NGỮ NGHĨA HỌC TÍNH TOÁN

BÀI 1 – PHÂN TÍCH CÚ PHÁP THÀNH TỐ VỚI THƯ VIỆN NLTK

NGUYỄN TRỌNG CHỈNH

TRÌNH BÀY

1. GOOGLE COLAB
2. THƯ VIỆN NLTK
3. TẬP LUẬT SẢN SINH
4. BIỂU DIỄN CÂY CÚ PHÁP
5. BÀI THỰC HÀNH

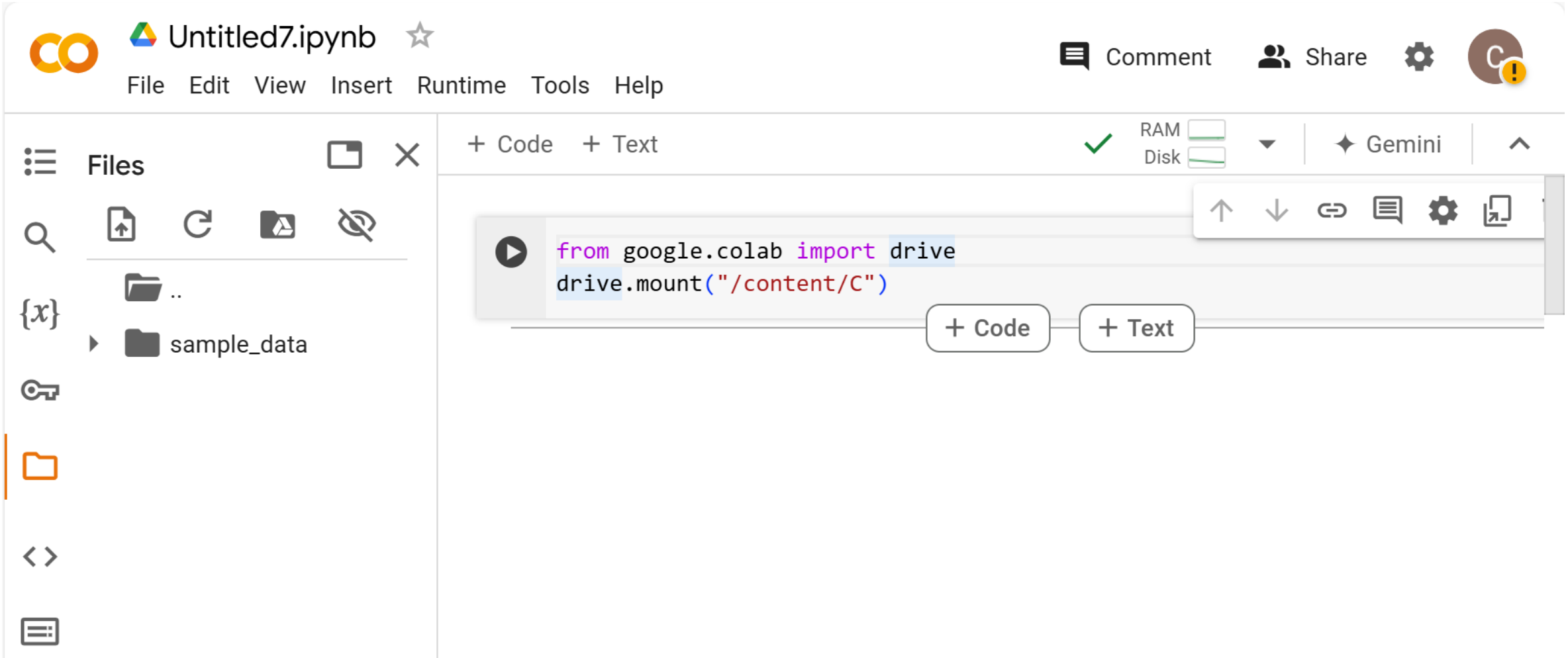
1. GOOGLE COLAB

1. GOOGLE COLAB

https://colab.research.google.com/

- Là một máy tính ảo:
  - ➢ Hệ điều hành Linux.
  - ➢ Có thể sử dụng thêm GPU (Tesla T4)
- Ưu điểm:
  - ➢ Dễ dàng cộng tác
  - ➢ Mỗi máy ảo dành riêng cho một dự án
- Nhược điểm: Giới hạn thời gian sử dụng

1. GOOGLE COLAB



1. GOOGLE COLAB

- Lệnh sử dụng trong Shell
  - !mkdir DIR
  - !pip install PACK
  - !unzip DATA.zip -d DIR
- Lệnh sử dụng trong Python
  - import PACK
  - class CLS(BASE_CLS):
  - …

1. GOOGLE COLAB

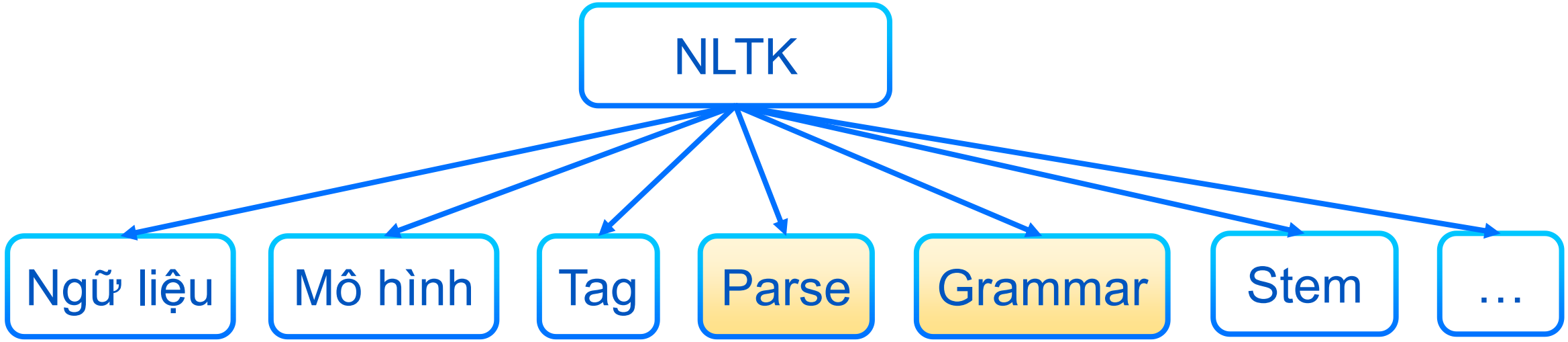
- Một môi trường Python cho mỗi máy ảo
  - Trạng thái của môi trường Python giống nhau ở tất cả code cell
  - Sử dụng như đang dùng trên máy tính cá nhân.

2. THƯ VIỆN NLTK

2. THƯ VIỆN NLTK

https://www.nltk.org/

- Thư viện xử lý ngôn ngữ tự nhiên:



2. THƯ VIỆN NLTK

Phân tích cú pháp với thuật toán Top-down.

```
import nltk
from nltk.grammar import CFG
from nltk.parse import  RecursiveDescentParser
G = CFG.fromstring(P)
ps = RecursiveDescentParser(G)
results = ps.parse(sent)
print(results[0].pformat())
```

2. THƯ VIỆN NLTK

Phân tích cú pháp với thuật toán quy hoạch động Earley.

```
import nltk
from nltk.grammar import CFG
from nltk.parse import  EarleyChartParser
G = CFG.fromstring(P)
ps = EarleyChartParser(G)
results = ps.parse(sent)
print(results[0].pformat())
```

2. THƯ VIỆN NLTK

Các thao tác trên cây cú pháp.

from nltk.tree import Tree

Tạo cây cú pháp từ dạng dấu ngoặc tròn (pformat):

Tree.fromstring(s)

Trích tất cả luật sản sinh đã được áp dụng cho cây:

tobj.productions()

Lấy danh sách cây con: tobj.subtrees()

Lấy các nút lá: tobj.leaves()

3. TẬP LUẬT SẢN SINH

3. TẬP LUẬT SẢN SINH

Các bước xây dựng tập luật sản sinh:

1. Chọn một tập câu.
2. Phân tích cú pháp thủ công:
    - Tham khảo từ điển. Từ điển tiếng Việt VLSP:

      https://vlsp.hpda.vn/demo/?page=vcl
    - Áp dụng các luật văn phạm.
3. Tổng hợp luật sản sinh.

3. TẬP LUẬT SẢN SINH

Ví dụ: Tập câu thu thập được gồm một câu

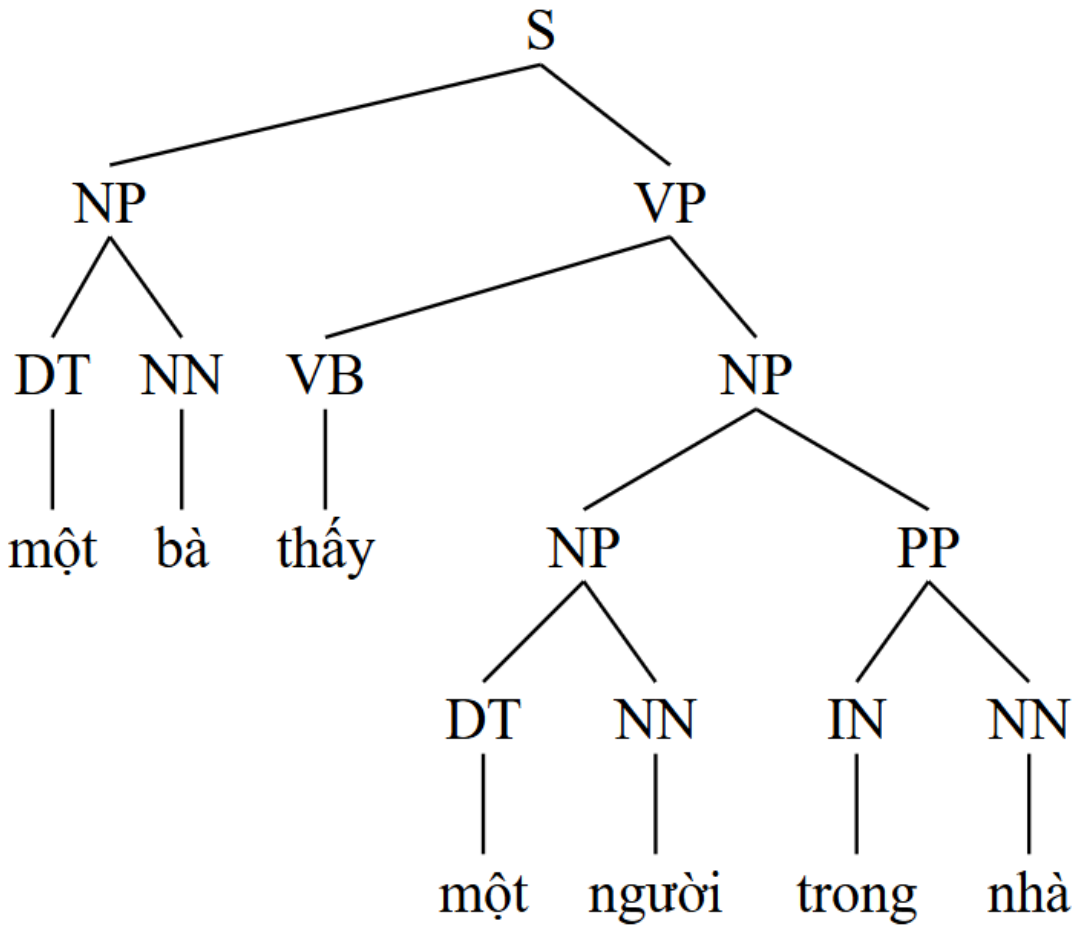
Một bà thấy một người trong nhà

Phân tích cú pháp:

- Sử dụng từ điển tiếng Việt
- Các quy tắc thành lập câu, danh ngữ, động ngữ và giới ngữ

3. TẬP LUẬT SẢN SINH

Kết quả phân tích cú pháp



3. TẬP LUẬT SẢN SINH

Tổng hợp luật sản sinh

S -> NP VP

VP -> VB NP

NP -> NP PP

VB -> 'thấy’

DT -> 'một’

VP -> VB

NP -> DT NN

PP -> IN NN

NN -> 'bà' | 'người' | 'nhà’

IN -> 'trong'

3. TẬP LUẬT SẢN SINH

Khử đệ quy trái, nếu dùng thuật toán Top-down

S -> NP VP

VP -> VB NP

NP -> NP1 PP

PP -> IN NN

VB -> 'thấy'

DT -> 'một'

VP -> VB

NP -> DT NN

NP1 -> DT NN

NN -> 'bà' | 'người' | 'nhà'

IN -> 'trong'

4. BIỂU DIỄN CÂY CÚ PHÁP

4. BIỂU DIỄN CÂY CÚ PHÁP

1. Cài đặt thư viện svgling

```
!pip install svgling
```

2. Sử dụng hàm display của thư viện IPython

```
from IPython import display
```

3. Phân tích cú pháp của một câu
4. Hiển thị cây cú pháp với hàm display.

4. BIỂU DIỄN CÂY CÚ PHÁP

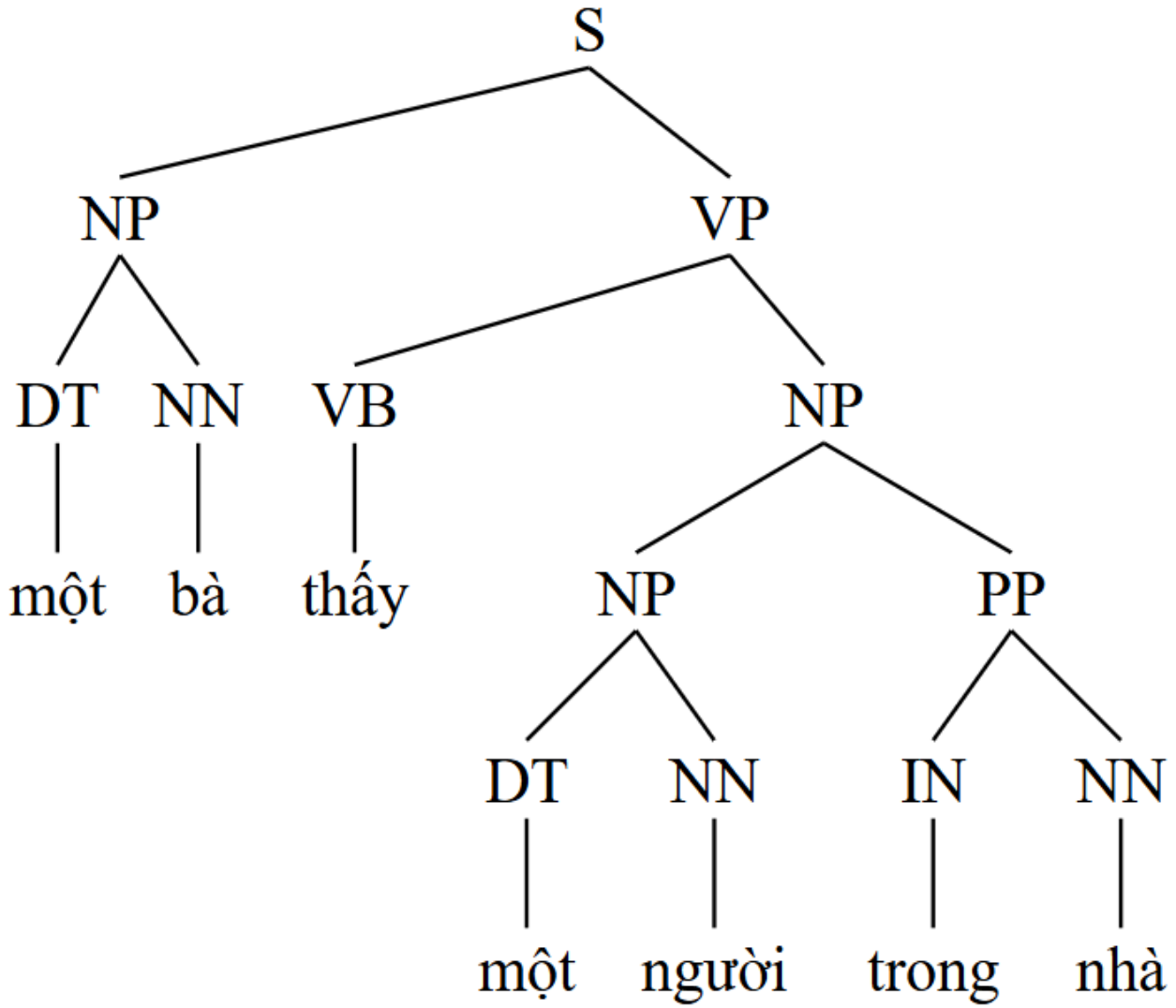
Ví dụ: hiển thị kết quả phân tích cú pháp câu bên dưới với thuật toán Earley.

Một bà thấy một người trong nhà.

```
s = 'một bà thấy một người trong nhà'.split()
for tree in ps.parse(s):
    display(tree)
```

4. BIỂU DIỄN CÂY CÚ PHÁP

Kết quả:



5. BÀI THỰC HÀNH

5. BÀI THỰC HÀNH

Cho tập câu D như sau:

1. Học sinh phải chăm chỉ.
2. Một học sinh không làm bài.
3. Rất nhiều học sinh thích môn tiếng Anh.
4. Học sinh nghe giảng bài trong lớp
5. Mỗi học sinh là một cây xanh trong vườn tri thức.

5. BÀI THỰC HÀNH

Yêu cầu:

1) Tổng hợp tập luật sản sinh P cho tập câu D.
2) Hiển thị kết quả phân tích cú pháp tự động của các câu trong tập D, theo tập luật sản sinh P với Python.
3) Nhận xét kết quả phân tích cú pháp tự động với cây cú pháp được phân tích thủ công.