

Algorytmy i struktury danych – podstawowe informacje

Zawartość

Złożoność obliczeniowa	3
Szacowanie wydajności.....	3
Przykładowe zadanie	3
Algorytmy sortowania	4
Sortowanie bąbelkowe (bubble sort)	4
Zasada działania	4
Implementacja:	5
Rozrysowanie	5
Sortowanie przez wstawianie (karciane, insert sort)	6
Zasada działania	6
Implementacja	6
Rozrysowanie	6
Sortowanie przez wybieranie (selection sort).....	7
Zasada działania	7
Implementacja	7
Sortowanie przez zliczanie (counting sort)	7
Zasada działania	7
Implementacja	8
Sortowanie przez scalanie (merge sort).....	8
Zasada działania	8
Implementacja	8
Rozrysowanie	9
Szybkie sortowanie (quick sort)	10
Zasada działania	10
Implementacja	11
Rozrysowanie	11
Podsumowanie.....	13
Tabela złożoności (szeregowanie algorytmów)	13
Algorytmy stabilne	13
Algorytmy niestabilne	13
Definicje	13

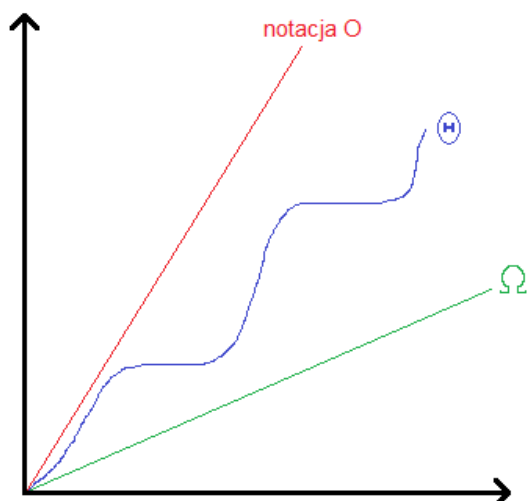
Sortowanie w miejscu.....	13
Sortowanie stabilne	13
Złożoności obliczeniowe	13
Stos	13
Odwrotna notacja polska na przykładach.....	14
Kolejka	14
Haszowanie	15
Haszowanie liniowe i obsługa kolizji.....	15
Haszowanie łańcuchowe	16
Haszowanie kwadratowe	16
Pojęcia związane z haszowaniem.....	17
Drzewo	17
Przeszukiwanie.....	18
Podstawowe pojęcia	18
Grafy.....	18
Graf skierowany (zorientowane)	18
Graf nieskierowany (niezorientowane)	18
Macierze sąsiedztwa.....	18
Listy sąsiedztwa	18
Macierze incydencji	19
Przeszukiwanie BFS (w szerz)	19
Przeszukiwanie DFS (w głąb)	19
Złożoność pamięciowa grafu.....	19
Kompresja danych.....	19
Podstawowe dane	19
Algorytm Huffmana	19
Średnia długość słowa	20
Entropia	21
Redundancja	21
Kompresja RLE.....	22
LZ77	22
Stringologia:	23
Długość Hamminga	23
Bit parzystości	23
Kontrola parzystości.....	23
LCS – Longest common subsequence.....	24
Odległość Levenshteina (edycyjna)	26

Ważne!

gddy $\Leftrightarrow \Leftrightarrow$

Złożoność obliczeniowa

Szacowanie wydajności



Rys. 1 szacowanie wydajności

Wzór do szacowania Θ [teta]

$$0 \leq c_1 * g(n) \leq f(n) \leq c_2 * g(n)$$

$n_0 > 0$ – ilość danych $\in N$ uint

$c_1, c_2 > 0$ – parametry $\in R$ float

Przykładowe zadanie

Udowodnij używając definicji, że $\frac{1}{2}n^2 - 3n = \Theta(n^2)$

$$f(n) = \frac{1}{2}n^2 - 3n$$

$$g(n) = n^2$$

$$0 \leq c_1 * n^2 \leq \frac{1}{2}n^2 - 3n \leq c_2 * n^2 \mid : n^2$$

$$c_1 \leq \frac{1}{2} - \frac{3}{n} \leq c_2$$

Liczenie ograniczenia górnego:

$$c_2 = \lim_{n \rightarrow \infty} \left(\frac{1}{2} - \frac{3}{n} \right) = \frac{1}{2}$$

Liczenie ograniczenia dolnego:

Szukamy najmniejszego naturalnego n , dla którego równanie jest większe od 0.

1) metoda: podstawianie po kolei

$$n_1 = \frac{1}{2} - \frac{3}{1} < 0$$

...

$$n_7 = \frac{1}{2} - \frac{3}{7} > 0$$

Więc $n_0 = 7$

2) metoda: przyrównanie do zera

$$\frac{1}{2} - \frac{3}{n} = 0$$

$$\frac{1}{2} = \frac{3}{n}$$

$$n = 6$$

Zawsze dodajemy do wyniku 1. Więc $n_0 = 7$

$$c_1 = \frac{1}{2} - \frac{3}{7}$$

$$c_1 = \frac{7}{14} - \frac{6}{14}$$

$$c_1 = \frac{1}{14}$$

Odp.: Udowodniłem, że funkcja $f(n)$ jest ograniczona funkcją $\Theta(n^2)$ dla $c_1 = \frac{1}{14}$ $c_2 = \frac{1}{2}$ $n_0 = 7$

Algorytmy sortowania

Sortowanie bąbelkowe (bubble sort)

Zasada działania

Porównuje sąsiadujące ze sobą elementy, aż cała tablica nie będzie posortowana.

- Złożoność obliczeniowa: $O(n^2)$
- Złożoność pamięciowa: $O(1)$
- Zalety: sortowanie w miejscu, stabilny, łatwy w implementacji
- Wady: powolny

Implementacja:

```
void bubbleSort(int *A, int n)
{
    do
    {
        for(int i = 0; i < n - 1; i++)
        {
            if(A[i] > A[i + 1])
            {
                swap(&A[i], &A[i + 1]);
            }
        }
        --n;
    }
    while(n > 1)
}
```

Rys. 2 sortowanie bąbelkowe

Rozrysowanie

Dla danych 8 4 5 60 22 30 66 50

0) 8 4 5 60 22 30 66 50

1) 4 8 5 22 60 30 66 50

2) 4 5 8 22 60 30 66 50

3) 4 5 8 22 60 30 66 50

4) 4 5 8 22 60 30 66 50

5) 4 5 8 22 30 60 66 50

6) 4 5 8 22 30 60 66 50

7) 4 5 8 22 30 60 50 66 – po pierwszym przejściu przez wszystkie elementy

0) 4 5 8 22 30 60 50 66

1) 4 5 8 22 30 60 50 66

2) 4 5 8 22 30 60 50 66

3) 4 5 8 22 30 60 50 66

4) 4 5 8 22 30 60 50 66

5) 4 5 8 22 30 60 50 66

6) 4 5 8 22 30 50 60 66 – po drugim przejściu (wersja którą mamy się nauczyć w tym miejscu by się nie zatrzymała tylko sprawdziła wszystkie kolejne przestawienia pomimo tego, że już widzimy iż jest posortowany)

Sortowanie przez wstawianie (karciane, insert sort)

Zasada działania

Łap po kolei elementy i przesuwaj je tak długo w lewo aż nie znajdziesz elementu od niej większego (mniejszego).

- Złożoność obliczeniowa: $O(n^2)$
- Złożoność pamięciowa: $O(1)$
- Zalety: stabilny, szybki dla małej ilości danych, szybki dla wstępnie posortowanych danych
- Wady: powolny dla rzeczywistych danych (średniego przypadku)

Implementacja

```
void InsertSort(int *A, int n)
{
    int klucz;
    for(int i = 1; i < n; ++i)
    {
        klucz = A[i];
        int j = i - 1;
        while(j >= 0 && A[j]>klucz)
        {
            A[j + 1] = A[j];
            --j;
            A[j + 1] = klucz;
        }
    }
}
```

Rys. 3 sortowanie przez wstawianie

Rozrysowanie

Dla danych 6 5 3 1 8 7 2 4

0) 6 5 3 1 8 7 2 4

1) 5 6 3 1 8 7 2 4

2) 3 5 6 1 8 7 2 4

3) 1 3 5 6 8 7 2 4

4) 1 3 5 6 8 7 2 4

5) 1 3 5 6 7 8 2 4

6) 1 2 3 5 6 7 8 4

7) 1 2 3 4 5 6 7 8

Sortowanie przez wybieranie (selection sort)

Zasada działania

Znajdź najmniejszy element i zamień go z tym na początku (ważne zamień a nie przestawiaj), a potem wyszukuj od kolejnego elementu który jest traktowany jako pierwszy. Często mylony z sortowaniem przez wstawianie można sobie skojarzyć, że wybieramy sobie najmniejszy lub największy element żeby go przestawić na początek.

- Złożoność obliczeniowa: $O(n^2)$
- Złożoność pamięciowa: $O(1)$
- Zalety: sortowanie w miejscu, łatwy w implementacji
- Wady: niestabilny, powolny

Implementacja

```
void SelectionSort(int n, int *A)
{
    int key;
    for(int i = 0; i < n; i++)
    {
        key = i;
        for(int j = i + 1; j < n; j++) //miniumum
        {
            if(A[j] < A[key]) key = j;
        }
        swap(&A[key], &A[i]);
    }
}
```

Rys. 4 sortowanie przez wybieranie

Sortowanie przez zliczanie (counting sort)

Zasada działania

Zapisuje do dodatkowej tablicy ilość wystąpień poszczególnych wartości. Np. pod indeksem 2 jest ilość wystąpień liczby 2.

- Złożoność obliczeniowa: $O(n + k)$
- Złożoność pamięciowa: $O(n + k)$
gdzie k – rozpiętość danych (różnica pomiędzy najmniejszą a największą wartością danych)
- Zalety: szybki, dobry dla danych z dużą ilością powtórzeń, stabilny
- Wady: wykorzystywana dodatkowa pamięć (nie jest w miejscu), słaby dla zróżnicowanych danych

Implementacja

```
void CountingSort(int *A, int *B, int n)
{
    int k; //ilość różnych elementów
    int *temp; //tablica z ilością elementów o danej wartości
    ...
    for(int i = 0; i < k; ++i) temp[i] = 0; //zerowanie tablicy
    for(int i = 0; i < n; ++i) ++temp[A[i]]; //zliczanie
    for(int i = 1; i < k; ++i) temp[i] += temp[i - 1]; //zapisanie pozycji
    for(int i = n - 1; i >= 0; --i) B[--temp[A[i]]] = A[i];
}
```

Rys. 5 sortowanie przez zliczanie

UWAGA: powyższa implementacja nie zawiera funkcji alokujących pamięć.

Sortowanie przez scalanie (merge sort)

Zasada działania

Dzieli tablice na pół tak długo aż nie zostanie jeden element i potem łączy od dołu sortując poszczególne podtablice używając funkcji do scalania porostowych ciągów.

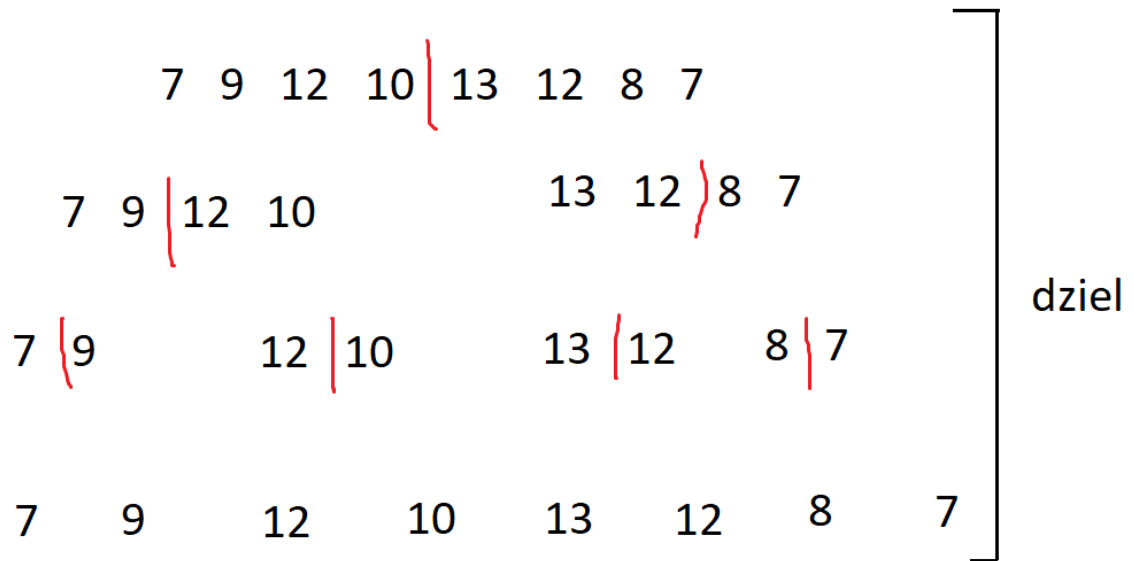
- Złożoność obliczeniowa: $O(n * \log n)$
- Złożoność pamięciowa: $O(n)$
- Zalety: stabilny, prosty, szybki także w praktyce
- Wady: potrzeba dodatkowej pamięci

Implementacja

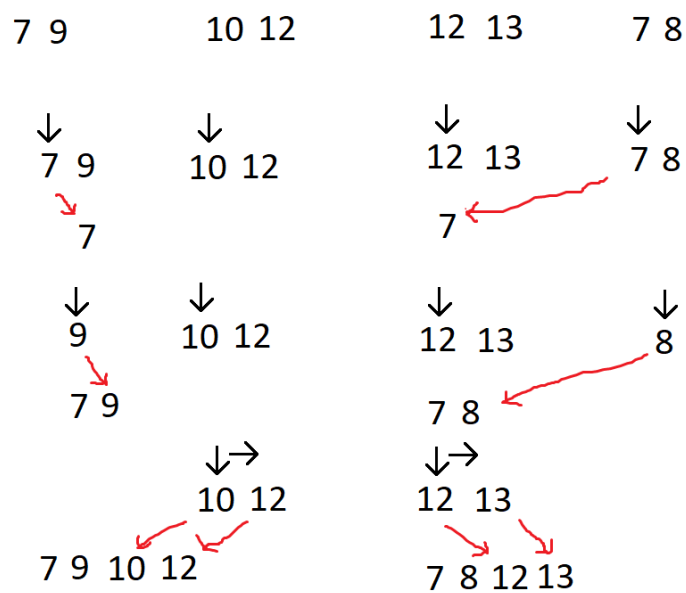
```
void mergesort(int pocz, int kon)
{
    int sr;
    if(pocz < kon)
    {
        sr = (pocz + kon) / 2;
        mergesort(pocz, sr); // Dzielenie lewej czesci
        mergesort(sr + 1, kon); // Dzielenie prawej czesci
        merge(pocz, sr, kon); /* laczenie czesci lewej i prawej*/
    }
}
```

Rys. 6 sortowanie przez łączenie

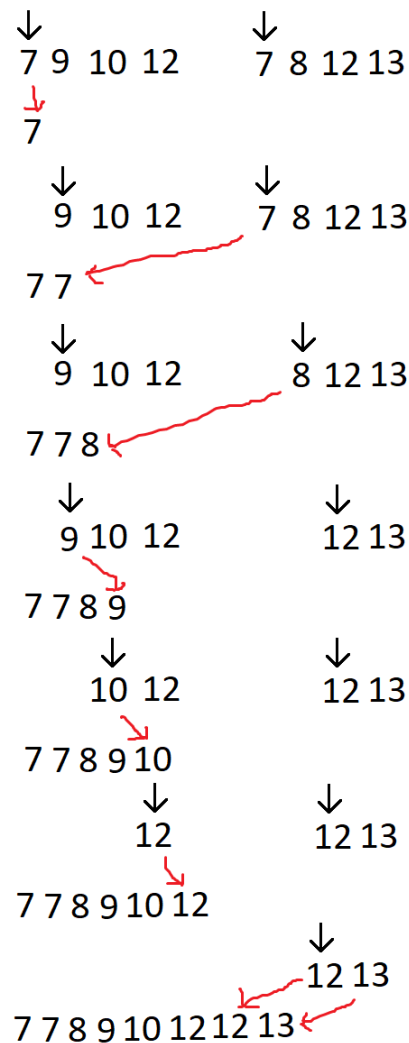
Rozrysowanie



Rys. 7 faza dzielenia w algorytmie sortowania przez scalanie



Rys. 8 faza łączenia w algorytmie przez scalanie



Rys. 9 faza łączenia w algorytmie przez scalanie ostatnie łączenie

Szybkie sortowanie (quick sort)

Zasada działania

Wybieramy arbitralnie jakiś punkt i przerzucamy na lewo mniejsze, a na prawo większe bądź równe od danego elementu. Potem wchodzimy do każdej z podzielonych części i powtarzamy aż nie zostanie tylko jeden element.

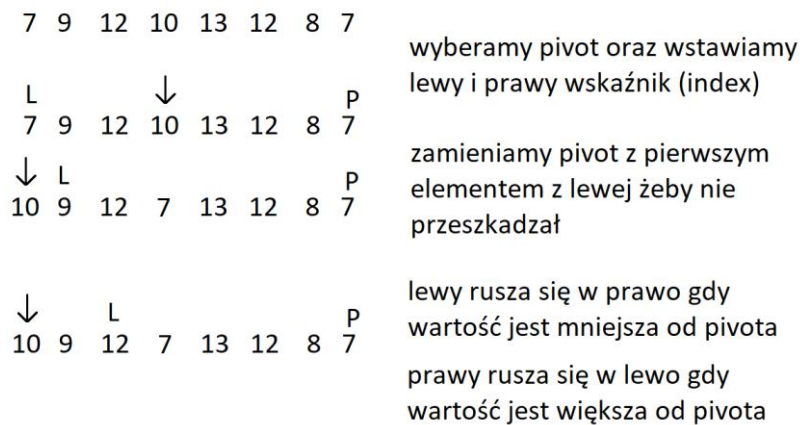
- Złożoność obliczeniowa: $O(n * \log n)$
- Złożoność pamięciowa: $O(1)$
- Zalety: szybki także w praktyce, sortuje w miejscu
- Wady: w pesymistycznych sytuacjach złożoność obliczeniowa $\Theta(n^2)$, niestabilny

Implementacja

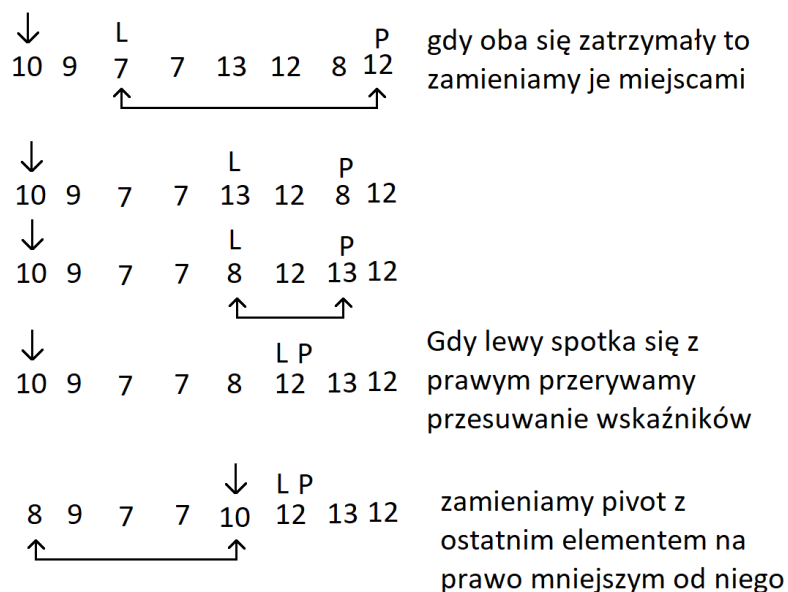
```
void quicksort(int tablica[], int p, int r)
{
    int q;
    if(p < r)
    {
        q = partition(tablica, p, r);
        //dzielimy tablice na dwie czesci, q oznacza punkt podzialu
        quicksort(tablica, p, q);
        //wywołujemy rekurencyjnie quicksort dla pierwszej czesci tablicy
        quicksort(tablica, q + 1, r);
        // wywołujemy rekurencyjnie quicksort dla drugiej czesci tablicy
    }
}
```

Rys. 10 implementacja sortowania szybkiego

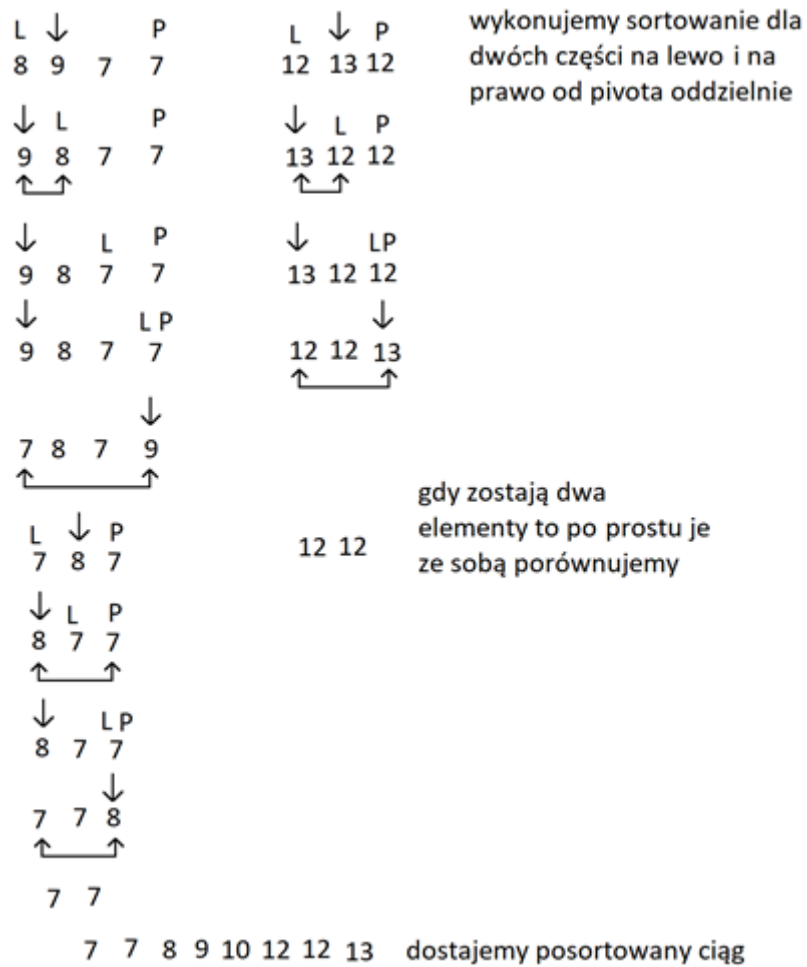
Rozrysowanie



Rys. 11 sortowanie szybki graficzne rozwiązanie cz1



Rys. 12 sortowanie szybki graficzne rozwiązanie cz2



Rys. 13 sortowanie szybkie graficzne rozwiązanie cz3

Podsumowanie

Tabela złożoności (szeregowanie algorytmów)

Nazwa	optymistyczna	typowa	pesymistyczna	stabilność	w miejscu	Pamięć
Bąbelkowe	$O(n^2)$	$O(n^2)$	$O(n^2)$	TAK	TAK	$O(1)$
Wstawianie	$O(n)$	$O(n^2)$	$O(n^2)$	TAK	TAK	$O(1)$
Wybieranie	$O(n^2)$	$O(n^2)$	$O(n^2)$	NIE	TAK	$O(1)$
Zliczanie	$O(k + n)$	$O(k + n)$	$O(k + n)$	TAK	NIE	$O(n + k)$
Scalanie	$O(n \cdot \log n)$	$O(n \cdot \log n)$	$O(n \cdot \log n)$	TAK	NIE	$O(n)$
Szybkie	$O(n \cdot \log n)$	$O(n \cdot \log n)$	$O(n^2)$	NIE	TAK	$O(1)$

Algorytmy stabilne

- bąbelkowe
- przez wstawianie
- przez scalanie
- przez zliczanie
- kubelkowe
- pozycyjne

Algorytmy niestabilne

- przez wybieranie
- Shella
- szybkie
- przez kopcowanie

Definicje

Sortowanie w miejscu

Algorytm nie potrzebuje dodatkowej pamięci na obliczenia. Ewentualnie niewielką stałą niezależną od ilości danych.

Sortowanie stabilne

Gdy wśród elementów do sortowania występują takie, które się powtarzają to po sortowaniu jest zachowana ich pierwotna kolejność. Czyli ta która wystąpiła jako pierwsza nadal będzie pierwsza.

Złożoności obliczeniowe

Każdy kolejny algorytm jest słabszy od poprzedniego. Należy czytać od góry do dołu.

- 1
- $\log \log n$
- $\sqrt{\log n}$
- $\frac{\log n}{\log \log n}$
- $\log n$
- $\log^2 n$
- $\log^3 n$
- $n^{0,01}$
- $n^{0,1}$
- $n^{0,1} \log n$
- $\frac{n}{\log n}$
- $\frac{n}{\log \log n}$
- n
- $n \log \log n$
- $n \log n$
- $n^{\frac{3}{2}}$
- $\frac{n^2}{\log n}$
- n^2
- n^3
- $n^{\log n}$
- 2^n
- $16^{\frac{n}{2}}$
- 10^n
- $n!$
- n^n
- $n^{n!}$

Stos

Stos jest interfejsem, w którym mamy określone funkcję. Możemy podglądać tylko element, który położyliśmy, jako pierwszy.

Operacje:

- `push()` – dokładamy nowy element do stosu.
- `pop()` – która pozwala na zdjęcie i odczytanie ostatnio położonego elementu.
- `isEmpty()` – sprawdza czy stos jest pusty
- `pick()` – pobranie wartości bez ściągania jej ze stosu. (opcjonalnie)
- `size()` – zwraca rozmiar stosu. (opcjonalnie)

Nie mylić stosu z listą czy kolejką. Stos może być zrealizowany na dowolnej strukturze danych (tablica, tablica dynamiczna, lista dwukierunkowa) i może być dla dowolnego typu (int, float char).

Odwrotna notacja polska na przykładach

1) $AB*CD*E/-$

$(A*B)-((C*D)/E)$

2) $6\ 5*3\ 7+8\ 4-2/*+$

Gdy natrafiamy na liczbę dokładamy ją do stosu, gdy na trafimy na działanie ściągamy dwie liczby i wykonujemy na nich działanie, a wynik wrzucamy na stos.

Lp.	wejście	stos
1	6	6
2	5	6 5
3	*	30
4	3	30 3
5	7	30 3 7
6	+	30 10
7	8	30 10 8
8	4	30 10 8 4
9	-	30 10 4
10	2	30 10 4 2
11	/	30 10 2
12	*	30 20
13	+	50

$$(6 * 5) + ((3 + 7) * ((8 - 4) / 2)) = 50$$

Kolejka

Kolejka podobnie jak stos nie jest strukturą danych tylko interfejsem. Z kolejki możemy, jeśli jest jednostronna to zabierać i odczytywać element, który został położony, jako pierwszy, a dokładamy elementy na sam koniec. W dwustronnej możemy zabierać elementy jeszcze z góry, ale nie ze środka.

Operacje:

- `Init(q)` – opróżnia kolejkę i przygotowuje pamięć
- `isEmpty(q)` – sprawdza czy kolejka jest pusta
- `Enqueue(q, x)` – dokłada element **x** na koniec kolejki **q**
- `Dequeue(q)` – zabiera element z początku kolejki

Haszowanie

Haszowanie liniowe i obsługa kolizji

Mamy tablicę 8-elementową:

13	0	7	6	8	10	12	5
----	---	---	---	---	----	----	---

Zakładamy że funkcja haszująca to $f(n) = (2n + 1) \% 10$. Zwraca ona wartości od 0 do 9, więc tworzymy 10-elementową tablicę haszującą. Zwrócona wartość to nr docelowego indeksu w tablicy.

$$f(13) = (2 \cdot 13 + 1) \% 10 = 27 \% 10 = 7$$

							13		
0	1	2	3	4	5	6	7	8	9

$$f(0) = (2 \cdot 0 + 1) \% 10 = 1 \% 10 = 1$$

	0						13		
0	1	2	3	4	5	6	7	8	9

$$f(7) = 15 \% 10 = 5$$

	0				7		13		
0	1	2	3	4	5	6	7	8	9

$$f(6) = 13 \% 10 = 3$$

	0		6		7		13		
0	1	2	3	4	5	6	7	8	9

$$f(8) = 17 \% 10 = 7$$

Występuje konflikt, więc przesuwamy się w prawo tak długo aż znajdziemy wolne miejsce.

	0		6		7		13	8	
0	1	2	3	4	5	6	7	8	9

$$f(10) = 21 \% 10 = 1$$

Występuje konflikt, więc przesuwamy się w prawo tak długo aż znajdziemy wolne miejsce.

	0	10	6		7		13	8	
0	1	2	3	4	5	6	7	8	9

$$f(12) = 25 \% 10 = 5$$

Występuje konflikt, więc przesuwamy się w prawo tak długo aż znajdziemy wolne miejsce.

	0	10	6		7	12	13	8	
0	1	2	3	4	5	6	7	8	9

$$f(5) = 11 \% 10 = 1$$

Występuje konflikt, więc przesuwamy się w prawo tak długo aż znajdziemy wolne miejsce.

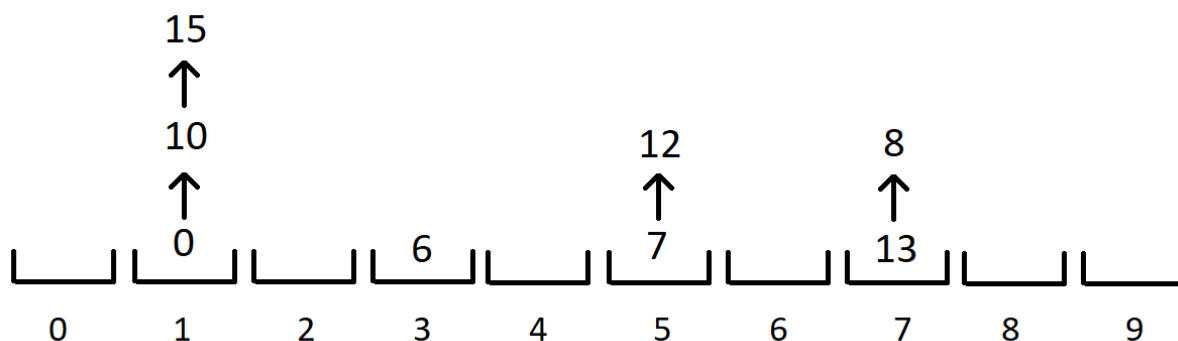
	0	10	6	5	7	12	13	8	
0	1	2	3	4	5	6	7	8	9

Haszowanie łańcuchowe

Dla ułatwienia i pokazania różnic użyjemy tablicy z poprzedniego zadania

13	0	7	6	8	10	12	15
7	1	5	3	7	1	5	1

Drugi rząd to już obliczone pozycje do tablicy za pomocą tego samego wzoru co w poprzednim przykładzie $f(n) = (2n + 1) \% 10$



Haszowanie łańcuchowe zamiast nie jest tablicą przechowującą dane wartości tylko tablicą list.

Te które były jako pierwsze są niżej na liście. Nie ma różnicy czy narysujemy to pionowo (tak jak tu) czy poziomo (jak było na prezentacji).

Haszowanie kwadratowe

Jeszcze raz rozważmy te same dane dla tej samej funkcji haszującej $f(n) = (2n + 1) \% 10$.

13	0	7	6	8	10	12	15
7	1	5	3	7	1	5	1

Rozważmy liczby 0, 10 oraz 15 ponieważ mają kolizje na pozycji 1.

Zastąpmy naszą liczbę 1 w funkcji przez i : $f(n) = (2n + i) \% 10$

Przy każdej kolizji używamy następującego wzoru $f(n) = (2n + i^2) \% 10$ poczym wykonujemy $++i$

Każda następna kolizja jest przesuwana o kwadrat danej liczby, a nie o jeden jak w liniowym

$$(2n + 1^2) \% 10$$

$$(2n + 2^2) \% 10$$

$$(2n + 3^2) \% 10$$

	0		6		7		13		
0	1	2	3	4	5	6	7	8	9

Pierwsza kolizja to 8 bo chcemy ją postawić na indeks 7 który jest zajęty przez liczbę 10.

$$(2 * 8 + 2^2) \% 10 = (16 + 4) \% 10 = 0$$

8	0		6		7		13		
0	1	2	3	4	5	6	7	8	9

Druga kolizja to 10.

$$(2 * 10 + 2^2) \% 10 = (20 + 4) \% 10 = 4$$

8	0		6	10	7		13		
0	1	2	3	4	5	6	7	8	9

Trzecia kolizja to 12.

$$(2 * 12 + 2^2) \% 10 = (24 + 4) \% 10 = 8$$

8	0		6	10	7		13	12	
0	1	2	3	4	5	6	7	8	9

Czwarta kolizja to 15.

$$(2 * 15 + 2^2) \% 10 = (30 + 4) \% 10 = 4$$

Tu następuje kolejna kolizja więc dodajemy do i jeszcze 1.

$$(2 * 15 + 3^2) \% 10 = (30 + 9) \% 10 = 9$$

8	0		6	10	7		13	12	15
0	1	2	3	4	5	6	7	8	9

Pojęcia związane z haszowaniem

- $O(1)$ wyszukiwanie w optymistycznym przypadku
- $O(n/m + 1)$ wyszukiwanie w średnim przypadku
- n – ilość elementów w ciągu m – długość haszowanej tablicy
- $O(n)$ wyszukiwanie w pesymistycznym przypadku
- $\alpha = \frac{n}{m}$ współczynnik przeładowania (load factor)
- $\frac{1}{2}(1 + 1/(1 - \alpha)^2)$ oczekiwana liczba prób do znalezienia oczekiwanej wartości

Drzewo

- Drzewo (wolne): graf spójny acykliczny nieskierowany. Jeśli graf jest nieskierowany i acykliczny, ale niekoniecznie spójny, to nazywamy go lasem.
- Drzewo z korzeniem (drzewo ukorzenione): drzewo wolne z jednym wyróżnionym wierzchołkiem, tzw. korzeniem (oznaczamy go zwykle przez r).

- Drzewo uporządkowane: drzewo ukorzenione, w którym dzieci każdego wężła są uporządkowane.
- Regularne drzewo binarne: każdy węzeł ma 0 lub 2 dzieci.
- Pełne drzewo binarne: wszystkie liście są na tej samej głębokości h (dla pewnego h), a wszystkie węzły wewnętrzne mają stopień 2. (Innymi słowy, pełne drzewo binarne jest regularne.)

Przeszukiwanie

INORDER	PREORDER	POSTORDER
<ul style="list-style-type: none"> • Lewe dziecko • Węzeł • Prawe dziecko 	<ul style="list-style-type: none"> • Węzeł • Lewe dziecko • Prawe dziecko 	<ul style="list-style-type: none"> • Lewe dziecko • Prawe dziecko • Węzeł

Podstawowe pojęcia

- **Liść** - Węzeł drzewa, z którego nie wychodzi żadna krawędź. Nie ma dzieci
- **Korzeń** - Węzeł początkowy drzewa znajdujący się, jako jedyny na wysokości 0 i nie ma rodzica
- **Bracia** – Węzły wychodzące z tego samego rodzica
- **Wysokość** - Odległość między korzeniem, a najodleglejszym liściem. Do wysokości drzewa nie wlicza się korzenia (jest na głębokości 0).
- **Głębokość** - Odległość między korzeniem, a konkretnym wierzchołkiem
- **Drzewo binarne** - Drzewo, w którym stopień każdego wierzchołka jest nie większy od 3.
- **Drzewo BST (przeszukiwań binarnych)** - To takie drzewo, w którym prawe dzieci są większe od rodzica, a lewe mniejsze od niego.

Grafy

Graf skierowany (zorientowane)

To para $G = (V, E)$ zbiorów skończonych gdzie V to zbiór węzłów (wierzchołków), a E to zbiór uporządkowanych par wierzchołków (krawędzi, łuków, pętli).

Graf nieskierowany (niezorientowane)

To para $G = (V, E)$ zbiorów skończonych gdzie V to zbiór węzłów (wierzchołków), a E to zbiór dwuelementowych zbiorów wierzchołków (krawędzi, łuków, pętli).

Macierze sąsiedztwa

Dwuwymiarowa (V^2), jest jedna dla całego grafu. Wypełniona 0 (brak połączenia) lub 1 (połączenie).

- Dla grafów skierowanych: diagonalna nie musi być z samych 0.
- Dla grafów nieskierowanych: diagonalna jest z samych 0, oraz jest symetryczna względem diagonalnej.

Diagonalna to ta przekątna od góry po lewo do dołu po prawo.

Listy sąsiedztwa

Jednowymiarowa (V), przechowuje informację o tym, do których wierzchołków prowadzi droga z danego wierzchołka. Jest listą wskaźników.

Macierze incydencji

Dwuwymiarowa ($V \times E$) prostokątna, niesymetryczna i przyjmuje wartości -1, 0, 1 oraz 2.

-1 oznacza krawędź w grafie skierowanym w przeciwnym kierunku do wierzchołka (wychodzi z niego), 0 to brak połączenia, 1 to krawędź wchodząca do danego wierzchołka, a 2 to pętla.

Przeszukiwanie BFS (w szerz)

- wrzucamy wierzchołek startowy do kolejki
- sprawdzamy wszystkich sąsiadów i wrzucamy do kolejki każdego, który nie został jeszcze odwiedzony.
- po dorzuceniu wierzchołków wierzchołka na dole kolejki wyrzucamy go

Lub

- Na początku wszystkie wierzchołki są białe.
- Wierzchołki szare są przechowywane w kolejce FIFO.
- Po jej opuszczeniu kolorujemy je na czarno.

Przeszukiwanie DFS (w głąb)

- badamy wszystkie krawędzie ostatnio odwiedzonego wierzchołka v
- gdy wszystkie krawędzie v są zbadane przechodzimy wierzchołka v , z którego przeszliśmy do obecnego
- kontynuujemy do momentu, gdy wszystkie wierzchołki osiągalne z wierzchołka startowego zostaną odwiedzone

Złożoność pamięciowa grafu

Złożoność pamięciowa grafu jest $\leq V^2$

- **Graf spójny** - gdy do każdego wierzchołka prowadzi krawędź i da się dojść od niego do każdego innego. Graf spójny może być słabo lub silnie spójny.
- **Graf słabo spójny** - gdy do każdego wierzchołka prowadzi krawędź i da się dojść od niego do każdego innego z tym, że gdy graf jest skierowany to by trzeba było iść pod prąd.
- **Graf silnie spójny** - gdy zawsze da się dojść z każdego do dowolnego innego wierzchołka uwzględniając kierunki dróg.

Kompresja danych

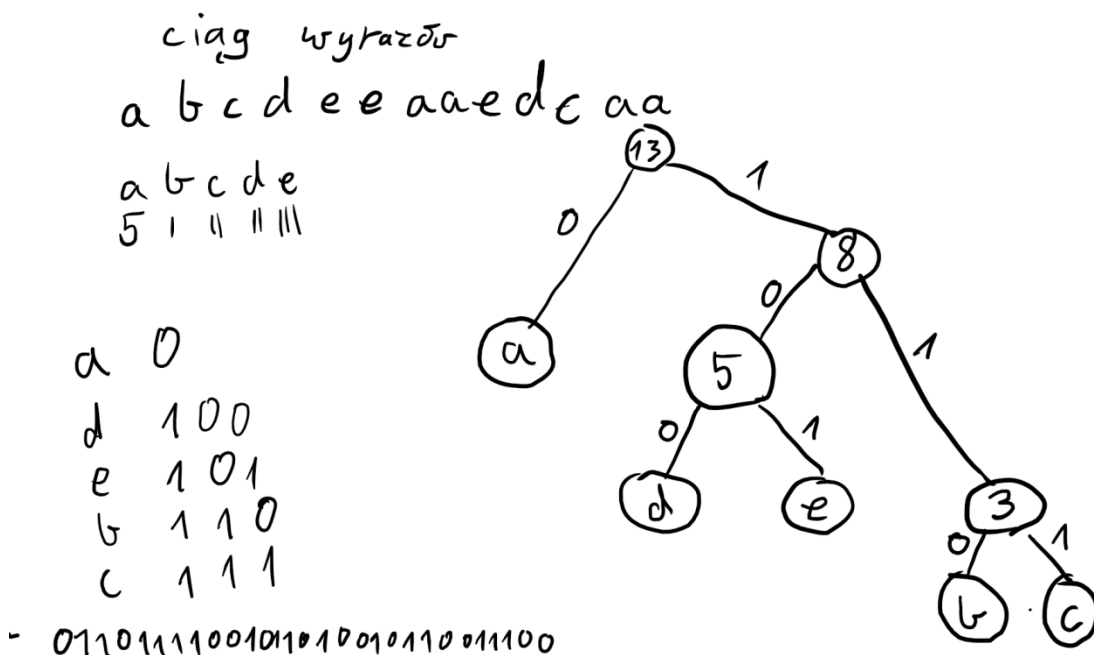
Podstawowe dane

KOMPRESJA BEZSTRATNA (dane są w pełni zgodne z początkowymi)	KOMPRESJA STRATNA (część informacji jest tracona)
<ul style="list-style-type: none">• Tekst• Programy• Bazy danych	<ul style="list-style-type: none">• Muzyka• Filmy• Zdjęcia

Algorytm Huffmana

Jest to algorytm kompresji danych, który polega na jak największym skróceniu danych które najczęściej występują.

W dowolnym ciągu liczymy ilość wystąpienia poszczególnych znaków i budujemy na ich podstawie drzewo tak by na samym dole znajdowały się najrzadziej występujące, a im wyżej tym częściej występujące.



Typowa litera zajmuje pamięć 8 bitów czyli 1 bajt. Zauważmy, że tu poszczególne litery zajmują od 1 do 3 bitów, a kod jest jednoznacznie odczytywalny, np. jeśli litera zaczyna się od 0 to od razu wiemy, że to **a**, zaś gdy od 11 to będzie to **b** lub **c**.

Algorytm bardzo dobrze działa dla danych o **małej ilości różnych znaków** w tym przypadku było tylko 5 różnych liter. Do algorytmu jest potrzebna dodatkowa pamięć przetrzymująca drzewo, oprócz samej sekwencji.

Średnia długość słowa

ciąg wyrazów
 $T[n] = \{a, b, c, d, e, e, a, a, e, d, c, a, a\}$

a b c d e
5 1 1 1 1

C_j

a 0 $\rightarrow 1$
d 1 0 0 $\rightarrow 3$
e 1 0 1 $\rightarrow 3$
b 1 1 0 $\rightarrow 3$
c 1 1 1 $\rightarrow 3$

$p(a) = \frac{5}{13}$
 $p(b) = \frac{1}{13}$
 $p(c) = \frac{2}{13}$
 $p(d) = \frac{2}{13}$
 $p(e) = \frac{3}{13}$

Średnia dł. słowa
 $\sum_{i=1}^n p(i) |C_i| =$
 $= \frac{5}{13} \cdot 1 + \frac{1}{13} \cdot 3 + \frac{2}{13} \cdot 3 + \frac{2}{13} \cdot 3 + \frac{3}{13} \cdot 3 =$
 $= \frac{1}{13} (5 + 3 + 6 + 6 + 9) = \underline{\underline{2.3}}$

- 01101111001011010010110011100

Entropia

entropia

a b a a b c a d

$n=8$

a b c d
4 2 1 1

$$\begin{aligned} P(a) &= \frac{4}{8} \\ P(b) &= \frac{2}{8} \\ P(c) &= \frac{1}{8} \\ P(d) &= \frac{1}{8} \end{aligned}$$

$$H(S) = - \sum_{i=0}^{s-1} P(i) \cdot \log_2 P(i) = \sum_{i=0}^{s-1} P(i) \cdot \log_2 \left(\frac{1}{P(i)} \right) =$$

$$= \frac{4}{8} \cdot \log_2 \frac{1}{\frac{4}{8}} + \frac{2}{8} \cdot \log_2 \frac{1}{\frac{2}{8}} + \frac{1}{8} \cdot \log_2 \frac{1}{\frac{1}{8}} + \frac{1}{8} \cdot \log_2 \frac{1}{\frac{1}{8}} = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \left(\frac{1}{8} \cdot 3 \right) \cdot 2 =$$

$$= 1 \frac{3}{4}$$

Entropia jest miarą rozrzutu danych. Określa ona czy warto kompresować badane dane. Im mniejsza entropia, tym lepiej dane powinny się skompresować.

Redundancja

Redundancja

$$\underbrace{\sum_{i=1}^n P(i) |C(i)|}_{\text{średnia dł. słowa}} - \underbrace{\sum_{i=1}^n P(i) \cdot \log_2 \frac{1}{P(i)}}_{\text{entropia}} \geq 0$$

Redundancja to średnia nadwyżka ponad entropię. Jest miarą przeciwną do entropii jest to wartość tego jak dużo jest nadmiarowych danych. Im wyższa redundancja, tym lepiej dane powinny się skompresować.

Kompresja RLE

aaaabbaabcaaaa

RLE z flagą
#ba b#3a bc #4a

- specjalny dowolny znak
po # piszemy ilość
wystąpienia słowa, a następnie
to słowo. np. #6a

ilość - wystąpienia słowa

RLE bez flagi
aa h baa 1 bca a2

jeżeli litera się powtarza
2 razy, to cyfra za nią
oznacza ile jeszcze
takich samych znaków
powinno wystąpić.

może przyjmować wartości
od 0 do n
np. aabba => aa 0 ba

Run-length encoding to algorytm kompresji bezstratnej gdzie znaki zapisujemy zliczając ich ilość i pisząc np. 10g zamiast gggggggggg.

W praktyce mało przydatny. Dobry dla danych, w których jest dużo powtarzających się **znaków następujących po sobie**.

LZ77

Algorytm kompresji, w którym wykorzystujemy powtarzający się ciąg, i zamiast go pisać drugi raz wskazujemy gdzie on wystąpił i jakiej jest długości.

Przykład

- Wczoraj wieczorem Hamming grał w golfa z **Hamming**way'em
- Wczoraj wieczorem Hamming grał w golfa z <24, 7>way'em


Oznacza to, że 24 znaki wcześniej jest siedmioliterowe powtórzenie.

Dobry dla długich ciągów znaków, które się powtarzają. Przydatny także w praktyce.

Stringologia:

Długość Hamminga

odległość Hamminga

1^o $\begin{matrix} 00000000 \\ 00000001 \end{matrix} = 1$ 

2^o $\begin{matrix} 10101011 \\ 11001011 \end{matrix} = 5$ - bity, które się różnią

3^o $\begin{matrix} 110011101 \\ 000111010 \end{matrix} = 6$ ilość bitów, które się różnią w słowach tej samej długości.

Bit parzystości

By sprawdzić czy dane są poprawne często dodaje się do nich sztuczne dane które pokazują, że Informacje są zgodne z tym co się przesłało.

Na początku lub na końcu dodaje się jeden bit którego wartość oznacza ilość jedynek w całym pozostałym ciągu.

0 to nieparzysta ilość jedynek, 1 to parzysta ilość jedynek.

Kontrola parzystości

kontrola parzystości wg. kodu Hamming

Pozycja bitu	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Bit parzystości (p), danych (d)	p1	p2	d1	p4	d2	d3	d4	p8	d5	d6	d7	d8	d9	d10	d11	p16	d12	d13	d14	d15
Sekwencja sprawdzanych bitów	p1	x	x		x	x		x		x		x		x		x		x		
	p2		x	x			x	x			x	x			x	x			x	x
	p4				x	x	x	x				x	x	x	x					x
	p8							x	x	x	x	x	x	x	x					
	p16															x	x	x	x	x

11101111010001

p1 1
p2 1
p4 1
p8 1

10110110110001

p1 1
p2 0
p4 1
p8 0

Gdy wysyłamy dane przez sieć i nie chcemy by nam jakieś dane się zgubiły lub co gorsze przekłamały.

Do policzenia ilości bitów potrzebnych do korekcji bierzemy ilość wartości, które chcemy wysłać.
Liczbę 7 można zapisać jako $2^r - 1 = 7$ gdy $r = 3$, bo $2^3 - 1 = 8 - 1 = 7$

$$7 - r = 3$$

Więc potrzeba 3 bitów, by można było skorygować.

Liczbę 9 można zapisać jako $2^r - 1 = 9$ gdy $r = 5$, bo $2^5 - 1 = 16 - 1 = 15$

$$9 - r = 4$$

Więc potrzeba 4 bitów, by można było skorygować. W tym przypadku trzeba dać trochę nadmiaru.

LCS – Longest common subsequence

Najdłuższa wspólna podsekwencja. Algorytm jest stosowany np.: w repozytoriach do sprawdzania różnic między poszczególnymi wersjami plików.

LCS

		-1	0	1	2	3	4	5
		G	R	A	T	K	A	
-1		0	0	0	0	0	0	0
0	K	0	0	0	0	1	1	
1	R	0	0	1	1	1	1	
2	A	0	0	1	2	2	2	
3	T	0	0	1	2	3	3	
4	K	0	0	1	2	3	4	
5	I	0	0	1	2	3	4	4

if $y[j] == x[i]$

then $T[i][j] = T[i-1][j-1] + 1$

else

$\text{MAX}(T[i-1][j], T[i][j-1])$

wynik - ilość wspólnych znaków

or odpowiedniej kolejności

nie muszą być obok siebie!!

		-1	0	1	2	3	4	5
		M	A	L	A	R	Z	
-1		0	0	0	0	0	0	0
0	M	0	1	1	1	1	1	
1	V	0	1	1	1	1	1	
2	R	0	1	1	1	2	2	
3	A	0	1	2	2	2	2	
4	R	0	1	2	2	2	3	
5	Z	0	1	2	2	2	3	4

		A	M	L	A	R	Z	
		0	0	0	0	0	0	
M		0	0	1	1	1	1	
V		0	0	1	1	1	1	
R		0	0	1	1	2	2	
A		0	1	1	1	2	2	
R		0	1	1	1	2	3	
Z		0	1	1	1	2	3	4

wszystkie możliwości ciągów po 2 kolejne znaki

1° 2° 3° 4°

(niekoniecznie obok siebie)

		A	L	A	Z	R	M	
		0	0	0	0	0	0	
M		0	0	0	0	0	0	1
V		0	0	0	0	0	0	1
R		0	0	0	0	0	1	1
A		0	1	1	1	1	1	1
R		0	1	1	1	1	2	2
Z		0	1	1	1	2	2	3

		A	L	A	R	Z	M	
		0	0	0	0	0	0	
M		0	0	0	0	0	0	1
V		0	0	0	0	0	0	1
R		0	0	0	1	1	1	
A		0	1	1	1	1	1	
R		0	1	1	1	2	2	
Z		0	1	1	1	2	3	4

Odległość Levenshteina (edycyjna)

odległość LEVENSHTEINA

if ($x[i] == y[j]$)
 then
 $T[i][j] = T[i-1][j-1]$
 else
 $T[i][j] =$
 $= \min(T[i][j-1],$
 $T[i-1][j],$
 $T[i-1][j-1]) + 1$

y

	L	E	V	E	N	S	H	T	E	I	M	A	T
L	0	1	2	3	4	5	6	7	8	9	10	11	12
E	1	0	1	2	3	4	5	6	7	8	9	10	11
V	2	1	1	2	3	4	5	6	7	8	9	10	11
E	3	2	1	2	2	3	4	5	6	7	8	9	10
N	4	3	2	2	3	3	4	5	6	7	8	9	10
S	5	4	3	3	3	4	4	5	6	7	8	9	10
H	6	5	4	4	4	5	5	6	7	8	9	10	11
T	7	6	5	5	5	6	6	7	8	9	10	11	12
E	8	7	6	6	6	7	7	8	9	10	11	12	13
I	9	8	7	7	7	8	8	9	10	11	12	13	14
M	10	9	8	8	8	9	9	10	11	12	13	14	15
A	11	10	9	9	9	10	10	11	12	13	14	15	16
T	12	11	10	10	10	11	11	12	13	14	15	16	17

Mohrasiński

	L	E	E	D
L	0	1	2	3
E	1	0	1	2
E	2	1	0	1
D	3	2	1	0

	L	E	D
L	0	1	2
E	1	0	1
A	2	1	0
D	3	2	1

	T	A	K
N	0	1	2
T	1	1	2
E	2	2	3
E	3	3	3

Służy do wyznaczania różnic między poszczególnymi wyrazami

Mówi nam o tym ile minimalnie znaków musimy **usunąć**, **przestawić** lub **zastąpić**, aby jeden wyraz zamienić w drugi.

Algorytm ten może służyć jako podpowiedzi w słownikach np.: gdy zrobimy jakąś literówkę zjemy jakąś jedną literę możemy sobie obliczyć do jakich wyrazów ze słownika jest podobny ten, w którym się pomyliliśmy.