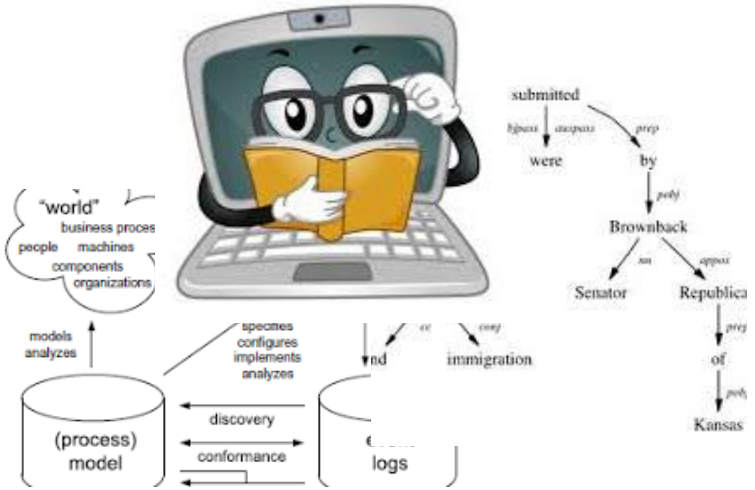
ĐẠI HỌC QUỐC GIA TPHCM
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

XỬ LÝ NGÔN NGỮ TỰ NHIÊN

CHƯƠNG III

NGÔN NGỮ HÌNH THỨC



Nguyễn Trọng Chỉnh
chinhnt@uit.edu.vn

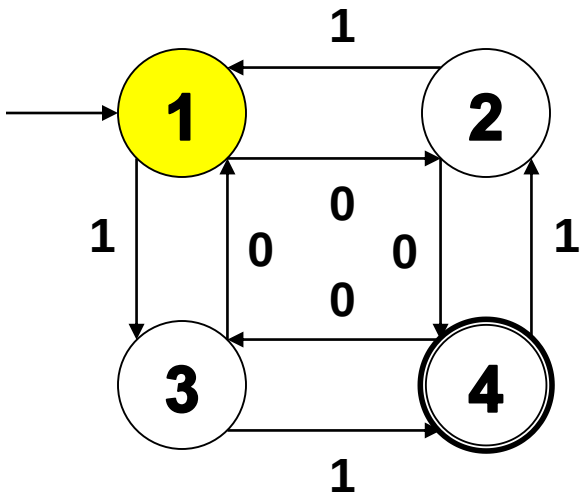
NGÔN NGỮ HÌNH THỨC

- KHÁI NIỆM
- BIỂU THỨC CHÍNH QUY
- **PHÂN TÍCH HÌNH THÁI**
- CHUẨN HÓA VĂN BẢN
- EDIT DISTANCE

III. PHÂN TÍCH HÌNH THÁI

❖FINITE STATE AUTOMATON

Máy chuyển đổi trạng thái hữu hạn (Finite State Automaton – FSA, hoặc FSM) là mô hình tính toán của máy tính.

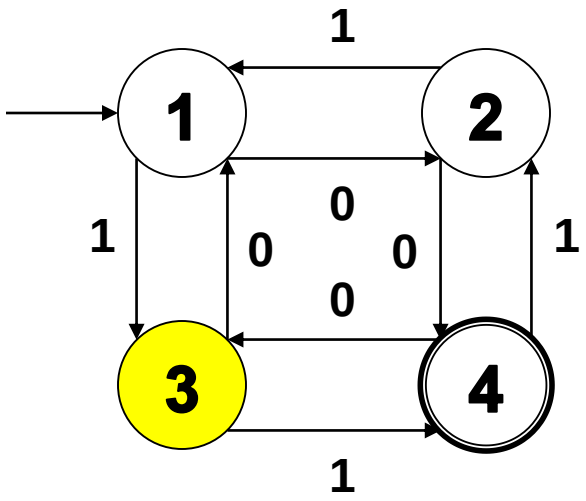


1
0
0
1

III. PHÂN TÍCH HÌNH THÁI

❖FINITE STATE AUTOMATON

Máy chuyển đổi trạng thái hữu hạn (Finite State Automaton – FSA, hoặc FSM) là mô hình tính toán của máy tính.

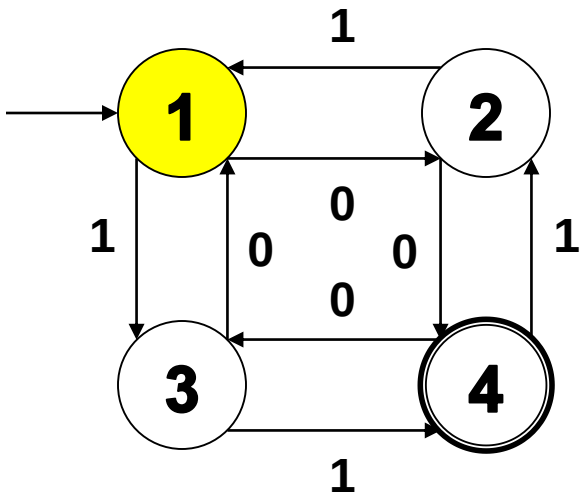


0
0
1

III. PHÂN TÍCH HÌNH THÁI

❖FINITE STATE AUTOMATON

Máy chuyển đổi trạng thái hữu hạn (Finite State Automaton – FSA, hoặc FSM) là mô hình tính toán của máy tính.



0
1

III. PHÂN TÍCH HÌNH THÁI

❖FINITE STATE AUTOMATON

Có hai dạng FSA là:

- Deterministic Finte Automaton (DFA – Máy chuyển đổi trạng thái hữu hạn đơn định)
- Non-deterministic Finite Automaton (NFA – Máy chuyển đổi trạng thái hữu hạn không đơn định)

III. PHÂN TÍCH HÌNH THÁI

❖DETERMINISTIC FINITE AUTOMATON

DFA được định nghĩa là một bộ năm

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

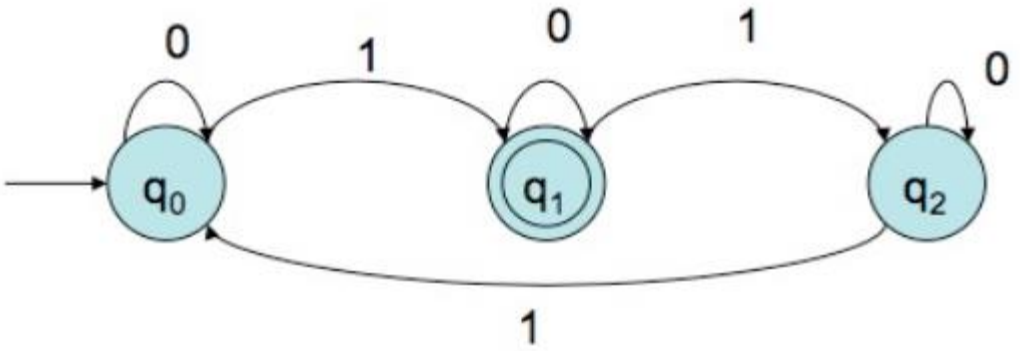
Trong đó

- 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. $F \subset Q$.

III. PHÂN TÍCH HÌNH THÁI

❖DETERMINISTIC FINITE AUTOMATON

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

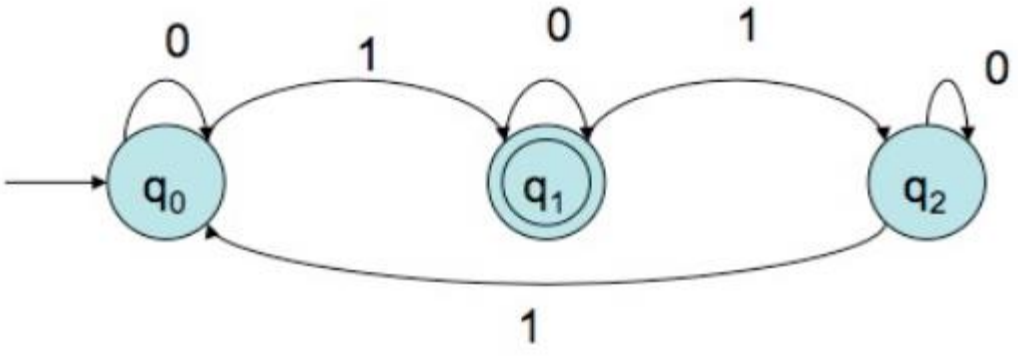


- $Q = \{q_0, q_1, q_2\}$, $\Sigma = \{0, 1\}$
- δ là tập hợp các cung có hướng,
- trạng thái bắt đầu và kết thúc lần lượt là q_0, q_1

III. PHÂN TÍCH HÌNH THÁI

❖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, 001110010110 1

III. PHÂN TÍCH HÌNH THÁI

❖DETERMINISTIC FINITE AUTOMATON

DFA có thể được biểu diễn dưới dạng bảng

- Các cột là đầu vào tại thời điểm đang xét
- Các dòng là các trạng thái tại thời điểm đang xét
- Giá trị tại ô A_{ij} là kết quả chuyển trạng thái khi máy đang ở trạng thái q_i nhận được đầu vào là Σ_j

VD: Biểu diễn dạng bảng của M là

δ	0	1
q_0	q_0	q_1
q_1	q_1	q_2
q_2	q_2	q_0

III. PHÂN TÍCH HÌNH THÁI

❖DETERMINISTIC FINITE AUTOMATON

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.

III. PHÂN TÍCH HÌNH THÁI

❖DETERMINISTIC FINITE AUTOMATON

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

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

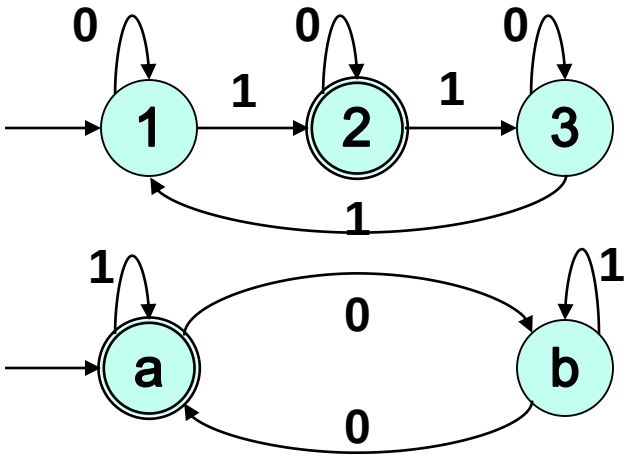
- Trạng thái bắt đầu: $q = (q_1, q_2)$
- Tập trạng thái kết thúc: tùy thuộc vào phép toán

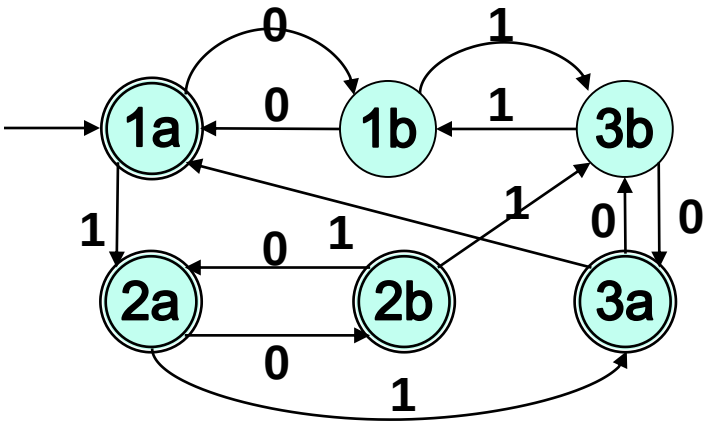
III. PHÂN TÍCH HÌNH THÁI

❖DETERMINISTIC FINITE AUTOMATON

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



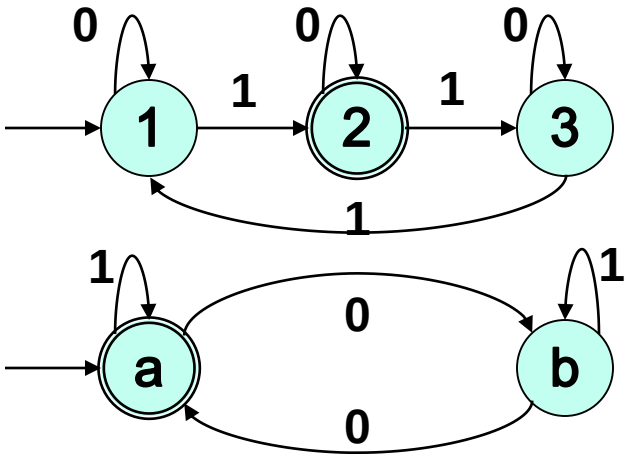


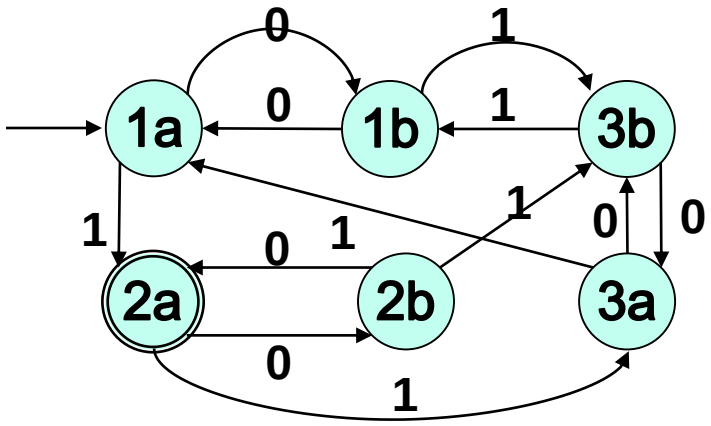
III. PHÂN TÍCH HÌNH THÁI

❖DETERMINISTIC FINITE AUTOMATON

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



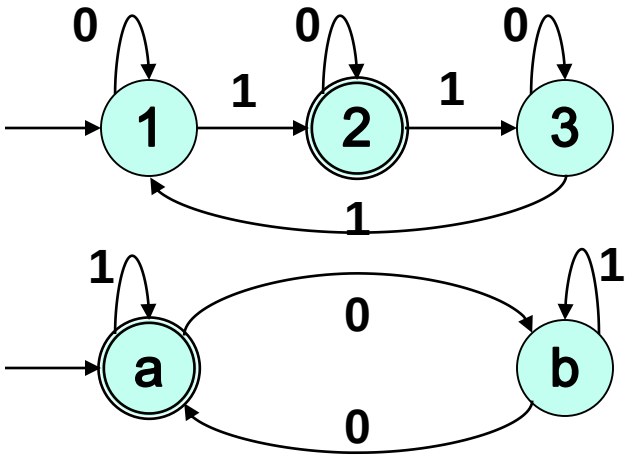


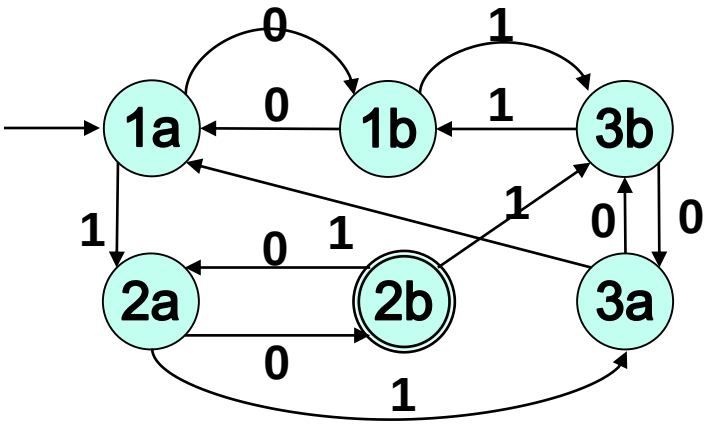
III. PHÂN TÍCH HÌNH THÁI

❖DETERMINISTIC FINITE AUTOMATON

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





III. PHÂN TÍCH HÌNH THÁI

❖DETERMINISTIC FINITE AUTOMATON

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)

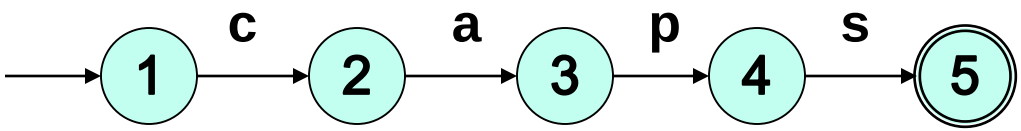
III. PHÂN TÍCH HÌNH THÁI

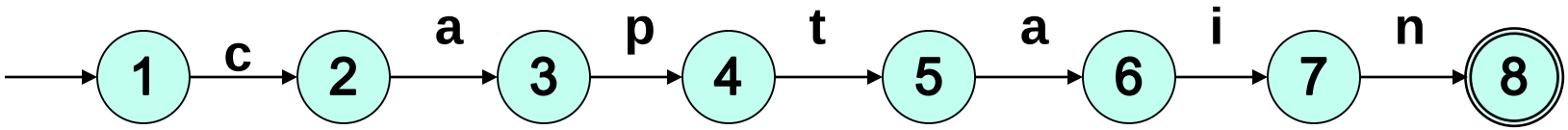
- DETERMINISTIC FINITE AUTOMATON

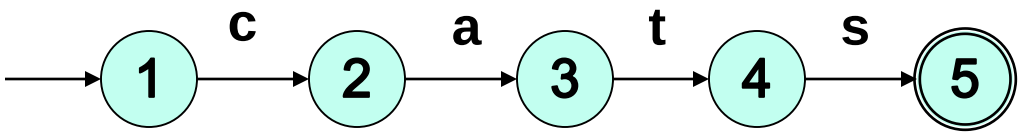
Ví dụ:

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

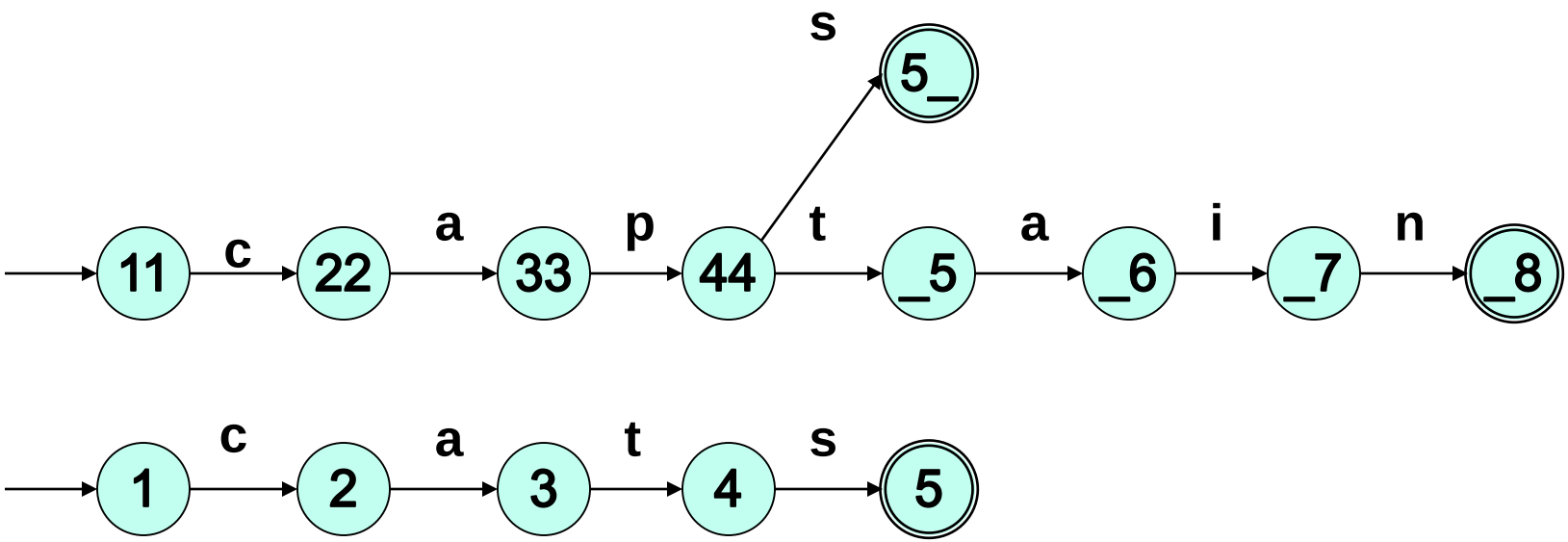
III. PHÂN TÍCH HÌNH THÁI



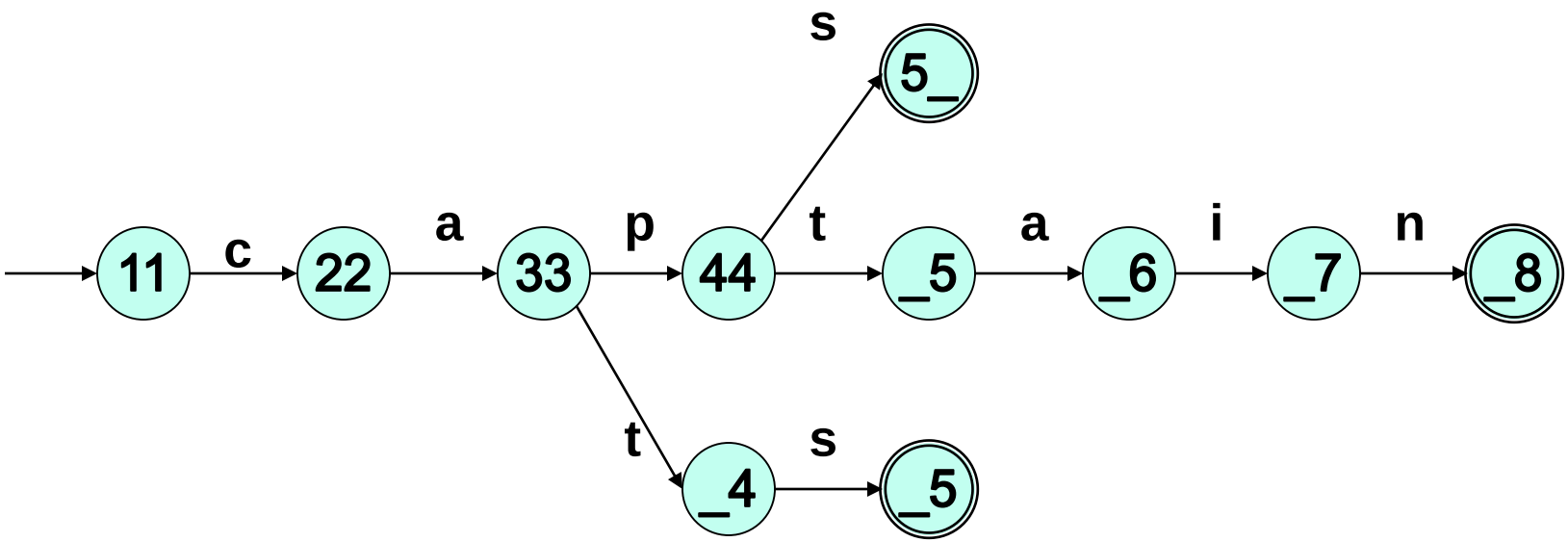




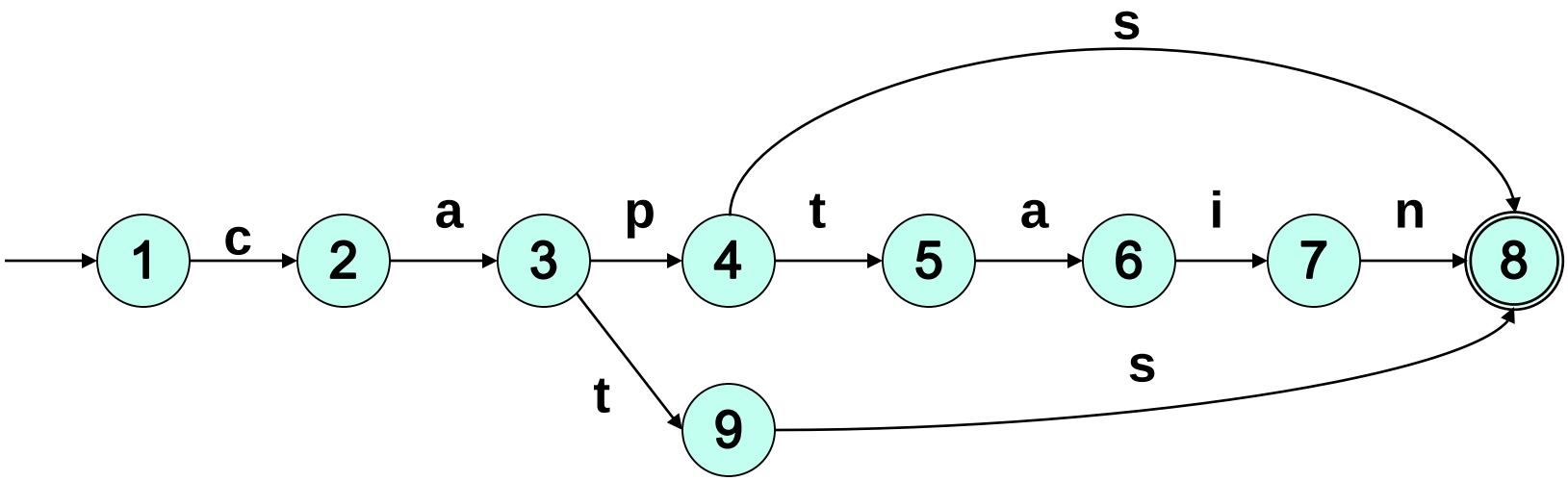
III. PHÂN TÍCH HÌNH THÁI



III. PHÂN TÍCH HÌNH THÁI



III. PHÂN TÍCH HÌNH THÁI



III. PHÂN TÍCH HÌNH THÁI

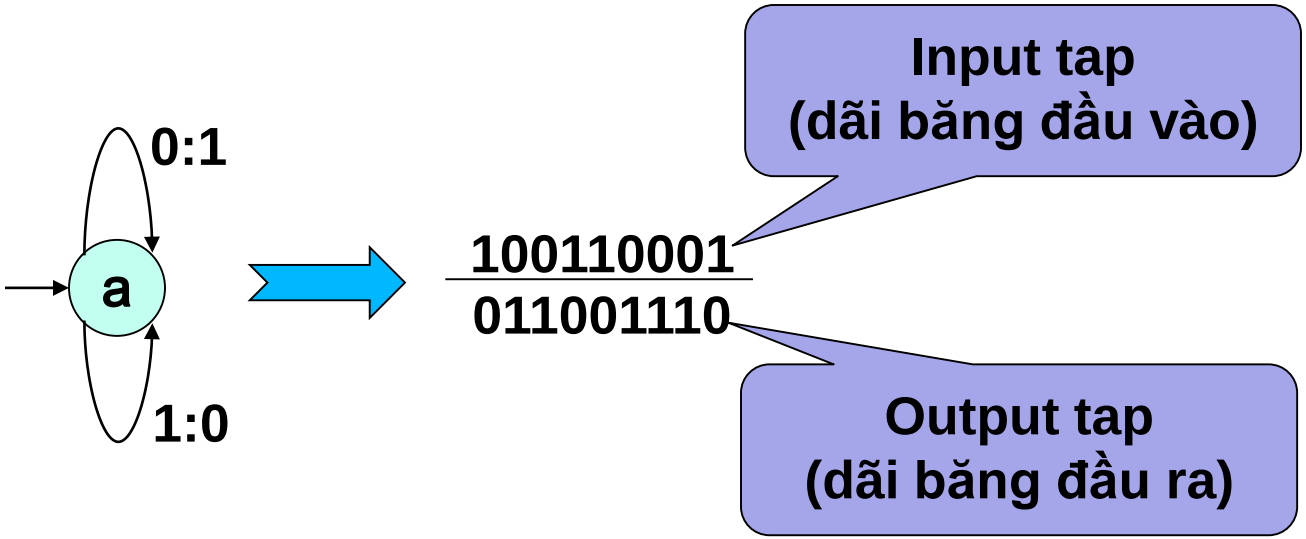
❖FINITE STATE TRANDUCER

- Một Transducer (bộ chuyển đổi) là một ánh xạ giữa hai tập biểu diễn.
- Một Finite State Transducer – FST (bộ chuyển đổi trạng thái hữu hạn) là một dạng FSA cho phép ánh xạ giữa hai tập ký hiệu. Trong đó mỗi cung được đánh dấu bằng một cặp ký hiệu cho biết ánh xạ giữa chúng.

III. PHÂN TÍCH HÌNH THÁI

❖FINITE STATE TRANDUCER

Ví dụ:



- 0:1 có nghĩa: đọc ký hiệu 0 trên input tap và ghi ra ký hiệu 1 trên output tap.
- 1:0 có nghĩa: đọc ký hiệu 1 trên input tap và ghi ra ký hiệu 0 trên output tap.

III. PHÂN TÍCH HÌNH THÁI

❖FINITE STATE TRANDUCER

FST được định nghĩa là một bộ bảy

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

Trong đó

- Q là tập các trạng thái.
- Σ và Δ là bộ ký tự đầu vào và đầu ra.
- δ là hàm chuyển đổi trạng thái
- σ là hàm ra, trả về ký tự tương ứng với một trạng thái và giá trị đầu vào.
- q_0 là trạng thái bắt đầu, $q_0 \in Q$.
- F là tập các trạng thái kết thúc. $F \subset Q$.

III. PHÂN TÍCH HÌNH THÁI

❖FINITE STATE TRANDUCER

- Mỗi FST là một quan hệ chính quy (Regular Relation). Một quan hệ chính quy là một tập các cặp chuỗi.
- FST có thể đoán nhận một cặp chuỗi có thuộc quan hệ chính quy tương ứng hay không
- FST có thể đoán nhận một chuỗi có thuộc một ngôn ngữ chính quy hay không và trả về chuỗi phân tích nếu đúng.

III. PHÂN TÍCH HÌNH THÁI

❖PHÂN TÍCH HÌNH THÁI

Phân tích hình thái là nhận diện xem một chuỗi có phải là một từ hay không, và nếu có thì cấu tạo của nó gồm những hình vị gì, có chức năng ngữ pháp như thế nào.

Ví dụ:

unstoppable	→ un – stop – able	→ J
leaves	→ leaf – s	→ N+PL

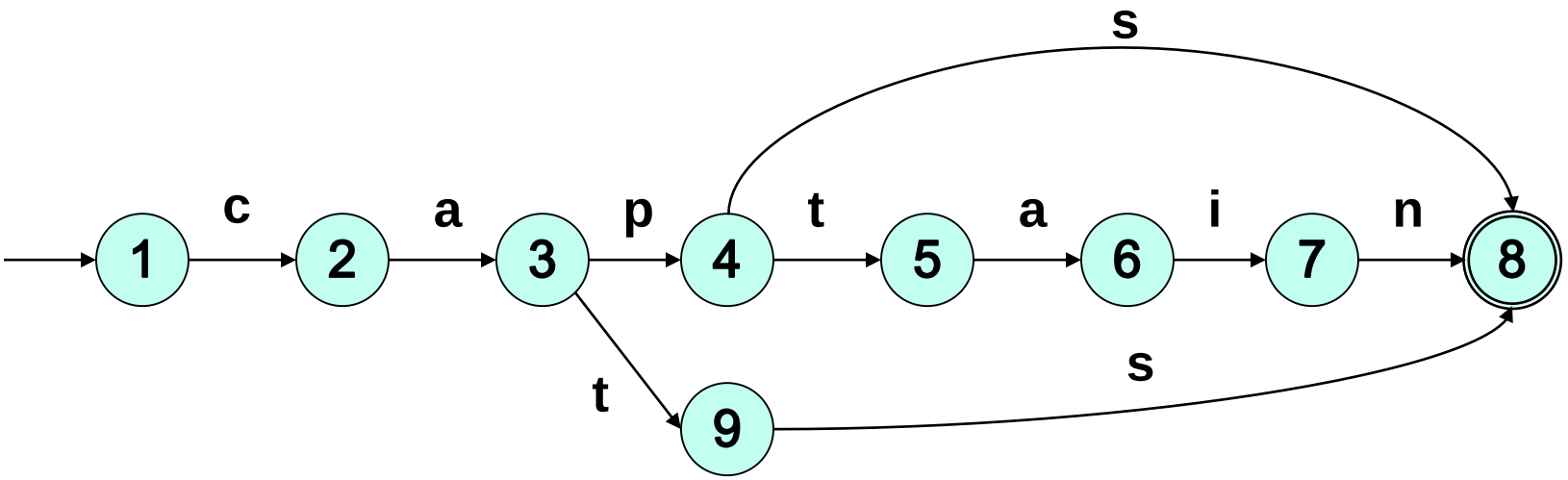
III. PHÂN TÍCH HÌNH THÁI

❖PHÂN TÍCH HÌNH THÁI

Bài toán nhận diện từ vựng: cho một chuỗi s, cho biết s có phải là một từ hay không?

→ Dùng Finite State Automata

(Nhận diện các từ caps, cats, captain)



III. PHÂN TÍCH HÌNH THÁI

❖PHÂN TÍCH HÌNH THÁI

Bài toán phân tích từ vựng: cho một chuỗi s, cho biết các chức năng ngữ pháp của s

→ Dùng Finite State Tranducer (dùng # thay cho ε)

(Ví dụ: danh từ cat, cats, cap, caps, child, children)