Wyszukiwanie wzorca w tekście

Będziemy rozpatrywać 4 metody wyszukiwania wzorca w tekście a same:

- Naiwny – działa w ten sposób ze po prostu w podwójnie pętli sprawdzamy wzorce od pierwszego symbola , z kolejnymi symbolami w tekście i jeśli wszystkie symbole wzorca po koleje występują w tekście to miejsce występowania wzorca znaleziono.

- Rabina-Karpa - działa w ten sposób ze wyszukuje wzorzec w tekście za pomocy haszowania

Przypadek pesymistyczny dla tej metody będzie w tym przypadku jeżeli hasz suma kolejnego podejścia w tekście będzie równa hasz sumie wzorca ale przy sprawdzaniu symbol po symbole wzorzec będzie różny od kolejnego podejścia, i będzie jeszcze gorzej w przypadku długiego wzorca i jeśli w tym wzorcu symbole które znajdują na końcu wzorca będzie różne, a wszystkie pozostałe równe.

Przypadek optymistyczny będzie w przypadku jeśli w tekście przy kolejnej hasz sumie która będzie równa hasz sumie wzorca to wzorce będzie równy kolejnej próbie, i nie będzie przypadków w których hasz suma w tekście równa hasz sumie wzorca, i przy sprawdzaniu wszystkie symbole będzie różne.

- Knutha-Morrisa-Pratta - algorytm wykorzystuje fakt, że w przypadku wystąpienia niezgodności ze wzorcem, sam wzorzec zawiera w sobie informację pozwalającą określić, gdzie powinna się zacząć kolejna próba dopasowania, pomijając ponowne porównywanie już dopasowanych znaków. Dzięki temu właściwy algorytm działa w czasie liniowym względem długości przeszukiwanego tekstu i wzorca

W przypadku gdy wzorzec będzie różny od kolejnego podejścia i będzie omijany elementy w tekście takiej długości jak wzorzec, czyli nie będziemy sprawdzać już sprawdzonych elementów to będzie przypadek optymistyczny. Jeśli sytuacja będzie odwrotna, czyli wzorzec będzie takiego formatu ze przy kolejnym podejściu będzie omijane 1-2 symbole, to będzie sytuacja pesymistyczna.

- Boyera-Moora - algorytm wykonuje wstępne przetwarzanie, dla którego wyszukany jest ciąg (szablon), ale nie jest przeszukiwany w (tekście). Jest zatem odpowiedni dla aplikacji, w których obraz jest znacznie krótszy niż tekst lub gdzie jest przechowywany w wielu zapytaniach. Algorytm Boyera-Moore'a wykorzystuje informacje uzyskane podczas etapu wstępnego przetwarzania do pomijania części tekstu, co prowadzi do zmniejszenia współczynnika stałego niż w przypadku wielu innych algorytmów ciągu wyszukiwania.

Algorytm Boyera Moore'a dla „dobrych” danych jest bardzo szybki, a prawdopodobieństwo wystąpienia „złych” danych jest bardzo małe. Dlatego optymalne jest w większości przypadków, gdy nie jest możliwe wstępne przetworzenie tekstu, w którym przeprowadzane jest wyszukiwanie. O ile w krótkich tekstach zysk nie uzasadnia wstępnych obliczeń.

Sprawdzimy ilość wykonanych porównań w przypadku losowym (wybieramy fragment tekstu o odpowiedniej długości rozpoczynając od losowo wybranej pozycji).Dla uzyskania dokładniejszych wyników będziemy 100 raz losować wzorzec i go wyszukiwać dla wszystkich metod.

Długość tekstu ~ 88000

W tej tabeli będziemy wyszukiwać wzorzec w kodzie dwójkowym, czyli nasz tekst składa się z 2 symbolów (0 i 1)

długość tekstu 88720 symbolów.

	3	5	10	50	1000	5000
Naive	175866	189582	207428	208934	210271	216076
Knuth-Morris-Pratt	107578	115369	116486	116488	118051	117893
Rabina-Karpa	129524	118898	107680	91390	90778	90683
Boyer-Moor	125930	152200	164793	177101	179685	171474

Jak można zobaczyć w powyższej tabeli, w tekście który składa się z alfabeta 2 symbolowego metoda Rabina-Karpa jest najlepsza.

W tej tabeli będziemy wyszukiwać wzorzec w fragmencie DNA, czyli nasz tekst składa się z 4 symbolów (A, C, T, G)

długość tekstu 88836 symbolów.

	3	5	10	50	1000	5000
Naive	117435	116924	117636	117945	123967	130150
Knuth-Morris-Pratt	108298	109423	110035	109772	108720	107067
Rabina-Karpa	93485	90371	90086	90414	92283	91513
Boyer-Moor	57754	46387	39779	31410	35414	42980

Przy wyszukiwaniu w tekście który składa się z alfabeta 4 symbolowego najlepsza metoda jest Boyer-Moor

W tej tabeli będziemy wyszukiwać wzorzec w fragmencie powieści, czyli nasz tekst składa się z symbolów angielskiego języka, w którym występuje litery od [a-z] [A-Z] znaki interpunkcyjne i liczby

długość tekstu 88481 symbolów.

	3	5	10	50	1000	5000
Naive	95906	94334	95517	95678	97255	100327
Knuth-Morris-Pratt	94806	93419	94351	94740	95007	94053
Rabina-Karpa	90008	89504	89501	89426	89425	89371
Boyer-Moor	35559	22791	13048	5061	3316	6471

Przy wyszukiwaniu w tekście który składa się z liter od [a-z] [A-Z] znaków interpunkcyjne i liczb najlepsza metoda jest Boyer-Moor