Câu 1)

1) Cho văn phạm G như sau:

S → A A → aA A → bB

B → bB B → A A → ε

a) Phân tích cú pháp của câu sau và vẽ cây dẫn xuất

- "aaab"
- "abbaba"

b) Cho biết G có phải là văn phạm chính quy hay không? Vì sao? Vẽ DFA tương ứng với G nếu G là chính quy.

c) Nếu G là văn phạm chính quy, cho biết biểu thức chính quy tương ứng.

Bài tập này nhằm lưu ý các bạn các vấn đề:

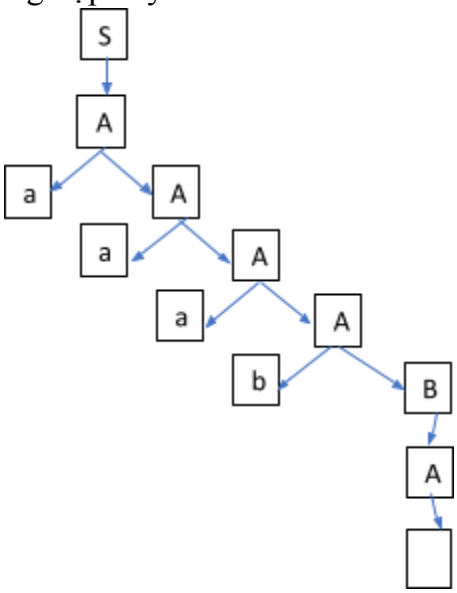
- Văn phạm tuyến tính, trong đó có các luật sản sinh.
- Cần biết cách áp dụng tập luật sản sinh để phân tích một chuỗi hoặc xây dựng tập luật sản sinh để có thể tạo ra các chuỗi theo yêu cầu

a) Để phân tích cú pháp có thể dùng một trong 2 chiến lược topdown và bottom up. Giả sử mình sử dụng chiến lược topdown

Phân tích chuỗi aaab

Chuỗi dẫn xuất:	Luật sử dụng
S	S->A
A	A->aA
aA	A->aA
aaA	A->aA
aaabB	A->bB
aaabA	B->A
aaab	A->ε

Cây dẫn xuất trong trường hợp này là:



b) **Để biết G có phải là văn phạm chính quy hay không, cần phải xét đặc điểm của văn phạm chính quy. Văn phạm chính quy, hay văn phạm loại 0 là văn phạm mà mọi luật sản sinh của nó có dạng Y → α hoặc Y → αX trong đó α là ký hiệu kết thúc và Y,X là ký hiệu phi kết thúc. Vậy, G chính là văn phạm chính quy.**

Một số loại văn phạm khác:

❖VĂN PHẠM CỦA NGÔN NGỮ

Một văn phạm G là một bộ (N, Σ, P, S) trong đó

- N là tập các từ vựng phụ trợ, gọi là ký hiệu không kết thúc (non-terminal)
- Σ là tập các từ của ngôn ngữ, gọi là ký hiệu kết thúc (terminal) và N ∩ = ∅.
- P là tập các luật văn phạm, gọi là luật sản sinh (production)
- S là yếu tố nguyên thủy của ngữ pháp, S ∈ N, là điểm khởi đầu cho các sản sinh trong P.

❖PHÂN LOẠI VĂN PHẠM

Phân cấp văn phạm của Chomsky gồm các kiểu văn phạm như sau:

- Văn phạm không hạn chế (unrestricted), loại 0, nếu mọi sản sinh đều có dạng $\alpha \to \beta$ với $\beta \in (N \cup \Sigma)^*$, $\alpha \in (N \cup \Sigma)^+$
- Văn phạm cảm ngữ cảnh (context-sensitive), loại 1, nếu mọi sản sinh đều có dạng $\gamma_1 X \gamma_2 \to \gamma_1 \alpha \gamma_2$ với X, α, γ_1, $\gamma_2 \in (N \cup \Sigma)^+$

❖VĂN PHẠM CỦA NGÔN NGỮ

Ví dụ, cho văn phạm G = (N, Σ, P, S) với:

- N = {A, B}
- Σ = {a, b}
- P = {S → AB, A→Aa, A → a, B → Bb, B → b}

Khi đó, các sau câu thỏa văn phạm G:

abbbbb aaaab aaabbbbb

❖PHÂN LOẠI VĂN PHẠM

- Văn phạm phi ngữ cảnh (context-free), loại 2, nếu mọi sản sinh đều có dạng $X \to \alpha$ với $\alpha \in (N \cup \Sigma)^+$, $X \in N$
- Văn phạm tuyến tính (linear), loại 3, nếu mọi sản sinh đều có dạng $X \to \alpha$ và $X \to \alpha Y$ với X, Y ∈ N và α ∈ Σ. Văn phạm tuyến tính được biểu diễn bằng DFA và được gọi là văn phạm chính quy

Theo phân loại này thì:

Loại 3 ⊂ Loại 2 ⊂ Loại 1 ⊂ Loại 0

Máy chuyển đổi trạng thái đơn định FSA có 2 dạng: DFA (máy chuyển đổi tt đơn định) & NFA (máy chuyển đổi trạng thái không đơn định)

❖DETERMINISTIC FINITE AUTOMATON

DFA được định nghĩa là một bộ năm

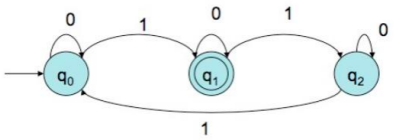
$$M = \{Q, \Sigma, \delta, q_0, F\}$$

Trong đó

- Q là tập hữu hạn các trạng thái.
- Σ là bộ ký tự.
- δ là hàm chuyển đổi trạng thái trả về trạng thái sẽ được chuyển đến, δ là đơn ánh.
- q_0 là trạng thái bắt đầu duy nhất, $q_0 \in Q$.
- F là tập các trạng thái kết thúc. F ⊂ Q.

❖DETERMINISTIC FINITE AUTOMATON

Một DFA M có thể được biểu diễn bằng một đồ thị như sau:



- $Q = \{q_0, q_1, q_2\}$, $\Sigma = \{0, 1\}$
- δ là tập hợp các cung có hướng,
- trạng thái bắt đầu và kết thúc lần lượt là q_0, q_1

Các phép toán trên DFA

Tích của hai DFA

Cho $M_1 = \{Q_1, \Sigma, \delta_1, q_1, F_1\}$, $M_2 = \{Q_2, \Sigma, \delta_2, q_2, F_2\}$ là hai DFA, tích của M_1 và M_2, ký hiệu $M_1 \times M_2$, được định nghĩa như sau:

- Tập trạng thái $Q = Q_1 \times Q_2$
- Bộ ký tự Σ
- Hàm chuyển trạng thái

$$\delta((q_1,q_2), a) = (\delta_1(q_1,a), \delta_2(q_2,a))$$

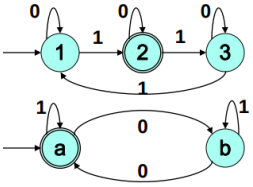
- Trạng thái bắt đầu: $q = (q_1, q_2)$
- Tập trạng thái kết thúc: tùy thuộc vào phép toán

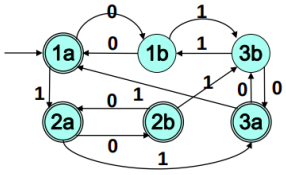
Phép lấy phần bù của DFA

Cho M là DFA, phần bù của M, ký hiệu $\overline{M}$, là một DFA có được bằng cách đổi các trạng thái kết thúc trong M thành trạng thái không kết thúc và các trạng thái không kết thúc thành trạng thái kết thúc.

Phép hợp

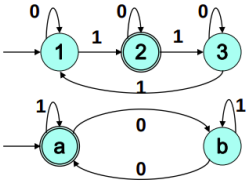
Cho M_1 và M_2 là hai DFA, M là hợp của M_1 và M_2, ký hiệu $M = M_1 \cup M_2$, nếu $M=M_1 \times M_2$ và tập các trạng thái kết thúc trong M là những bộ có chứa ít nhất một trạng thái kết thúc trong M_1 hoặc M_2

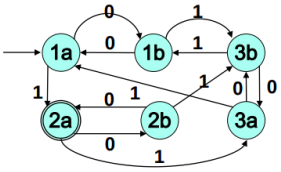




Phép giao

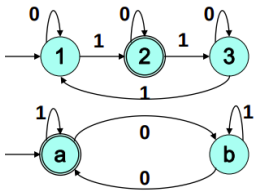
Cho M_1 và M_2 là hai DFA, M là giao của M_1 và M_2, ký hiệu $M = M_1 \cap M_2$, nếu $M=M_1 \times M_2$ và tập các trạng thái kết thúc trong M là những bộ chứa cả hai trạng thái kết thúc trong M_1 và M_2

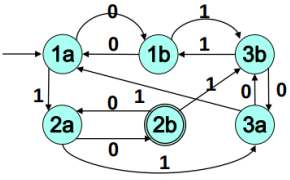




Phép hiệu

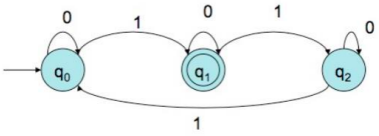
Cho M_1 và M_2 là hai DFA, M là hiệu của M_1 và M_2, ký hiệu $M = M_1 \setminus M_2$, nếu $M=M_1 \times M_2$ và tập các trạng thái kết thúc trong M là những bộ chứa trạng thái kết thúc của M_1 nhưng không chứa trạng thái kết thúc của M_2





❖DETERMINISTIC FINITE AUTOMATON

Một chuỗi được gọi là được M đoán nhận (recognize) nếu M đạt trạng thái kết thúc khi duyệt toàn bộ chuỗi



Cho biết M như trên có đoán nhận được các chuỗi sau hay không: 01010100, 010010011, 0011100101101

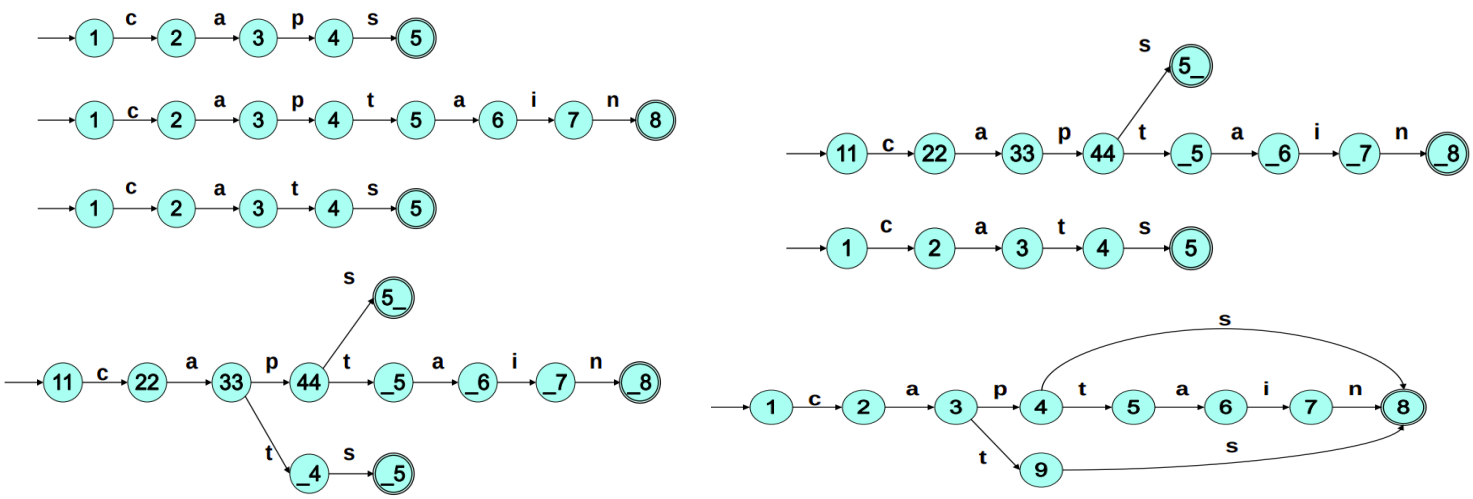
Xây dựng DFA

DFA đoán nhận một tập L các chuỗi có thể được xây dựng qua 3 bước:

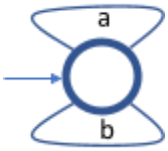
- Tách tập L thành các tập hợp con, xây dựng DFA cho từng tập hợp.
- Hợp các DFA của các tập con.
- Tối tiểu DFA (gộp các trạng thái thừa)

Ví dụ:

Xây dựng DFA cho L = {caps, captain, cats}



DFA tương ứng trong trường hợp này là:



Khi vẽ DFA các bạn cần xét: nếu có chuỗi rỗng thì trạng thái đầu tiên cũng là trạng thái kết thúc. Nếu có lặp thì sẽ có vòng.

c) Biểu thức chính quy tương ứng là $(a^*b^*)^*$

Câu 2) Bài tập này lưu ý các bạn các vấn đề:

- Vẽ DFA, các phép toán trên DFA.
- Biểu thức chính quy (regular expression) và sự tương ứng giữa DFA và biểu thức chính quy.

2) Cho hai ngôn ngữ L_1, L_2 như sau:

$L_1 = \{(a^m b^n)^i\}$ $L_2 = \{(ab)^n\}$

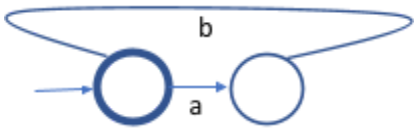
a) Vẽ DFA tương ứng với văn phạm của từng ngôn ngữ L_1, L_2

b) Vẽ DFA của ngôn ngữ L được xác định bởi

- $L_1 + L_2$
- $L_1 - L_2$

Cho ví dụ một câu có độ dài hơn 5 tương ứng với từng trường hợp

a) **DFA1 của L1 chính là DFA của câu 1c, DFA2 của L2 như sau:**



b) **Ngôn ngữ L1+L2 có DFA tương ứng là DFA1∪DFA2 và ngôn ngữ L1-L2 có** DFA tương ứng là DFA1\DFA2. Các bạn xem lại phép hợp DFA và phép lấy phần bù DFA trong slide để xác định kết quả nhé.

Cho ví dụ một câu có độ dài lớn hơn 5 trong trường hợp L1+L2 thì có thể lấy một câu trong L1 hoặc trong L2. Chẳng hạn ababab
Trường hợp L1-L2 thì chỉ lấy những câu trong L1 mà không có trong L2. Chẳng hạn aaaaaaab.

Câu 3) Bài tập này lưu ý các bạn các vấn đề:

3) Cho các câu sau:
- Lan day con sáo hót
- Lan và ban học bài

a) Vẽ cây cú pháp và xác định tập luật sản sinh của cả hai câu trên

b) Dựa trên kết quả xác định từ loại ở câu a), xây dựng mô hình Hidden Markov bậc 1 sử dụng phương pháp smoothing là Laplace + 1 và xác định từ loại cho câu sau:

Hoa day Lan học

c) Dùng thuật toán Earley, phân tích cú pháp của câu sau theo tập luật sản sinh đã được xác định ở câu a)

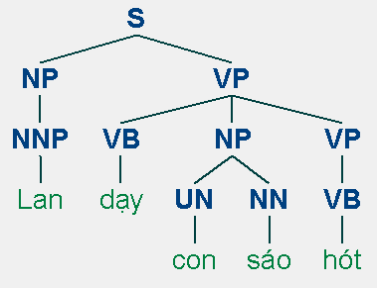
Lan và ban học bài

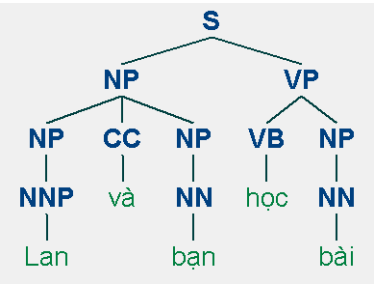
d) Xác định văn pham PCFG từ hai cây cú pháp đã xác định ở câu a) với phương pháp smoothing là Laplace + 1 và phân tích cú pháp cho câu:

Hoa day Lan học

- Gán nhãn từ loại, cú pháp thủ công cho các câu tiếng Việt.
- Huấn luyện mô hình HMM từ ngữ liệu gán nhãn từ loại.
- Áp dụng thuật toán Viterbi để tính toán xác suất và tìm chuỗi nhãn tốt nhất.
- Xác định tập luật sản sinh từ cây cú pháp.
- Dùng các thuật toán CKY, Earley để phân tích cú pháp.
- Xác định tập luật sản sinh có xác suất từ các cây cú pháp.
- Dùng các thuật toán CKY, Earley để phân tích cú pháp có xác suất.
    a) Để vẽ cây cú pháp, các bạn cần nhớ các quy tắc cú pháp tiếng Việt đã được trình bày trong slide bài giảng chương 5 hoặc các bạn có thể tham khảo trong giáo trình Ngôn ngữ học máy tính của PGS. TS. Nguyễn Tuấn Đăng. Sách này có ở thư viện trường. Các bạn phải lưu ý kiến thức từ vựng của mình nhé.
    Các từ như con, chiếc, cái, … là danh từ đơn vị, ký hiệu là UN. Lưu ý từ con trong con gà, con vịt chứ không phải từ xưng hô trong quan hệ con, cháu với ông bà, cha mẹ.

Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coordinating conjunction	*and, but, or*	PDT	predeterminer	*all, both*	VBP	verb non-3sg present	*eat*
CD	cardinal number	*one, two*	POS	possessive ending	*'s*	VBZ	verb 3sg pres	*eats*
DT	determiner	*a, the*	PRP	personal pronoun	*I, you, he*	WDT	wh-determ.	*which, that*
EX	existential 'there'	*there*	PRP$	possess. pronoun	*your, one's*	WP	wh-pronoun	*what, who*
FW	foreign word	*mea culpa*	RB	adverb	*quickly*	WP$	wh-possess.	*whose*
IN	preposition/ subordin-conj	*of, in, by*	RBR	comparative adverb	*faster*	WRB	wh-adverb	*how, where*
JJ	adjective	*yellow*	RBS	superlatv. adverb	*fastest*	$	dollar sign	$
JJR	comparative adj	*bigger*	RP	particle	*up, off*	#	pound sign	#
JJS	superlative adj	*wildest*	SYM	symbol	+,%, &	‘	left quote	‘ or “
LS	list item marker	*1, 2, One*	TO	“to”	*to*	”	right quote	’ or ”
MD	modal	*can, should*	UH	interjection	*ah, oops*	(	left paren	[, (, {, <
NN	sing or mass noun	*llama*	VB	verb base form	*eat*	)	right paren	],), }, >
NNS	noun, plural	*llamas*	VBD	verb past tense	*ate*	,	comma	,
NNP	proper noun, sing.	*IBM*	VBG	verb gerund	*eating*	.	sent-end punc	. ! ?
NNPS	proper noun, plu.	*Carolinas*	VBN	verb past part.	*eaten*	:	sent-mid punc	: ; ... – -





Tập luận sản sinh:

S NP VP

NP NNP

NP UN NN

NP NP CC NP

NP NN

VP VB NP VP

VP VB

VP VB NP

NNP Lan

NN bạn

NN sáo

NN bài

VB dạy

VB học

VB hót

CC và

UN con

b) Dữ liệu gán nhãn từ loại của hai câu đã cho:

Lan/NNP dạy/VB con/UN sáo/NN hót/VB
Lan/NNP và/CC bạn/NN học/VB bài/NN
Dữ liệu huấn luyện:
$ Lan/NNP dạy/VB con/UN sáo/NN hót/VB
$ Lan/NNP và/CC bạn/NN học/VB bài/NN
Ma trận chuyển trạng thái **A** đã áp dụng Laplace smooth

	VB	NN	NNP	UN	CC
$	0.14	0.14	0.43	0.14	0.14
VB	0.14	0.29	0.14	0.29	0.14
NN	0.43	0.14	0.14	0.14	0.14
NNP	0.29	0.14	0.14	0.14	0.29
UN	0.17	0.33	0.17	0.17	0.17
CC	0.17	0.33	0.17	0.17	0.17

Ma trận thể hiện (phát xạ) **B** đã áp dụng Laplace smooth. Chú ý, từ Hoa không có trong ngữ liệu huấn luyện nên mình xem nó như là một từ UNKNOW. ở đây mình ghi luôn chữ Hoa để dễ theo dõi.

	Lan	dạy	con	sáo	hót	và	bạn	học	bài	Hoa (UNK)
VB	0.08	0.15	0.1	0.1	0.15	0.1	0.1	0.15	0.08	0.08
NN	0.08	0.08	0.1	0.2	0.08	0.1	0.2	0.08	0.15	0.08
NNP	0.25	0.08	0.1	0.1	0.08	0.1	0.1	0.08	0.08	0.08
UN	0.09	0.09	0.2	0.1	0.09	0.1	0.1	0.09	0.09	0.09
CC	0.09	0.09	0.1	0.1	0.09	0.2	0.1	0.09	0.09	0.09

Kết quả tính toán xác suất các trường hợp theo thuật toán Viterbi. Các bạn trình bày bảng tính toán các xác suất và có các mũi tên cho thấy các đường đi theo thuật toán Viterbi theo dạng bên dưới. Phần tính toán các bạn tự làm để ước tính thời gian làm của mình nhé.

	Lan	Dạy	Hoa	Học

VB				
NN				
NNP				
UN				
CC				

c) Thuật toán Earley phân tích câu "Lan và bạn học bài"

CHART 0

1	γ □ * S	[0,0]	
2	S □ * NP VP	[0,0]	
3	NP □ * NNP	[0,0]	
4	NP □ * NNP	[0,0]	
5	NP □ * UN NN	[0,0]	
6	NP □ * NP CC NP	[0,0]	
7	NP □ *NN	[0,0]	

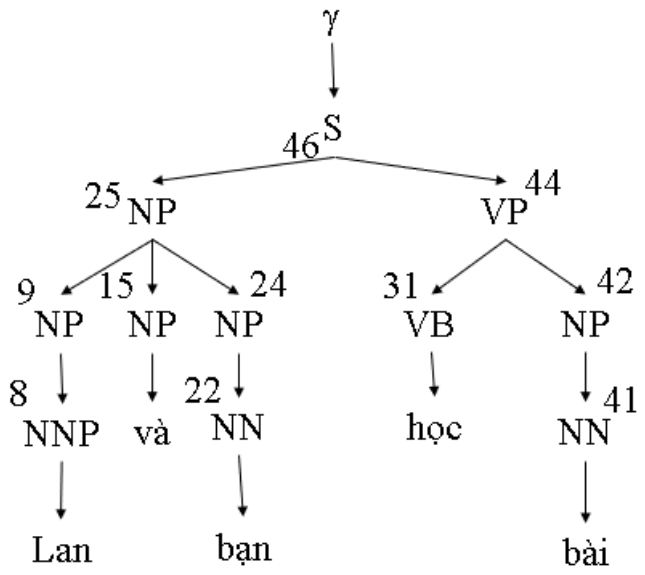
CHART 1

8	NNP □ Lan *	[0,1]	
9	NP □ NNP *	[0,1]	8
10	NP □ NP * CC NP	[0,1]	9
11	S □ NP * VP	[0,1]	
12	VP □* VB NP VP	[1, 1]	
13	VP □ * VB	[1,1]	

14	VP * VB NP	[1,1]	
	CHART 2		
15	CC và *	[1, 2]	
16	NP NP CC * NP	[0, 2]	9,15
17	NP * NNP	[2,2]	
18	NP * NNP	[2,2]	
19	NP * UN NN	[2,2]	
20	NP * NP CC NP	[2,2]	
21	NP * NN	[2,2]	
	CHART 3		
22	NN bạn *	[2, 3]	
24	NP NN *	[2, 3]	22
25	NP NP CC NP *	[0, 3]	9,15,24
26	NP NP *CC NP	[2, 3]	24
27	S NP * VP	[0, 3]	25
28	VP * VB NP VP	[3, 3]	
29	VP * VB	[3, 3]	
30	VP * VB NP	[3, 3]	
	CHART 4		

31	VB học *	[3, 4]	
32	VP VB * NP VP	[3, 4]	31
33	VP VB *	[3, 4]	31
34	VP VB * NP	[3, 4]	31
35	S NP VP *	[0, 4]	25,33
36	γ S *	[0, 4]	35
37	NP * NNP	[4, 4]	
38	NP * UN NN	[4, 4]	
39	NP * NP CC NP	[4, 4]	
40	NP * NN	[4, 4]	
	CHART 5		
41	NN bài *	[4, 5]	
42	NP NN *	[4, 5]	41
43	VP VB NP * VP	[3, 5]	31, 42
44	VP VB NP *	[3, 5]	31,42
45	NP NP * CC NP	[4, 5]	44
46	S NP VP *	[0, 5]	25,44
47	γ S *	[0, 5]	46

Kết quả:



d) Xác định văn phạm PCFG từ hai cây cú pháp đã có. Lưu ý chỉ áp dụng Laplace smoothing cho các luật sản sinh ký hiệu kết thúc để tính toán cho từ không xuất hiện (UNK) trong ngữ liệu huấn luyện, UNK ở đây là từ Hoa. (Không cần trình bày cột count, Laplace và sum, chỉ cần cột P

Luật	count	sum	Laplace	sum	P
S NP VP	2	2	3	3	1
NP NNP	2	6	3	10	0.3
NP UN NN	1		2		0.2
NP NP CC NP	1		2		0.2
NP NN	2		3		0.3
VP VB NP VP	1	3	2	6	0.3333
VP VB	1		2		0.3333
VP VB NP	1		2		0.3333
NNP Lan	2	2	3	4	0.75
NNP Hoa	0		1		0.25

NN → bạn	1	3	2	7	0.2857
NN → sáo	1		2		0.2857
NN → bài	1		2		0.2857
NN → Hoa	0		1		0.1429
VB → dạy	1	3	2	7	0.2857
VB → học	1		2		0.2857
VB → hót	1		2		0.2857
VB → Hoa	0		1		0.1429
CC → và	1	1	2	3	0.6667
CC → Hoa	0		1		0.3333
UN → con	1	1	2	3	0.6667
UN → Hoa	0		1		0.3333

Từ tập luật sản sinh có xác suất này, các bạn dùng thuật toán CKY để chọn cây cú pháp có xác suất cao nhất. Việc tính CKY này các bạn tự làm để ước tính thời gian của mình.