ALGORITMO DE AHO-CORASICK

Yan Soares Couto

Orientadora: Cristina Gomes Fernandes

2017

Instituto de Matemática e Estatística

DEFINIÇÃO

String $S[1..|S|]$: vetor em que cada elemento é de um alfabeto Σ finito.

Em geral, $\Sigma = \{a, b, \ldots, z\}$.

DEFINIÇÃO

String $S[1..|S|]$: vetor em que cada elemento é de um alfabeto Σ finito.

Em geral, $\Sigma = \{a, b, \ldots, z\}$.

"abracadabra"

Substring de S: subvetor de S, por exemplo, "cadab".

KMP

BORDAS

Borda de uma string: Maior prefixo que é também sufixo da string.

BORDAS

Borda de uma string: Maior prefixo que é também sufixo da string.

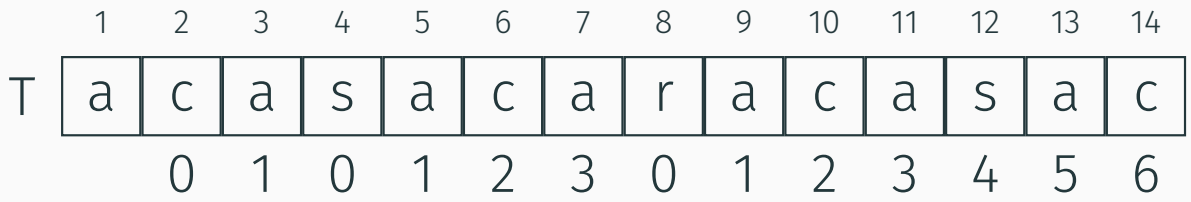
$$\text{abracadabra} \Rightarrow \text{abra}$$

$$\text{casa} \Rightarrow \varepsilon$$

$$\text{botobot} \Rightarrow \text{bot}$$

BORDAS

KMP calcula o tamanho da borda de todo prefixo da string.



TRIES

DEFINIÇÃO

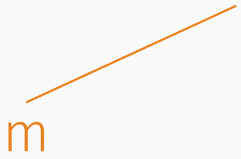
Trie: árvore enraizada que armazena um conjunto de strings.

Strings são representadas como caminhos a partir da raiz.

CONSTRUÇÃO

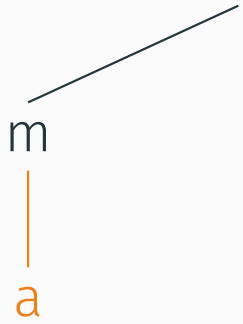
Adicionando “mata”.

CONSTRUÇÃO



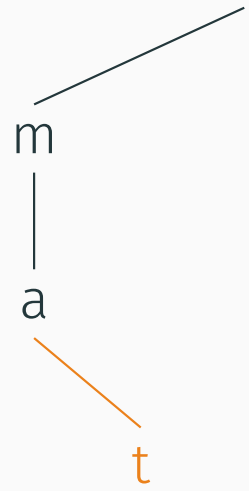
Adicionando “mata”.

CONSTRUÇÃO



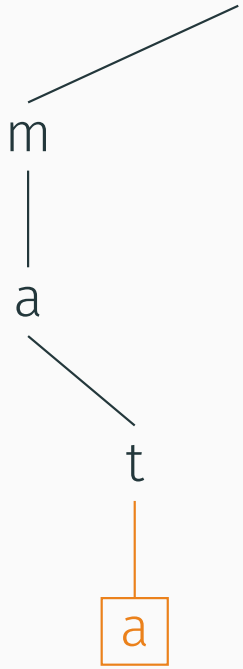
Adicionando “mata”.

CONSTRUÇÃO



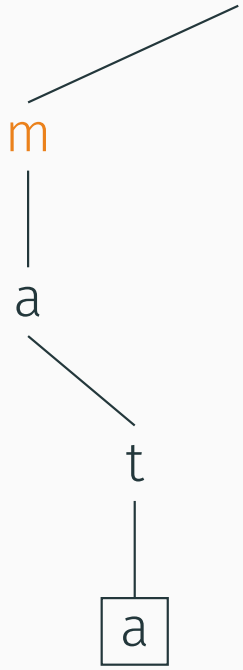
Adicionando “mata”.

CONSTRUÇÃO



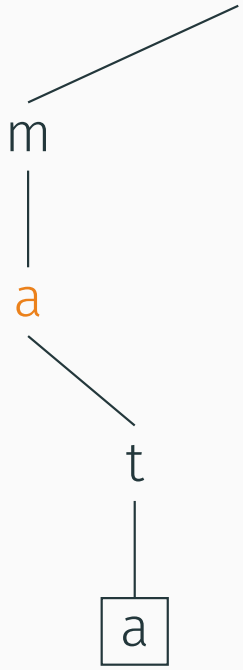
Adicionando “mata”.

CONSTRUÇÃO



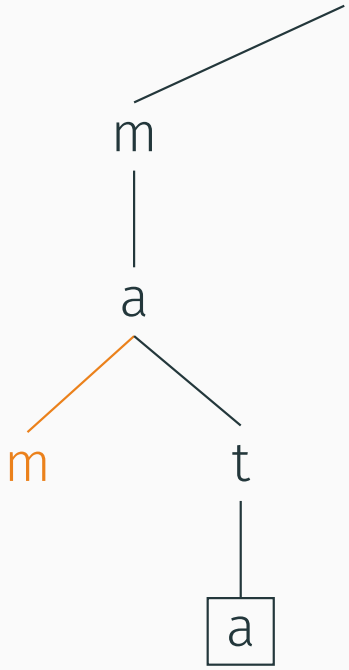
Adicionando “mamata”.

CONSTRUÇÃO



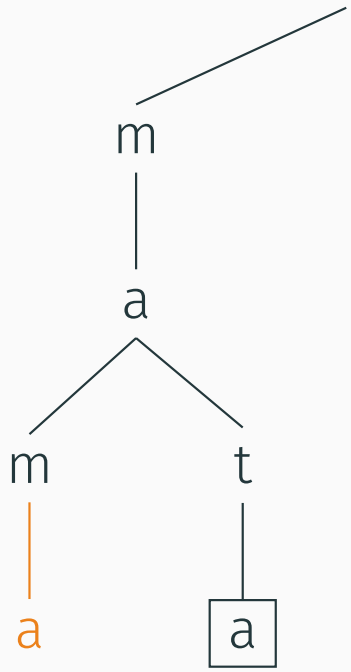
Adicionando “mamata”.

CONSTRUÇÃO



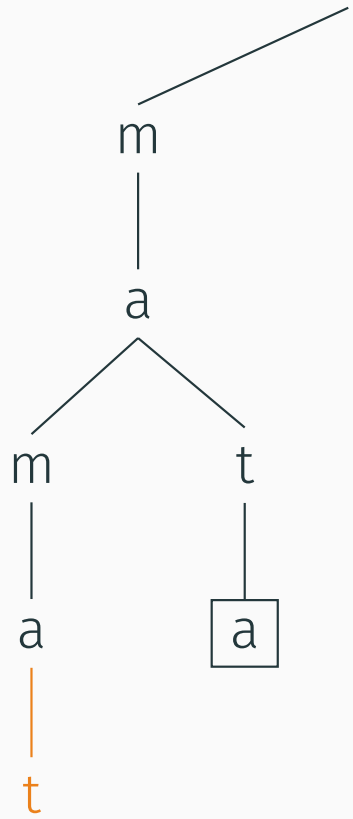
Adicionando “mamata”.

CONSTRUÇÃO



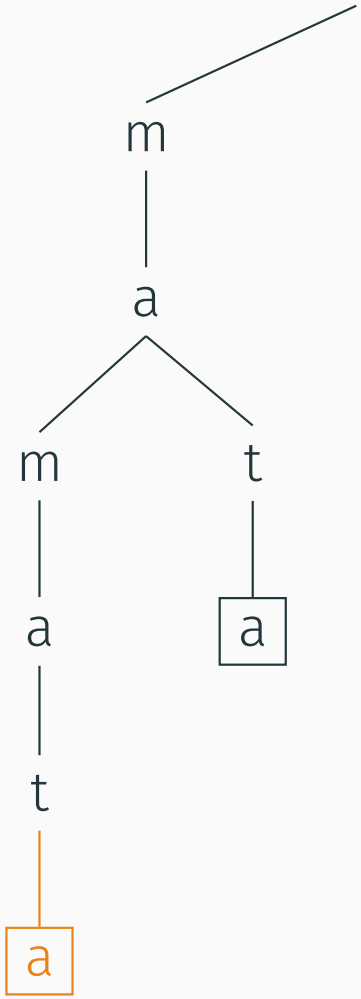
Adicionando “mamata”.

CONSTRUÇÃO



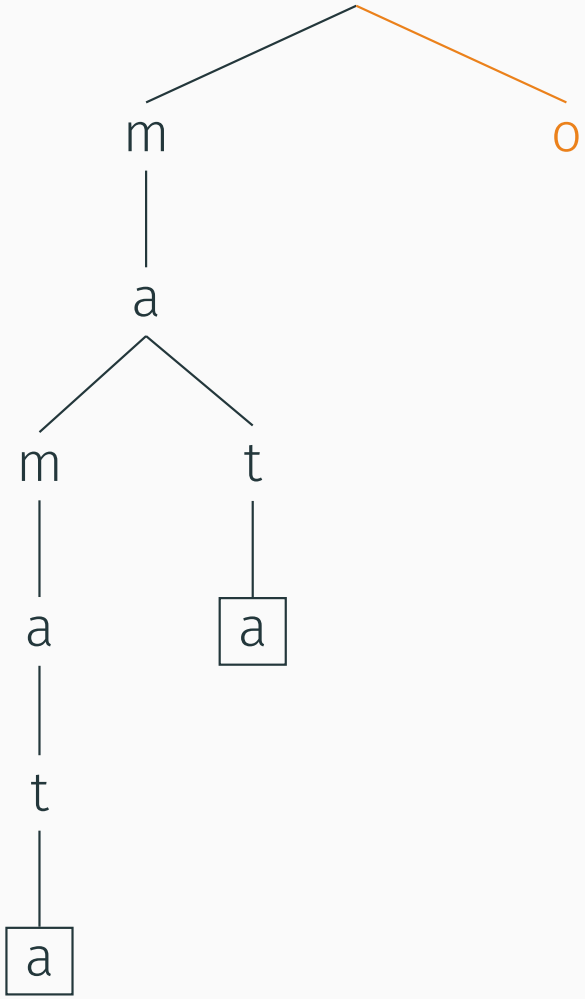
Adicionando “mamata”.

CONSTRUÇÃO



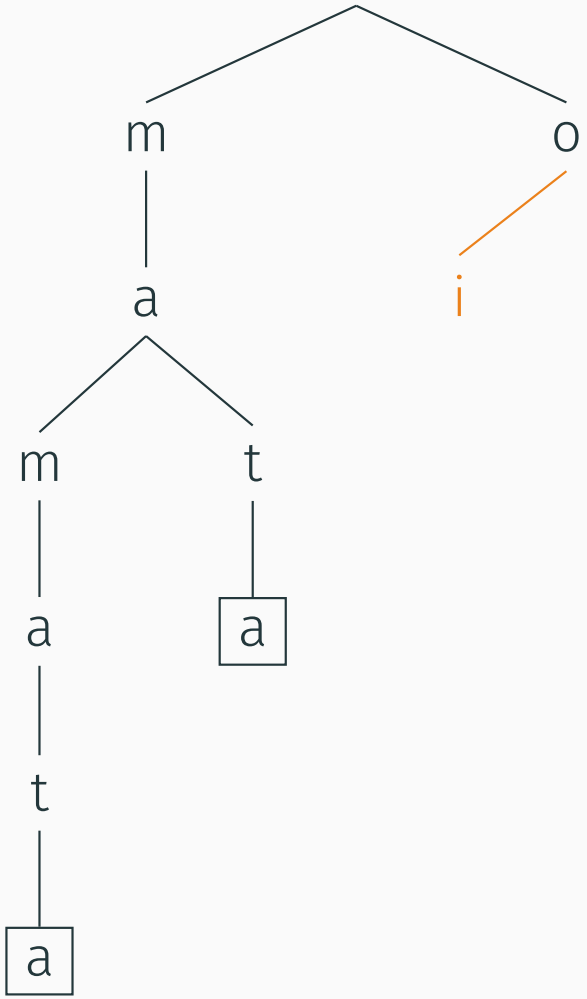
Adicionando “mamata”.

CONSTRUÇÃO



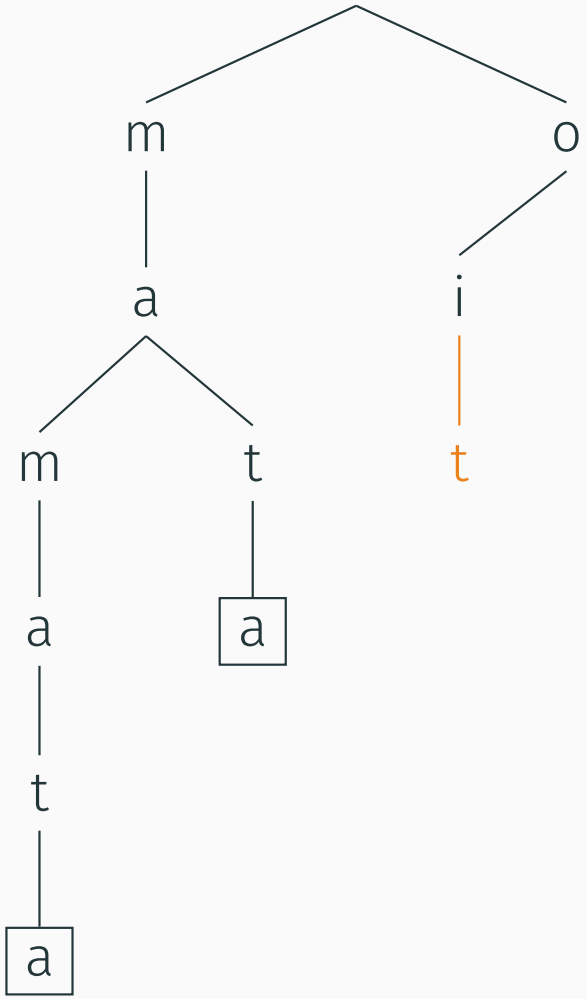
Adicionando “oito”.

CONSTRUÇÃO



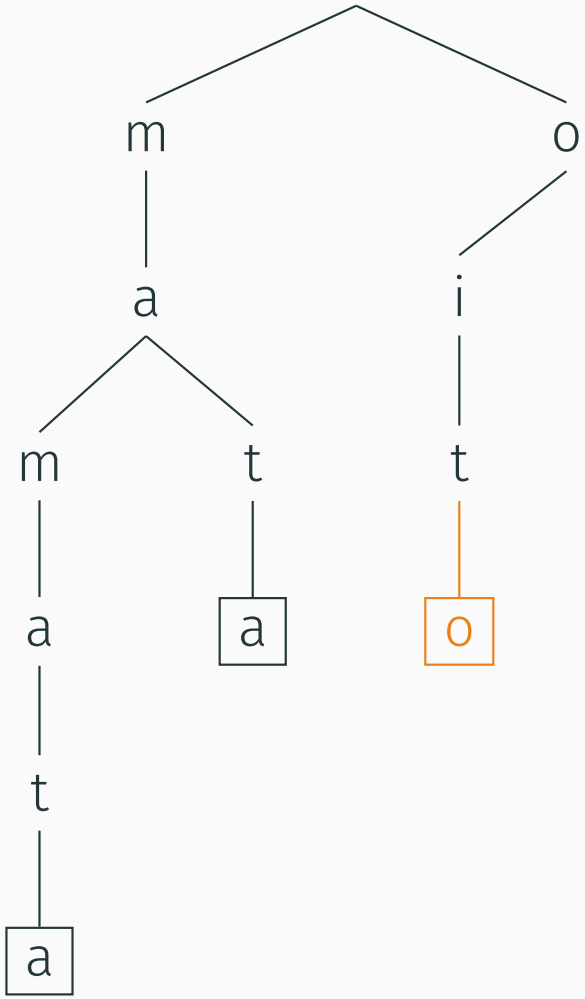
Adicionando “oito”.

CONSTRUÇÃO



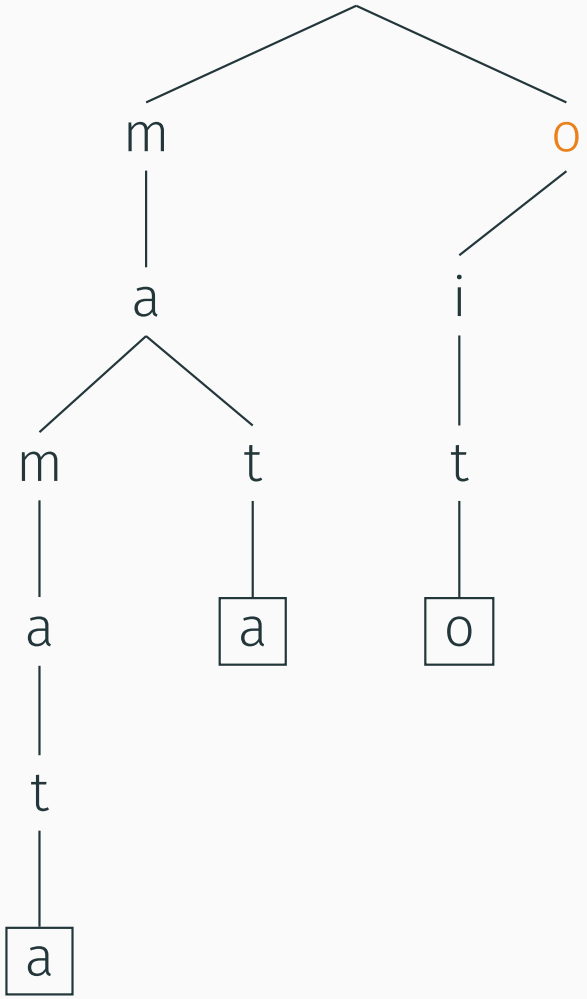
Adicionando “oito”.

CONSTRUÇÃO



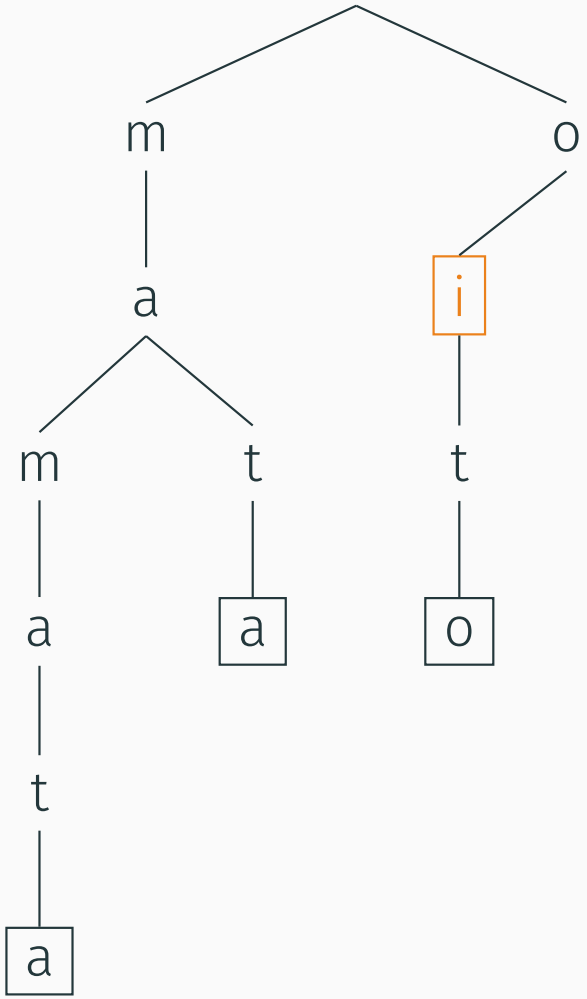
Adicionando “oito”.

CONSTRUÇÃO



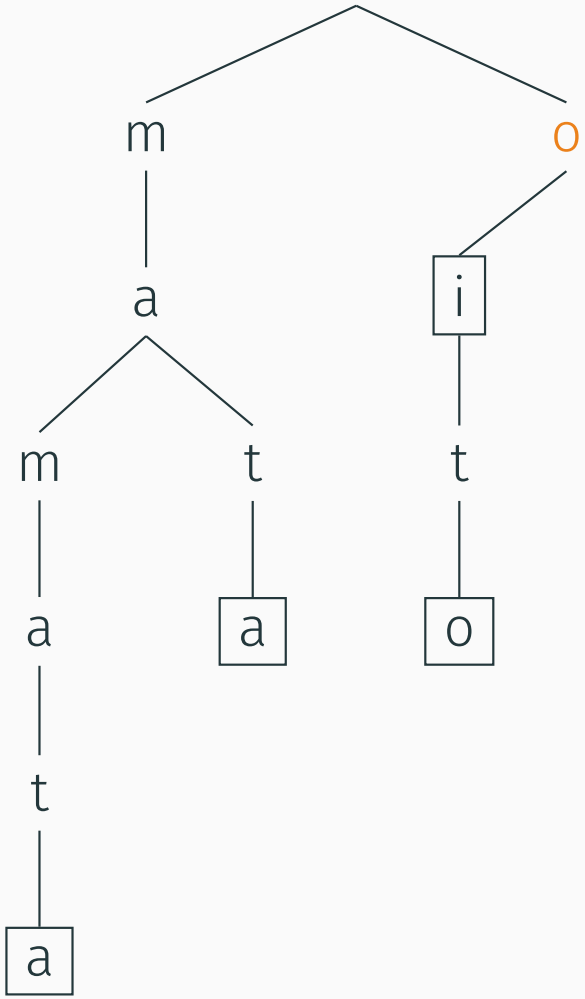
Adicionando “oi”.

CONSTRUÇÃO



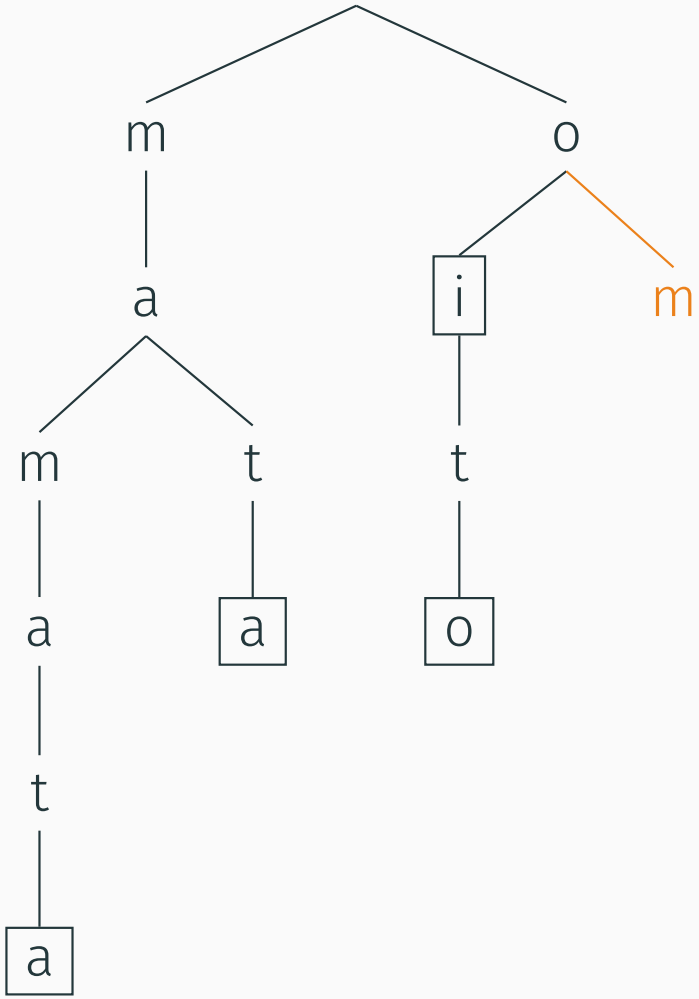
Adicionando “oi”.

CONSTRUÇÃO



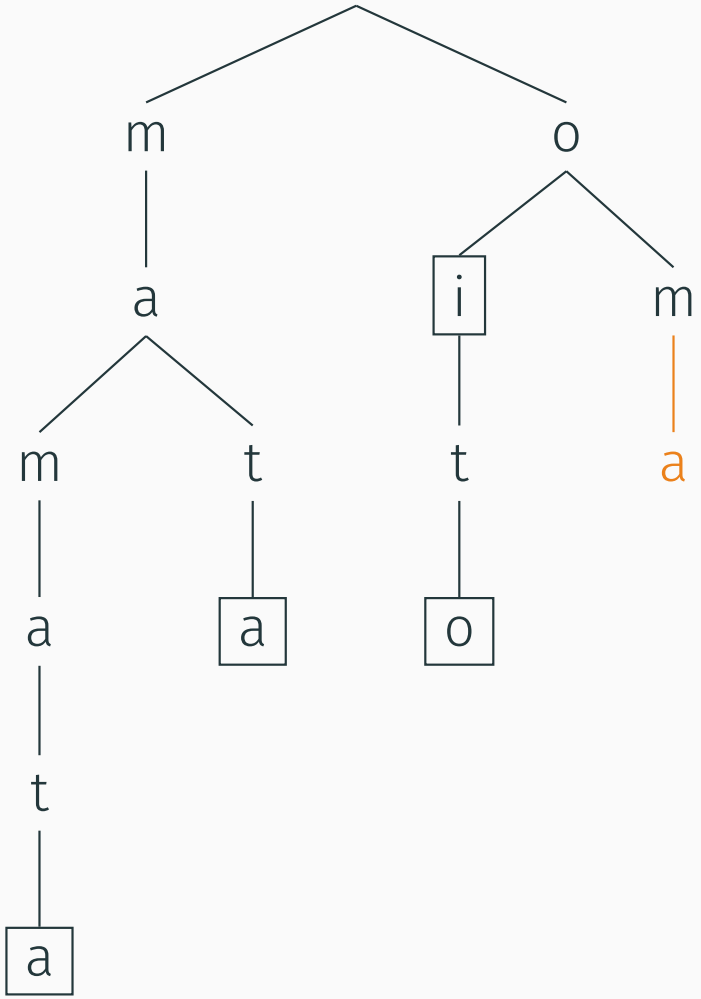
Adicionando “omar”.

CONSTRUÇÃO



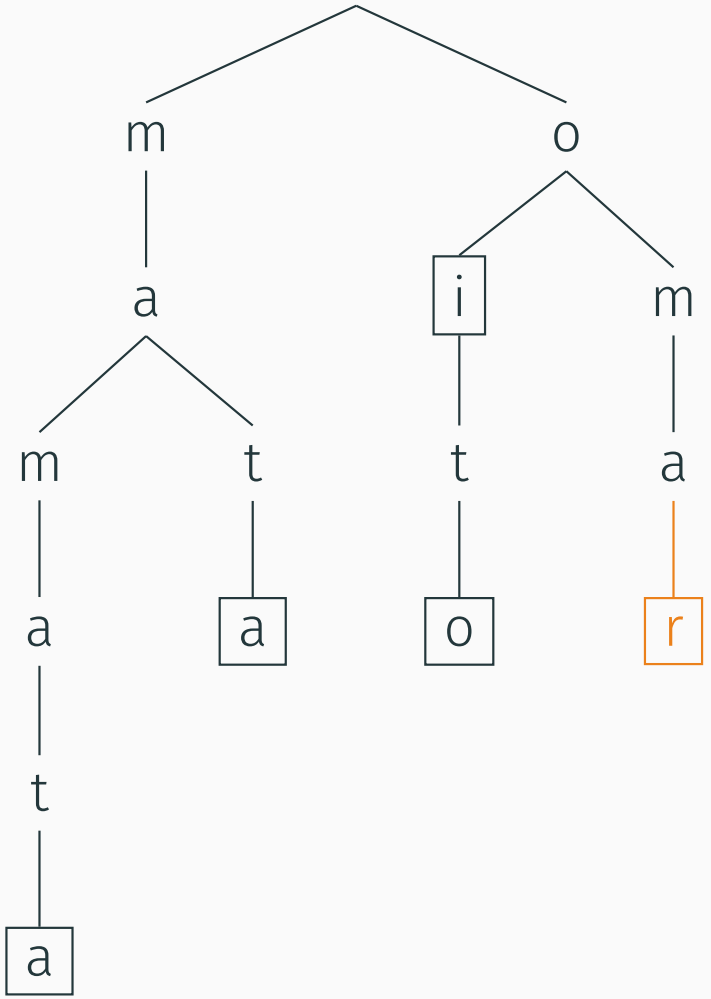
Adicionando “omar”.

CONSTRUÇÃO



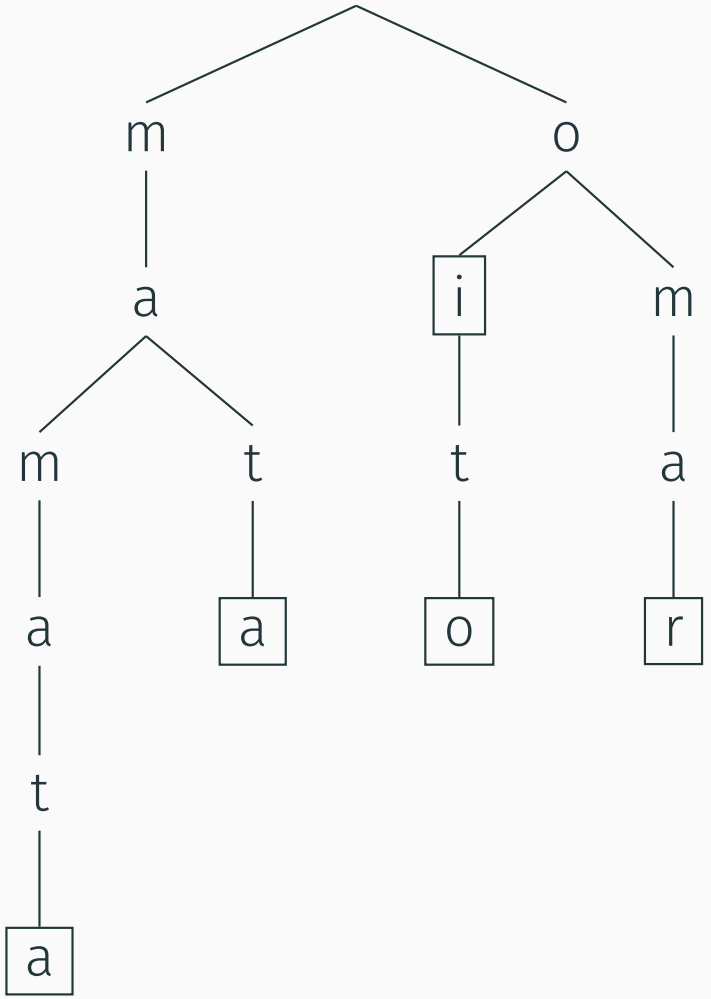
Adicionando “omar”.

CONSTRUÇÃO



Adicionando “omar”.

CONSTRUÇÃO



Adicionando “omar”.

USOS

Construir uma trie para $\mathcal{S} = \{S_1, \ldots, S_k\}$ consome tempo $\mathcal{O}(\sum_{i=1}^{k} |S_i|)$.

Com esta trie, podemos realizar:

CONTAINS(S) Determina se $S \in \mathcal{S}$.

LCP(S) Determina o maior prefixo comum de S com alguma string de $\mathcal{S}$.

Consumo de tempo: $\mathcal{O}(|S|)$.

AHO-CORASICK

Problema: Determine todas as ocorrências de todas as strings de $\mathcal{S} = \{S_1, \ldots, S_k\}$ em T.

Problema: Determine todas as ocorrências de todas as strings de $\mathcal{S} = \{S_1, \ldots, S_k\}$ em T.

Para cada sufixo $T[i \ldots |T|]$, usando uma trie, determinamos as strings de $\mathcal{S}$ que ocorrem **no início** de $T[i \ldots |T|]$.

Isso leva tempo $\mathcal{O}(|T|^2)$.

LINKS DE FALHA

No KMP: função prefixo guarda, para cada i, o comprimento do maior prefixo de T que é sufixo próprio de T[1 . . i].

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
T	a	c	a	s	a	c	a	r	a	c	a	s	a	c
		0	1	0	1	2	3	0	1	2	3	4	5	6

No KMP: função prefixo guarda, para cada i, o comprimento do maior prefixo de T que é sufixo próprio de T[1 . . i].

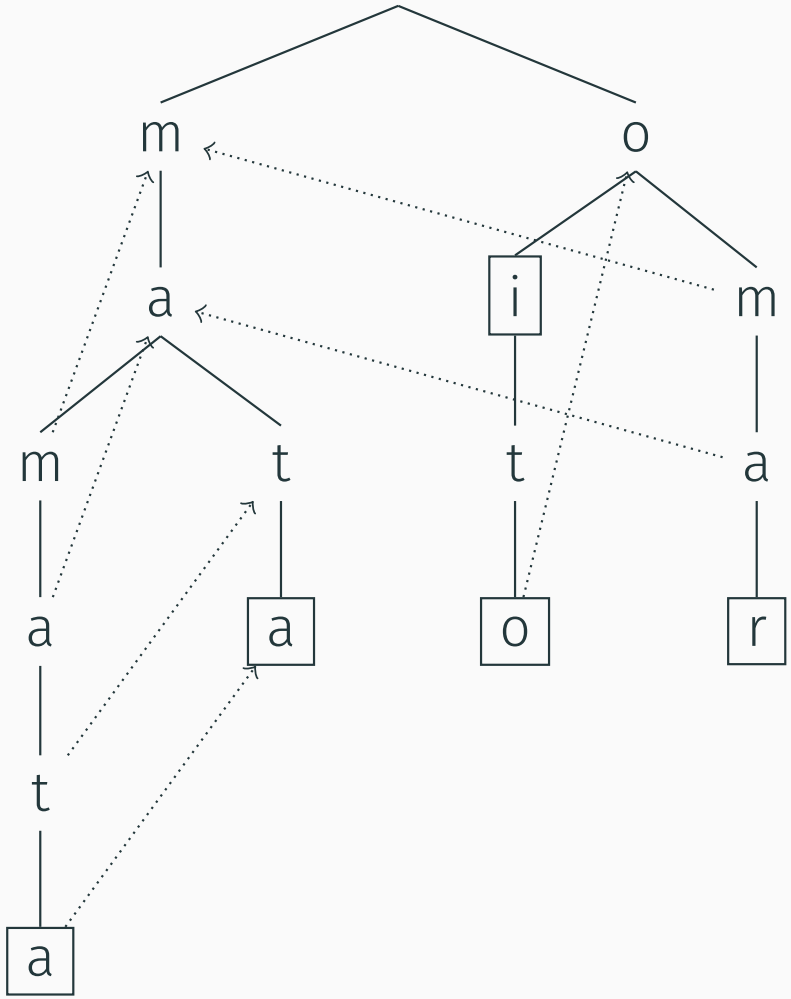
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
T	a	c	a	s	a	c	a	r	a	c	a	s	a	c
		0	1	0	1	2	3	0	1	2	3	4	5	6

Em Tries: para cada nó v, guarda o nó mais profundo cuja string seja um sufixo próprio da string de v.

link de falha

LINKS DE FALHA

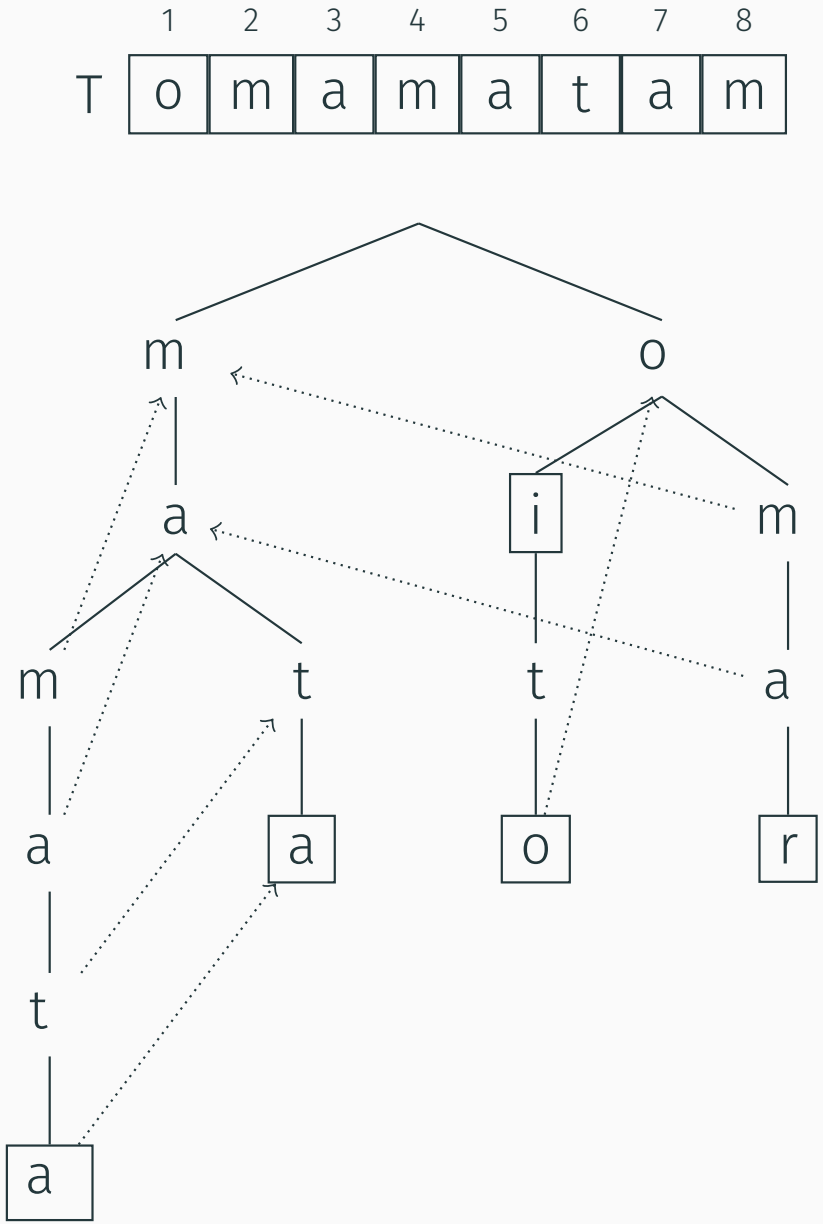
Para cada nó v, guardar o nó mais profundo cuja string seja um sufixo próprio da string de v.



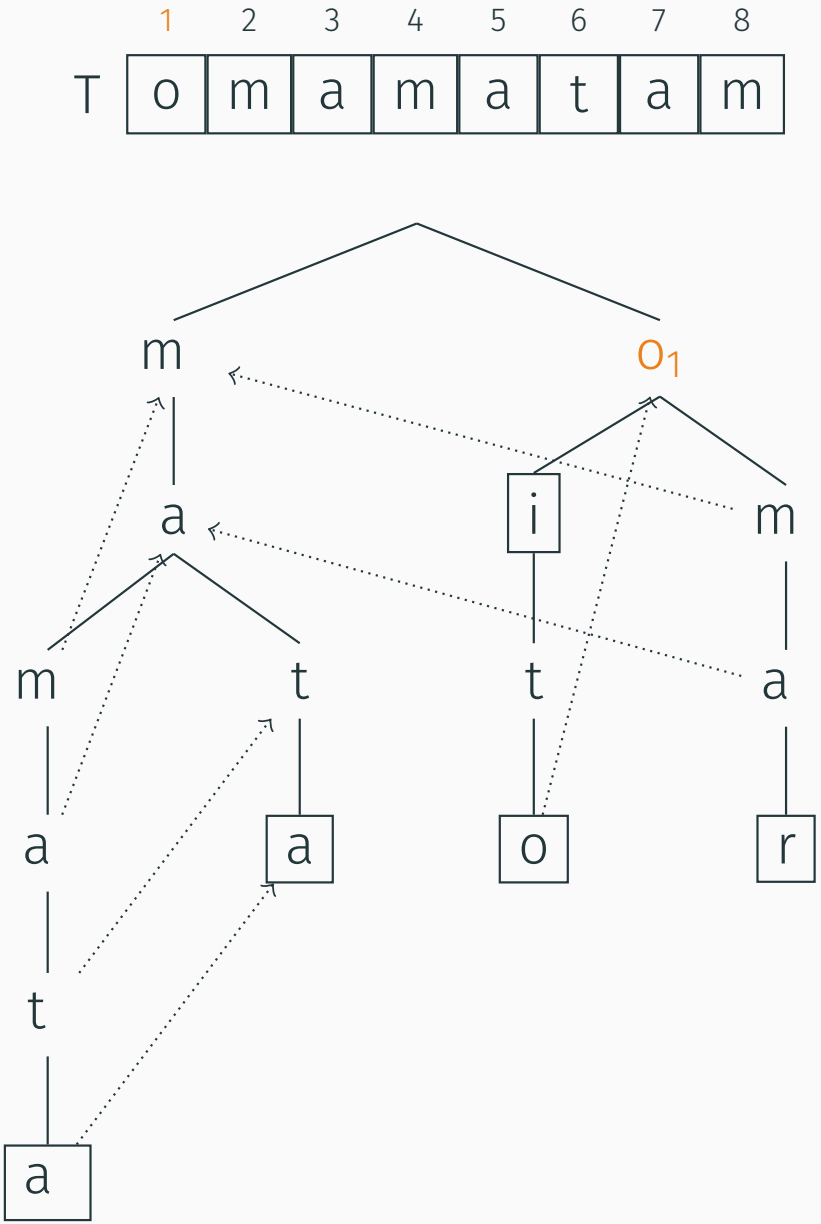
ALGORITMO

Para cada prefixo T[1 . . i], determinar o maior prefixo S[1 . . j] de alguma string de $\mathcal{S}$ que é sufixo de T[1 . . i].

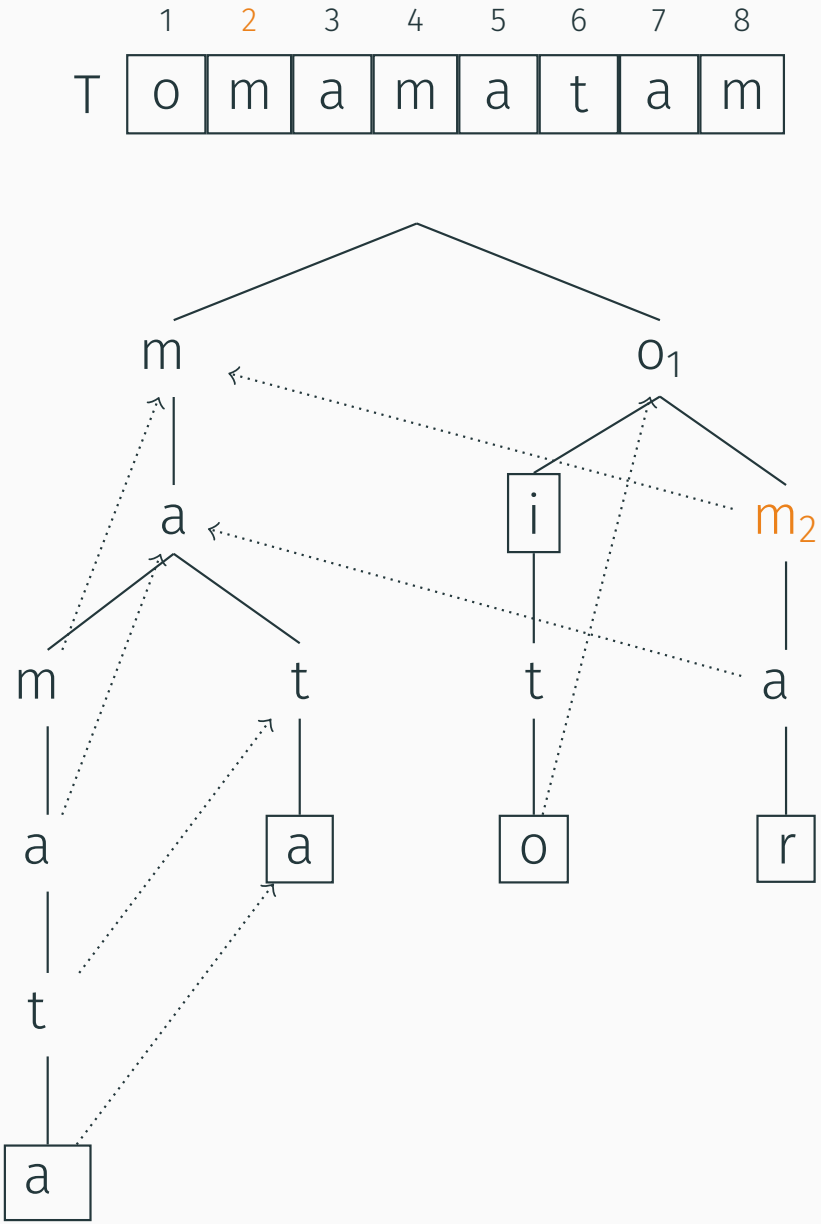
ALGORITMO



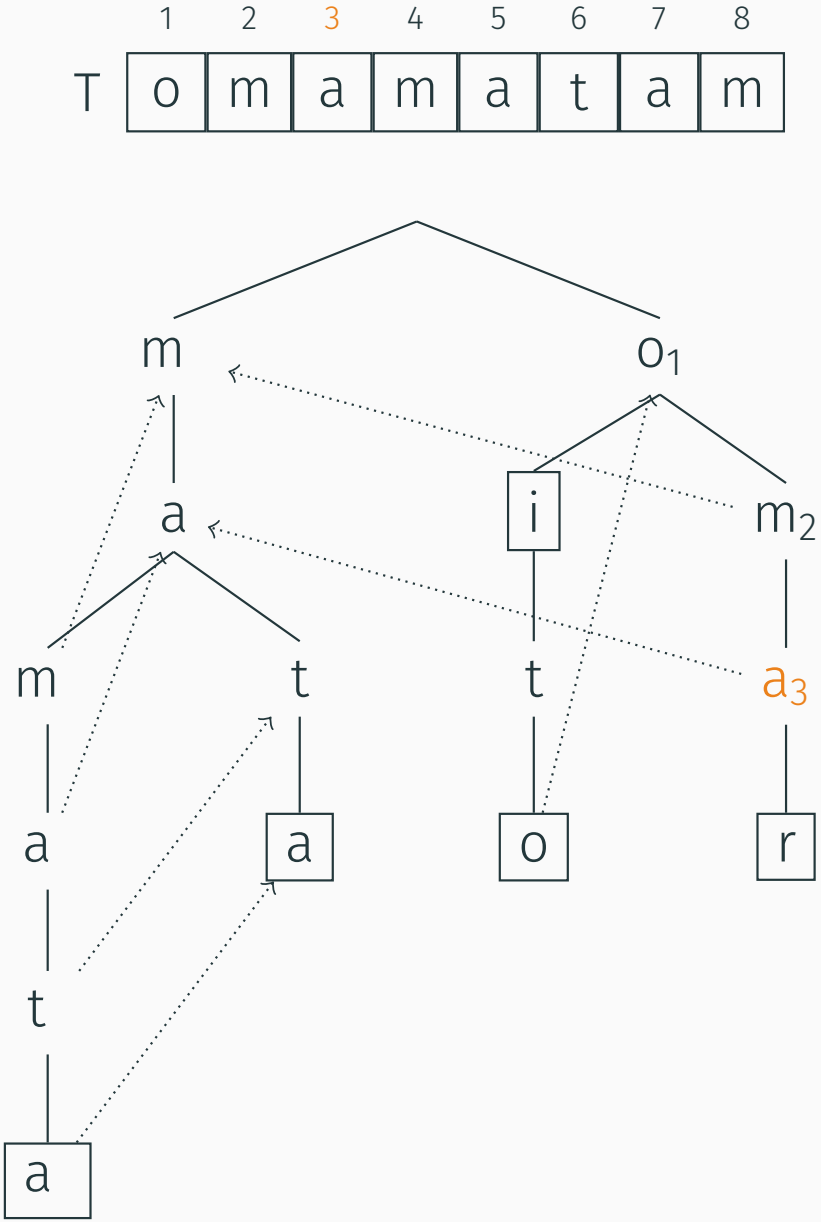
ALGORITMO



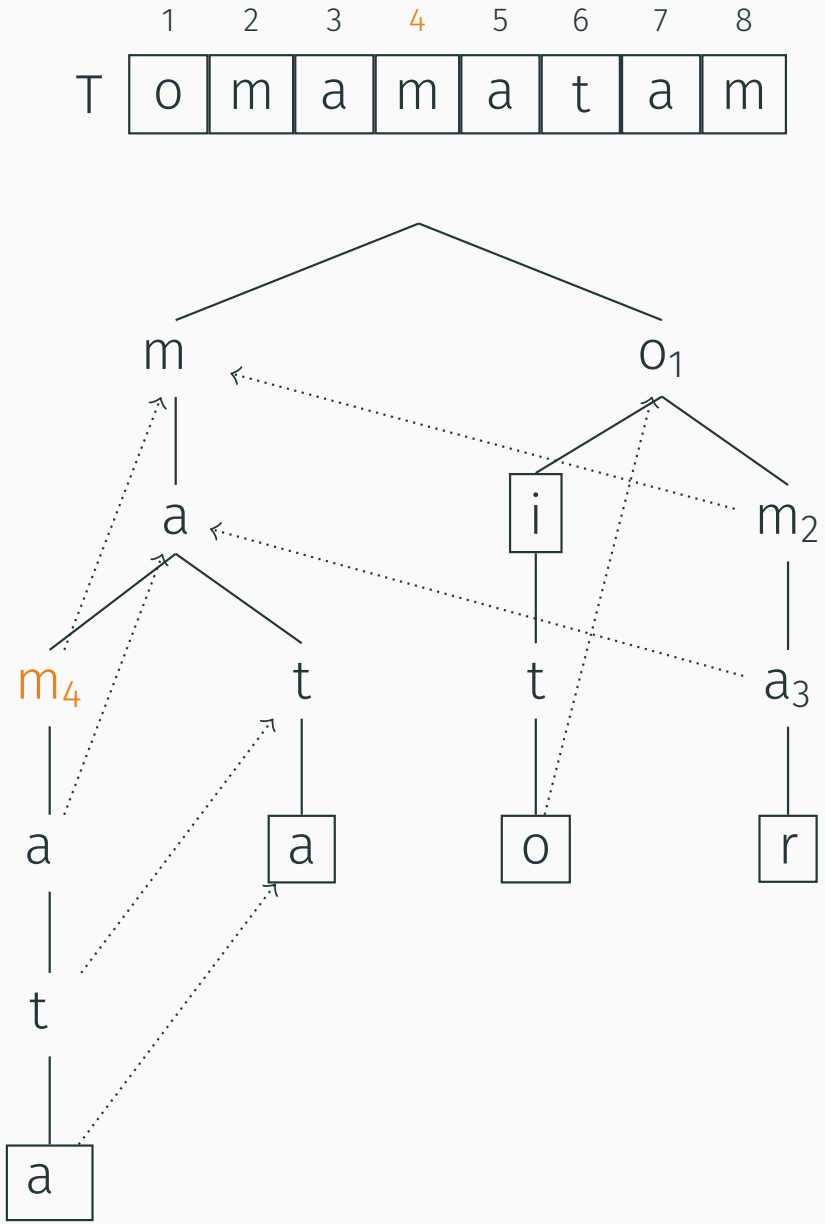
ALGORITMO



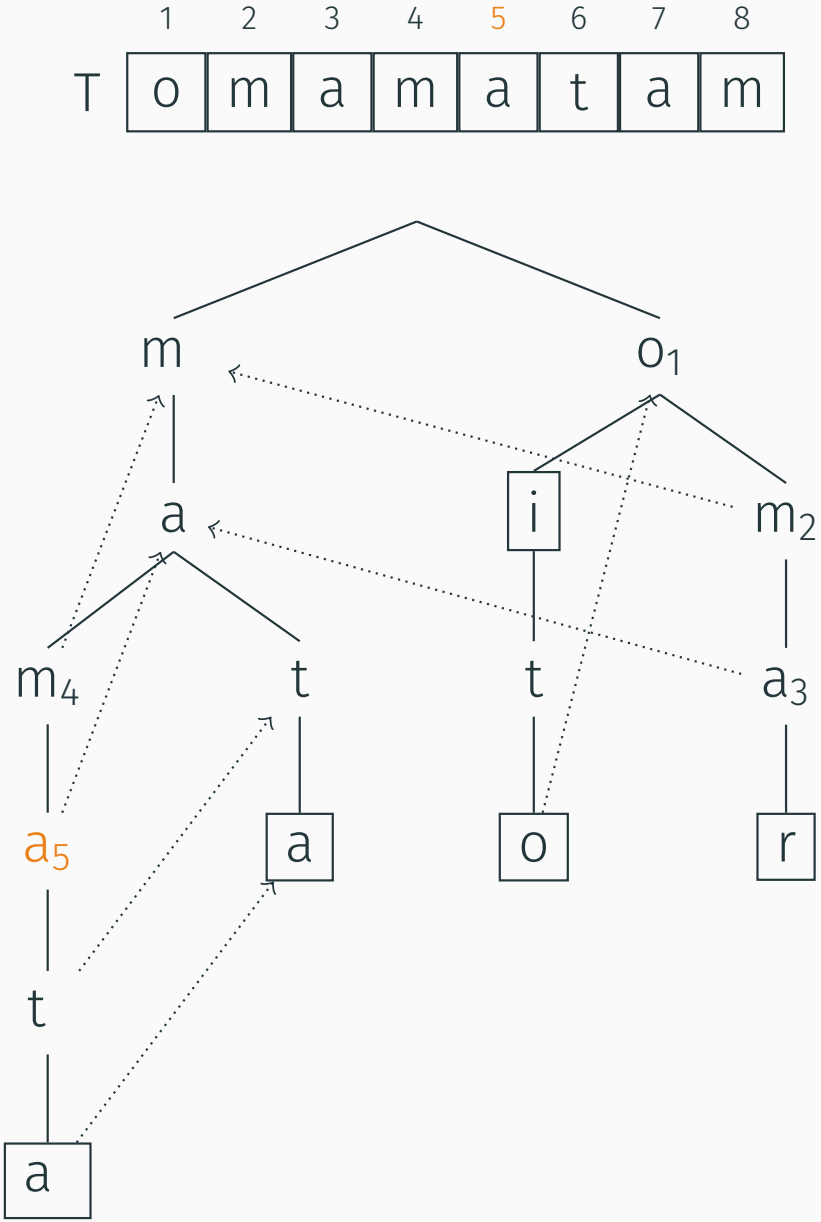
ALGORITMO



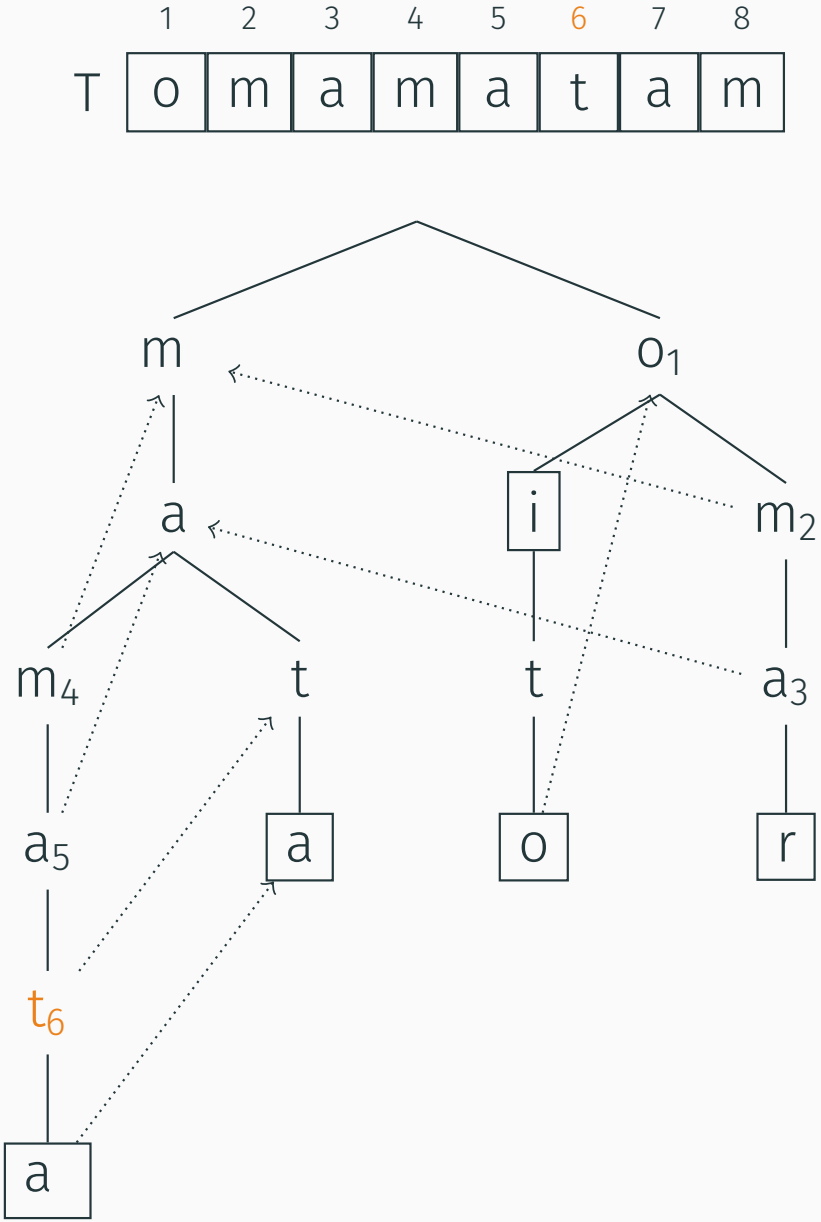
ALGORITMO



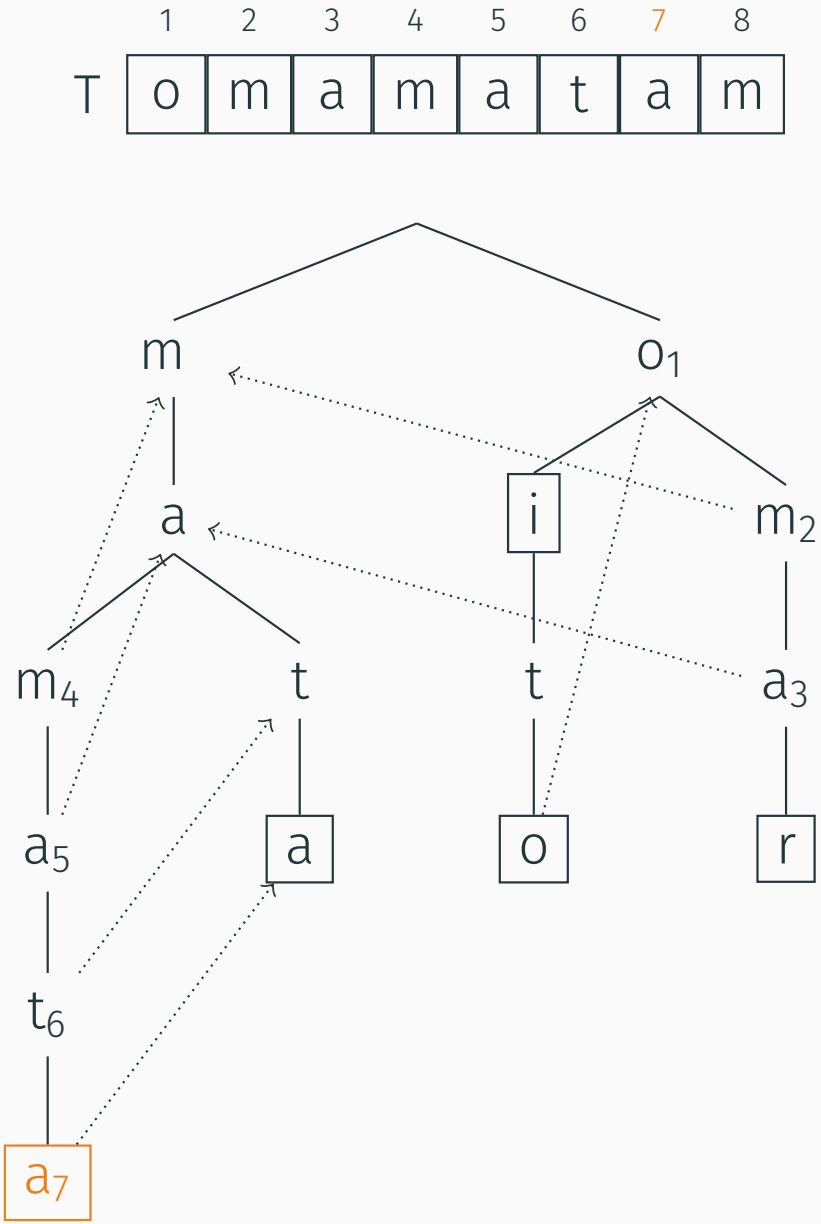
ALGORITMO



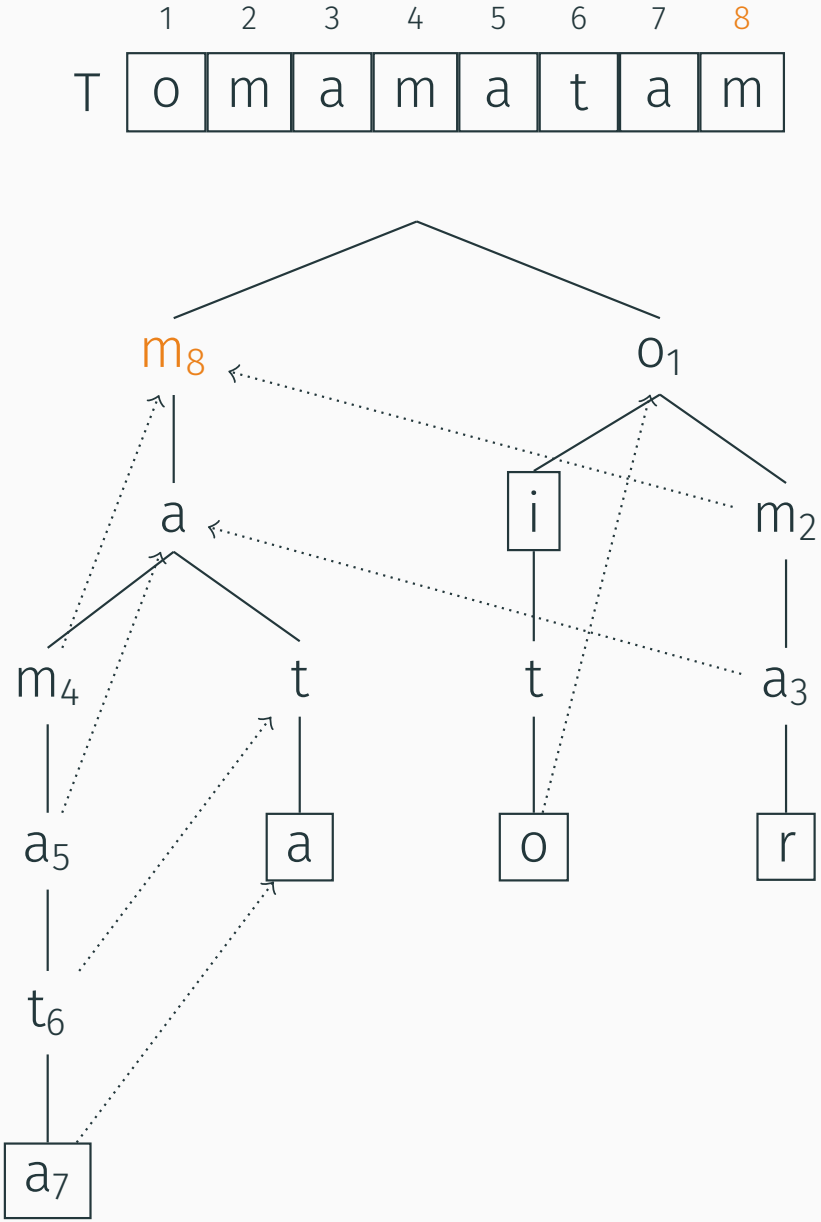
ALGORITMO



ALGORITMO



ALGORITMO



CONSIDERAÇÕES FINAIS

Uma string de $\mathcal{S}$ pode ser sufixo próprio de S[1 . . j]!

link de ocorrência de v: vértice que representa a maior string de $\mathcal{S}$ que é sufixo próprio da string de v.

CONSIDERAÇÕES FINAIS

Uma string de $\mathcal{S}$ pode ser sufixo próprio de $S[1..j]$!

link de ocorrência de v: vértice que representa a maior string de $\mathcal{S}$ que é sufixo próprio da string de v.

O algoritmo pode ser implementado em tempo $\mathcal{O}(\sum_{i=1}^{k} |S_i||\Sigma| + |T| + x)$, onde x é o número de ocorrências.

Perguntas?