ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

BÁO CÁO ĐỒ ÁN

XỬ LÝ NGÔN NGỮ TỰ NHIÊN

ĐỀ TÀI

GÁN NHÃN TỪ LOẠI TIẾNG VIỆT

Giảng viên hướng dẫn: ThS. Nguyễn Trọng Chỉnh

Sinh viên thực hiện:		Tỉ lệ đóng góp
Mai Duy Ngọc	20520654	100%
Trần Đăng Khoa	20520589	100%
Đào Danh Đăng Phụng	20520699	100%

TP. Hồ Chí Minh, tháng 1 năm 2023

LỜI CẢM ƠN

Lời cảm ơn của sinh viên gửi đến thầy cô giảng viên.

Thành phố Hồ Chí Minh, tháng 1 năm 2023

Nhóm sinh viên thực hiện

NHẬN XÉT CỦA GIẢNG VIÊN

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

MỤC LỤC

DANH MỤC BẢNG

DANH MỤC HÌNH ẢNH

PHẦN MỞ ĐẦU

1. Lý do chọn đề tài

Part of speech (POS) tagging là một trong những phương pháp xử lí quan trọng và là nền tảng của xử lí ngôn ngữ tự nhiên. Với sự ngày càng phát triển của ngôn ngữ, mỗi từ trong câu có thể gắn với nhiều từ loại và việc hiểu đúng nghĩa của từ đó phụ thuộc vào ta có xác định chính xác từ loại của từ đó hay không. Nhận thấy tầm quan trọng của việc phải xác định đúng từ loại trong câu, vì thế nhóm sinh viên đã lựa chọn đề tài này làm báo cáo cuối kỳ đồng thời trong tương lai sẽ tiếp tục phát triển đề tài này ở môn học tiếp theo.

2. Mục tiêu nghiên cứu

Đề tài nghiên cứu nhằm giải quyết 2 mục tiêu sau:

- Thực hiện tách từ trên bộ dữ liệu đã thu thập.
- Xây dựng mô hình gán nhãn từ loại.

3. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu: Thực hiện trên cơ sở ngôn ngữ là tiếng Việt.

Phạm vi nghiên cứu: Bộ dữ liệu được sưu tầm từ các bài báo, đoạn văn ngắn, các câu nói nổi tiếng trên mạng xã hội, lời bài hát.

4. Phương pháp nghiên cứu

Tách từ: Sử dụng thuật toán So khớp cực đại để tách từ.

Gán nhãn từ loại: Kết hợp mô hình HMM (Hidden Markov Model) cùng với với thuật toán Viterbi.

5. Cấu trúc báo cáo

Chương 1: Giới Thiệu

Chương 2: Dữ Liệu Và Bộ Từ Điển

Chương 3: Tách Từ

Chương 4: Ngữ Liệu

Chương 5: Gán Nhãn Từ Loại

Chương 6: Kết Luận

Chương 1: Giới Thiệu

1.1. Bài toán tách từ tiếng Việt

1.1.1. Khái quát tiếng Việt

1.1.1.1. Tiếng (tiếng một)

Là đơn vị cơ bản trong tiếng Việt, được phát ra một hơi, nghe thành một tiếng và có mang một thanh điệu nhất định. Mỗi tiếng được viết rời theo khoảng trắng hay các dấu ngắt.

1.1.1.2. Từ

Có hai quan điểm:

- Mọi tiếng đều là từ.
- Tiếng chưa hẳn là từ (đa số):
  - Tiếng là đơn vị nhỏ hơn hoặc bằng hình vị: đo đỏ, lác đác,…
  - Hình vị cấu tạo nên từ.

Tiêu chuẩn của từ:

- Về hình thức:
  - Tính cố định: không thể chêm-xen được.
  - Tính độc lập: khả năng kết hợp tự do hay hạn chế.
  - Tính từ loại và quan hệ cú pháp
- Về nội dung:
  - Chức năng định danh.
  - Biểu thị khái niệm.
  - Ý nghĩa biểu niệm.
  - Hoàn chỉnh về nghĩa.

Các dạng từ:

- Từ đơn (từ đơn tiết). Ví dụ: nhà, trời, mây, …
- Từ ghép gồm:
  - Từ ghép đẳng lập. Ví dụ: ăn ở, đi lại, …
  - Từ ghép chính phụ. Ví dụ: tàu hỏa, sân bay, …
- Từ láy.
- Từ ngẫu hợp:
  - Thuần Việt: bồ câu, bồ hòn, …
  - Gốc Hán: mâu thuẫn, hàn lâm, …
  - Từ vay mượn: a-xít (acid), xà phòng (sapoo)

1.1.1.3. Nhận diện từ ghép

Đặc điểm của từ ghép:

- Thành tố trực tiếp của từ ghép đa số là những tiếng không độc lập.

- Khó tách rời trong những văn cảnh bình thường.
- Nghĩa của từ ghép hiếm khi được suy ra từ nghĩa của các thành tố trực tiếp.

Từ ghép khác với thành ngữ, quán ngữ, tục ngữ.

1.1.1.4. Từ loại tiếng Việt

Gồm các từ loại là: Danh từ, động từ, tính từ, số từ, lượng từ, phó từ, đại từ, chỉ từ, quan hệ từ, trợ từ, thán từ và tình thái từ.

1.1.2. Tổng quan về bài toán tách từ tiếng Việt

Tách từ là một quá trình xử lí nhằm mục đích xác định ranh giới của các từ trong câu văn, cũng có thể nói tách từ chính là quá trình xác định từ đơn, từ ghép, … có trong câu. Đây là bước cơ bản trước khi gán nhãn từ loại hay các vấn đề xử lí ngôn ngữ phức tạp khác.

1.1.3. Cách tiếp cận bài toán

Hiện tại có một số cách tiếp cận như sau:

- Ghép cực đại: Đặt các từ vào câu sao cho phủ hết được câu đó, thoả mãn một số heuristic nhất định. Phương pháp này các ưu điểm là rất nhanh, nhưng có rất nhiều hạn chế, ví dụ như độ chính xác thấp, không xử lý được những từ không có trong từ điển.
- Luật: Xây dựng tập luật bằng tay hoặc tự động để phân biệt các cách kết hợp được phép và không được phép.
- Đồ thị hoá: Xây dựng một đồ thị biểu diễn câu và giải bài toán tìm đường đi ngắn nhất trên đồ thị.
- Máy học: Coi như bài toán gán nhãn chuỗi. Cách này được sử dụng trong JVNSegmenter, Đông du.
- Dùng mô hình ngôn ngữ: Cho trước một số cách tách từ của toàn bộ câu, một mô hình ngôn ngữ có thể đánh giá được cách nào có khả năng cao hơn. Đây là cách tiếp cận của vnTokenizer.

Trong phạm vi đề tài này, nhóm nghiên cứu sẽ tiếp cận bài toán theo phương pháp ghép cực đại vì phương pháp này tương đối đơn giản, phù hợp với tài nguyên lẫn tài nguyên của nhóm nghiên cứu.

1.2. Bài toán gán nhãn từ loại tiếng Việt

1.2.1. Tổng quan về bài toán gán nhãn từ loại tiếng Việt

Gán nhãn từ loại tiếng Việt là việc xác định chức năng ngữ pháp của một từ trong câu, từ đó làm rõ nghĩa của từ, tránh các trường hợp nhập nhằng cũng như thể hiện khả năng kết hợp của các từ trong câu.

1.2.2. Cách tiếp cận bài toán

Có hai tiếp cận chính cho bài toán gán nhãn từ loại:

- Tiếp cận có hướng dẫn: Bộ gán nhãn theo hướng dẫn có đặc thù là dựa trên kho ngữ liệu đã được gán nhãn cho việc tạo ra các công cụ được sử dụng cho quá trình gán nhãn: từ điển bộ gán nhãn, các tần suất từ/nhãn, các xác suất chuỗi nhãn, tập các luật. Đây cũng là hướng tiếp cận chính của đề tài này trong việc gán nhãn từ loại.
- Tiếp cận không hướng dẫn: Các mô hình không hướng dẫn không yêu cầu kho ngữ liệu đã gán nhãn nhưng lại sử dụng các thuật toán tính toán phức tạp để tự động xây dựng các nhóm từ (nghĩa là xây dựng các tập nhãn).

Chương 2: Dữ Liệu Và Bộ Từ Điển

2.1. Dữ liệu

2.1.1. Tổng quan về bộ dữ liệu

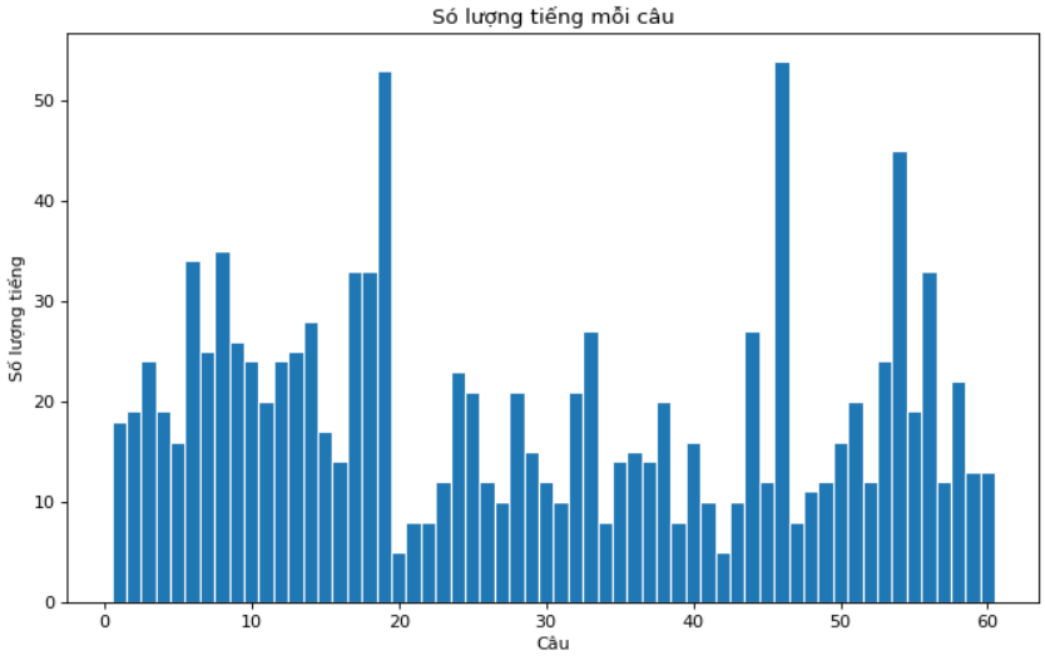
Bộ dữ liệu được thu thập từ nhiều nguồn khác nhau với đa dạng các chủ đề. Để đảm bảo tính đa dạng và bao quát của bộ dữ liệu, nhóm nghiên cứu đã sưu tầm từ các trang báo, lời bài hát, câu trả lời ứng xử của thí sinh trong các cuộc thi và một số câu để đánh giá trường hợp nhập nhằng.

```
Từ những vết sẹo của bạo lực gia đình, đã có lúc, tôi đau đớn, tôi gục ngã.
Nhưng điều quan trọng nhất là tôi tìm thấy sức mạnh nội tại và tôi chiến đấu vì nó.
Ai trong chúng ta cũng có những cuộc chiến cá nhân nhưng điều quan trọng nhất là tin vào giá trị của bản thân.
Đội vương miện chính mình, viết tiếp cuộc đời bằng trang hành trình nhiệt huyết, bản lĩnh cá nhân.
Mỗi chúng ta là một chiến binh và hãy chiến đấu vì giấc mơ của mình.
Sân nhà mình hồi ấy có rộng mấy đâu, chỉ có khoảng trời là lồng lộng phía trên đầu, nhưng đã đi hết cả tuổi thần tiên rồi, sao tôi vẫn
còn nhớ tiếc.
Tháng năm vừa rồi bánh cam đã lọt vào top 30 món rán ngon nhất thế giới do hãng tin danh tiếng CNN bình chọn.
Hồi đó má hay cho tiền dằn túi thì mình lại dằn vặt, vì mình không biết nên chọn giữa ổ bánh mì đầy thịt hay là một ly trà sữa thơm
ngon đầy thạch.
Người ta ngây ngất trước sự hào nhoáng, mê mẩn trước sự bóng bẩy nhưng chỉ rơi nước mắt trước sự giản dị tự đáy lòng.
Vô cảm, thờ ơ hay ích kỷ là cảm xúc, mà cảm xúc thì tồn tại trong tất cả, không riêng gì giới trẻ.
Điều quan trọng là thay vì chỉ trích giới trẻ, hãy đặt câu hỏi tại sao họ lại như vậy?
Hãy giáo dục, cho họ thấy rằng gia đình là nơi nung nấu và nguồn cội giúp họ ngày càng được dâng trào hơn.
Hiện nay có rất nhiều sinh viên ra trường nhưng không xin được việc làm, trong khi đó nhiều doanh nghiệp khó tuyển nhân sự.
Vấn đề này nằm ở việc chúng ta làm thế nào trong tương lai, nhiều hơn là đổ lỗi cho bất kỳ cá nhân, tổ chức nào. *
Kết thúc lượt của bạn và buộc người chơi kế tiếp phải bốc thêm 2 lượt nữa.
Đánh lá này xuống để cướp một lá ngẫu nhiên của người chơi khác.
Tôi chọn cách trò chuyện với thế giới chung quanh bằng những trang viết, bỗng thấy thật dễ chịu khi suy nghĩ, niềm vui, nỗi buồn của
mình được mọi người chia sẻ.
Tại UIT, em có cơ hội để hòa nhập vào một môi trường năng động, sáng tạo và chuyên nghiệp, được học tập từ các thầy cô giàu kinh
nghiệm và tận tâm.
Môn học nhằm cung cấp cho sinh viên một số kiến thức nhập môn của chuyên ngành xử lý ngôn ngữ tự nhiên, bao gồm những nội dung chính
về: văn phạm phi ngữ cảnh, văn phạm DCG, cài đặt và giải thích cơ chế xử lý văn phạm DCG trên Prolog, FSA.
Học sinh học sinh học.
Cuối cùng, tôi không thể thốt lên lời.
Em sai rồi, anh xin lỗi em đi.
```

Hình 1. Bộ dữ liệu

2.1.2. Phân tích bộ dữ liệu

Bộ dữ liệu được thu thập với tổng cộng 60 câu, với số lượng các tiếng mỗi câu là khác nhau.



Hình 2. Số lượng tiếng mỗi câu

Trong đó, 2 câu có số lượng tiếng ít nhất là 5 tiếng:

- Học sinh học sinh học.
- Quả thơm đó thơm thật.

Và 1 câu có số lượng tiếng nhiều nhất là 54 tiếng:

- Triết học là bộ môn nghiên cứu về những vấn đề chung và cơ bản của con người, thế giới quan và vị trí của con người trong thế giới quan, những vấn đề có kết nối với chân lý, sự tồn tại, kiến thức, giá trị, quy luật, ý thức, và ngôn ngữ.

Ngoài ra, các câu còn chứa nhiều từ viết tắt, không phải là tiếng Việt,… như UIT, FSA, CNN,…

2.3. Bộ từ điển

Vì mục đích thống nhất kết quả của 2 bài toán, nhóm chỉ sử dụng duy nhất một bộ từ điển đã qua xử lý cho phù hợp với nhu cầu riêng cho từng bài toán. Bộ từ điển được nhóm tham khảo trên internet, với số lượng các từ là 55070. Bộ từ điển bao gồm số, các kí hiệu đặc biệt, tiếng Việt và tiếng nước ngoài.

```
thư thái
thụ thai
thu thanh
thủ thành
thu thập
thú thật
thủ thế
thủ thỉ
thú thiệt
thư thoại
thủ thư
thủ thuật
thú thực
thư tịch
thủ tiêu
thư tín
thư tín điện tử
thư tín dụng
thú tính
thụ tinh
thụ tinh nhân tạo
```

Hình 3. Bộ từ điển dành cho tách từ

```
thư_thái
thụ_thai
thu_thanh
thủ_thành
thu_thập
thú_thật
thủ_thế
thủ_thỉ
thú_thiệt
thư_thoại
thủ_thư
thủ_thuật
thú_thực
thư_tịch
thủ_tiêu
thư_tín
thư_tín_điện_tử
thư_tín_dụng
thú_tính
thụ_tinh
thụ_tinh_nhân_tạo
```

Hình 4. Bộ từ điển dành cho gán nhãn

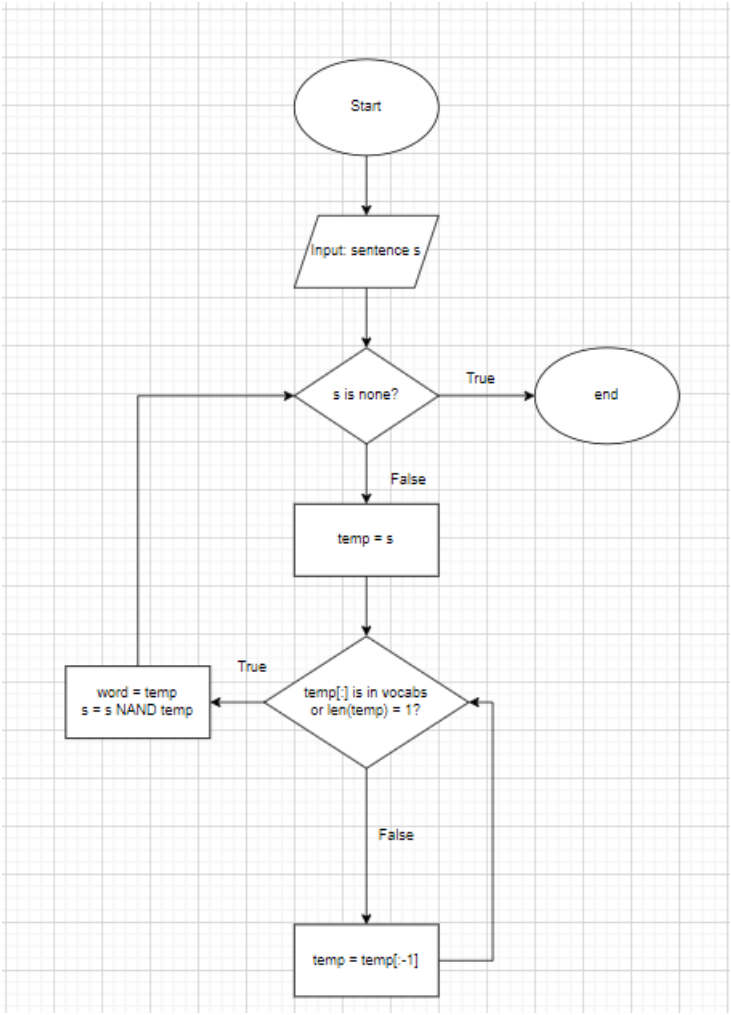
Chương 3: Tách Từ

3.1. Thuật toán so khớp cực đại

3.1.1. Giới thiệu thuật toán

Đây là một trong những phương pháp tiếp cận dựa trên từ điển. Ý tưởng của thuật toán dựa trên tư tưởng tham lam. Từ một bộ từ điển sẵn có, thuật toán sẽ bắt đầu so khớp tất cả các tiếng từ trái sang phải. Tiếng đầu tiên dài nhất và khớp với bộ từ điển sẽ được tính là một từ. Thuật toán sẽ dừng lại khi ta xét hết tất cả các tiếng trong câu.

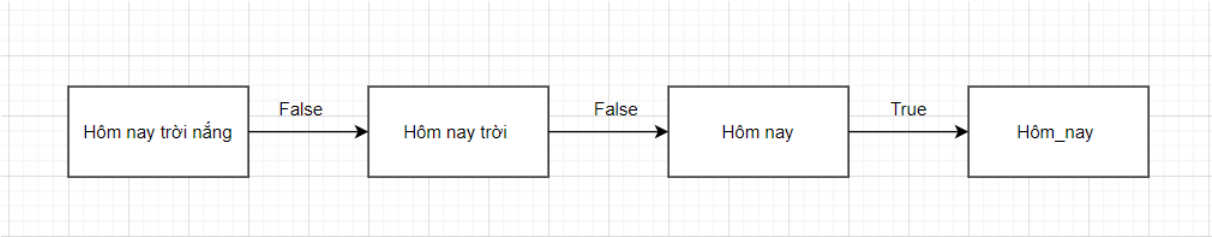
3.1.2. Lưu đồ thuật toán



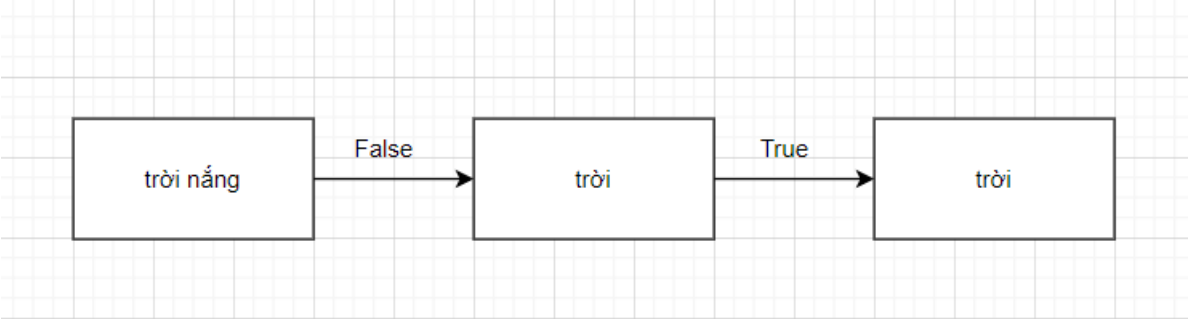
Hình 5. Lưu đồ thuật toán so khớp cực đại

3.1.3. Ví dụ

Xét đầu vào của thuật toán là câu: Hôm nay trời nắng



Hình 6. Lần xét đầu tiên của thuật toán



Hình 7. Lần xét thứ hai của thuật toán

Output sau 2 lần lặp lại của thuật toán sẽ là: Hôm_nay trời nắng

3.1.4. Đánh giá thuật toán

3.1.3.1. Ưu điểm

Thuật toán đơn giản, dễ cài đặt, không yêu cầu vào dữ liệu đầu vào. Thời gian tính toán nhanh chóng, không tốn nhiều tài nguyên.

3.1.3.2. Nhược điểm

Phụ thuộc hoàn toàn vào bộ từ điển và chưa thể giải quyết được các trường hợp nhập nhằng.

```
check = tokenizer("Học sinh học sinh học",dict)
check
```

```
['Học_sinh', 'học_sinh', 'học']
```

Hình 8. Trường hợp nhập nhằng

3.1.4 So sánh kết quả với thư viện VnCoreNLP

Nhóm nghiên cứu đã sử dụng thư viện VnCoreNLP để thực hiện so sánh với thuật toán đã xây dựng.

```
Tokenizing của VNCorelp: [['Em', 'sai', 'rồi', ',', 'anh', 'xin_lỗi', 'em', 'đi', '.']]
Tokenizing của thuật toán: ['Em', 'sai', 'rồi', ',', 'anh', 'xin_lỗi', 'em', 'đi', '.']
```

Hình 9. Một số kết quả so sánh giữa hai phương pháp

Để đánh giá được khách quan hơn, nhóm nghiên cứu đã thực hiện tách từ thủ công trên tập dữ liệu, lấy đó làm chuẩn để đối sánh hai phương pháp.

	Longest Matching	VnCoreNLP
Accuracy	0.892683	0.88499
Precision	0.790262	0.775281
Recall	0.796226	0.781132
True Positive	211	207
True Negative	704	701
False Positive	56	60
False Negative	54	58

Hình 10. Tổng hợp kết quả tách từ

Từ bảng so sánh kết quả trên, nhóm nhận định rằng thuật toán So khớp cực đại mà nhóm cài đặt cho kết quả tương đối ổn định. Lí do là bộ từ điển khá rộng, tuy nhiên, do dữ liệu chỉ có 60 câu nên có thể bộ dữ liệu chưa bao quát được tất cả trường hợp.

Chương 4: Ngữ Liệu

4.1. Tập nhãn

Về nguyên tắc, các thông tin về từ có thể được chứa trong nhãn từ loại bao gồm: thông tin hình thái, từ loại cơ sở, thông tin ngữ nghĩa, … Tuy nhiên, trong phạm vi đề tài này, nhóm nghiên cứu chỉ xây dựng tập nhãn từ loại chỉ chứa thông tin về từ loại cơ sở mà không chứa các thông tin khác. Dưới đây là tập nhãn dựa trên nhãn từ loại trong từ điển VCL.

STT	Nhãn	Tên	Ví dụ
1	A	Tính từ	xấu, đẹp, …
2	C	Liên từ	tuy nhiên, do đó, …
3	CH	Dấu câu	. ! ? …
4	D	Định từ	những, cái, và, …
5	E	Giới từ	trên, dưới, …
6	I	Thán từ	chao ôi, ôi, …
7	M	Số từ	một, hai, …
8	N	Danh từ	bàn, ghế, …
9	Nc	Danh từ chỉ loại	con, cái, …
10	Np	Danh từ riêng	Thủ Đức, Hồ Chí Minh, …
11	P	Đại từ	chúng tôi, tôi, …
12	R	Phụ từ	đang, sẽ, …
13	V	Động từ	đi, ăn, …
14	X	Các từ không thể phân loại	
15	Z	Yếu tố cấu tạo từ	bất, vô, …

Bảng 1. Danh mục nhãn từ loại

4.2. Tạo ngữ liệu

Như đã đề cập, vì đảm bảo tính chính xác của bộ ngữ liệu để cài đặt và so sánh, nhóm nghiên cứu đã thực hiện xây dựng bộ ngữ liệu hoàn toàn thủ công. Tức từ dữ liệu thu thập ban đầu, nhóm thực hiện tách từ, sau đó đưa dữ liệu đã tách từ vào và tiếp tục gán nhãn cho các từ đã tách.

Từ E-những D-vết Nc-sẹo N-của E-bạo_lực N-gia_đình N-, CH-đã R-có V-lúc N-, CH-tôi P-đau_đớn A-, CH-tôi P-gục_ngã V-. CH

Nhưng C-điều N-quan_trọng A-nhất A-là V-tôi P-tìm V-thấy V-sức_mạnh N-nội_tại N-và C-tôi P-chiến_đấu V-vì E-nó P-. CH-

Ai P-trong E-chúng_ta P-cũng R-có V-những D-cuộc_chiến N-cá_nhân N-nhưng C-điều N-quan_trọng A-nhất R-là V-tin V-vào E-giá_trị N-của E-bản_thân N-. CH

Đội V-vương_miện N-chính I-mình P-, CH-viết_tiếp V-cuộc_đời N-bằng E-trang_hành_trình N-nhiệt_huyết A-, CH-bản_lĩnh N-cá_nhân N-. CH

Mỗi D-chúng_ta P-là V-một M-chiến_binh N-và C-hãy R-chiến_đấu V-vì E-giấc_mơ N-của E-mình P-. CH

Sân_nhà N-mình P-hồi_ấy P-có V-rộng A-mấy D-đâu P-, CH-chỉ R-có V-khoảng A-trời N-là C-lồng_lộng A-phía N-trên E-đầu N-, CH-nhưng C-đã R-đi V-hết R-cả I-tuổi N-thần_tiên A-rồi R-, CH-sao P-tôi P-vẫn R-còn R-nhớ_tiếc A-. CH

Tháng N-năm M-vừa_rồi X-bánh_cam N-đã R-lọt V-vào E-top N-30 M-món_rán N-ngon A-nhất A-thế_giới N-do E-hãng_tin N-danh_tiếng N-CNN Np-bình_chọn V-. CH

Hồi_đó P-má N-hay C-cho V-tiền N-dằn_túi V-thì C-mình P-lại R-dằn_vặt V-,CH-vì E-mình P-không R-biết V-nên C-chọn V-giữa E-ổ_bánh_mì N-đầy A-thịt N-hay_là C-một M-ly_trà_sữa N-thơm_ngon A-đầy A-thạch N-. CH

Người_ta P-ngây_ngất V-trước E-sự N-hào_nhoáng A,-mê_mẩn V-trước E-sự N-bóng_bẩy A-nhưng C-chỉ R-rơi V-nước_mắt N-trước E-sự N-giản_dị A-tự P-đáy_lòng N-. CH

Vô_cảm V-, CH-thờ_ơ V-hay C-ích_kỷ V-là V-cảm_xúc N-, CH-mà C-cảm_xúc N-thì C-tồn_tại V-trong E-tất_cả P-, CH-không R-riêng A-gì P-giới_trẻ N-. CH

Hình 11. Ngữ liệu thủ công

Tiếp theo, nhóm thực hiện phân chia bộ ngữ liệu thành 2 tập train và test với tỉ lệ 40 câu /20 câu).

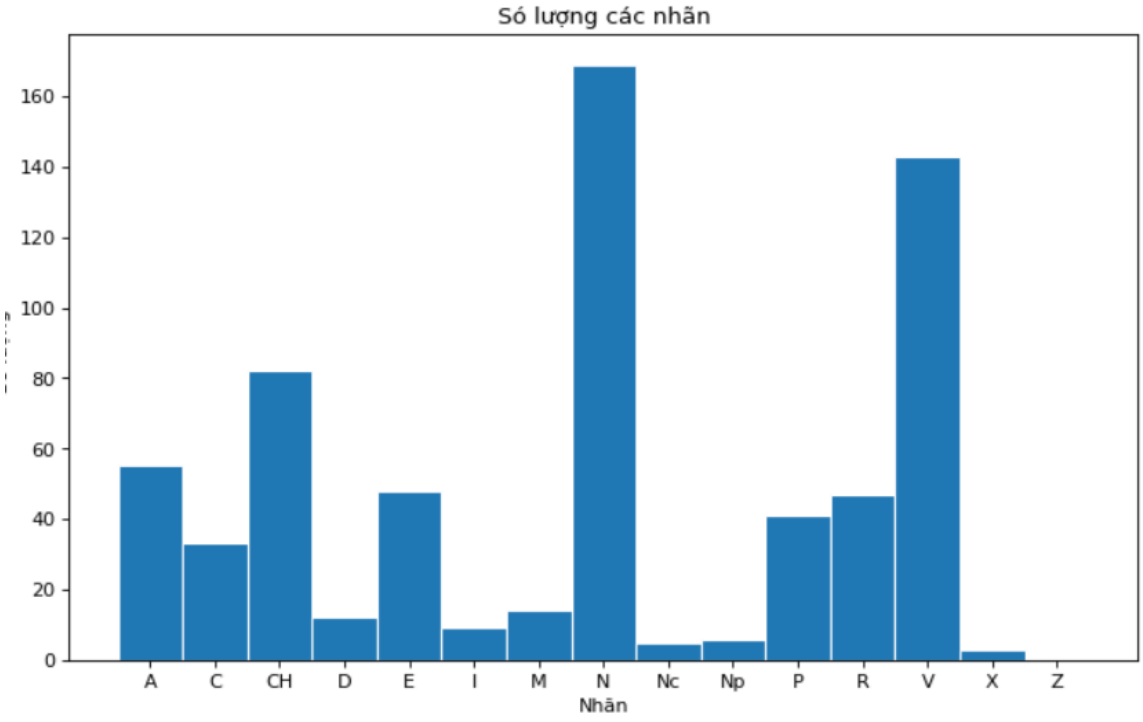
4.2.1. Tập train

Với tập train, sau khi chia tỉ lệ, nhóm đã thực hiện phân tích để nắm bắt thông tin:

- Có tổng cộng 682 từ trong tập train.
- 33 từ không có trong bộ từ điển.

```
Số từ của tập train không có trong từ điển là:  33
{'DCG',
 'FSA',
 'Prolog',
 'UIt',
 'bánh_cam',
 'bên_ngoài',
 'bẩm',
 'câu_hỏi',
 'dâng_trào',
 'giấc_mơ',
 'giới_trẻ',
 'gục_ngã',
 'hãng_tin',
 'hòa_nhập',
 'hồi_đó',
 'hồi_ấy',
 'lao_ra',
 'ly_trà_sữa',
 'lúa_nếp',
 'lớp_lớp',
 'món_rán',
 'ngước_nhìn',
 'nhiều_hơn',
 'thơm_ngon',
 'thầm_nghĩ',
 'thế_này',
 'trang_hành_trình',
 'tới_đâu',
 'viết_tiếp',
 'văn_phạm',
 'đáy_lòng',
 'đổ_lỗi',
 'ổ_bánh_mì'}
```

Hình 12. Các từ "lạ" trong tập train



Hình 13. Số lượng các nhãn trong tập train

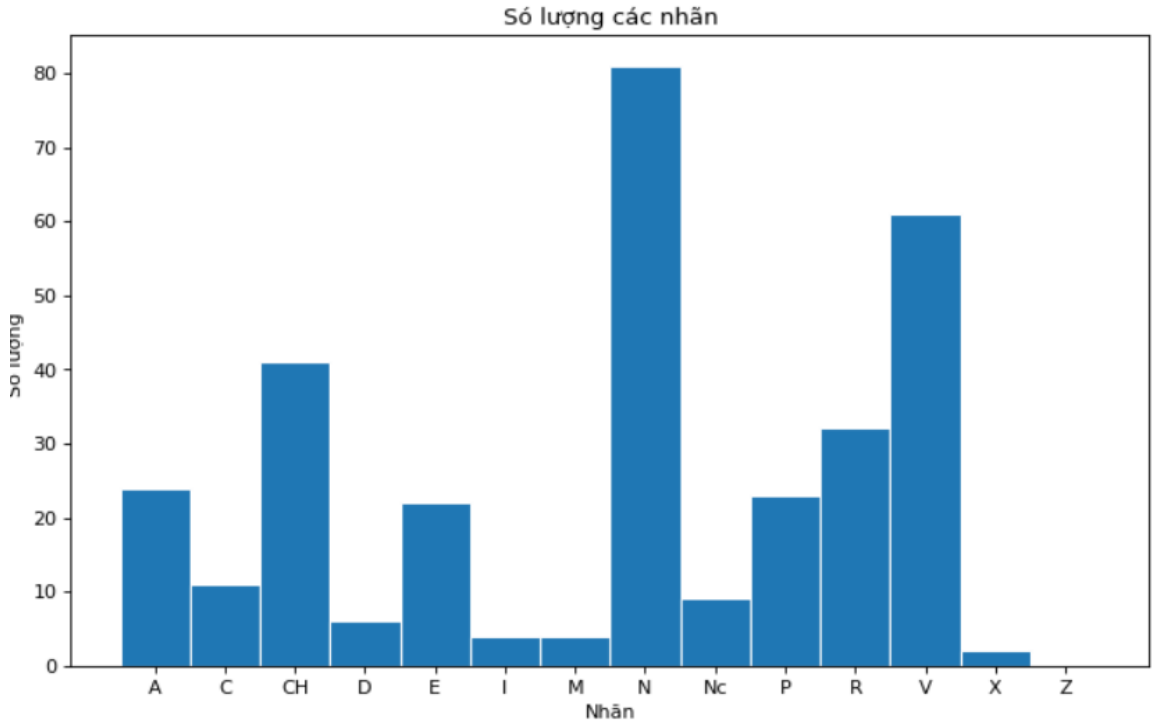
4.2.2. Tập test

Tương tự như tập train, nhóm cũng đã phân tích thông tin của tập test:

- Có tổng cộng 334 từ.
- 6 từ không có trong bộ từ điển.

```
Số từ của tập test không có trong từ điển là:  6
{'khả', 'matcha', 'ngang_nhau', 'sinh_ra', 'thế_giới_quan_duy_vật', 'ánh_mắt'}
```

Hình 14. Các từ "lạ" trong tập test



Hình 15. Số lượng các nhãn trong tập test

Chương 5: Gán Nhãn Từ Loại

5.1. Mô hình Hidden Markov Model

5.1.1. Markov Chain

Markov Chain (Xích Markov), hay Visible Markov Model, là một dạng FSA được dùng để mô hình hóa xác suất của các biến ngẫu nhiên có quan hệ với nhau theo dạng chuỗi.

Một Markov Chain M là một bộ trong đó:

- Q là tập các trạng thái
- δ là hàm chuyển đổi trạng thái có trọng số. Trọng số trong δ là xác suất chuyển trạng thái tương ứng.
- q0 là trạng thái bắt đầu

Xác định Markov chain:

- Q là các biến ngẫu nhiên tạo thành chuỗi
- δ là ma trận xác suất chuyển trạng thái từ q_i sang q_j ($q_i, q_j \in Q$). Xác suất này được tính theo mô hình bậc 1 (theo giải thiết $p(x_n | x_1..x_{n-1})$ như sau:

$$\delta q_i q_j = \frac{count(qi,qj)}{count(qi)}$$

- q_0 là một ký hiệu khởi đầu của tất cả các chuỗi

5.1.2. Giới thiệu Hidden Markov Model

Mô hình Markov ẩn (Hidden Markov Model – HMM):

- Mô hình xác suất
- Tính toán trên dữ liệu dạng chuỗi
- Mô hình Markov ẩn có dạng tương tự mô hình Markov rõ trong đó:
- Đỉnh trong đồ thị trạng thái được thay bằng trạng thái ẩn
- Trạng thái quan sát được được quyết định bởi trạng thái ẩn

Các thành phần của mô hình HMM:

- S = {s1 , s2 , ... , sn } là tập các trạng thái ẩn
- Thái đặc biệt s0 là trạng thái bắt đầu.
- K = {k1 , k2 , ... , km} là tập các giá trị quan sát
- A = {aij}, (i,j = 1...n) là ma trận chuyển trạng thái, trong đó aij là xác suất chuyển từ trạng thái si sang trạng thái sj .
- B = {bij} (i=1...n, j=1..m) là ma trận emisson (thể hiện), trong đó bij là xác suất trạng thái ẩn si thể hiện bằng giá trị quan sát kj .

Một số giả thiết trong HMM:

- Sự độc lập của các kết quả quan sát được: Kết quả quan sát chỉ phụ thuộc vào trạng thái ẩn hiện tại, không phụ thuộc vào các trạng thái ẩn trước đó. $p(ot | x1 ...xT , o1 ..oT) = p(ot | xt)$

- Giới hạn kinh nghiệm: trạng kế tiếp phụ thuộc vào trạng thái hiện tại và một số hữu hạn k các trạng thái trước đó: p(xT |x1 ..xT-1) = p(xT |xT-kxT-1)

Bậc của HMM: là số trạng thái ẩn trước đó được dùng để tính toán đến xác suất của trạng thái kế tiếp:

- Bậc 1: p(xT |x1 ...xT-1) = p(xT |xT-1)
- Bậc 2: p(xT |x1 ...xT-1) = p(xT |xT-2xT-1)

Trong quá trình tính toán xác suất, nhóm gặp phải một số trường hợp xác suất bằng 0 và điều đó ảnh hưởng đến các kết quả khác. Để khắc phục nhược điểm, nhóm sử dụng phương pháp Laplace smoothing.

Laplace smoothing là một kĩ thuật làm mịn dữ liệu dạng phân loại, một giá trị nhỏ gọi là pseudo-count sẽ được thêm vào thay đổi xác suất đầu ra, có hiểu đơn giản là cộng thêm vào cả tử lẫn mẫu số để giá trị luôn khác 0. Kết quả là sẽ không còn xác suất nào bằng không nữa.

$$P(x_i|y) = \frac{x_i + \alpha}{N + \alpha d}$$

Trong đó:

- α thường là số dương, bằng 1.
- αd được cộng vào mẫu để đảm bảo $\sum i = 1 d P(x_i|y) = 1$

5.1.3. Ma trận chuyển trạng thái

5.1.3.1. Giới thiệu

Trước khi bắt đầu dự đoán nhãn của từ, nhóm đã thêm kí hiệu <s> vào tập nhãn mang ý nghĩa là nhãn bắt đầu một câu.

Nhắc lại ý nghĩa của ma trận chuyển trạng thái, tức là mỗi ô trong ma trận chứa xác suất chuyển từ trạng thái này sang trạng thái kia.

	<s>	A	C	CH	D	E	I	M	N	Nc	Np	P	R	V	X	Z
<s>	1.0	0.000985	0.000985	0.000985	0.000985	0.986207	0.000985	0.000985	0.000985	0.000985	0.000985	0.000985	0.000985	0.000985	0.000985	0.000985
A	1.0	0.071427	0.053575	0.303508	0.017870	0.107132	0.035723	0.000018	0.160689	0.000018	0.000018	0.053575	0.107132	0.089280	0.000018	0.000018
C	1.0	0.088226	0.000029	0.000029	0.000029	0.000029	0.058827	0.029428	0.205821	0.000029	0.000029	0.117625	0.176422	0.323416	0.000029	0.000029
CH	1.0	0.036591	0.073170	0.000012	0.024398	0.060977	0.000012	0.012205	0.353606	0.036591	0.012205	0.134134	0.085362	0.158520	0.012205	0.000012
D	1.0	0.000077	0.000077	0.000077	0.000077	0.000077	0.000077	0.076911	0.691587	0.076911	0.000077	0.153746	0.000077	0.000077	0.000077	0.000077
E	1.0	0.000020	0.000020	0.000020	0.081628	0.020422	0.000020	0.020422	0.612078	0.000020	0.040824	0.142834	0.020422	0.061226	0.000020	0.000020
I	1.0	0.099950	0.000100	0.499351	0.000100	0.000100	0.000100	0.000100	0.099950	0.000100	0.000100	0.099950	0.000100	0.199800	0.000100	0.000100
M	1.0	0.000067	0.000067	0.000067	0.000067	0.000067	0.000067	0.000067	0.799267	0.133267	0.000067	0.000067	0.000067	0.000067	0.066667	0.000067
N	1.0	0.164697	0.082352	0.158815	0.005888	0.058824	0.023533	0.017651	0.123524	0.000006	0.023533	0.029415	0.047061	0.252925	0.005888	0.005888
Nc	1.0	0.000166	0.000166	0.000166	0.000166	0.000166	0.000166	0.000166	0.665170	0.000166	0.000166	0.000166	0.000166	0.332668	0.000166	0.000166
Np	1.0	0.000143	0.000143	0.570349	0.000143	0.142694	0.000143	0.000143	0.000143	0.000143	0.000143	0.000143	0.000143	0.285246	0.000143	0.000143
P	1.0	0.047626	0.000024	0.142830	0.000024	0.071427	0.000024	0.000024	0.071427	0.000024	0.000024	0.095228	0.214233	0.333238	0.023825	0.000024
R	1.0	0.083328	0.041674	0.124982	0.041674	0.000021	0.020848	0.041674	0.041674	0.000021	0.000021	0.000021	0.041674	0.562345	0.000021	0.000021
V	1.0	0.076388	0.062500	0.124994	0.020838	0.152769	0.006951	0.041669	0.277756	0.000007	0.000007	0.027782	0.062500	0.145825	0.000007	0.000007
X	1.0	0.000249	0.000249	0.000249	0.000249	0.000249	0.000249	0.000249	0.498381	0.000249	0.000249	0.249315	0.000249	0.249315	0.000249	0.000249
Z	1.0	0.000985	0.000985	0.000985	0.000985	0.000985	0.000985	0.000985	0.986207	0.000985	0.000985	0.000985	0.000985	0.000985	0.000985	0.000985

Hình 16. Ma trận chuyển trạng thái dựa vào tập train

Ma trận trong hình 16 đã được tính toán với tham số smoothing nhằm làm mịn ma trận, tránh trường hợp tồn tại một xác suất bằng 0. Giả sử ở ô [1,1] sẽ mang ý nghĩa có 0.071427 khả năng từ nhãn A sẽ tiếp tục chuyển sang nhãn A. Hay ở ô [1,2] mang ý nghĩa có 0.053575 khả năng từ nhãn A sẽ chuyển sang nhãn C.

5.1.3.2. Cài đặt

Giả sử câu đầu vào là: "Từ/E những/D vết/Nc sẹo/N của/E bạo_lực/N gia_đình/N ,/CH đã/R có/V lúc/N ,/CH tôi/P đau_đớn/A ,/CH tôi/P gục_ngã/V ./CH". Dựa vào ý nghĩa của ma trận chuyển trạng thái, nhóm nghiên cứu đã đề xuất xây dựng ma trận bằng 4 bước:

Bước 1: Tạo ma trận giá trị mặc định là 0 với kích thước ma trận là m x m (m là kích thước tập nhãn)

	<s>	E	D	Nc	N	R	V	A	P	CH
<s>	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
E	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
D	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Nc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
N	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
R	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
V	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CH	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Hình 17. Bước 1 trong xây dựng ma trận chuyển trạng thái

Bước 2: Duyệt các cặp từ và cộng một đơn vị tương ứng với nhãn trong ma trận. Xác suất ở cột <s> trong ma trận là không cần thiết vì theo quy ước <s> là kí hiệu bắt đầu câu, do đó sẽ không tồn tại một nhãn nào có thể chuyển qua nhãn <s>. Vì thế ở cột <s> sẽ đại diện cho tổng nhãn tương ứng.

	Tổng	E	D	Nc	N	R	V	A	P	CH
<s>	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
E	2.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
D	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
Nc	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
N	4.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	2.0
R	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
V	2.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
A	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
P	2.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0
CH	3.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	2.0	0.0

Hình 18. Bước 2 trong xây dựng ma trận chuyển trạng thái

Bước 3: Cộng hệ số smoothing (0.001) vào các phần tử để xác suất khác 0.

	Tổng	E	D	Nc	N	R	V	A	P	CH
<s>	1.009	1.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
E	2.009	0.001	1.001	0.001	1.001	0.001	0.001	0.001	0.001	0.001
D	1.009	0.001	0.001	1.001	0.001	0.001	0.001	0.001	0.001	0.001
Nc	1.009	0.001	0.001	0.001	1.001	0.001	0.001	0.001	0.001	0.001
N	4.009	1.001	0.001	0.001	1.001	0.001	0.001	0.001	0.001	2.001
R	1.009	0.001	0.001	0.001	0.001	0.001	1.001	0.001	0.001	0.001
V	2.009	0.001	0.001	0.001	1.001	0.001	0.001	0.001	0.001	1.001
A	1.009	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	1.001
P	2.009	0.001	0.001	0.001	0.001	0.001	1.001	1.001	0.001	0.001
CH	3.009	0.001	0.001	0.001	0.001	1.001	0.001	0.001	2.001	0.001

Hình 19. Bước 3 trong xây dựng ma trận chuyển trạng thái

Bước 4: Tính xác suất bằng cách lấy các phần tử chia cho giá trị tổng tương ứng.

	Tổng	E	D	Nc	N	R	V	A	P	CH
<s>	1.0	0.992071	0.000991	0.000991	0.000991	0.000991	0.000991	0.000991	0.000991	0.000991
E	1.0	0.000498	0.498258	0.000498	0.498258	0.000498	0.000498	0.000498	0.000498	0.000498
D	1.0	0.000991	0.000991	0.992071	0.000991	0.000991	0.000991	0.000991	0.000991	0.000991
Nc	1.0	0.000991	0.000991	0.000991	0.992071	0.000991	0.000991	0.000991	0.000991	0.000991
N	1.0	0.249688	0.000249	0.000249	0.249688	0.000249	0.000249	0.000249	0.000249	0.499127
R	1.0	0.000991	0.000991	0.000991	0.000991	0.000991	0.992071	0.000991	0.000991	0.000991
V	1.0	0.000498	0.000498	0.000498	0.498258	0.000498	0.000498	0.000498	0.000498	0.498258
A	1.0	0.000991	0.000991	0.000991	0.000991	0.000991	0.000991	0.000991	0.000991	0.992071
P	1.0	0.000498	0.000498	0.000498	0.000498	0.000498	0.498258	0.498258	0.000498	0.000498
CH	1.0	0.000332	0.000332	0.000332	0.000332	0.332669	0.000332	0.000332	0.665005	0.000332

Hình 20. Bước 4 trong xây dựng ma trận chuyển trạng thái

5.1.4. Ma trận thể hiện

5.1.4.1. Giới thiệu

Ma trận thể hiện mang ý nghĩa mỗi phần tử trong ma trận sẽ cho thấy khả năng mà một từ mang xác suất tương ứng.

	từ\n	những\n	--unk--\n
<s>	0.000018	0.000018	0.000018
A	0.000009	0.000009	0.036022
C	0.000011	0.000011	0.033693
CH	0.000007	0.000007	0.000007
D	0.000015	0.073468	0.014705
E	0.038445	0.000010	0.009619
I	0.000015	0.000015	0.030751
M	0.000014	0.000014	0.014286
N	0.000004	0.000004	0.133296
Nc	0.000016	0.000016	0.032766
Np	0.000016	0.000016	0.096681
P	0.000010	0.000010	0.041218
R	0.000010	0.000010	0.019414
V	0.000005	0.000005	0.070332
X	0.000017	0.000017	0.000017
Z	0.000018	0.000018	0.000018

Hình 21. Ma trận thể hiện dựa vào tập train

Hình 21 chỉ là một phần của ma trận thể hiện. Thực tế, ma trận thể hiện có kích thước là cả một tập từ điển, ví dụ ở từ 'những' sẽ có 0.000009 khả năng mang nhãn A hay có 0.073468 khả năng mang nhãn D. Do đây là phương pháp dựa trên từ điển, nên chắc chắn sẽ có những từ "lạ" mà nhóm đã đề cập sẽ xuất hiện mà không có trong bộ từ điển. Giải pháp sẽ là quy các từ đó thành một từ cụ thể đó là --unk--. Để tăng độ chính xác của ma trận thể hiện, ta có thể chia nhỏ --unk-- thành --unk_N--, --unk_V-- ,…

5.1.4.2. Cài đặt

Cùng với câu đầu vào như trên, nhóm sẽ tiếp tục đề xuất các bước để hoàn thành ma trận thể hiện như sau:

Bước 1: Khởi tạo ma trận thể hiện giá trị 0 với kích thước là tập nhãn x bộ từ điển. Ta có thể thấy từ "gục_ngã" không được tìm thấy trong bộ từ điển, do đó ta thay thế bằng --unk--. Ta sẽ thêm một cột làm cột tổng các nhãn.

	từ\n	những\n	vết\n	sẹo\n	của\n	bạo_lực\n	gia_đình\n	,\n	đã\n	có\n	lúc\n	đau_đớn\n	--unk--\n	.\n	Tổng
<s>	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
C	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CH	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
D	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
E	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
N	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Nc	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Np	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
R	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
V	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Z	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Hình 22. Bước 1 trong xây dựng ma trận thể hiện

Bước 2: Duyệt lần lượt các từ trong câu và cộng 1 đơn vị vào nhãn tương ứng.

	từ\n	những\n	vết\n	sẹo\n	của\n	bạo_lực\n	gia_đình\n	,\n	đã\n	có\n	lúc\n	đau_đớn\n	--unk--\n	.\n	Tổng
<s>	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
C	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CH	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	1.0	4.0
D	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
E	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	2.0
I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
N	0.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	4.0
Nc	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
Np	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
R	0.0	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	1.0
V	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	2.0
X	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Z	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Hình 23. Bước 2 trong xây dựng ma trận thể hiện

Bước 3: Cộng hệ số smoothing (0.001) vào các phần tử để xác suất khác 0.

	từ\n	những\n	vết\n	sẹo\n	của\n	bạo_lực\n	gia_đình\n	,\n	đã\n	có\n	lúc\n	đau_đớn\n	--unk--\n	.\n	Tổng
<s>	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	55.071
A	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	1.001	0.001	0.001	56.071
C	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	55.071
CH	0.001	0.001	0.001	0.001	0.001	0.001	0.001	3.001	0.001	0.001	0.001	0.001	0.001	1.001	59.071
D	0.001	1.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	56.071
E	0.001	0.001	0.001	0.001	1.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	1.001	0.001	57.071
I	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	55.071
M	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	55.071
N	0.001	0.001	0.001	1.001	0.001	1.001	1.001	0.001	0.001	0.001	1.001	0.001	0.001	0.001	59.071
Nc	0.001	0.001	1.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	56.071
Np	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	55.071
P	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	57.071
R	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	1.001	0.001	0.001	0.001	0.001	0.001	56.071
V	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	1.001	0.001	0.001	1.001	0.001	57.071
X	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	55.071
Z	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	55.071

Hình 24. Bước 3 trong xây dựng ma trận thể hiện

Bước 4: Tính xác suất.

	từ\n	những\n	vết\n	sẹo\n	của\n	bạo_lực\n	gia_đình\n	,\n	đã\n	có\n	lúc\n	đau_đớn\n	--unk--\n	.\n	Tổng
<s>	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	1.0
A	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.017852	0.000018	0.000018	1.0
C	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	1.0
CH	0.000017	0.000017	0.000017	0.000017	0.000017	0.000017	0.000017	0.050803	0.000017	0.000017	0.000017	0.000017	0.000017	0.016946	1.0
D	0.000018	0.017852	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	1.0
E	0.000018	0.000018	0.000018	0.000018	0.017540	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.017540	0.000018	1.0
I	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	1.0
M	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	1.0
N	0.000017	0.000017	0.000017	0.016946	0.000017	0.016946	0.016946	0.000017	0.000017	0.000017	0.016946	0.000017	0.000017	0.000017	1.0
Nc	0.000018	0.000018	0.017852	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	1.0
Np	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	1.0
P	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	1.0
R	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.017852	0.000018	0.000018	0.000018	0.000018	0.000018	1.0
V	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.017540	0.000018	0.000018	0.017540	0.000018	1.0
X	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	1.0
Z	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	0.000018	1.0

Hình 25. Bước 4 trong xây dựng ma trận thể hiện

5.2. Thuật toán Viterbi

Nhóm nghiên cứu tiến hành kết hợp thuật toán Viterbi với mô hình Hidden Markov. Thuật toán Viterbi được xây dựng dựa trên ý tưởng của quy hoạch động. Dựa vào thuật toán, ta sẽ tìm được chuỗi trạng thái ẩn tốt nhất cho mô hình.

5.2.1. Khởi tạo

Bước đầu tiên ta sẽ khởi tạo 2 ma trận xác suất (probs) và ma trận đường đi (paths):

- Ma trận paths: ma trận giúp tìm đường đi tốt nhất. Ở bước này ta chỉ khởi tạo ma trận paths với giá trị mặc định là 0.
- Ma trận probs: thể hiện khả năng đi từ nhãn của từ này sang nhãn của từ kế cạnh. Ở bước khởi tạo, mặc định giá trị của ma trận sẽ là 0 ngoại trừ cột đầu tiên. Vì với giả định rằng từ đầu tiên sẽ là nhãn bắt đầu câu (<s>) chuyển sang.

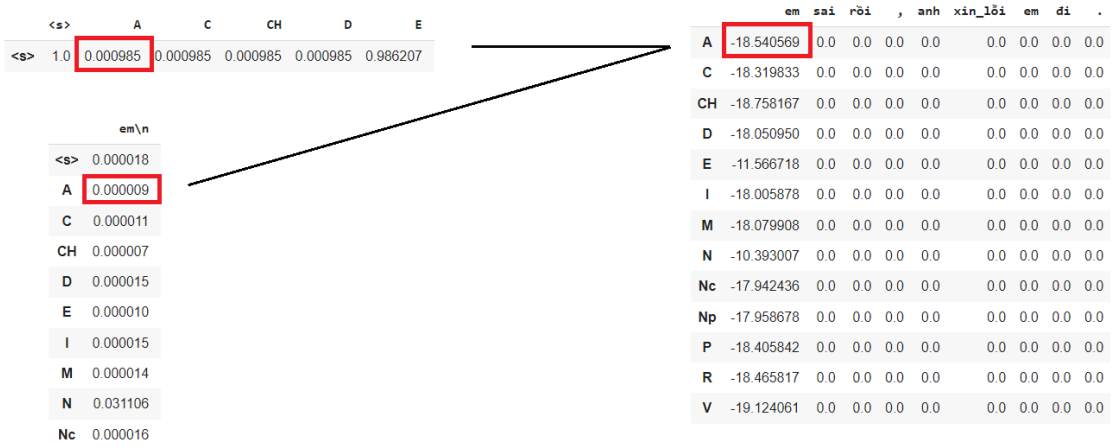
	em	sai	rồi	,	anh	xin_lỗi	em	đi	.
A	-18.540569	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
C	-18.319833	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CH	-18.758167	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
D	-18.050950	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
E	-11.566718	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
I	-18.005878	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M	-18.079908	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
N	-10.393007	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Nc	-17.942436	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Np	-17.958678	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P	-18.405842	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
R	-18.465817	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
V	-19.124061	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X	-17.909139	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Z	-17.857018	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Hình 26. Ma trận probs ở bước khởi tạo

Ở hình 26 trên, với từ "em" sẽ có 15 khả năng mang các nhãn từ A đến Z. Xác suất từ <s> chuyển sang từ "em" mang nhãn A là -18.540569, hoặc từ <s> chuyển sang từ "em" mang nhãn V là -19.124061.

Làm rõ ma trận probs:

Vì ma trận probs sẽ cho thấy xác suất chuyển nhãn giữa hai từ kế cạnh nhau. Lưu ý ma trận probs khác hoàn toàn với ma trận chuyển trạng thái, ma trận chuyển trạng thái chỉ quan tâm đến nhãn này chuyển sang nhãn khác, ví dụ từ A -> C, còn ma trận probs quan tâm đến cả việc nhãn đó là của từ gì, ví dụ "em" - A -> "sai" - C. Do đó ma trận probs sẽ được tính bằng tích xác suất của ma trận chuyển trạng thái với ma trận thể hiện tương ứng.

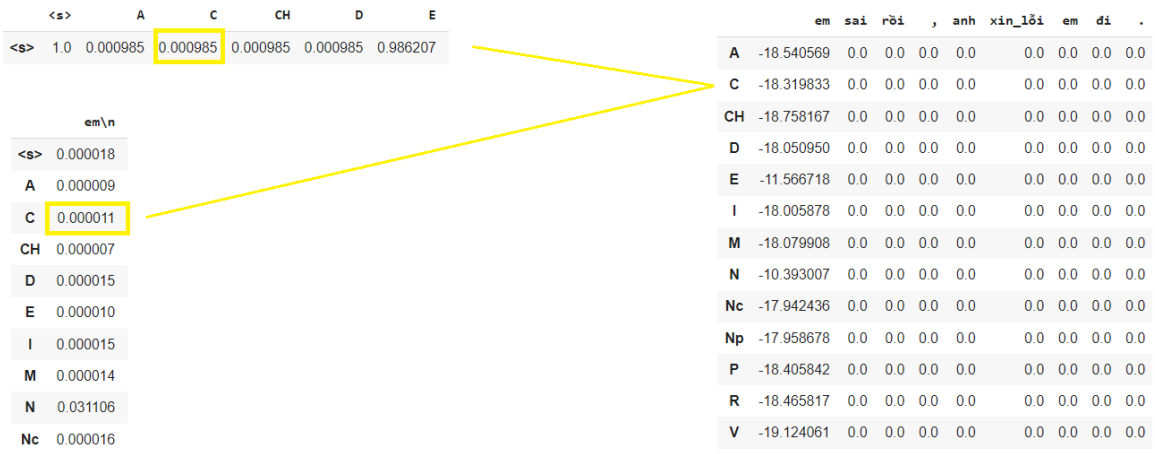


	<s>	A	C	CH	D	E
<s>	1.0	0.000985	0.000985	0.000985	0.000985	0.986207

	em\n
<s>	0.000018
A	0.000009
C	0.000011
CH	0.000007
D	0.000015
E	0.000010
I	0.000015
M	0.000014
N	0.031106
Nc	0.000016

	em	sai	rồi	,	anh	xin_lỗi	em	đi	.
A	-18.540569	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
C	-18.319833	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CH	-18.758167	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
D	-18.050950	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
E	-11.566718	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
I	-18.005878	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M	-18.079908	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
N	-10.393007	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Nc	-17.942436	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Np	-17.958678	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P	-18.405842	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
R	-18.465817	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
V	-19.124061	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Hình 27. Cách tính ma trận probs từ <s> sang A



	<s>	A	C	CH	D	E
<s>	1.0	0.000985	0.000985	0.000985	0.000985	0.986207

	em\n
<s>	0.000018
A	0.000009
C	0.000011
CH	0.000007
D	0.000015
E	0.000010
I	0.000015
M	0.000014
N	0.031106
Nc	0.000016

	em	sai	rồi	,	anh	xin_lỗi	em	đi	.
A	-18.540569	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
C	-18.319833	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CH	-18.758167	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
D	-18.050950	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
E	-11.566718	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
I	-18.005878	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M	-18.079908	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
N	-10.393007	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Nc	-17.942436	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Np	-17.958678	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P	-18.405842	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
R	-18.465817	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
V	-19.124061	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Hình 28. Cách tính ma trận probs từ <s> sang C

Vậy tại sao tích hai xác suất lại là số âm? Dễ nhận thấy, xác suất của hai ma trận đều rất nhỏ bé hơn 0, nên nếu ta nhân hai xác suất thì kết quả sẽ vô cùng nhỏ gây khó khăn cho việc tính toán. Thế nên, để tránh việc đó, ta sẽ sử dụng tổng của ln() để thay thế. Do đó, thông thường kết quả càng gần 1 sẽ càng chính xác, nhưng trong trường hợp này thì xác suất càng gần 0 sẽ chính xác hơn. Công thức tổng quát sẽ như sau:

$$probs[\text{ nhãn, từ }] = \ln(A[\text{ 0, nhãn }]) + \ln(B[\text{ từ, nhãn }])$$

Trong đó: A là ma trận chuyển trạng thái, B là ma trận thể hiện.

5.2.2. Forward

Bước tiếp theo sẽ là Forward, ở giai đoạn này, ta sẽ tiếp tục hoàn thành ma trận probs, đồng thời sẽ đánh dấu đường đi ở ma trận paths.

Như hình 26, ta sẽ tiếp tục tìm xác suất cho từ "sai". Cũng như "em", từ "sai" sẽ có khả năng mang 15 nhãn từ A đến Z. Với từ "sai" mang nhãn A, sẽ tồn tại 15 trường hợp từ "em" mang các nhãn chuyển trạng thái sang, ví dụ: em (A) -> sai (A), em (C) -> sai (A),…, em (Z) -> sai (A). Trong 15 trường hợp đó, ta sẽ phải chọn trường hợp nào là khả thi nhất thỏa mãn 2 điều kiện sau:

- Phải phụ thuộc vào nhãn của tất cả các từ đứng trước nó. Ví dụ ta đang xét 2 chuỗi nhãn sau A -> ? (1) và V -> ? (2), nếu chỉ dựa vào ma trận chuyển trạng thái và ma trận thể hiện thì có thể dựa vào (2) ta sẽ tìm được nhãn cần tìm là N. Vậy là hoàn toàn sai vì ta đã xét thiếu một khả năng, đó chính là từ đứng trước liệu sẽ mang nhãn A hay V? Vì vậy ta phải thêm một tham số để đánh giá sự khả thi của từ đứng trước, đó chính là giá trị của cột trước từ đang xét trong ma trận probs, với từ "sai" thì đó là cột "em".
- Khả năng đó phải là một xác suất lớn nhất.

Dựa vào 2 điều kiện trên, ta rút ra công thức tổng quát như sau:

probs[nhãn, từ] = max (probs[15 nhãn, từ - 1] + ln(A[0, nhãn]) + ln(B[từ, nhãn])

Một điều thú vị ở đây đó chính là công thức ở bước khởi tạo chính là công thức tổng quát trên nhưng đã được rút gọn, vì ta đã giả định ở trên đó chính là từ đầu câu luôn là bắt đầu bởi nhãn <s> nên probs[15 nhãn, từ - 1] = 1, vì thế ở trên ta lượt bỏ luôn probs[15 nhãn, từ - 1].

	em	sai	rồi	,	anh	xin_lỗi	em	đi	.
A	-18.540569	-16.905823	-31.162823	-37.876829	-41.271466	-45.124199	-52.559809	-56.542900	-63.285862
C	-18.319833	-24.286954	-31.229686	-38.348991	-40.357750	-45.596574	-52.539723	-57.015275	-63.265777
CH	-18.758167	-24.068543	-29.933694	-26.345585	-49.495766	-45.378163	-52.284966	-56.796865	-52.519718
D	-18.050950	-25.200606	-32.058750	-38.080108	-41.187147	-47.965821	-53.369231	-59.384522	-64.095285
E	-11.566718	-24.779037	-30.692345	-44.283787	-40.695678	-46.088657	-51.801611	-57.507359	-62.527664
I	-18.005878	-25.225583	-24.412275	-38.727684	-48.743477	-46.535204	-54.422105	-57.953905	-65.148158
M	-18.079908	-25.587212	-38.304343	-38.109066	-41.908752	-46.896833	-52.705208	-58.315534	-63.431262
N	-10.393007	-24.381785	-31.058278	-39.039531	-31.702627	-46.118115	-43.121328	-57.536816	-62.701190
Nc	-17.942436	-31.115103	-38.396292	-44.261443	-40.673334	-54.766062	-59.705912	-65.740403	-71.124614
Np	-17.958678	-25.178383	-38.875232	-44.660149	-41.787522	-46.488004	-60.323171	-57.906705	-71.741872
P	-18.405842	-24.995989	-31.315695	-38.198558	-39.837699	-46.712074	-53.436524	-58.130775	-64.162577
R	-18.465817	-24.992500	-23.773935	-38.494975	-40.349608	-46.302121	-52.685708	-57.720822	-63.411761
V	-19.124061	-23.969087	-31.523222	-36.550992	-40.388878	-38.369953	-52.496717	-49.096006	-63.222770
X	-17.909139	-26.514389	-38.825694	-44.610610	-41.737983	-47.824010	-60.591378	-59.242711	-71.960174
Z	-17.857018	-26.462268	-38.773572	-44.558489	-48.594616	-47.771888	-61.181999	-59.190590	-71.908053

Hình 29. Ma trận probs sau khi hoàn thiện

Với ma trận probs trên, ta sẽ chỉ tìm được nhãn của từ cuối cùng trong câu, đó chính là dấu “.” có xác suất là -52.519718 mang nhãn CH (xác suất lớn nhất). Nhưng nếu chỉ có vậy, ta chắc chắn sẽ quên mất nhãn nào của từ “đi” sẽ dẫn đến probs[CH,.] là max. Vì thế, ta cần một ma trận để đánh dấu đường đi, đó chính là ma trận paths.

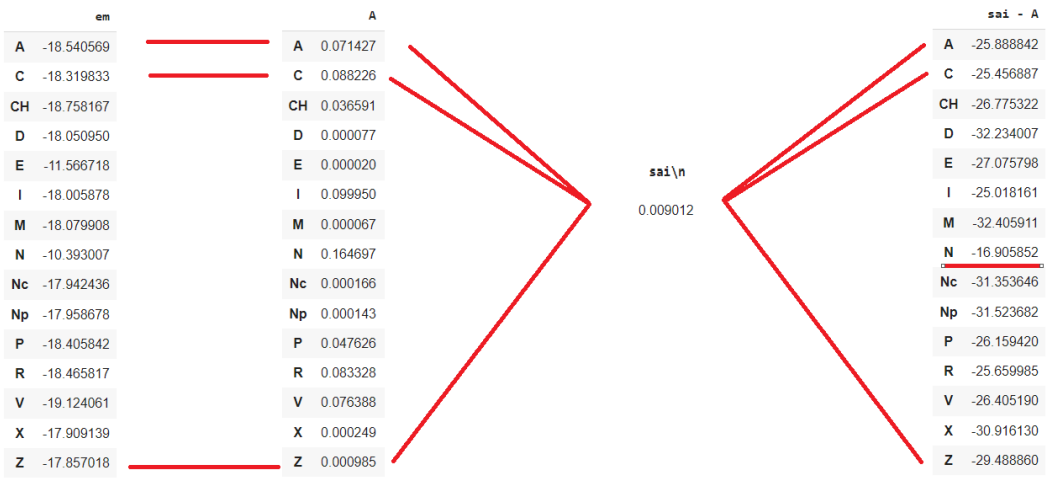
	em	sai	rồi	,	anh	xin_lỗi	em	đi	.
A	0	8	1	12	3	8	13	8	13
C	0	8	1	12	3	8	13	8	13
CH	0	8	1	6	3	8	13	8	13
D	0	5	1	12	3	8	13	8	13
E	0	8	1	3	3	8	13	8	13
I	0	8	1	12	3	8	13	8	13
M	0	8	13	12	3	8	13	8	13
N	0	5	1	6	3	8	13	8	13
Nc	0	7	3	3	3	8	3	7	3
Np	0	8	1	6	3	8	5	8	5
P	0	5	1	6	3	8	13	8	13
R	0	8	1	12	3	8	13	8	13
V	0	8	1	12	3	8	13	8	13
X	0	8	1	6	3	8	7	8	13
Z	0	8	1	6	3	8	13	8	13

Hình 30. Ma trận paths

Nguyên tắc đánh dấu đường đi rất đơn giản, ta sẽ ghi nhận vị trí nhãn của từ đứng trước nó vào chính nó ở ma trận paths. Ví dụ, nếu nhãn V của từ "đi" dẫn đến probs[CH, .] là max, vậy thì ở ở paths[CH, .] ta sẽ đánh dấu là 13, tức là vị trí của nhãn V.

Ví dụ cụ thể bước Forward

Ở ví dụ này, ta đi tìm giá trị probs nhãn A của từ "sai.



Hình 31. Hiện thực hóa công thức tổng quát

Từ hình 31 có thể thấy, ta sẽ lần lượt cộng giá trị các nhãn của từ đứng trước nó là từ "em" trong ma trận probs với xác suất chuyển trạng thái tương ứng, với em(A) thì ta cộng xác suất chuyển từ nhãn A sang A, với em(C) thì ta cộng xác suất chuyển từ nhãn C sang A,… tức là ta tìm mọi khả năng của từ em chuyển sang nhãn A. Tiếp đến ta sẽ tiếp tục cộng 15 kết quả đó với xác suất từ em mang nhãn A trong ma trận thể hiện. Cuối cùng ta sẽ được 15 giá trị như hình trên, như đã nói, ta chỉ chọn trường hợp nào khả thi nhất, nghĩa là xác suất cao nhất. Vậy ta chọn nhãn N với giá trị là -16.905852, vì nhãn N nằm ở vị trí thứ 8 trong danh sách các nhãn, nên đồng thời, ta cũng sẽ ghi nhận giá trị 8 trong ma trận paths.

5.2.3. Backward

Dựa vào 2 ma trận probs và paths đã hoàn thiện, ta sẽ dự đoán được nhãn của tất cả các từ trong câu.

Theo như công thức tổng quát ở phần Forward thì xác suất của ta sẽ là cộng dồn tất cả các từ trong câu lại với nhau. Do đó, ở cuối câu, vị trí nào có giá trị xác suất lớn nhất trong ma trận probs thì sẽ là điểm bắt đầu ta truy ngược lại trong ma trận paths. Vì vậy theo hình 26, ta sẽ bắt đầu đi ở nhãn CH của dấu ".".

	em	sai	rồi	,	anh	xin_lỗi	em	đi	.
A	0	8	1	12	3	8	13	8	13
C	0	8	1	12	3	8	13	8	13
CH	0	8	1	6	3	8	13	8	13
D	0	5	1	12	3	8	13	8	13
E	0	8	1	3	3	8	13	8	13
I	0	8	1	12	3	8	13	8	13
M	0	8	13	12	3	8	13	8	13
N	0	5	1	6	3	8	13	8	13
Nc	0	7	3	3	3	8	3	7	3
Np	0	8	1	6	3	8	5	8	5
P	0	5	1	6	3	8	13	8	13
R	0	8	1	12	3	8	13	8	13
V	0	8	1	12	3	8	13	8	13
X	0	8	1	6	3	8	7	8	13
Z	0	8	1	6	3	8	13	8	13

Hình 32. Giá trị khởi đầu

Tại vị trí nhãn CH đã được gạch đỏ ở hình trên có số 13, tức là nhãn của từ "đi" sẽ là giá trị nhãn thứ 13 (nhãn V). Tương tự, ở vị trí nhãn R cột từ "đi" có số 8, thì từ "em" trước nó sẽ mang nhãn giá trị thứ 8 (nhãn N). Cứ đi ngược như vậy, ta sẽ tìm được nhãn của tất cả các từ trong câu

	em	sai	rồi	,	anh	xin_lỗi	em	đi	.
A	0	8	1	12	3	8	13	8	13
C	0	8	1	12	3	8	13	8	13
CH	0	8	1	6	3	8	13	8	13
D	0	5	1	12	3	8	13	8	13
E	0	8	1	3	3	8	13	8	13
I	0	8	1	12	3	8	13	8	13
M	0	8	13	12	3	8	13	8	13
N	0	5	1	6	3	8	13	8	13
Nc	0	7	3	3	3	8	3	7	3
Np	0	8	1	6	3	8	5	8	5
P	0	5	1	6	3	8	13	8	13
R	0	8	1	12	3	8	13	8	13
V	0	8	1	12	3	8	13	8	13
X	0	8	1	6	3	8	7	8	13
Z	0	8	1	6	3	8	13	8	13

Hình 33. Quá trình truy xuất

5.3 Đánh giá

5.3.1. Ưu điểm

Mô hình Hidden Markov được xây dựng dựa trên cơ sở toán học vững chắc, cài đặt tương đối đơn giản nếu đã nắm vững ý tưởng thuật toán. Dễ dàng theo dõi và quan sát từng bước của thuật toán.

5.3.2. Nhước điểm

Mô hình cần rất nhiều tham số và tốn nhiều tài nguyên. Mô hình cũng phụ thuộc khá nhiều vào từ điển, đặc biệt là các từ mới chưa có trong bộ từ điển.

5.3.3. So sánh với thư viện VnCoreNLP

Nhóm nghiên cứu đánh giá mô hình HMM và VnCoreNLP trên tập test của nhóm. Kết quả so sánh như hai hình bên dưới.

	precision	recall	f1-score	support
A	0.54	0.28	0.37	25
C	0.60	1.00	0.75	12
CH	0.89	1.00	0.94	42
D	0.45	0.71	0.56	7
E	0.68	0.91	0.78	23
I	0.80	0.80	0.80	5
M	0.80	0.80	0.80	5
N	0.71	0.51	0.60	82
Nc	1.00	0.10	0.18	10
P	0.58	0.62	0.60	24
R	0.47	0.67	0.55	33
V	0.61	0.68	0.64	62
X	0.00	0.00	0.00	3
Z	0.00	0.00	0.00	1
accuracy			0.65	334

Hình 34. Kết quả đánh giá HMM

	precision	recall	f1-score	support
A	0.83	0.76	0.79	25
C	0.90	0.75	0.82	12
CH	1.00	1.00	1.00	42
Cc	0.00	0.00	0.00	0
D	0.00	0.00	0.00	7
E	0.96	0.96	0.96	23
I	0.00	0.00	0.00	5
L	0.00	0.00	0.00	0
M	1.00	1.00	1.00	5
N	0.84	0.89	0.86	82
Nb	0.00	0.00	0.00	0
Nc	0.00	0.00	0.00	10
P	0.89	1.00	0.94	24
R	1.00	0.73	0.84	33
T	0.00	0.00	0.00	0
V	0.76	0.94	0.84	62
X	1.00	1.00	1.00	3
Z	0.00	0.00	0.00	1
accuracy			0.84	334

Hình 35. Kết quả đánh giá VnCoreNLP

Có thể thấy, VnCoreNLP cho kết quả chính xác hơn rất nhiều so với phương pháp của nhóm. Lí do chính là do bộ dữ liệu train của nhóm quá ít khiến mô hình dẫn đến tình trạng bị overfitting. Để khắc phục tình trạng này nhóm sẽ thực hiện mở rộng bộ dữ liệu để đa dạng các nhãn, đồng thời, tìm thêm giải pháp để khắc phục hiện tượng những từ mới không tồn tại trong bộ từ điển.

Chương 6: Kết Luận

Trong đề tài này, nhóm đã áp dụng các kiến thức về xử lý ngôn ngữ tự nhiên để thực hiện tách từ bằng phương pháp So khớp cực đại. Đồng thời, nhóm cũng thực hiện tách từ và gán nhãn từ loại thủ công để tạo ngữ liệu cho mô hình Hidden Markov.

Thuật toán So khớp cực đại nhóm nhận xét là khá đơn giản và dễ cài đặt. Tuy nhiên, đi kèm với điều đó là thuật toán vẫn còn chưa tốt, nhất là trong trường hợp xảy ra nhập nhằng và việc phân loại các danh từ riêng.

Với mô hình Hidden Markov kết hợp cùng thuật toán Viterbi, nhóm nhận định đây là hướng tiếp cận khá ổn định và có cơ sở toán học vững chắc. Nhóm cũng đã bước đầu xây dựng thành công được mô hình, tuy nhiên, do hạn chế về mặt dữ liệu nên kết quả mô hình vẫn chưa đạt như kỳ vọng của nhóm. Hướng phát triển của nhóm sẽ là thu thập nhiều dữ liệu hơn, đa dạng các nhãn từ, không để dữ liệu bị mất cân bằng từ loại.

Bên cạnh đó, nhóm cũng sẽ tiếp tục tìm hiểu và phát triển một số phương pháp khác như TBL, Support Vector Machines, … Vì đã có nền tảng trong việc phát triển đề tài này, nên nhóm đã có nền tảng để triển khai một số hướng tiếp cận khác trong vấn đề phân nhãn từ loại.

Tài Liệu Tham Khảo

[1] https://www.thuvientailieu.vn/tai-lieu/xay-dung-treebank-tieng-viet-57885/?fbclid=IwAR1Vt6SxsPB30eSLhduY6sziXqqr6r1ckKUewZlXDXxgI18SZm5rnj6Fj4o

[2] https://github.com/vncorenlp/VnCoreNLP

[3] http://viet.jnlp.org/kien-thuc-co-ban-ve-xu-ly-ngon-ngu-tu-nhien/thuat-toan-tach-tu-tokenizer/thuat-toan-tach-tu

[4] https://timoday.edu.vn/bai-toan-tach-tu-tieng-viet/

[5] https://repository.vnu.edu.vn/bitstream/VNU_123/8188/1/01050000498.pdf