**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KHOA HỌC MÁY TÍNH**

**BÁO CÁO MÔN: XỬ LÝ NGÔN NGỮ TỰ NHIÊN NÂNG CAO**

**ĐỀ TÀI: XÂY DỰNG BỘ PHÂN TÍCH CÚ PHÁP THEO PCFG CHO TẬP CÂU**

Lớp:  **CS222.K11**

Giảng Viên : **Nguyễn Trọng Chỉnh**

Thành viên:

1. Võ Huyền Mỹ Trâm-16521538
2. Nguyễn Minh Cảnh-16521493
3. Trần Thiện Trong-16521541

TP.HCM, ngày tháng năm

Mục lục

* Hợp đồng nhóm 2
* Tài liệu tham khảo 35

**Chương 1: Giới thiệu đề tài**

* 1. POS-TAGGING là gì 6
  2. Quy trình thực hiện 7

**Chương 2: Chuẩn bị dữ liệu**

* 1. Yêu cầu đề tài 8
  2. Quy trình xây dựng ngữ liệu 8
  3. Gán nhãn dữ liệu 19

**Chương 3: Giải quyết bài toán**

* 1. Giới thiệu về hidden makov model 21
  2. Giới thiệu thuật toán viterbi 27
  3. Giới thiệu về maximun entropy 29

**Chương 4: Cài đặt chương trình**

* 1. Cài đặt chương trình 31
  2. Demo chương trình 34
  3. Tổng kết 35

# PHÂN CÔNG CÔNG VIỆC VÀ ĐÁNH GIÁ CÁC THÀNH VIÊN

# 

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | HỌ TÊN | MSSV | CÔNG VIỆC | ĐÁNH GIÁ |
| 1 | Nguyễn Minh Cảnh | 16521493 | -Tham gia soạn bộ dữ liệu ( gồm 35 câu)  -Tiền xử lý dữ liệu (phát hiện từ ghép, gán nhãn).  -Vẽ treebank 35 câu trên.  -Lập trình phần từ cây treebank sang biểu diễn tập luật.  -Viết báo cáo chương 3. | -Tham gia đầy đủ các buổi học nhóm.  -Hoàn thành nhiệm vụ được giao.  -Có tinh thần trách nhiệm. |
| 2 | Võ Huyền Mỹ Trâm | 16521538 | -Tham gia soạn bộ dữ liệu (gồm 35 câu).  -Tiền xử lý dữ liệu (phát hiện từ ghép, gán nhãn).  -Vẽ treebank 35 câu trên.  -Viết báo cáo chương 1, 6. | -Tham gia đầy đủ các buổi học nhóm.  -Hoàn thành nhiệm vụ được giao.  -Có tinh thần trách nhiệm. |
| 3 | Trần Thiện Trong | 16521541 | -Tham gia soạn bộ dữ liệu (gồm 35 câu).  -Tiền xử lý dữ liệu (phát hiện từ ghép, gán nhãn).  -Vẽ treebank 35 câu trên.  -Lập trình phần từ tập luật sang prolog.  -Viết báo cáo chương 2, 4, tổng hợp báo cáo. | -Tham gia đầy đủ các buổi học nhóm.  -Hoàn thành nhiệm vụ được giao.  -Có tinh thần trách nhiệm. |

Nhìn chung tất cả các thành viên đều có đóng góp cho đề tài như nhau, hợp tác vui vẻ và làm việc rất tích cực.

# CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

* 1. Giới thiệu bài toán:

-Thu thập 100 câu (tiếng Việt hoặc tiếng Anh) với chiều dài mỗi câu từ 8 đến 15 từ đơn.

-Xây dựng treebank bằng cách phân tích cú pháp 100 câu này, lưu thành file text theo cách biểu diễn của Penn Treebank (dạng dấu ngoặc tròn).

-Xác định tập luật sản sinh PCFG cho treebank vừa tạo.

-Viết chương trình phân tích cú pháp với tập luật PCFG đã xác định

-Dùng Stanford Parser để huấn luyện một bộ phân tích cú pháp dựa trên treebank đã tạo.

-Đánh giá và so sánh kết quả phân tích cú pháp do chương trình tự cài đặt và chương trình được huấn luyện theo Stanford Parser. Tập câu dùng để đánh giá gồm 30 câu được chọn tùy ý từ treebank đã tạo.

* 1. Ứng dụng:

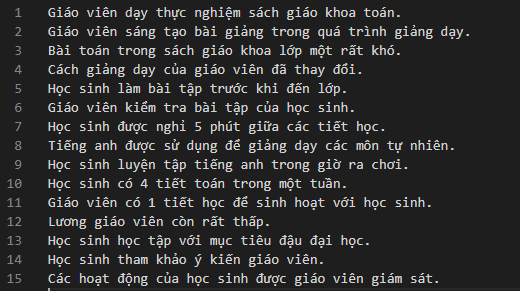
Treebank thường được dùng để xây dựng các hệ phân tích cú pháp chất lượng cao. Các hệ phân tích cú pháp này lại được sử dụng trong các ứng dụng quan trọng như truy vấn thông tin, dịch máy, v.v

# CHƯƠNG 2: CÁC BƯỚC CHUẨN BỊ DỮ LIỆU

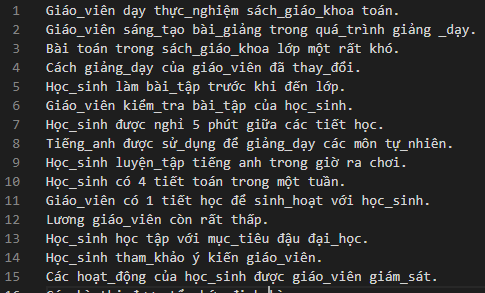
2.1. Thu thập dữ liệu:

Để giải quyết bài toán này, trước tiên cần có một lượng dữ liệu cụ thể. Nhóm chúng em đã thu thập dữ liệu bằng cách mỗi thành viên trong nhóm tự chuẩn bị 35 câu trong cùng một chủ đề là giáo dục. Lí do nhóm em chọn trong cùng một chủ đề nhằm mục đích hạn chế số lượng từ vựng giúp cho tập luật ít hơn để dễ dàng kiểm soát lỗi khi làm sai.

Dưới đây là hình ảnh của một đoạn dữ liệu được thu thập về:



Ở trên là dữ liệu thô được tổng hợp từ các thành viên trong nhóm. Trước khi sử dụng dữ liệu này gán nhãn nhóm em đã xử lí các trường hợp từ ghép để giúp cho kết quả phân tích tốt hơn. Do lượng dữ liệu không quá nhiều nên nhóm đã thực hiện việc xử lí từ ghép bằng tay, có nghĩa là tự xác định đâu là tự ghép trong câu, và có sử dụng công cụ VLSP để hỗ trợ những trường hợp nhập nhằng. Vì là thực hiện bằng tay công đoạn này nên độ chính xác về từ ghép gần như hoàn toàn chính xác. Hình bên dưới là dữ liệu sau khi được xử lí về từ ghép.

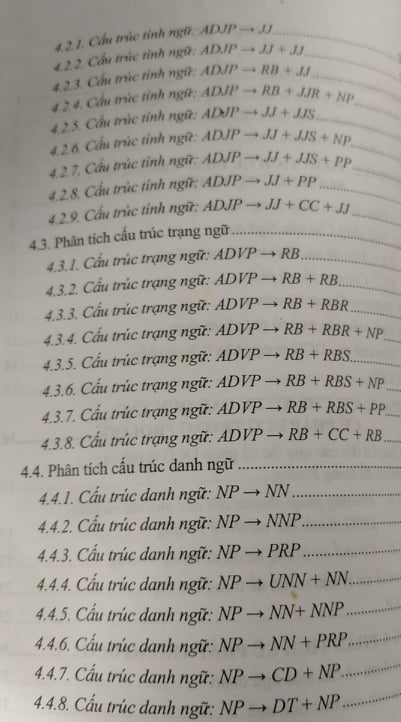


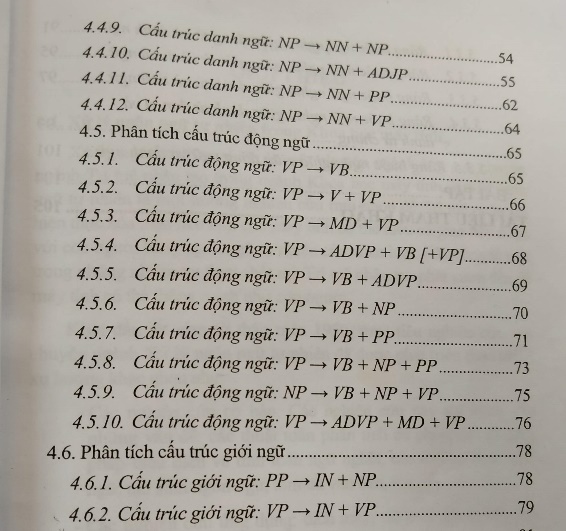
2.2. Gán nhãn dữ liệu:

Để thực hiện công đoạn gán nhãn dự liệu cho chính xác và nhanh chống, nhóm em đã đưa ra một bảng định nghĩa, quy tắc gán nhãn bên dưới:

|  |  |  |  |
| --- | --- | --- | --- |
| Nhãn | Tên | Chức năng | Ví dụ |
| NN | Danh từ chung | Là những từ chỉ sự vật (người, động vật, hiện tượng, khái niệm), vị trí (địa điểm, hướng trong không gian), đơn vị (đơn vị đo lường)[4] | Học\_sinh, giáo\_viên, trường, lớp,... |
| NNP | Danh từ riêng | Là tên riêng của một sự vật (tên người, tên địa phương, tên địa danh,..)[5] | Trong, Cảnh, Trâm,... |
| NP | Danh ngữ | **là** cụm từ tự do có quan hệ chính phụ và có **danh** từ làm thành tố trung tâm. dạng đầy đủ, **danh ngữ** gồm ba phần: phần trung tâm, phần phụ trước và phần phụ sau.[6] | Học\_sinh ấy, hoa và các bạn,... |
| VB | Động từ | Là những [từ](https://vi.wikipedia.org/wiki/T%E1%BB%AB) ([thành phần câu](https://vi.wikipedia.org/w/index.php?title=Th%C3%A0nh_ph%E1%BA%A7n_c%C3%A2u&action=edit&redlink=1)) dùng để biểu thị hoạt động (*chạy*, *đi*, *đọc*), [trạng thái](https://vi.wikipedia.org/w/index.php?title=Tr%E1%BA%A1ng_th%C3%A1i&action=edit&redlink=1) (*tồn tại*, *ngồi*).[7] | Đi, học, dạy, sáng\_tạo,... |
| VP | Động ngữ | Là cụm từ tự do có quan hệ chính phụ và có động từ làm thành tố trung tâm.[8] | Vừa đi học, dạy thực\_nghiệm,... |
| IN | Giới từ | Giới từ ([tiếng Anh](https://vi.wikipedia.org/wiki/Ti%E1%BA%BFng_Anh): preposition) là một [bộ phận lời nói](https://vi.wikipedia.org/w/index.php?title=B%E1%BB%99_ph%E1%BA%ADn_l%E1%BB%9Di_n%C3%B3i&action=edit&redlink=1) giới thiệu một [giới ngữ](https://vi.wikipedia.org/w/index.php?title=Gi%E1%BB%9Bi_ng%E1%BB%AF&action=edit&redlink=1).[9] | Vào, ở, cho, của,.. |
| PP | Giới ngữ | Là cụm từ tự do có quan hệ chính phụ và có giới từ làm thành tố trung tâm. | Trong lớp, giữa các tiết học, vào giảng\_dạy,.. |
| JJ | Tính từ | Là từ mà vai trò cú pháp chính của nó dùng để xác định một [danh từ](https://vi.wikipedia.org/wiki/Danh_t%E1%BB%AB) hoặc [đại từ](https://vi.wikipedia.org/wiki/%C4%90%E1%BA%A1i_t%E1%BB%AB), đưa thêm thông tin về [referent](https://vi.wikipedia.org/w/index.php?title=Referent&action=edit&redlink=1) của danh từ hoặc đại từ (referent là đối tượng hoặc ý tưởng mà từ hay đoạn văn hướng đến).[10] | Khó, thấp, giỏi, mềm,... |
| ADJP | Tính ngữ | Là cụm từ tự do có quan hệ chính phụ và có tính từ làm thành tố trung tâm. | Rất khó, rất thấp, luôn khó khăn,.. |
| RB | Trạng từ | Là những từ chỉ cách thức, thời gian, tuần suất, nơi chốn, mức độ, số lượng, nghi vấn, liên hệ,..[11] | Sẽ, sớm, rất, đã |
| ADVP | Trạng ngữ | Là cụm từ tự do có quan hệ chính phụ và có trạng từ làm thành tố trung tâm. | Định\_kỳ hằng\_tháng, rất lưu loát,... |
| CC | Liên từ | Là từ dùng để liên kết hai từ, hai cụm từ hoặc 2 mệnh đề lại với nhau. | Và, với,... |
| CD | Số từ (đếm dược) | Là những từ dùng để chỉ [số](https://vi.wikipedia.org/wiki/S%E1%BB%91) lượng và thứ tự của sự vật nào đó. | Một, năm, hai\_mươi,... |
| DET | Lượng từ(không đếm được | Là những [từ loại](https://vi.wikipedia.org/wiki/T%E1%BB%AB_lo%E1%BA%A1i) dùng để chỉ [số](https://vi.wikipedia.org/wiki/S%E1%BB%91) lượng và thứ tự của sự vật nào đó. | Các, nhiều, mỗi,... |
| MD | Động từ tình thái | Là những từ dùng để miêu tả trạng thái hoặc tình cảm của con người. Nó còn được hiểu như điều kiện để tồn tại sự vật nào đó. | Cần, phải,... |

Bên cạnh tập nhãn, việc xác định luật cũng rất quan trọng. Vì vậy nhóm em đã căn cứ vào tập luật trong sách tham khảo (Xử lí ngôn ngữ tự nhiên) của thầy Nguyễn Tuấn Đăng





-> Kết quả thực hiện gán nhãn:

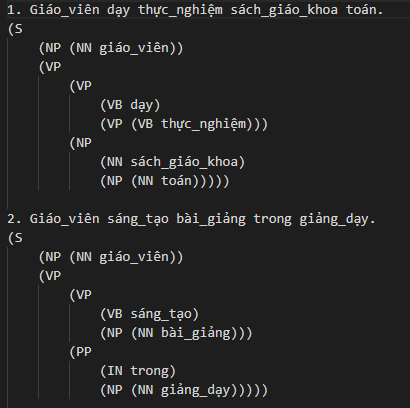
2.3. Xây dựng Treebank:

2.3.1. Treebank là gì?

Trong ngôn ngữ học, treebank là một văn bản phân tích cú pháp chú thích cấu trúc câu cú pháp hoặc ngữ nghĩa. Việc xây dựng các bộ phận được phân tích cú pháp vào đầu những năm 1990 đã cách mạng hóa ngôn ngữ học tính toán, được hưởng lợi từ dữ liệu thực nghiệm quy mô lớn.[3]

2.3.2. Ví dụ:

Dưới đây là ảnh minh họa về treebank nhóm em đã thực hiện:



2.4. Độ đồng thuận:

2.4.1. Hệ số đo đồng thuận:

**Hệ cố Cohen’s kappa (Cohen's kappa coefficient)** là một thống kê giúp đo đạc độ đồng thuận giữa các thành phần định tính (phân loại). Nó thường được xem là phương pháp hiệu quả hơn tính toán phần trăm đồng đồng thuận, vì k thường xem xét độ đồng thuận ngẫu nhiên. Phép đo Cohen’s kappa tính toán độ đồng thuận giữa hai biến đánh giá được dùng để phân loại N đối tượng thành C tập loại trừ lẫn nhau.

Hệ số Cohen’s kappa được định nghĩa và tính toán bởi công thức sau:

k=(p0−pe)/(1−pe)=1−((1−po)/(1−p))

2.4.2. Kết quả đo trong dữ liệu của bài toán này:

# CHƯƠNG 3: GIẢI QUYẾT BÀI TOÁN

# CHƯƠNG 4: ĐÁNH GIÁ VÀ SO SÁNH KẾT QUẢ PHÂN TÍCH

# CHƯƠNG 5: CÀI ĐẶT CHƯƠNG TRÌNH

## 4.1. Cài đặt chương trình:

Bước 1: Clone project tại link: [https://github.com/minhcanh198/VietnameseTreeBank](https://github.com/minhcanh198/VietnameseTreeBank?fbclid=IwAR2JRXL_EwtA3_z260Z3cHeaJ5sgTqxnpH2q0MiTrQMmB16roCHOotUJjtk)

Bước 2: Run file **toJson.py** bằng lệnh **python toJson.py**

Bước 3: Run file **treebank.py** bằng lệnh **python treebank.py**, copy kết quả in ra màn hình bỏ vào file **output.txt**

Bước 4: Run file **prolog.py** bằng lệnh **python prolog.py**, copy kết quả in ra màn hình bỏ vào file **prolog.txt**, xử lý đệ quy trái ở đây. Sau khi xử lí đệ quy trái thì ra được kết quả code của chương trình prolog

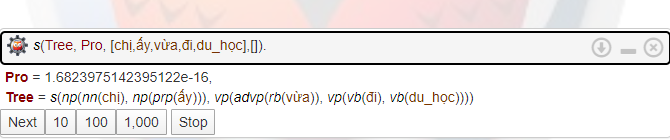
## 4.2. Test chương trình:

Bước 1: Có 2 IDE có thể chạy chương trình prolog:

* + Truy cập đường dẫn: <https://swish.swi-prolog.org/>, bấm vào nút program
  + Sử dụng app desktop, link download: <https://www.swi-prolog.org/Download.html>

Bước 2: Nhập đoạn chương trình kết quả bên trên và thực hiện lệnh truy vấn

Vd:



# CHƯƠNG 6: MỘT SỐ THÁCH THỨC

* Nhập nhằng ngữ nghĩa giữa các từ trong câu.

-

# TÀI LIỆU THAM KHẢO

[1] Sách tham khảo Xử lý ngôn ngữ tự nhiên, tác giả: Nguyễn Tuấn Đăng

[2] Slide tham khảo môn Xử lý ngôn ngữ tự nhiên nâng cao của thầy Nguyễn Trọng Chỉnh

[3] <https://en.wikipedia.org/wiki/Treebank>

[4] <https://ngnnghc.wordpress.com/tag/danh-t%E1%BB%AB-chung/>

[5] <http://tieuhoc.daytot.vn/thuat-ngu/Lop-4/DANH-TU-13.html>

[6] 1084-1-2110-1-10-20160519.pdf

[7] <https://vi.wikipedia.org/wiki/%C4%90%E1%BB%99ng_t%E1%BB%AB>

[8] 1488-1-2914-1-10-20160725.pdf

[9] <https://vi.wikipedia.org/wiki/Gi%E1%BB%9Bi_t%E1%BB%AB>

[10] <https://vi.wikipedia.org/wiki/T%C3%ADnh_t%E1%BB%AB>

[11] <https://vi.wikipedia.org/wiki/Tr%E1%BA%A1ng_t%E1%BB%AB>