**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 10 tháng 01 năm 2020

Mục lục

[PHÂN CÔNG CÔNG VIỆC VÀ ĐÁNH GIÁ CÁC THÀNH VIÊN 3](#_Toc29839591)

[I. CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI 4](#_Toc29839592)

[**1.** **Giới thiệu bài toán:** 4](#_Toc29839593)

[**2.** **Ứng dụng:** 4](#_Toc29839594)

[II. CHƯƠNG 2: CÁC BƯỚC CHUẨN BỊ DỮ LIỆU 5](#_Toc29839595)

[**1. Thu thập dữ liệu:** 5](#_Toc29839596)

[**2. Gán nhãn dữ liệu:** 6](#_Toc29839597)

[**3. Xây dựng Treebank:** 9](#_Toc29839598)

[**3.1. Treebank là gì?** 9](#_Toc29839599)

[**3.2. Ví dụ:** 9](#_Toc29839600)

[**4. Độ đồng thuận:** 9](#_Toc29839601)

[**4.1. Hệ số đo đồng thuận:** 9](#_Toc29839602)

[**4.2. Hệ số Cohen’s Kappa của bài toán này:** 10](#_Toc29839603)

[**4.3. Độ đồng thuận về cây cú pháp:** 10](#_Toc29839604)

[III. CHƯƠNG 3: GIẢI QUYẾT BÀI TOÁN 10](#_Toc29839605)

[**1.** **Xây dựng CFG từ file text PennTreeBank:** 10](#_Toc29839606)

[**2.** **Xây dựng văn phạm PCFG:** 11](#_Toc29839607)

[**3.** **Kết quả xử lý:** 12](#_Toc29839608)

[IV. CHƯƠNG 4: ĐÁNH GIÁ VÀ SO SÁNH KẾT QUẢ PHÂN TÍCH 20](#_Toc29839609)

[V. CHƯƠNG 5: CÀI ĐẶT CHƯƠNG TRÌNH 20](#_Toc29839610)

[**1.** **Cài đặt chương trình:** 20](#_Toc29839611)

[**2.** **Test chương trình:** 21](#_Toc29839612)

[VI. CHƯƠNG 6: MỘT SỐ THÁCH THỨC 21](#_Toc29839613)

[VII. TÀI LIỆU THAM KHẢO 22](#_Toc29839614)

# 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

## **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.

## **Ứng dụng:**

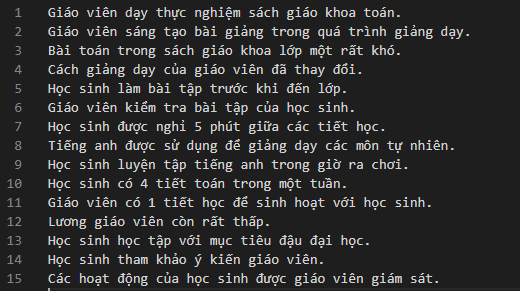
Treebank thường được dùng để xây dựng các hệ phân tích cú pháp chất lượng cao. Các hệ phân tích cú pháp này lại được sử dụng trong các ứng dụng quan trọng như truy vấn thông tin, dịch máy, v.v

# CHƯƠNG 2: CÁC BƯỚC CHUẨN BỊ DỮ LIỆU

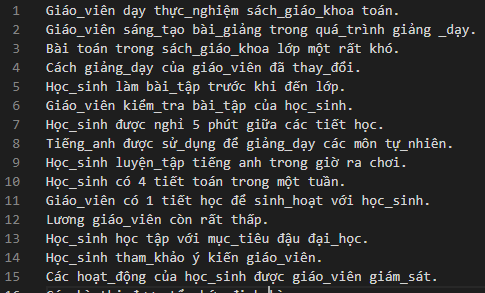
## **1. Thu thập dữ liệu:**

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 nhóm tổng hợp:



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 lý 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 là rất cao. Hình bên dưới là dữ liệu sau khi được xử lý về từ ghép:

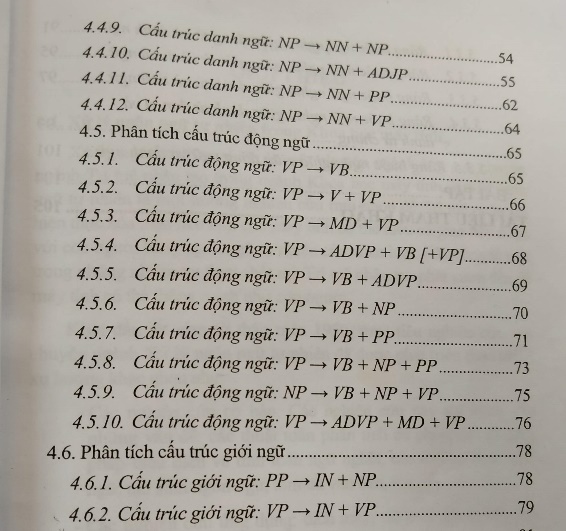
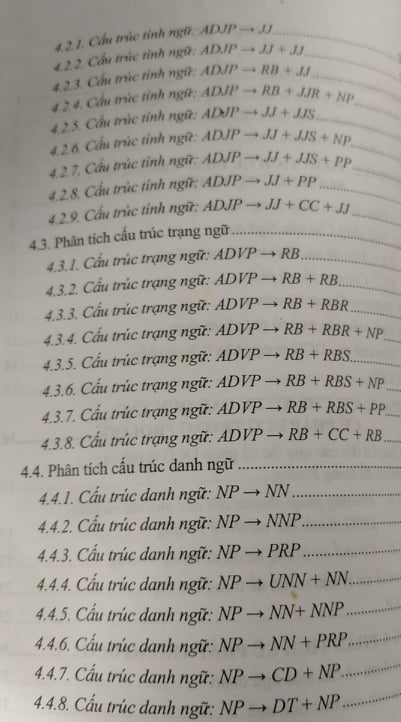


## **2. Gán nhãn dữ liệu:**

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:

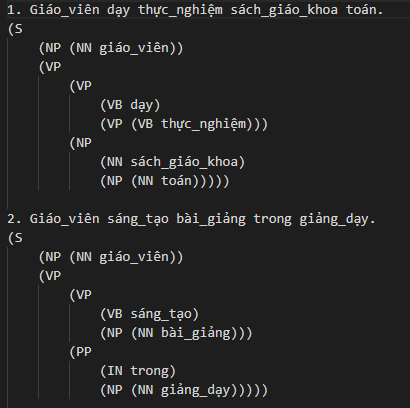
## **3. Xây dựng Treebank:**

### **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]

### **3.2. Ví dụ:**

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



## **4. Độ đồng thuận:**

### **4.1. Hệ số đo đồng thuận:**

**Hệ số 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:

Trong đó:

* là giá trị đồng thuận quan sát được giữa các biến đánh giá
* là xác suất giả định của khả năng đồng thuận

### **4.2. Hệ số Cohen’s Kappa của bài toán này:**

Độ đồng thuận Kappa được tính trên 10 câu trong tập data do cả 3 người cùng thực hiện gán nhãn cho 10 câu đó:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **TRONG – CẢNH** | **TRONG – TRÂM** | **CẢNH – TRÂM** |
|  | 94% | 91% | 91% |

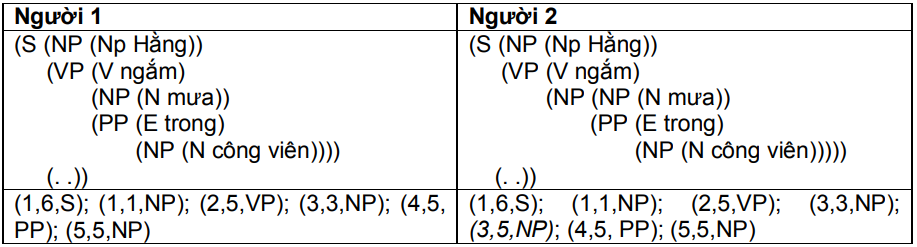
### **4.3. Độ đồng thuận về cây cú pháp:**

Công thức:

Trong đó:

* C1 là số thành phần cú pháp trong kết quả gán nhãn của người thứ nhất
* C2 là số thành phần cú pháp trong kết quả gán nhãn của người thứ hai
* C là số thàng phần cú pháp giống nhau

Ví dụ: ta có câu: “Hằng ngắm mưa trong công viên”



Ta có:

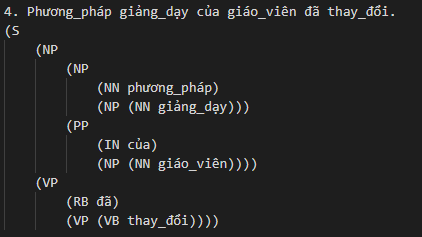
Theo đó, ta có:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **TRONG – CẢNH** | **TRONG – TRÂM** | **CẢNH – TRÂM** |
|  | 91% | 83% | 81% |

# CHƯƠNG 3: GIẢI QUYẾT BÀI TOÁN

## **Xây dựng CFG từ file text PennTreeBank:**

Sau quá trình gán nhãn và xây dựng PennTreebank, ta thu được một file text chứa 100 cây cú pháp, hình minh họa cây số 4 trong tập data:



Việc ta cần làm là xử lý cây cú pháp này để sinh ra tập luật 🡪 xây dựng văn phạm CFG 🡪 xây dựng văn phạm PCFG 🡪 xây dựng bộ luật cho prolog.

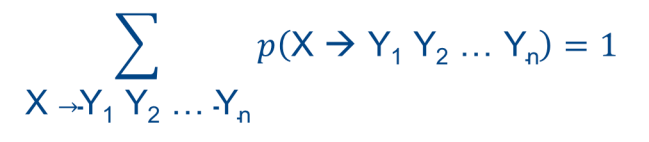
Để giải quyết vấn đề trên, thay vì phải phụ thuộc vào các phần mềm, công cụ có sẵn, nhóm đã cố gắng parse file text bằng cách sử dụng cấu trúc **Tree** từ thư viện **nltk** của **python**. Trong đó, node gốc là tất cả các nhãn *Non-terminal* và *terminal*, node lá là *từ vựng*. Sau khi áp dụng cho câu trên ta sẽ được văn phạm CFG sau:

* S 🡪 NP VP
* NP 🡪 NP PP
* NP 🡪 NN NP
* NP 🡪 NN
* VP 🡪 RB VP
* VP 🡪 VB
* PP 🡪 IN NP
* NN 🡪 phương\_pháp
* NN 🡪 giảng\_dạy
* NN 🡪 giáo\_viên
* VB 🡪 thay\_đổi
* RB 🡪 đã
* IN 🡪 của

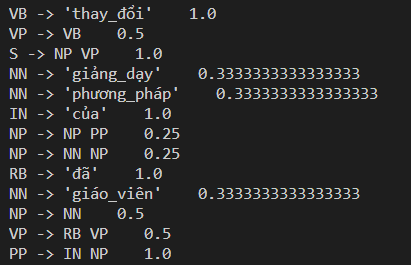
## **Xây dựng văn phạm PCFG:**

Định nghĩa văn phạm PCFG:

* Gồm một văn phạm CFG, .
* Một tham số cho mỗi luật sản sinh. được được coi là xác suất chọn luật sản sinh sao cho với mỗi thì:



Áp dụng định nghĩa trên vào câu “*Phương\_pháp giảng\_dạy của giáo\_viên đã thay\_đổi*”, ta được văn phạm PCFG sau:



Từ đó, ta xây dựng PCFG cho toàn bộ tập data.

## **Kết quả xử lý:**

* Văn phạm PCFG:

S -> NP VP 1.0

NP -> NNP 0.05142857142857143

NP -> UNN NP 0.02

NP -> NN 0.5228571428571429

NP -> NN ADJP 0.025714285714285714

NP -> NN CD 0.002857142857142857

NP -> NP CC NP 0.008571428571428572

NP -> NN VP 0.014285714285714285

NP -> PRP 0.02857142857142857

NP -> CD NP 0.017142857142857144

NP -> JJ NP 0.005714285714285714

NP -> NN PP 0.02

NP -> NP VP 0.014285714285714285

NP -> NN PRP 0.025714285714285714

NP -> NN NN 0.008571428571428572

NP -> NP PP 0.017142857142857144

NP -> NN NP 0.13714285714285715

NP -> NP PRP 0.005714285714285714

NP -> DET NP 0.06285714285714286

NP -> NP NP 0.002857142857142857

NP -> NP ADJP 0.008571428571428572

VP -> VB 0.04132231404958678

VP -> VB CC VB 0.004132231404958678

VP -> VP PP 0.16942148760330578

VP -> VB ADJP 0.03305785123966942

VP -> UNN NP 0.004132231404958678

VP -> VB ADVP 0.049586776859504134

VP -> VP ADVP 0.01652892561983471

VP -> VB VB 0.008264462809917356

VP -> MD VP 0.02066115702479339

VP -> VB PP 0.049586776859504134

VP -> RB VP 0.2024793388429752

VP -> RB VB 0.02066115702479339

VP -> VB NP 0.29338842975206614

VP -> VB VP 0.028925619834710745

VP -> VP NP 0.03305785123966942

VP -> VP VP 0.02066115702479339

VP -> ADVP VP 0.004132231404958678

PP -> IN NP 0.9090909090909091

PP -> RB VP 0.015151515151515152

PP -> IN VP 0.07575757575757576

ADJP -> JJ 0.6896551724137931

ADJP -> RB ADJP 0.3103448275862069

ADVP -> RB 0.85

ADVP -> RB ADVP 0.15

NN -> 'thể\_dục' 0.0037313432835820895

VB -> 'gặp' 0.015267175572519083

NN -> 'hôm\_nay' 0.0037313432835820895

NN -> 'phút' 0.007462686567164179

VB -> 'giúp' 0.007633587786259542

RB -> 'thường' 0.023809523809523808

NN -> 'an\_toàn' 0.0037313432835820895

CD -> 'một' 0.42857142857142855

NN -> 'truyện' 0.0037313432835820895

NN -> 'mâu\_thuẫn' 0.0037313432835820895

RB -> 'đang' 0.14285714285714285

NN -> 'kỳ' 0.0037313432835820895

NN -> 'thầy' 0.0037313432835820895

NN -> 'mọi\_người' 0.007462686567164179

VB -> 'trang\_điểm' 0.007633587786259542

VB -> 'đậu' 0.007633587786259542

NN -> 'cô' 0.007462686567164179

JJ -> 'chăm\_chỉ' 0.045454545454545456

NN -> 'sân\_trường' 0.0037313432835820895

VB -> 'chuẩn\_bị' 0.007633587786259542

NN -> 'bảo\_vệ' 0.0037313432835820895

VB -> 'sử\_dụng' 0.015267175572519083

UNN -> 'cuốn' 0.125

NN -> 'bóng' 0.007462686567164179

DET -> 'mọi' 0.045454545454545456

RB -> 'khó\_khăn' 0.011904761904761904

VB -> 'tích\_cực' 0.007633587786259542

MD -> 'cần' 0.6

NN -> 'sân' 0.011194029850746268

NN -> 'tháng' 0.007462686567164179

NN -> 'gái' 0.0037313432835820895

NN -> 'cuộc\_sống' 0.0037313432835820895

NN -> 'học\_phí' 0.0037313432835820895

NN -> 'quê' 0.0037313432835820895

JJ -> 'thấp' 0.045454545454545456

VB -> 'làm' 0.03816793893129771

NN -> 'kỹ\_thuật' 0.0037313432835820895

VB -> 'hoàn\_thành' 0.007633587786259542

NN -> 'bài\_tập' 0.022388059701492536

VB -> 'viểt' 0.007633587786259542

VB -> 'tốn' 0.007633587786259542

NN -> 'tiếng\_anh' 0.03731343283582089

NN -> 'ý\_kiến' 0.0037313432835820895

NN -> 'kỹ\_năng' 0.011194029850746268

VB -> 'tham\_khảo' 0.007633587786259542

VB -> 'luyện\_tập' 0.007633587786259542

NN -> 'ngày' 0.0037313432835820895

NNP -> 'tết' 0.16666666666666666

NN -> 'sinh\_viên' 0.026119402985074626

VB -> 'đánh\_giá' 0.007633587786259542

RB -> 'bị' 0.047619047619047616

NN -> 'hiệu\_trưởng' 0.0037313432835820895

VB -> 'nói' 0.007633587786259542

NN -> 'môn' 0.022388059701492536

NN -> 'sách' 0.0037313432835820895

PRP -> 'nay' 0.09523809523809523

NN -> 'học\_sinh' 0.13805970149253732

JJ -> 'sắp\_tới' 0.045454545454545456

NN -> 'niềm\_tin' 0.0037313432835820895

VB -> 'đá' 0.007633587786259542

RB -> 'mới' 0.023809523809523808

NN -> 'giảng\_dạy' 0.011194029850746268

VB -> 'cảm\_ơn' 0.007633587786259542

VB -> 'sáng\_tạo' 0.007633587786259542

NN -> 'phụ\_huynh' 0.0037313432835820895

VB -> 'phát\_triển' 0.007633587786259542

RB -> 'trước\_khi' 0.011904761904761904

NN -> 'du\_học\_sinh' 0.0037313432835820895

RB -> 'to' 0.011904761904761904

IN -> 'trong' 0.24615384615384617

NN -> 'trường\_học' 0.0037313432835820895

NN -> 'kỳ\_thi' 0.0037313432835820895

IN -> 'giữa' 0.015384615384615385

NN -> 'công\_nhân\_viên' 0.0037313432835820895

VB -> 'ghi' 0.007633587786259542

NN -> 'mọi\_chuyện' 0.0037313432835820895

CD -> 'năm' 0.14285714285714285

RB -> 'đã' 0.07142857142857142

NN -> 'nơi' 0.0037313432835820895

VB -> 'về' 0.007633587786259542

VB -> 'nở' 0.007633587786259542

NN -> 'người\_yêu' 0.0037313432835820895

VB -> 'tuân\_thủ' 0.007633587786259542

NN -> 'chú' 0.0037313432835820895

NN -> 'vệ\_sinh' 0.0037313432835820895

VB -> 'nghiệm\_thu' 0.007633587786259542

NN -> 'dự\_án' 0.007462686567164179

VB -> 'nghĩ' 0.007633587786259542

VB -> 'sinh\_hoạt' 0.007633587786259542

VB -> 'tụ\_tập' 0.007633587786259542

VB -> 'đọc' 0.007633587786259542

VB -> 'chia\_tay' 0.007633587786259542

RB -> 'nặng' 0.011904761904761904

RB -> 'rất' 0.10714285714285714

VB -> 'lấy' 0.007633587786259542

NN -> 'vấn\_đề' 0.0037313432835820895

NN -> 'văn' 0.0037313432835820895

NN -> 'nhóm' 0.0037313432835820895

VB -> 'liên\_quan' 0.007633587786259542

NN -> 'tuần' 0.007462686567164179

RB -> 'tốt\_đẹp' 0.011904761904761904

VB -> 'du\_học' 0.007633587786259542

JJ -> 'khó\_khăn' 0.09090909090909091

NN -> 'thứ\_hai' 0.0037313432835820895

NN -> 'hè' 0.007462686567164179

VB -> 'mưa' 0.007633587786259542

UNN -> 'một' 0.125

VB -> 'nghỉ' 0.015267175572519083

VB -> 'lựa\_chọn' 0.007633587786259542

VB -> 'cổ\_vũ' 0.007633587786259542

NN -> 'câu\_hỏi' 0.0037313432835820895

VB -> 'cập\_nhật' 0.007633587786259542

VB -> 'giúp\_đỡ' 0.007633587786259542

NN -> 'sáng' 0.0037313432835820895

NN -> 'bà' 0.0037313432835820895

VB -> 'nhờ' 0.007633587786259542

NN -> 'cỏ' 0.0037313432835820895

NN -> 'nội\_quy' 0.0037313432835820895

VB -> 'kết\_thúc' 0.007633587786259542

NN -> 'cổng' 0.0037313432835820895

VB -> 'đến' 0.015267175572519083

RB -> 'luôn' 0.023809523809523808

NN -> 'tự\_nhiên' 0.0037313432835820895

NNP -> 'nam' 0.3888888888888889

VB -> 'trộm' 0.007633587786259542

NN -> 'bài' 0.0037313432835820895

JJ -> 'lớn' 0.09090909090909091

VB -> 'học\_tập' 0.007633587786259542

PRP -> 'bạn' 0.047619047619047616

VB -> 'rớt' 0.007633587786259542

NN -> 'toán' 0.022388059701492536

NN -> 've\_sầu' 0.0037313432835820895

NNP -> 'lan' 0.16666666666666666

RB -> 'sẽ' 0.03571428571428571

VB -> 'là' 0.030534351145038167

NN -> 'lớp' 0.011194029850746268

IN -> 'ngoài' 0.03076923076923077

VB -> 'dọn' 0.007633587786259542

NN -> 'ngoại\_khóa' 0.0037313432835820895

PRP -> 'đó' 0.14285714285714285

DET -> 'mỗi' 0.045454545454545456

NNP -> 'phượng' 0.05555555555555555

DET -> 'vài' 0.045454545454545456

IN -> 'bằng' 0.046153846153846156

NN -> 'mơ\_ước' 0.0037313432835820895

NN -> 'sinh\_học' 0.0037313432835820895

RB -> 'sớm' 0.011904761904761904

VB -> 'tổ\_chức' 0.022900763358778626

NN -> 'áp\_lực' 0.0037313432835820895

VB -> 'đóng' 0.007633587786259542

NN -> 'đồng\_nghiệp' 0.0037313432835820895

NN -> 'bài\_giảng' 0.0037313432835820895

NN -> 'chiến\_thắng' 0.0037313432835820895

RB -> 'vụng' 0.011904761904761904

JJ -> 'giỏi' 0.045454545454545456

VB -> 'thi' 0.007633587786259542

NN -> 'đồng' 0.0037313432835820895

RB -> 'vừa' 0.011904761904761904

NN -> 'công\_ty' 0.007462686567164179

NN -> 'lá' 0.0037313432835820895

NN -> 'thi\_thử' 0.0037313432835820895

IN -> 'về' 0.015384615384615385

VB -> 'biến\_mất' 0.007633587786259542

DET -> 'nhiều' 0.22727272727272727

NN -> 'mùa' 0.007462686567164179

VB -> 'cho' 0.015267175572519083

VB -> 'gây' 0.007633587786259542

NN -> 'miền\_quê' 0.0037313432835820895

PRP -> 'chúng\_ta' 0.047619047619047616

RB -> 'cao' 0.011904761904761904

CC -> 'và' 0.75

CD -> 'hai\_mươi' 0.14285714285714285

RB -> 'sắp' 0.011904761904761904

RB -> 'quá' 0.011904761904761904

NN -> 'khuôn\_viên' 0.0037313432835820895

NN -> 'tiết\_học' 0.007462686567164179

NN -> 'kì\_thi' 0.0037313432835820895

NN -> 'giờ' 0.018656716417910446

NN -> 'nhà\_trường' 0.011194029850746268

NN -> 'tiệc' 0.0037313432835820895

JJ -> 'mới' 0.09090909090909091

NN -> 'chị' 0.0037313432835820895

IN -> 'để' 0.06153846153846154

NN -> 'thời\_gian' 0.0037313432835820895

NN -> 'điện\_tử' 0.0037313432835820895

UNN -> 'cây' 0.125

IN -> 'với' 0.1076923076923077

NN -> 'căn\_tin' 0.0037313432835820895

VB -> 'chờ' 0.007633587786259542

NN -> 'luật' 0.0037313432835820895

JJ -> 'mềm' 0.09090909090909091

NN -> 'đồng\_phục' 0.007462686567164179

NN -> 'đêm' 0.007462686567164179

JJ -> 'tốt' 0.045454545454545456

RB -> 'riêng' 0.011904761904761904

NN -> 'tuyển\_sinh' 0.0037313432835820895

VB -> 'thực\_nghiệm' 0.007633587786259542

UNN -> 'đợt' 0.125

NN -> 'pháo\_hoa' 0.007462686567164179

RB -> 'hầu\_như' 0.011904761904761904

NN -> 'phương\_pháp' 0.011194029850746268

RB -> 'hay' 0.047619047619047616

VB -> 'hỗ' 'trợ' 0.007633587786259542

NN -> 'mục\_tiêu' 0.0037313432835820895

IN -> 'ở' 0.06153846153846154

VB -> 'thiếu' 0.007633587786259542

NN -> 'đại\_học' 0.014925373134328358

NNP -> 'campuchia' 0.1111111111111111

VB -> 'kêu' 0.007633587786259542

VB -> 'hoạt\_động' 0.007633587786259542

DET -> 'các' 0.5909090909090909

VB -> 'học' 0.03816793893129771

PRP -> 'này' 0.09523809523809523

JJ -> 'dễ' 0.045454545454545456

NN -> 'tổn\_thương' 0.0037313432835820895

PRP -> 'nó' 0.047619047619047616

RB -> 'trễ' 0.011904761904761904

VB -> 'giảng\_dạy' 0.007633587786259542

IN -> 'của' 0.12307692307692308

MD -> 'phải' 0.4

VB -> 'ngắm' 0.015267175572519083

RB -> 'thuộc' 0.011904761904761904

VB -> 'cố\_gắng' 0.007633587786259542

UNN -> 'tỉ' 0.125

JJ -> 'nóng\_hổi' 0.045454545454545456

NN -> 'đầu\_vào' 0.0037313432835820895

VB -> 'kiểm\_tra' 0.015267175572519083

NN -> 'giải\_lao' 0.007462686567164179

VB -> 'hỗ\_trợ' 0.022900763358778626

NN -> 'đội\_tuyển' 0.0037313432835820895

NN -> 'học\_tập' 0.0037313432835820895

VB -> 'thay\_đổi' 0.022900763358778626

VB -> 'ăn' 0.007633587786259542

UNN -> 'bốn' 0.125

CD -> 'hai' 0.14285714285714285

RB -> 'lưu\_loát' 0.011904761904761904

NN -> 'hoa' 0.0037313432835820895

VB -> 'thích' 0.015267175572519083

IN -> 'đến' 0.015384615384615385

JJ -> 'thú\_vị' 0.045454545454545456

NN -> 'trường' 0.018656716417910446

VB -> 'dạy' 0.007633587786259542

NN -> 'bạn' 0.014925373134328358

VB -> 'động\_viên' 0.007633587786259542

NN -> 'bài\_toán' 0.0037313432835820895

NN -> 'tài\_liệu' 0.018656716417910446

VB -> 'thưởng' 0.007633587786259542

VB -> 'trang\_trải' 0.007633587786259542

NN -> 'đội\_bóng' 0.0037313432835820895

VB -> 'có' 0.03816793893129771

NN -> 'việc' 0.007462686567164179

VB -> 'ban\_hành' 0.007633587786259542

RB -> 'thêm' 0.011904761904761904

NN -> 'sách\_giáo\_khoa' 0.018656716417910446

NN -> 'tài\_chính' 0.0037313432835820895

NN -> 'lãnh\_đạo' 0.0037313432835820895

NN -> 'cây' 0.0037313432835820895

NN -> 'thực\_phẩm' 0.0037313432835820895

NN -> 'lương' 0.007462686567164179

VB -> 'phạt' 0.015267175572519083

NN -> 'nhau' 0.0037313432835820895

NN -> 'tiết' 0.0037313432835820895

JJ -> 'khó' 0.045454545454545456

NN -> 'giao\_thừa' 0.007462686567164179

PRP -> 'chúng\_tôi' 0.09523809523809523

NNP -> 'việt\_nam' 0.05555555555555555

CD -> 'mười' 0.14285714285714285

PRP -> 'ấy' 0.14285714285714285

NN -> 'tình\_yêu' 0.0037313432835820895

RB -> 'định\_kỳ' 0.023809523809523808

VB -> 'hỏng' 0.007633587786259542

UNN -> 'chiếc' 0.25

VB -> 'đón' 0.015267175572519083

VB -> 'đi' 0.030534351145038167

VB -> 'nói\_chuyện' 0.007633587786259542

JJ -> 'cuối' 0.09090909090909091

IN -> 'trước' 0.03076923076923077

IN -> 'vào' 0.16923076923076924

PRP -> 'tôi' 0.23809523809523808

RB -> 'khẩn\_trương' 0.011904761904761904

VB -> 'cấm' 0.007633587786259542

VB -> 'thảo\_luận' 0.007633587786259542

RB -> 'còn' 0.011904761904761904

IN -> 'cho' 0.07692307692307693

RB -> 'kịp\_thời' 0.011904761904761904

NN -> 'bằng\_khen' 0.0037313432835820895

NN -> 'trường\_hợp' 0.0037313432835820895

NN -> 'bàng' 0.0037313432835820895

PRP -> 'tới' 0.047619047619047616

RB -> 'gần' 0.011904761904761904

DET -> 'những' 0.045454545454545456

VB -> 'ra' 0.007633587786259542

JJ -> 'cao' 0.045454545454545456

NN -> 'xã\_hội' 0.0037313432835820895

NN -> 'con' 0.0037313432835820895

VB -> 'đưa' 0.007633587786259542

NN -> 'hoạt\_động' 0.007462686567164179

RB -> 'được' 0.16666666666666666

JJ -> 'lười\_biếng' 0.045454545454545456

VB -> 'giám\_sát' 0.007633587786259542

RB -> 'liên\_tục' 0.011904761904761904

NN -> 'chiều' 0.007462686567164179

NN -> 'giáo\_viên' 0.048507462686567165

NN -> 'đồ\_án' 0.007462686567164179

JJ -> 'xanh' 0.045454545454545456

NN -> 'điện\_thoại' 0.007462686567164179

VB -> 'nhận' 0.007633587786259542

RB -> 'hằng\_tháng' 0.011904761904761904

VB -> 'rụng' 0.007633587786259542

NN -> 'nữ\_sinh' 0.0037313432835820895

RB -> 'đều' 0.011904761904761904

NN -> 'mức\_giá' 0.0037313432835820895

VB -> 'trốn' 0.007633587786259542

VB -> 'đánh\_nhau' 0.007633587786259542

VB -> 'thì' 0.05343511450381679

VB -> 'tham\_gia' 0.007633587786259542

NN -> 'trời' 0.0037313432835820895

NN -> 'máy\_tính' 0.0037313432835820895

NN -> 'anh' 0.0037313432835820895

VB -> 'chơi' 0.007633587786259542

CC -> 'với' 0.25

NN -> 'học' 0.022388059701492536

VB -> 'cắt' 0.007633587786259542

NNP -> 'hoa' 0.05555555555555555

* Xử lý đệ quy trái cho chương trình prolog:

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

## Ưu điểm:

* Kết quả phân tích hợp lý về mặt cấu trúc và ngữ nghĩa

## Hạn chế:

* Chương trình prolog còn chạy chậm khi xử lý các câu dài, phức tạp

# CHƯƠNG 5: CÀI ĐẶT CHƯƠNG TRÌNH

## **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

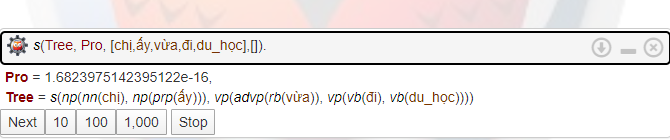
## **Test chương trình:**

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>