

Elementi di Bioinformatica

Gianluca Della Vedova

Univ. Milano-Bicocca
<https://gianluca.dellavedova.org>

23 ottobre 2023

Trie

Trie

Trie

Trie

Trie

Trie

Trie

Trie

- Albero

Trie

Trie

- Albero
- Query: parola \in dizionario

Trie

Trie

- Albero
- Query: parola \in dizionario
- archi etichettati

Trie

Trie

- Albero
- Query: parola \in dizionario
- archi etichettati
- Percorso radice-foglia = parola

Trie

Trie

- Albero
- Query: parola \in dizionario
- archi etichettati
- Percorso radice-foglia = parola

Dizionario

ABRACADABRA

ARRAY

ABRA

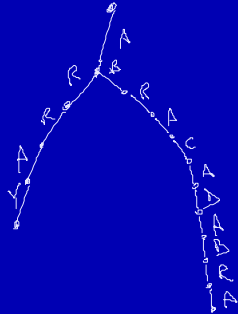
Trie

- Albero
- Query: parola \in dizionario
- archi etichettati
- Percorso radice-foglia = parola

ABRACADABRA

ARRAY

ABRA



Trie

Trie

Trie

Trie

Trie

Terminatore

\$ non appartiene all'alfabeto

Trie

Terminatore

\$ non appartiene all'alfabeto

Dizionario

ABRACADABRA\$

ARRAY\$

ABRA\$

Trie

Terminatore

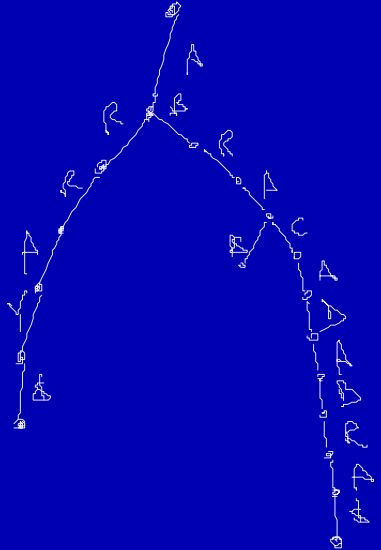
\$ non appartiene all'alfabeto

Dizionario

ABRACADABRA\$

ARRAY\$

ABRA\$



Suffix tree

Suffix tree

Suffix tree

Definizione

Suffix tree

Definizione

- Trie compatto di tutti i suffissi di T

Suffix tree

Definizione

- Trie compatto di tutti i suffissi di $T\$$
- Le etichette degli archi uscenti da x iniziano con simboli diversi

Suffix tree

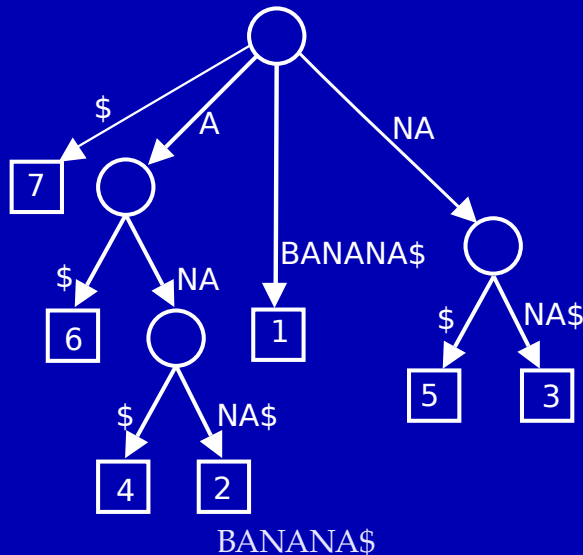
Definizione

- Trie compatto di tutti i suffissi di $T\$$
- Le etichette degli archi uscenti da x iniziano con simboli diversi
- suffissi \Leftrightarrow percorso radice-foglia

Suffix tree

Definizione

- Trie compatto di tutti i suffissi di T\$
- Le etichette degli archi uscenti da x iniziano con simboli diversi
- suffissi \Leftrightarrow percorso radice-foglia



Suffix tree 2: Definizione

- foglie etichettata con posizione inizio suffisso

Suffix tree 2: Definizione

- foglie etichettata con posizione inizio suffisso
- $\text{path-label}(x)$: concatenazione etichette

Suffix tree 2: Definizione

- foglie etichettata con posizione inizio suffisso
- $\text{path-label}(x)$: concatenazione etichette
- $\text{string-depth}(x)$: lunghezza $\text{path-label}(x)$

Suffix tree 2: Definizione

- foglie etichettata con posizione inizio suffisso
- $\text{path-label}(x)$: concatenazione etichette
- $\text{string-depth}(x)$: lunghezza $\text{path-label}(x)$
- Pattern matching = visita

Suffix tree 2: Definizione

- foglie etichettata con posizione inizio suffisso
- $\text{path-label}(x)$: concatenazione etichette
- $\text{string-depth}(x)$: lunghezza $\text{path-label}(x)$
- Pattern matching = visita

Problemi

Suffix tree 2: Definizione

- foglie etichettata con posizione inizio suffisso
- $\text{path-label}(x)$: concatenazione etichette
- $\text{string-depth}(x)$: lunghezza $\text{path-label}(x)$
- Pattern matching = visita

Problemi

- Spazio $O(n^2)$

Suffix tree 2: Definizione

- foglie etichettata con posizione inizio suffisso
- $\text{path-label}(x)$: concatenazione etichette
- $\text{string-depth}(x)$: lunghezza $\text{path-label}(x)$
- Pattern matching = visita

Problemi

- Spazio $O(n^2)$
- Puntatori al testo (posizioni)

Suffix tree 2: Definizione

- foglie etichettata con posizione inizio suffisso
- $\text{path-label}(x)$: concatenazione etichette
- $\text{string-depth}(x)$: lunghezza $\text{path-label}(x)$
- Pattern matching = visita

Problemi

- Spazio $O(n^2)$
- Puntatori al testo (posizioni)
- Spazio $20n$ bytes

Suffix array

Definizione

Suffix array

Definizione

- Array dei suffissi in ordine lessicografico

Suffix array

Definizione

- Array dei suffissi in ordine lessicografico
- Posizioni iniziali del suffisso nell'array

Suffix array

Definizione

- Array dei suffissi in ordine lessicografico
- Posizioni iniziali del suffisso nell'array
- Spazio $4n$ bytes

Suffix array

Definizione

- Array dei suffissi in ordine lessicografico
- Posizioni iniziali del suffisso nell'array
- Spazio $4n$ bytes
- $Lcp[i]$: lunghezza prefisso comune $SA[i]$, $SA[i + 1]$

Suffix array

Definizione

- Array dei suffissi in ordine lessicografico
- Posizioni iniziali del suffisso nell'array
- Spazio $4n$ bytes
- $Lcp[i]$: lunghezza prefisso comune $SA[i]$, $SA[i + 1]$

BANANA\$

i	1	2	3	4	5	6	7
SA	7	6	4	2	1	5	3
Lcp	0	1	3	0	0	2	-

Da Suffix tree a Suffix array

Da Suffix tree a Suffix array

Da Suffix tree a Suffix array

- Visita depth-first di ST

Da Suffix tree a Suffix array

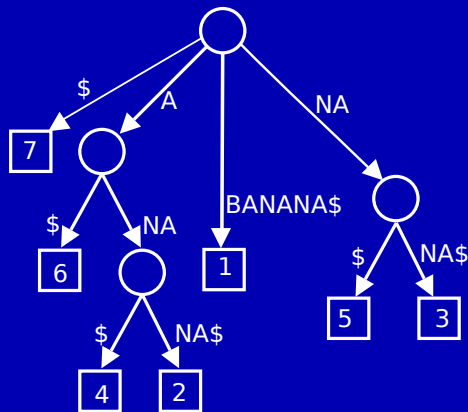
- Visita depth-first di ST
- archi uscenti di ogni nodo in ordine lessicografico

Da Suffix tree a Suffix array

- Visita depth-first di ST
- archi uscenti di ogni nodo in ordine lessicografico
- $Lcp[i] = \text{string-depth di } lca(i, i + 1)$

Da Suffix tree a Suffix array

- Visita depth-first di ST
- archi uscenti di ogni nodo in ordine lessicografico
- $Lcp[i] = \text{string-depth di } lca(i, i + 1)$



Da Suffix array a Suffix tree

Da Suffix array a Suffix tree

Da Suffix array a Suffix tree

- $Lcp = 0$: partizione SA

Da Suffix array a Suffix tree

- $Lcp = 0$: partizione SA
- corrispondono ai figli della radice

Da Suffix array a Suffix tree

- $Lcp = 0$: partizione SA
- corrispondono ai figli della radice
- ricorsione prendendo i numeri minimi

Da Suffix array a Suffix tree

- Lcp = 0: partizione SA
- corrispondono ai figli della radice
- ricorsione prendendo i numeri minimi

BANANA\$

i	0	1	2	3	4	5	6
SA	7	6	4	2	1	5	3
Lcp	0	1	3	0	0	2	-

Da Suffix array a Suffix tree

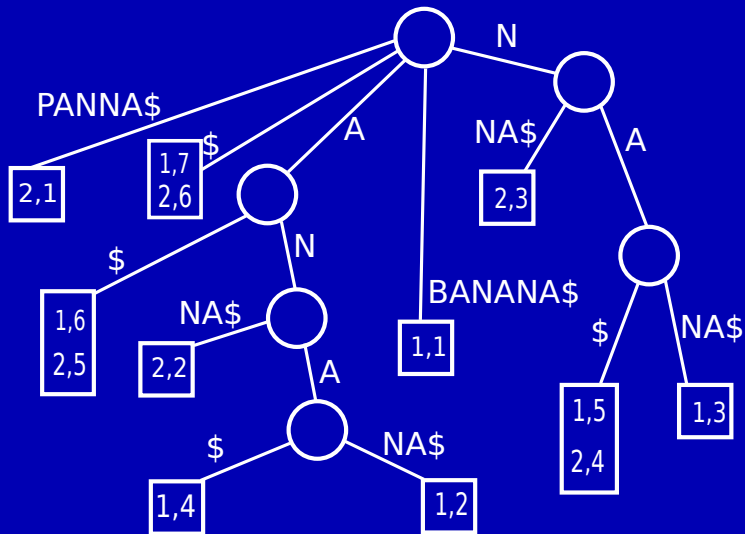
- Lcp = 0: partizione SA
- corrispondono ai figli della radice
- ricorsione prendendo i numeri minimi

BANANA\$

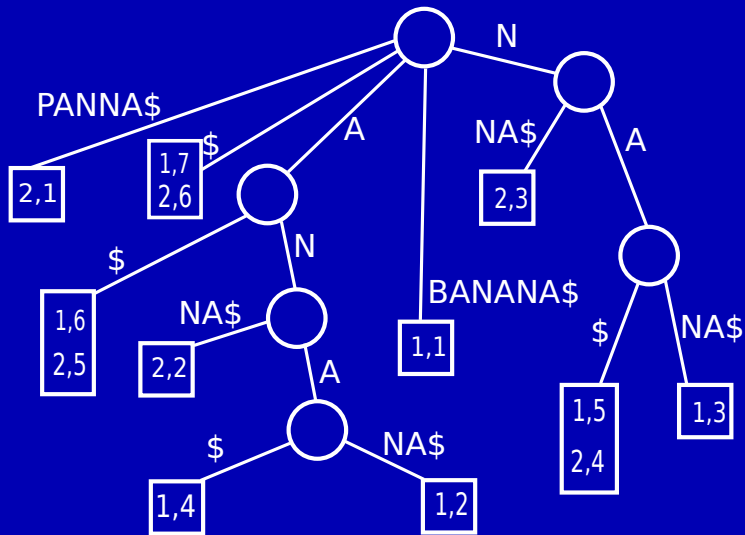
i	0	1	2	3	4	5	6
SA	7	6	4	2	1	5	3
Lcp	0	1	3	0	0	2	-

Suffix tree generalizzato

Suffix tree generalizzato



Suffix tree generalizzato



s_1 : BANANA\$

s_2 : PANNA\$

Sottostringa comune più lunga di due stringhe

Due stringhe s_1 e s_2

Sottostringa comune più lunga di due stringhe

Due stringhe s_1 e s_2

- Suffix tree generalizzato = insieme di stringhe

Sottostringa comune più lunga di due stringhe

Due stringhe s_1 e s_2

- Suffix tree generalizzato = insieme di stringhe
- $ST(s_1\$s_2\$)$

Sottostringa comune più lunga di due stringhe

Due stringhe s_1 e s_2

- Suffix tree generalizzato = insieme di stringhe
- $ST(s_1\$s_2\$)$
- Nodo x con foglie di s_1 e s_2

Sottostringa comune più lunga di due stringhe

Due stringhe s_1 e s_2

- Suffix tree generalizzato = insieme di stringhe
- $ST(s_1\$s_2\$)$
- Nodo x con foglie di s_1 e s_2
- Sottostringa di s_1 e s_2

Sottostringa comune più lunga di due stringhe

Due stringhe s_1 e s_2

- Suffix tree generalizzato = insieme di stringhe
- $ST(s_1\$s_2\$)$
- Nodo x con foglie di s_1 e s_2
- Sottostringa di s_1 e s_2
- $ST(s_1\$s_2\$)$

Sottostringa comune più lunga di due stringhe

Due stringhe s_1 e s_2

- Suffix tree generalizzato = insieme di stringhe
- $ST(s_1\$s_2\$)$
- Nodo x con foglie di s_1 e s_2
- Sottostringa di s_1 e s_2
- $ST(s_1\$s_2\$)$
- Max string-depth

Licenza d'uso

Quest'opera è soggetta alla licenza Creative Commons: Attribuzione-Condividi allo stesso modo 4.0. (<https://creativecommons.org/licenses/by-sa/4.0/>).

Sei libero di riprodurre, distribuire, comunicare al pubblico, esporre in pubblico, rappresentare, eseguire, recitare e modificare quest'opera alle seguenti condizioni:

- **Attribuzione** — Devi attribuire la paternità dell'opera nei modi indicati dall'autore o da chi ti ha dato l'opera in licenza e in modo tale da non suggerire che essi avallino te o il modo in cui tu usi l'opera.