viết sai, 1 là từ được viết đúng

```
0 0 0 0 0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 1 0 0
0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0
0 0 0 0 0 1 1 0 0 0 0 0 0 0 0
0 1 0 1 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 1 0 0 1 0 0 0 0 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0 0 1 0
0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 1 0
0 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0
0 1 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0
0 0 0 0 0 0 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0
1 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0
0 0 0 0 1 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0
0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0
1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 0 0
0 0 0 0 1 0 0 0 0 0 0 1 0 0
```

+ File thứ ba chưa các câu tương ứng nhưng đã được sửa các từ sai thành các từng đúng

```
Luật đã chỉ rõ, khi có sự tác động từ bên ngoài, trận đấu phải được dừng lại.
Tôi đã cố gắng để bảo vệ các trọng tài nhưng trong trường hợp này, tất cả mọi người sẽ cười vào mũi họ.
Nhiều khả năng, trọng tài sẽ bị treo còi trong loạt đấu cuối tuần này và một số trận sau đó.
Nhiều người bắt đầu lo ngại về việc sẽ có một đợt điều chỉnh sâu trong những tuần tới.
Có ba yếu tố làm cho ngưỡng này trở nên quan trọng trong tuần.
Với một sự hội tụ của khá nhiều yếu tố như vậy, khả năng phá vỡ ngưỡng chống đỡ này là khá thấp.
Trong phân tích kỹ thuật, các yếu tố trung và dài hạn luôn được ưu tiên hơn những sự điều chỉnh ngắn hạn.
Xu hướng chính của thị trường sẽ áp đảo những điều chỉnh thứ cấp.
Một tín hiệu như vậy thể hiện sự tích lũy lâu dài và chắc chắn của thị trường.
Hầu như tất cả các yếu tố ngắn hạn, trung hạn và dài hạn đều đang hỗ trợ rất tốt cho thị trường.
Thành công nhờ cộng sự tốt.
```

- Nguồn dữ liệu
    + Xây dựng trang web thu thập data: https://web-dataset.herokuapp.com/

4.2. Kết quả thử nghiệm

- Độ chính xác
    - Phát hiện lỗi: 91.88%
    - Sửa lỗi: 69.75%

5. Kết luận

Mô hình nhóm xây dựng có tỷ lệ phát hiện lỗi khá cao. Tuy nhiên, việc sửa lỗi vẫn còn nhiều khó khăn do người dùng gõ sai, do ngữ cảnh của câu. Ngoài ra, pha sửa lỗi không áp dụng các lỗi như là “Thay 1 ký tự bất kỳ bằng 1 ký tự gần đó trên bàn phím”, “Lỗi VNI”. Do đó , tỷ lệ sửa đúng chưa cao.

Hướng phát triển trong tương lai của nhóm. Tiến hành thu thập thêm nhiều dữ liệu để nâng cao tỷ lệ phát hiện lỗi. Ngoài ra, áp dụng thêm P của N-gram trong việc sửa lỗi có sử dụng ngữ cảnh

6. Tài liệu tham khảo

- Mô hình N-grams : https://web.stanford.edu/~jurafsky/slp3/3.pdf
- Dữ liệu PhoBert : https://drive.google.com/drive/folders/1En5uILZJUJrwa-BcFdwEWSWz60xrDeGY
- https://github.com/huynhnhathao/vietnamese_spelling_error_correction
- Zalo Tech, Tự động sửa lỗi chính tả tiếng Việt, đăng tải ngày 12/01/2018 trên medium.com
- Nguyen Thi Xuan Huong, Tran-Thai Dang, The-Tung Nguyen và Anh-CuongLe,
- University of Engineering and Technology Vietnam National University, Hanoi, “Using Large N-gram for Vietnamese Spell Checking”, 2015
- https://github.com/google/sentencepiece