Jak matematyka pomaga w wyszukiwanie wzorca

Artur Jeż

28 września 2011

Wiek nauki

Wiek nauki

- projekty nuklearne
- rozwój fizyki
- Internet
- digitalizacja danych
- poznanie genomu
- rozwój lotnictwa
- kryptografia

Wiek nauki

- projekty nuklearne
- rozwój fizyki
- Internet
- digitalizacja danych
- poznanie genomu
- rozwój lotnictwa
- kryptografia
- kryzys finansowy 2007–?

Wiek nauki

- projekty nuklearne
- rozwój fizyki
- Internet
- digitalizacja danych
- poznanie genomu
- rozwój lotnictwa
- kryptografia
- kryzys finansowy 2007–?

We wszystkich matematyka i informatyka były obecne.

Wiek nauki

- projekty nuklearne
- rozwój fizyki
- Internet
- digitalizacja danych
- poznanie genomu
- rozwój lotnictwa
- kryptografia
- kryzys finansowy 2007–?

We wszystkich matematyka i informatyka były obecne.

- jako podstawa lub ważne narzędzie
- w większości egalitarne

A Wy?

- Część z Was pozostanie w nauce — może będziecie uczestniczyć w równie wielkim przedsięwzięciu!

A Wy?

- Część z Was pozostanie w nauce — może będziecie uczestniczyć w równie wielkim przedsięwzięciu!
- Większość — chce po studiach mieć interesującą i dobrze płatną pracę

A Wy?

- Część z Was pozostanie w nauce — może będziecie uczestniczyć w równie wielkim przedsięwzięciu!
- Większość — chce po studiach mieć interesującą i dobrze płatną pracę

Bądźmy szczerzy, pewnie niewiele osób będzie uczestniczyć w czymś wielkim

A Wy?

- Część z Was pozostanie w nauce — może będziecie uczestniczyć w równie wielkim przedsięwzięciu!
- Większość — chce po studiach mieć interesującą i dobrze płatną pracę

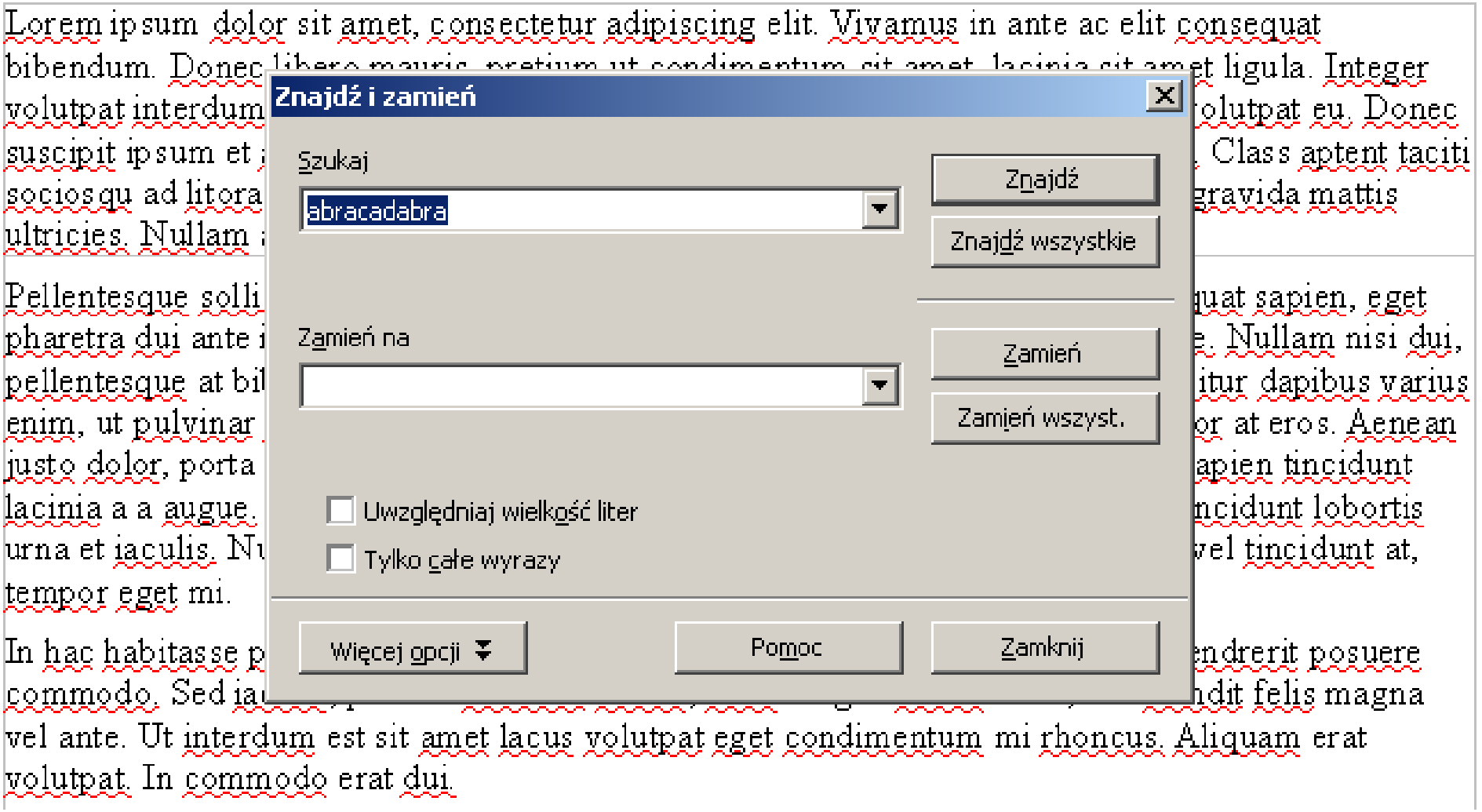
Bądźmy szczerzy, pewnie niewiele osób będzie uczestniczyć w czymś wielkim

Zmiany i zastosowania są wszechobecne. Każdy może się z nimi zetknąć. Nawet w rzeczach pozornie prostych jest naprawdę dużo nauki.

Przykład: wyszukiwanie wzorca

Przykład: wyszukiwanie wzorca

Czy zastanawialiście się kiedyś co tak naprawdę dzieje się po naciśnięciu Ctrl+F w Waszym ulubionym edytorze tekstu?


Znajdź i zamień
Szukaj
abracadabra
Znajdź
Znajdź wszystkie
Zamień na
Zamień
Zamień wszyst.
Uwzględniaj wielkość liter
Tylko całe wyrazy
Więcej opcji
Pomoc
Zamknij

```
#include<cstdio>
#inclue<vector>
using namespace std;
int main(){
  scanf("%d %d",&n,&m);
  for(int i=0;i<n;i++){
/scanf                                                    20,3          Góra
```

Wyszukiwanie wzorca

Wyszukiwanie wzorca

Zanim rozwiążemy, wypadałoby dokładnie zdefiniować.

Wyszukiwanie wzorca

Dane: tekst $t[1 \ldots n]$ i wzorzec $p[1 \ldots m]$.
Wynik:

- Czy wzorzec występuje w tekście?

Wyszukiwanie wzorca

Zanim rozwiążemy, wypadałoby dokładnie zdefiniować.

Wyszukiwanie wzorca

Dane: tekst $t[1 \ldots n]$ i wzorzec $p[1 \ldots m]$.
Wynik:

- Czy wzorzec występuje w tekście?
- Pozycja pierwszego wystąpienia
- Pozycje wystąpień.

Pierwszy pomysł

- przykładamy wzorzec w każdym możliwym miejscu;
- sprawdzamy litera po literze, czy się zgadza.

Pierwszy pomysł (niekoniecznie najlepszy)

- przykładamy wzorzec w każdym możliwym miejscu;
- sprawdzamy litera po literze, czy się zgadza.

Może się zdarzyć, że dla każdego przyłożenia sprawdzamy większość liter.

Pierwszy pomysł (niekoniecznie najlepszy)

- przykładamy wzorzec w każdym możliwym miejscu;
- sprawdzamy litera po literze, czy się zgadza.

Może się zdarzyć, że dla każdego przyłożenia sprawdzamy większość liter.

- przyłożeń: $\approx n - m + 1$
- sprawdzań: $\approx m/c$

Czas działania takiego algorytmu: $\Theta(nm)$.

Pierwszy pomysł (niekoniecznie najlepszy)

- przykładamy wzorzec w każdym możliwym miejscu;
- sprawdzamy litera po literze, czy się zgadza.

Może się zdarzyć, że dla każdego przyłożenia sprawdzamy większość liter.

- przyłożeń: $\approx n - m + 1$
- sprawdzań: $\approx m/c$

Czas działania takiego algorytmu: $\Theta(nm)$.

Jeśli $n = 10^8$ i $m = 10^4$, „chwilę" nam to zajmie*...

* 1 chwila = 60 mgnień oka = 17 minut

Drugi pomysł

- Przykładamy wzorzec w każdym możliwym miejscu.
- Sprytnie odrzucamy pozycje, które na pierwszy rzut oka nie rokują nadziei na sukces.
- Sprawdzamy litera po literze.

Drugi pomysł (trochę lepszy)

- Przykładamy wzorzec w każdym możliwym miejscu.
- Sprytnie odrzucamy pozycje, które na pierwszy rzut oka nie rokują nadziei na sukces.
- Sprawdzamy litera po literze.

Co to znaczy sprytnie?

aaaaaaaaaaaaaaaaaaaaaaaaa

aaaaabaaaaa

Drugi pomysł (trochę lepszy)

- Przykładamy wzorzec w każdym możliwym miejscu.
- Sprytnie odrzucamy pozycje, które na pierwszy rzut oka nie rokują nadziei na sukces.
- Sprawdzamy litera po literze.

Co to znaczy sprytnie? Myślmy o wzorcu jako o liczbie.

aaaaaaaaaaaaaaaaaaaaaaaaaa

aaaaabaaaaa

Drugi pomysł (trochę lepszy)

- Przykładamy wzorzec w każdym możliwym miejscu.
- Sprytnie odrzucamy pozycje, które na pierwszy rzut oka nie rokują nadziei na sukces.
- Sprawdzamy litera po literze.

Co to znaczy sprytnie? Myślmy o wzorcu jako o liczbie.

000000000000000000000000000

00000100000

Drugi pomysł (trochę lepszy)

- Przykładamy wzorzec w każdym możliwym miejscu.
- Sprytnie odrzucamy pozycje, które na pierwszy rzut oka nie rokują nadziei na sukces.
- Sprawdzamy litera po literze.

Co to znaczy sprytnie? Myślmy o wzorcu jako o liczbie.

000000000000000000000000000000

000001000000

- Chcemy sprawdzić czy zielona liczba jest taka sama jak czerwona.
- Obydwie mogą być bardzo długie!

Drugi pomysł (trochę lepszy)

- Przykładamy wzorzec w każdym możliwym miejscu.
- Sprytnie odrzucamy pozycje, które na pierwszy rzut oka nie rokują nadziei na sukces.
- Sprawdzamy litera po literze.

Co to znaczy sprytnie? Myślmy o wzorcu jako o liczbie.

000000000000000000000000000

00000100000

- Chcemy sprawdzić czy zielona liczba jest taka sama jak czerwona.
- Obydwie mogą być bardzo długie!
- Sprawdzimy, czy dają takie same reszty modulo p

Drugi pomysł (trochę lepszy)

- Przykładamy wzorzec w każdym możliwym miejscu.
- Sprytnie odrzucamy pozycje, które na pierwszy rzut oka nie rokują nadziei na sukces.
- Sprawdzamy litera po literze.

Co to znaczy sprytnie? Myślmy o wzorcu jako o liczbie.

$$0000000000000000000000000000$$

$$00000100000$$

- Chcemy sprawdzić czy zielona liczba jest taka sama jak czerwona.
- Obydwie mogą być bardzo długie!
- Sprawdzimy, czy dają takie same reszty modulo $p = 7$.

$$00000100000 = 5 \pmod 7$$

$$00000000000 = 0 \pmod 7$$

- Jeśli liczby są różne modulo 7 to oczywiście nie mogą być takie same!
- Jeśli zaś są równe. . . sprawdzamy je cyfra po cyfrze.

- Jeśli liczby są różne modulo 7 to oczywiście nie mogą być takie same!
- Jeśli zaś są równe. . . sprawdzamy je cyfra po cyfrze.

Dlaczego to ma sens?

Zamiast $p = 7$ wybierzmy $p = 10^9 + 7$. Szansa pomyłki jest mała.

- Jeśli liczby są różne modulo 7 to oczywiście nie mogą być takie same!
- Jeśli zaś są równe. . . sprawdzamy je cyfra po cyfrze.

Dlaczego to ma sens?

Zamiast $p = 7$ wybierzmy $p = 10^9 + 7$. Szansa pomyłki jest mała.

Bardziej formalnie

Wybieramy losowe p.
A z pewnych dodatkowych powodów warto, żeby była to liczba pierwsza.

- Jeśli liczby są różne modulo 7 to oczywiście nie mogą być takie same!
- Jeśli zaś są równe. . . sprawdzamy je cyfra po cyfrze.

Dlaczego to ma sens?

Zamiast $p = 7$ wybierzmy $p = 10^9 + 7$. Szansa pomyłki jest mała.

Bardziej formalnie

Wybieramy losowe p.
A z pewnych dodatkowych powodów warto, żeby była to liczba pierwsza.

Dlaczego pierwsze?

- $p = 10^n$ nie jest dobrym pomysłem;
- jak dłużej pomyśleć, to złożone liczby nie są dobre.

Dla każdego przyłożenia liczymy resztę z dzielenia przez p, co nie jest prostsze.

Dla każdego przyłożenia liczymy resztę z dzielenia przez p, co nie jest prostsze. A może jednak jest?

Dla każdego przyłożenia liczymy resztę z dzielenia przez p, co nie jest prostsze. A może jednak jest?

$$100010000\mathbf{1100000001}000000110$$

Dla każdego przyłożenia liczymy resztę z dzielenia przez p, co nie jest prostsze. A może jednak jest?

$$10001000011000000010000000110$$

Dla każdego przyłożenia liczymy resztę z dzielenia przez p, co nie jest prostsze. A może jednak jest?

$$1000100001\mathbf{10000000010}00000110$$

$$11000000001 = 4 \pmod 7$$

$$10000000010 = 0 \pmod 7$$

Dla każdego przyłożenia liczymy resztę z dzielenia przez p, co nie jest prostsze. A może jednak jest?

$$100010000\mathbf{11000000001}000000110$$

$$11000000001 = 4 \pmod 7$$

$$10000000010 = 0 \pmod 7$$

11000000001

Dla każdego przyłożenia liczymy resztę z dzielenia przez p, co nie jest prostsze. A może jednak jest?

$$1000100001\mathbf{1000000001}000000110$$

$$11000000001 = 4 \pmod 7$$

$$10000000010 = 0 \pmod 7$$

$$10000000000 = 4 \pmod 7$$

$$11000000001 - 10000000000$$

Dla każdego przyłożenia liczymy resztę z dzielenia przez p, co nie jest prostsze. A może jednak jest?

$$1000100001\mathbf{10000000010}00000110$$

$$11000000001 = 4 \pmod 7$$

$$10000000010 = 0 \pmod 7$$

$$10000000000 = 4 \pmod 7$$

$$(11000000001 - 10000000000) * 10$$

Dla każdego przyłożenia liczymy resztę z dzielenia przez p, co nie jest prostsze. A może jednak jest?

$$1000100001\mathbf{10000000010}00000110$$

$$11000000001 = 4 \pmod 7$$

$$10000000010 = 0 \pmod 7$$

$$10000000000 = 4 \pmod 7$$

$$(11000000001 - 10000000000) * 10 + 0$$

Dla każdego przyłożenia liczymy resztę z dzielenia przez p, co nie jest prostsze. A może jednak jest?

$$1000100001\mathbf{1000000010}00000110$$

$$11000000001 = 4 \pmod 7$$

$$10000000010 = 0 \pmod 7$$

$$10000000000 = 4 \pmod 7$$

$$10000000010 = (11000000001 - 10000000000) * 10 + 0$$

Dla każdego przyłożenia liczymy resztę z dzielenia przez p, co nie jest prostsze. A może jednak jest?

$$1000100001\mathbf{10000000010}00000110$$

$$11000000001 = 4 \pmod 7$$

$$10000000010 = 0 \pmod 7$$

$$10000000000 = 4 \pmod 7$$

$$10000000010 = (11000000001 - 10000000000) * 10 + 0 \pmod 7$$

Dla każdego przyłożenia liczymy resztę z dzielenia przez p, co nie jest prostsze. A może jednak jest?

$$100010000110000000010000000110$$

$$11000000001 = 4 \pmod 7$$

$$10000000010 = 0 \pmod 7$$

$$10000000000 = 4 \pmod 7$$

$$10000000010 = (11000000001 - 10000000000) * 10 + 0 \pmod 7$$

Nowa reszta

Można łatwo wyliczyć na podstawie:

- starej reszty,
- $10^m \pmod 7$.

To działa!

To podejście działa: jest mała szansa, że sprawdzimy litera po literze złe wystapienie. Oczekiwany czas działania (pierwsze wystąpienie): $\Theta(m+n)$.

To działa!

To podejście działa: jest mała szansa, że sprawdzimy litera po literze złe wystapienie. Oczekiwany czas działania (pierwsze wystąpienie): $\Theta(m+n)$.

Problemy

Sformułowania wybieramy losowe i szansa jest mała są niepokojące.

To działa!

To podejście działa: jest mała szansa, że sprawdzimy litera po literze złe wystapienie. Oczekiwany czas działania (pierwsze wystąpienie): $\Theta(m+n)$.

Problemy

Sformułowania wybieramy losowe i szansa jest mała są niepokojące. Słabo działa dla tekstu a^n i wzorca a^m.

To działa!

To podejście działa: jest mała szansa, że sprawdzimy litera po literze złe wystapienie. Oczekiwany czas działania (pierwsze wystąpienie): $\Theta(m+n)$.

Problemy

Sformułowania wybieramy losowe i szansa jest mała są niepokojące. Słabo działa dla tekstu a^n i wzorca a^m. Chcemy czegoś, co zawsze działa szybko!

To działa!

To podejście działa: jest mała szansa, że sprawdzimy litera po literze złe wystapienie. Oczekiwany czas działania (pierwsze wystąpienie): $\Theta(m+n)$.

Problemy

Sformułowania wybieramy losowe i szansa jest mała są niepokojące. Słabo działa dla tekstu a^n i wzorca a^m. Chcemy czegoś, co zawsze działa szybko!

Twierdzenie

Istnieje prosty algorytm, który znajduje pierwsze wystąpienie wzorca wykonując tylko $\Theta(n+m)$ operacji.

To działa!

To podejście działa: jest mała szansa, że sprawdzimy litera po literze złe wystapienie. Oczekiwany czas działania (pierwsze wystąpienie): $\Theta(m+n)$.

Problemy

Sformułowania wybieramy losowe i szansa jest mała są niepokojące. Słabo działa dla tekstu a^n i wzorca a^m. Chcemy czegoś, co zawsze działa szybko!

Twierdzenie

Istnieje prosty algorytm, który znajduje pierwsze wystąpienie wzorca wykonując tylko $\Theta(n+m)$ operacji.

Dowód.

. . . można zobaczyć na zajęciach z Algorytmów i Struktur Danych lub Algorytmów Tekstowych (w o wiele ciekawszej wersji). □

Świat nie jest taki prosty

Idealny świat

Rozwiązanie świetne do idealnego sferycznego świata.

Świat nie jest taki prosty

Idealny świat

Rozwiązanie świetne do idealnego sferycznego świata.

- Ile razy napisaliście „nei”?
- Ile razy nie wiedzieliście, czego dokładnie szukacie?

Świat nie jest taki prosty

Idealny świat

Rozwiązanie świetne do idealnego sferycznego świata.

- Ile razy napisaliście „nei”?
- Ile razy nie wiedzieliście, czego dokładnie szukacie?

Błędy

Czasami chcielibyśmy znaleźć nie tylko dokładne wystąpienie wzorca, ale także takie, w którym dopuszczamy trochę „przekłamań”.

Świat nie jest taki prosty

Idealny świat

Rozwiązanie świetne do idealnego sferycznego świata.

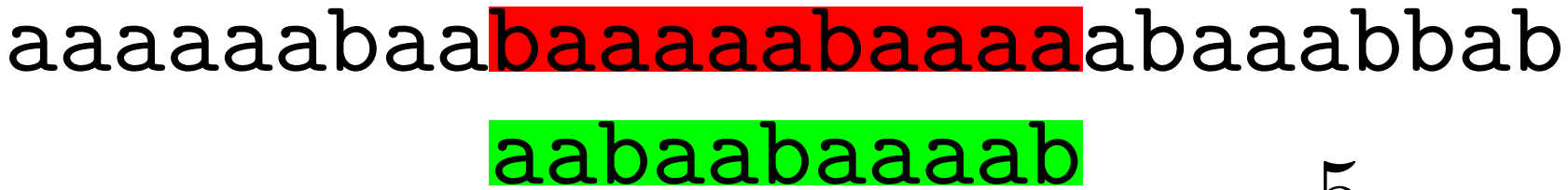
- Ile razy napisaliście „nei”?
- Ile razy nie wiedzieliście, czego dokładnie szukacie?

Błędy

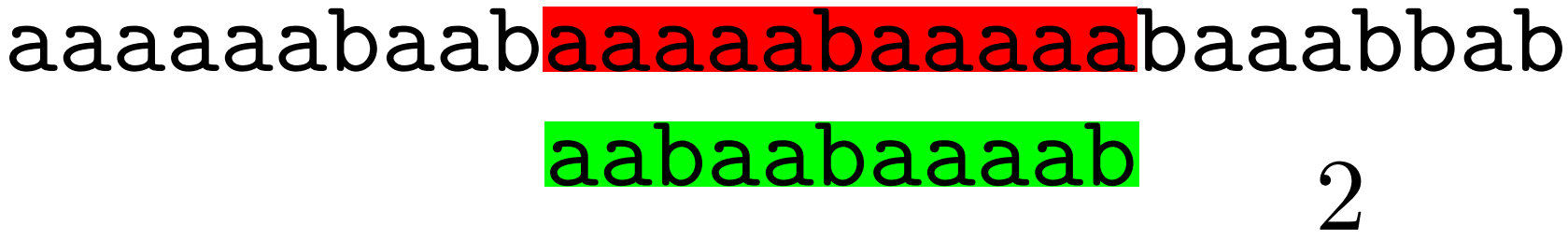
Czasami chcielibyśmy znaleźć nie tylko dokładne wystąpienie wzorca, ale także takie, w którym dopuszczamy trochę „przekłamań”.

Wyszukiwanie wzorca z błędami

Dane: tekst $t[1 \ldots n]$ i wzorzec $p[1 \ldots m]$.
Wynik: dla każdego przyłożenia wzorca liczba niezgodnych znaków.



5


aaaaaabaabaaaaaabaaaaabaaabbab
aabaabaaaab
2

aaaaaabaabaaaaabaaaaaabaaabbab

aabaabaaaab 3

Myślmy pozytywnie

Zgodności

Zamiast liczyć niezgodności, dla każdej litery osobno liczmy zgodności.

Myślmy pozytywnie

Zgodności

Zamiast liczyć niezgodności, dla każdej litery osobno liczmy zgodności.

Lekkie wprowadzenie

Tekst i wzorzec tej samej długości.

Myślmy pozytywnie

Zgodności

Zamiast liczyć niezgodności, dla każdej litery osobno liczmy zgodności.

Lekkie wprowadzenie

Tekst i wzorzec tej samej długości.

aaaaabaaaaa

aabaabaaaab

8

Myślmy pozytywnie

Zgodności

Zamiast liczyć niezgodności, dla każdej litery osobno liczmy zgodności.

Lekkie wprowadzenie

Tekst i wzorzec tej samej długości. Wektory 0-1.

Myślmy pozytywnie

Zgodności

Zamiast liczyć niezgodności, dla każdej litery osobno liczmy zgodności.

Lekkie wprowadzenie

Tekst i wzorzec tej samej długości. Wektory 0-1.

- Ponumerujmy cyfry wektorów dla wzorca i tekstu: $p_1[1\,..\,m]$, $t[1\,..\,m]$.
- Liczymy sumę $p_1 t_1 + p_2 t_2 + \ldots + p_m t_m$

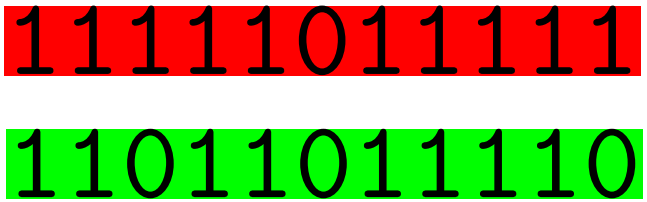
Myślmy pozytywnie

Zgodności

Zamiast liczyć niezgodności, dla każdej litery osobno liczmy zgodności.

Lekkie wprowadzenie

Tekst i wzorzec tej samej długości. Wektory 0-1.



8

- Ponumerujmy cyfry wektorów dla wzorca i tekstu: $p_1[1\,..\,m]$, $t[1\,..\,m]$.
- Liczymy sumę $p_1 t_1 + p_2 t_2 + \ldots + p_m t_m$
- ponumerujmy wzorzec odwrotnie: $p[m\,..\,1]$
- teraz liczymy: $t_1 p_m + t_2 p_{m-1} + \ldots + t_m p_1$

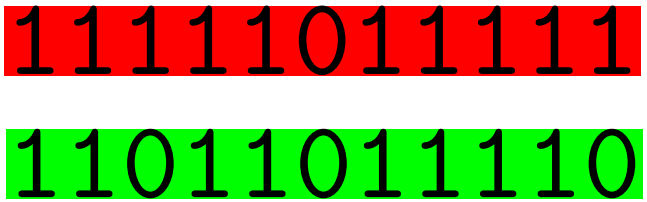
Myślmy pozytywnie

Zgodności

Zamiast liczyć niezgodności, dla każdej litery osobno liczmy zgodności.

Lekkie wprowadzenie

Tekst i wzorzec tej samej długości. Wektory 0-1.



- Ponumerujmy cyfry wektorów dla wzorca i tekstu: $p_1[1\,..\,m]$, $t[1\,..\,m]$.
- Liczymy sumę $p_1t_1 + p_2t_2 + \ldots + p_mt_m$
- ponumerujmy wzorzec odwrotnie: $p[m\,..\,1]$
- teraz liczymy: $t_1p_m + t_2p_{m-1} + \ldots + t_mp_1$
- współczynnik wielomianu (przy x^{m+1})

Traktujemy obydwa wektory jako wielomiany: $P_{\mathtt{a}}(x)$, $T_{\mathtt{a}}(x)$.
T ma współczynnik 1 przy $x^k \iff k$-ta litera t to a.
P ma współczynnik 1 przy $x^k \iff m-k$-ta litera p to a.

Traktujemy obydwa wektory jako wielomiany: $P_{\mathtt{a}}(x)$, $T_{\mathtt{a}}(x)$.
T ma współczynnik 1 przy $x^k \iff k$-ta litera t to a.
P ma współczynnik 1 przy $x^k \iff m-k$-ta litera p to a.

Mnożymy

Mnożymy wielomiany $P_{\mathtt{a}}(x)$ i $T_{\mathtt{a}}(x)$.

Traktujemy obydwa wektory jako wielomiany: $P_a(x)$, $T_a(x)$.
T ma współczynnik 1 przy $x^k \iff k$-ta litera t to a.
P ma współczynnik 1 przy $x^k \iff m-k$-ta litera p to a.

Mnożymy

Mnożymy wielomiany $P_a(x)$ i $T_a(x)$.

Przypomnienie: tekst i wzorzec tej samej długości

$t_1 p_m + t_2 p_{m-1} + \cdots + t_m p_1$, współczynnik przy x^{m+1}.

Traktujemy obydwa wektory jako wielomiany: $P_a(x)$, $T_a(x)$.
T ma współczynnik 1 przy $x^k \iff k$-ta litera t to a.
P ma współczynnik 1 przy $x^k \iff m-k$-ta litera p to a.

Mnożymy

Mnożymy wielomiany $P_a(x)$ i $T_a(x)$.

Przypomnienie: tekst i wzorzec tej samej długości

$t_1 p_m + t_2 p_{m-1} + \cdots + t_m p_1$, współczynnik przy x^{m+1}.

Ogólnie

Współczynnik przy x^{m+1+k} to dokładnie liczba zgodności przy przyłożeniu wzorca tak, że zaczyna się na k-tej literze tekstu.

Traktujemy obydwa wektory jako wielomiany: $P_a(x)$, $T_a(x)$.
T ma współczynnik 1 przy $x^k \iff k$-ta litera t to a.
P ma współczynnik 1 przy $x^k \iff m-k$-ta litera p to a.

Mnożymy

Mnożymy wielomiany $P_a(x)$ i $T_a(x)$.

Przypomnienie: tekst i wzorzec tej samej długości

$t_1 p_m + t_2 p_{m-1} + \cdots + t_m p_1$, współczynnik przy x^{m+1}.

Ogólnie

Współczynnik przy x^{m+1+k} to dokładnie liczba zgodności przy przyłożeniu wzorca tak, że zaczyna się na k-tej literze tekstu.

Zsumować po różnych literach.

No dobrze, ale jak mnożyć wielomiany?

No dobrze, ale jak mnożyć wielomiany?

Transformata Fouriera

Wielomiany można mnożyć korzystając z transformaty Fouriera.

No dobrze, ale jak mnożyć wielomiany?

Transformata Fouriera

Wielomiany można mnożyć korzystając z transformaty Fouriera.

Ta sama, która przedstawia funkcję jako sumę sinusów i cosinusów.

No dobrze, ale jak mnożyć wielomiany?

Transformata Fouriera

Wielomiany można mnożyć korzystając z transformaty Fouriera.

Ta sama, która przedstawia funkcję jako sumę sinusów i cosinusów.

Ale jak ją zrobić szybko?

No dobrze, ale jak mnożyć wielomiany?

Transformata Fouriera

Wielomiany można mnożyć korzystając z transformaty Fouriera.

Ta sama, która przedstawia funkcję jako sumę sinusów i cosinusów.

Ale jak ją zrobić szybko?

$\sqrt{-1}$ na ratunek!

Transformatę Fouriera można policzyć szybko, jeśli liczymy wartości $P_a(x)$ i $T_a(x)$ w zespolonych pierwiastkach z -1.

Pytania?