Wyszukiwanie wzorca w tekście

Algorytm Knutha-Morrisa-Pratta

Terminologia i przykład

Niech P będzie ciągiem znaków.

Wówczas P_i oznacza i elementowy prefiks ciągu P, czyli początkowy fragment P długości i.

Na przykład, jeżeli P = **aabaaaba** to

P_0 = ciąg pusty
P_1 = a
P_2 = aa
P_3 = aab
P_4 = aaba
ltd.

Prefiksem właściwym ciągu znaków *P* nazywamy każdy prefiks *P* różny od całego *P*

Ciąg znaków nazwiemy sufiksem *P* jeżeli stanowi on końcówkę ciągu *P,*

np. jeżeli

P = abcaabc

To sufiksami *P* są c, bc, abc, aabc, ...

Przykład 1.

Dany jest wzorzec P = **bbbabb**
Wypisać prefiks P_5 tego wzorca i wszystkie prefiksy właściwe prefiksu P_5. Który z tych prefiksów właściwych jest jednocześnie najdłuższym sufiksem P_5?

Rozwiązanie:

P_5 = **bbbab**
Prefiksy właściwe: **b**, **bb**, **bbb**, **bbba**
Najdłuższy sufiks: **b**

Funkcja prefiksowa *pi( )* dla wzorca *P* jest zdefinowana następująco

pi(k) = długość najdłuższego prefiksu właściwego ciągu P_k *,*
który jest jednocześnie sufiksem P_k

czyli

$pi(k) = max\{ i<k : P_i \text{ jest sufiksem } P_k \}$

np. jeżeli *P* = **bbbabb** to *pi(5) = 1*
bo najdłuższym prefiksem właściwym P_5 jest
Ciąg **b** o długości 1

Przykład 2.

Dany jest wzorzec P = **bbbabb**
Wypełnić tabelę zawierającą wartości funkcji prefiksowej dla wzorca P.

Rozwiązanie:

pozycja	1	2	3	4	5	6
znak wzorca	b	b	b	a	b	b
wartość f. prefiksowej	0	1	2	0	1	2

Przykład 3.

Dany jest wzorzec $P =$ **bbbabb**
i tekst $T =$ **bbbbabbcbab**

Jakie znaki tekstu i wzorca będą porównane ze sobą w czasie działania algorytmu naiwnego i algorytmu KMP?

P = **bbbabb**

T = **bbbbabbcbab**

Odpowiedź dla naiwnego (po lewej stronie litera tekstu, po prawej wzorca):

b-b

b-b

b-b

b-a (brak dopasowania, kontynuujemy od pozycji 2 tekstu i 1 wzorca)

b-b

b-b

b-b

b-b

a-a

b-b

b-b (jest dopasowanie, kontynuujemy od pozycji 3 tekstu i 1 wzorca)

P = **bbbabb**

T = **bbbbabbcbab**

Odpowiedź dla naiwnego (po lewej stronie litera tekstu, po prawej wzorca):

b-b

b-b

b-b

b-a (brak dopasowania, kontynuujemy od pozycji 2 tekstu i 1 wzorca)

b-b

b-b

b-b

b-b

a-a

b-b

b-b (jest dopasowanie, kontynuujemy od pozycji 3 tekstu i 1 wzorca)

P = **bbbabb**

T = **bbbbabbcbab**

b-b

b-b

a-b (brak dopasowania, kontynuujemy od pozycji 4 tekstu i 1 wzorca)

b-b

a-b (brak dopasowania, kontynuujemy od pozycji 5 tekstu i 1 wzorca)

a-b (brak dopasowania, kontynuujemy od pozycji 6 tekstu i 1 wzorca)

b-b

b-b

c-b (brak dopasowania, koniec – tekst za krótki)

P = **bbbabb**

T = **bbbbabbcbab**

Odpowiedź dla KMP (po lewej stronie litera tekstu, po prawej wzorca):

b-b

b-b

b-b

b-a (brak dopasowania, kontynuujemy od pozycji 4 tekstu i 3 wzorca bo *pi(3)*=*2* dla *P*

P = **bbbabb**

T = **bbbbabbcbab**

(kontynuujemy od pozycji 4 tekstu i 3 wzorca
bo *pi(3)=2* dla *P*)

b-b

a-a

b-b

b-b (jest dopasowanie, kontynuujemy od pozycji
7 tekstu i 3 wzorca bo *pi(6)=2* dla *P*)

P = **bbbabb**

T = **bbbbabbcbab**

c-b (brak dopasowania, koniec – tekst za krótki)