

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Delo z nizi

Matevž Jekovec

Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Programiranje v višji prestavi
29. junij 2017

Delo z nizi

Matevž
Jekovec

Uvod

▶ T — vhodni niz znakov oz. besedilo (ang. *text*)

Hramba
besedila

▶ n — dolžina niza (število znakov)

Iskanje vzorca
v besedilu

▶ Σ — abeceda

Podobnost
med nizi

▶ σ — velikost abecede

Vaje in naloge

¹Pozor: Dejanska velikost niza v pomnilniku je lahko drugačna (*null-delimited* nizi, različna kodiranje znakov)

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

- ▶ T — vhodni niz znakov oz. besedilo (ang. *text*)
- ▶ n — dolžina niza (število znakov)
- ▶ Σ — abeceda
- ▶ σ — velikost abecede
- ▶ Primer:
 - ▶ $T = \text{"BANANA"}$
 - ▶ $n^1 = 6$
 - ▶ $\Sigma = \{B, A, N\}$
 - ▶ $\sigma = 3$

¹Pozor: Dejanska velikost niza v pomnilniku je lahko drugačna (*null-delimited* nizi, različna kodiranje znakov)

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Oblike besedil:

- ▶ **Strukturirana:** Besedilo je sestavljeno iz več besed. Običajno tudi iščemo po besedah. (programska koda, angleščina, slovenščina)
- ▶ **Nestrukturirana:** Besedilo nima jasnih besed ali pa je veliko besed sestavljenk. (nemščina, človeški genom, glasba)

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Oblike besedil:

- ▶ Strukturirana: Besedilo je sestavljeno iz več besed. Običajno tudi iščemo po besedah. (programska koda, angleščina, slovenščina)
- ▶ Nestructurirana: Besedilo nima jasnih besed ali pa je veliko besed sestavljenk. (nemščina, človeški genom, glasba)

Problemi pri besedilih:

- ▶ Kako učinkovito shraniti besedilo?
- ▶ Kako učinkovito iskati po njem:
 - ▶ Ali se vzorec P (ang. *pattern*) nahaja v besedilu?
 - ▶ Kje in kolikokrat se vzorec ponovi?
 - ▶ Pri strukturiranih: Ali se katera koli beseda začne na P ? Ali je P koren katere koli besede? Ali je P pripona katere koli besede?

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Kako shraniti strukturirano besedilo tako, da bomo lahko hitro iskali po njem?

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Kako shraniti strukturirano besedilo tako, da bomo lahko hitro iskali po njem?

Ideja: Slovar, ki preslika iskan ključ na mesta pojavitev ključa v besedilu. Kaj točno uporabiti?

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Kako shraniti strukturirano besedilo tako, da bomo lahko hitro iskali po njem?

Ideja: Slovar, ki preslika iskan ključ na mesta pojavitev ključa v besedilu. Kaj točno uporabiti?

- ▶ Dvojiško iskalno drevo

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Kako shraniti strukturirano besedilo tako, da bomo lahko hitro iskali po njem?

Ideja: Slovar, ki preslika iskan ključ na mesta pojavitev ključa v besedilu. Kaj točno uporabiti?

- ▶ Dvojiško iskalno drevo
- ▶ Zgoščevalna tabela

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Kako shraniti strukturirano besedilo tako, da bomo lahko hitro iskali po njem?

Ideja: Slovar, ki preslika iskan ključ na mesta pojavitev ključa v besedilu. Kaj točno uporabiti?

- ▶ Dvojiško iskalno drevo
- ▶ Zgoščevalna tabela
- ▶ Številsko drevo z variantami

Delo z nizi

**Matevž
Jekovec**

Uvod

**Hramba
besedila**

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Ideja: Iskalno drevo, vozlišča hranijo nize, manjši niz je tisti, ki je po abecedi (*leksikografsko*) pred drugim.

Delo z nizi

Matevž
Jekovec

Uvod

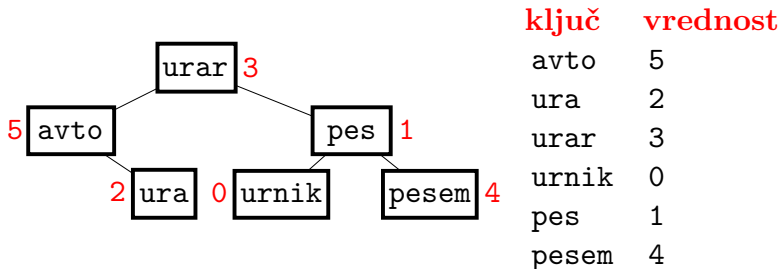
Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Ideja: Iskalno drevo, vozlišča hranijo nize, manjši niz je tisti, ki je po abecedi (*leksikografsko*) pred drugim.



Delo z nizi

Matevž
Jekovec

Uvod

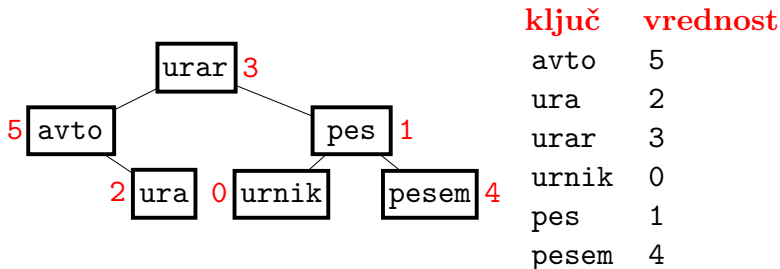
Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Ideja: Iskalno drevo, vozlišča hranijo nize, manjši niz je tisti, ki je po abecedi (*leksikografsko*) pred drugim.



Pozor: Če imamo nize že zložene v tabelo po abecedi, potem namesto iskanja po drevesu iščemo kar po tabeli — bisekcija!

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Zgoščeno vrednost izračunamo za vsako besedo v vhodnem besedilu. Nato vstavimo besedo kot običajni element (npr. število) v zgoščevalno tabelo.

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Zgoščeno vrednost izračunamo za vsako besedo v vhodnem besedilu. Nato vstavimo besedo kot običajni element (npr. število) v zgoščevalno tabelo.

Rabin-Karpova zgoščevalna funkcija:

$$H = c_1 a^{k-1} + c_2 a^{k-2} + \dots + c_k a^0$$

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Zgoščeno vrednost izračunamo za vsako besedo v vhodnem besedilu. Nato vstavimo besedo kot običajni element (npr. število) v zgoščevalno tabelo.

Rabin-Karpova zgoščevalna funkcija:

$$H = c_1 a^{k-1} + c_2 a^{k-2} + \dots + c_k a^0$$

Primer zgoščene vrednosti za $a = 2$: URAR, dolžina je $k = 4$
U je 20. črka, R je 17. črka, A pa 0. črka po abecedi:

$$20 \cdot 2^3 + 17 \cdot 2^2 + 0 \cdot 2^1 + 17 \cdot 2^0 = 160 + 68 + 17 = 245$$

Pozor: Če želimo hraniti več pojavitev niza, potrebujemo *multimap*.

Delo z nizi

Matevž
Jekovec

Uvod

**Hramba
besedila**

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Vzemimo velikost polja $M = 8$.

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Vzemimo velikost polja $M = 8$.

$$H(\text{"avto"}) = 136 \quad \text{mod } 8 = 0$$

$$H(\text{"ura"}) = 114 \quad \text{mod } 8 = 2$$

$$H(\text{"urar"}) = 245 \quad \text{mod } 8 = 5$$

$$H(\text{"urnik"}) = 534 \quad \text{mod } 8 = 6$$

$$H(\text{"pes"}) = 86 \quad \text{mod } 8 = 0$$

$$H(\text{"pesem"}) = 364 \quad \text{mod } 8 = 0$$

Zgoščevalna tabela nadal. (ang. Hash table)

Delo z nizi

Matevž
Jekovec

Uvod

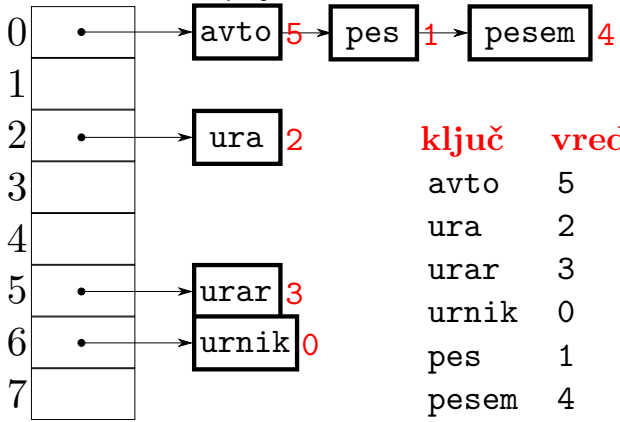
Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Vzemimo velikost polja $M = 8$.



ključ	vrednost
avto	5
ura	2
urar	3
urnik	0
pes	1
pesem	4

Delo z nizi

Matevž
Jekovec

Problem iskalnih dreves in zgoščevalne tabele: Ni enostavno iskati po začetkih besed (npr. *search-as-you-type*).

Uvod

**Hramba
besedila**

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Številsko drevo (ang. *trie*)

Delo z nizi

Matevž
Jekovec

Uvod

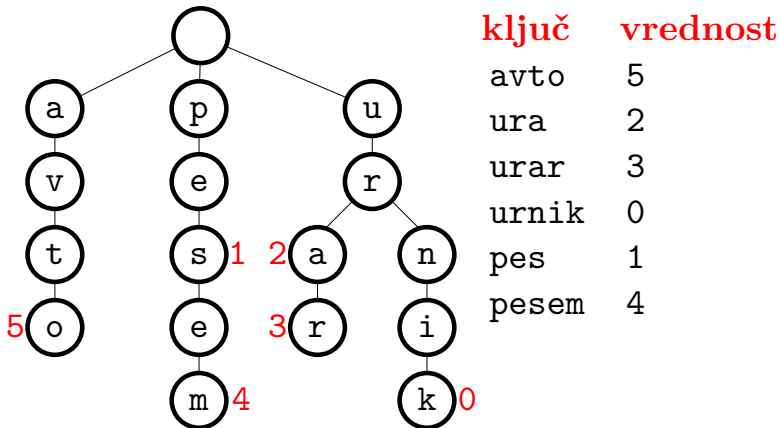
Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Problem iskalnih dreves in zgoščevalne tabele: Ni enostavno iskati po začetkih besed (npr. *search-as-you-type*).
Rešitev: Številsko drevo hrani besede od korena navzdol, razbite po črkah.



Delo z nizi

Matevž
Jekovec

Izboljšava: Namesto svojega vozlišča za vsako črko, vozlišča z enim naslednikom združimo.

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Delo z nizi

Matevž
Jekovec

Uvod

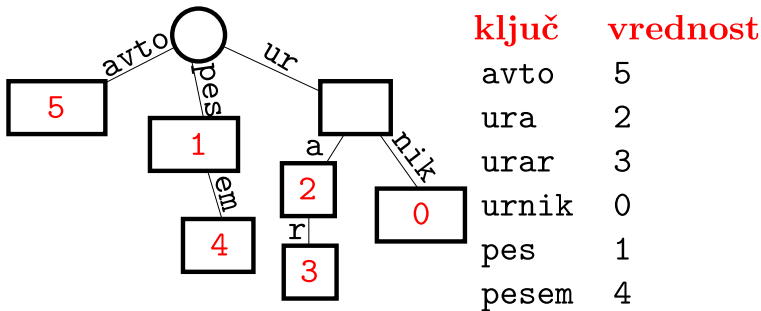
Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Izboljšava: Namesto svojega vozlišča za vsako črko, vozlišča z enim naslednikom združimo.



Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Kako predstaviti vozlišče?

```
1 class TrieNode {  
2     string label;  
3     void *value;  
4     ?     children;  
5 };
```

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Kako predstaviti vozlišče?

```
1 class TrieNode {  
2     string label;  
3     void *value;  
4     ?     children;  
5 };
```

Do otrok dostopamo prek “nekega” slovarja, npr.:

- ▶ povezane seznam,

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Kako predstaviti vozlišče?

```
1 class TrieNode {  
2     string label;  
3     void *value;  
4     ?     children;  
5 };
```

Do otrok dostopamo prek “nekega” slovarja, npr.:

- ▶ povezane seznam,
- ▶ polje kazalcev v velikosti abecede σ ,

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Kako predstaviti vozlišče?

```
1 class TrieNode {  
2     string label;  
3     void *value;  
4     ?     children;  
5 };
```

Do otrok dostopamo prek “nekega” slovarja, npr.:

- ▶ povezane seznam,
- ▶ polje kazalcev v velikosti abecede σ ,
- ▶ *trojiško* iskalno drevo.

Delo z nizi

Vozlišče ima kazalec na prvega otroka in na sorojenca.

Matevž
Jekovec

```
1 class TrieNode {
2     string    str;
3     void      *value;
4     TrieNode *child;
5     TrieNode *sibling;
6 };
```

Uvod

Hramba
besedilaIskanje vzorca
v besediluPodobnost
med nizi

Vaje in naloge

Delo z nizi

Vozlišče ima kazalec na prvega otroka in na sorojenca.

Matevž
Jekovec

```
1 class TrieNode {  
2     string    str;  
3     void      *value;  
4     TrieNode *child;  
5     TrieNode *sibling;  
6 };
```

Uvod

Hramba
besedila

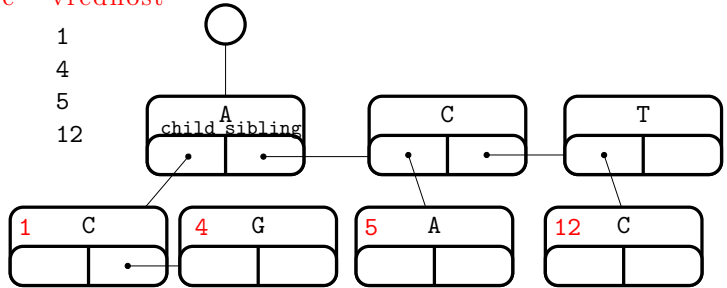
Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

ključ vrednost

AC	1
AG	4
CA	5
TC	12



Delo z nizi

Matevž
Jekovec

Vsako vozlišče vsebuje polje σ kazalcev na otroke. Do otroka dostopamo neposredno z indeksom njegove črke.

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

```
1 class TrieNode {
2     string    str;
3     void      *value;
4     TrieNode children[ALPHABET];
5 };
```

Delo z nizi

Matevž
Jekovec

Vsako vozlišče vsebuje polje σ kazalcev na otroke. Do otroka dostopamo neposredno z indeksom njegove črke.

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

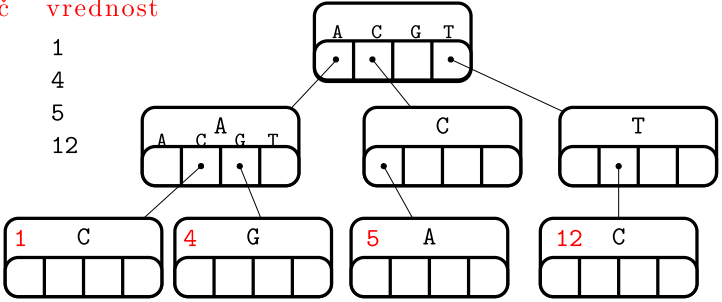
Podobnost
med nizi

Vaje in naloge

```
1 class TrieNode {  
2     string    str;  
3     void      *value;  
4     TrieNode children[ALPHABET];  
5 };
```

ključ vrednost

AC	1
AG	4
CA	5
TC	12



Delo z nizi

Matevž
Jekovec

Trojiško iskalno drevo (ang. *Ternary Search Trie*) je kompromis obeh prejšnjih metod: hranimo le kazalce na otroke, ki obstajajo, iskanje pa poteka z bisekcijo.

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

```

1 class TrieNode {
2     string    str;
3     void     *value;
4     TrieNode *eq;
5     TrieNode *ls;
6     TrieNode *gt;
7 };
  
```

Delo z nizi

Matevž
Jekovec

Trojiško iskalno drevo (ang. *Ternary Search Trie*) je kompromis obeh prejšnjih metod: hranimo le kazalce na otroke, ki obstajajo, iskanje pa poteka z bisekcijo.

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

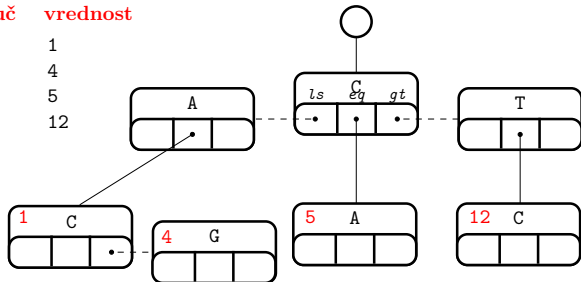
Vaje in naloge

```

1 class TrieNode {
2     string    str;
3     void      *value;
4     TrieNode *eq;
5     TrieNode *ls;
6     TrieNode *gt;
7 };
    
```

ključ **vrednost**

AC	1
AG	4
CA	5
TC	12



Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Kako iščemo?

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Kako iščemo? Odvisno od scenarija!

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Kako iščemo? Odvisno od scenarija!

- ▶ Vzorec ves čas isti, besedilo se menja.
- ▶ Besedilo ves čas isto, vzorci se menjajo.

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Kako iščemo? Odvisno od scenarija!

- ▶ Vzorec ves čas isti, besedilo se menja.
- ▶ Besedilo ves čas isto, vzorci se menjajo.

V prvem primeru porabimo nekaj časa, da spravimo vzorec v obliko, s katero bi hitreje iskali po besedilu.

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Kako iščemo? Odvisno od scenarija!

- ▶ Vzorec ves čas isti, besedilo se menja.
- ▶ Besedilo ves čas isto, vzorci se menjajo.

V prvem primeru porabimo nekaj časa, da spravimo vzorec v obliko, s katero bi hitreje iskali po besedilu.

V drugem primeru najprej poindeksiramo besedilo, kar traja nekaj časa, vendar bo iskanje po kazalu (in ne neposredno po besedilu) potem bistveno hitreje.

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

T: NA DREVESU RASTEJO BANANE.

P: BANANA

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

T: NA DREVESU RASTEJO BANANE.

P: BANANA

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

T: NA DREVESU RASTEJO BANANE.

P: BANANA

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

T: NA DREVESU RASTEJO BANANE.

P: BANANA

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

T: NA DREVESU RASTEJO BANANE.

P: BANANA

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

T: NA DREVESU RASTEJO BANANE.

P: BANANA

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

T: NA DREVESU RASTEJO BANANE.

P: BANANA

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

T: NA DREVESU RASTEJO BANANE.

P: BANANA

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

T:NA DREVESU RASTEJO BANANE.
P: . . .

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

T:NA DREVESU RASTEJO BANANE.

P: BANANA

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

T:NA DREVESU RASTEJO BANANE.

P:

BANANA

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

T:NA DREVESU RASTEJO BANANE.

P: BANANA

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

T:NA DREVESU RASTEJO BANANE.

P: BANANA

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

T:NA DREVESU RASTEJO BANANE.

P: BANANA

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

T: NA DREVESU RASTEJO **BANANE**.

P: **BANANA**

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

T: NA DREVESU RASTEJO **BANANE**.

P: **BANANA**

Vaja: Koliko časa potrebujemo?

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

T: NA DREVESU RASTEJO **BANANE.**

P: **BANANA**

Vaja: Koliko časa potrebujemo?

Vsaj n primerjav znakov, če se vedno že prvi znak okna ne ujema in $O(nm)$, če je vsak m -ti znak besedila zgrešitev. npr.

$T = \text{AAABAAABAAABAAAB}$ in $P = \text{AAAA}$.

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Za podan vzorec zgradimo *deterministični končni avtomat*, v katerega bomo usmerili besedilo. Če se avtomat ustavi, smo na tem mestu besedila našli vzorec.

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

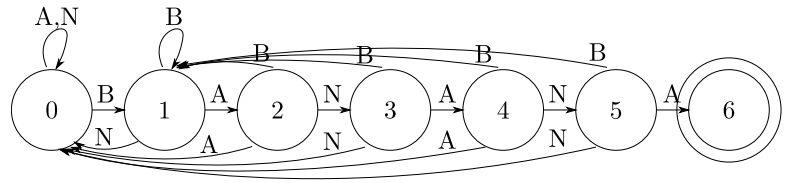
Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Za podan vzorec zgradimo *deterministični končni avtomat*, v katerega bomo usmerili besedilo. Če se avtomat ustavi, smo na tem mestu besedila našli vzorec.

Primer avtomata za vzorec BANANA:



Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

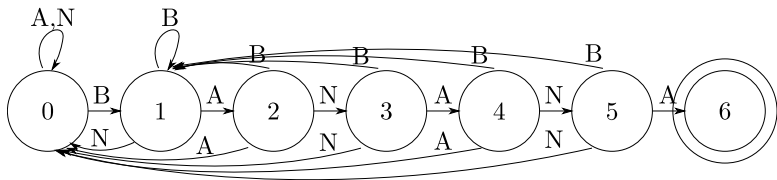
Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Za podan vzorec zgradimo *deterministični končni avtomat*, v katerega bomo usmerili besedilo. Če se avtomat ustavi, smo na tem mestu besedila našli vzorec.

Primer avtomata za vzorec BANANA:



Časovna zahtevnost iskanja z avtomatom je $O(n)$.

Delo z nizi

Matevž
Jekovec

Algoritem KMP:

Uvod

Hramba besedila

Iskanje vzorca v besedilu

Podobnost med nizi

Vaje in naloge

```
1 // input: pattern P of length m from alphabet SIGMA
2 // output: deterministic finite automaton as 2-d
   array dfa of dimensions SIGMA*m
3
4 dfa[P[0]][0] = 1;
5
6 for (int X=0, i=1; i<m; i++) {
7     for (int c=0; c<SIGMA; c++) {
8         dfa[c][i] = dfa[c][X];
9     }
10
11     dfa[P[i]][i] = i+1;
12     X = dfa[P[i]][X];
13 }
```

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Gradnja avtomata za vzorec BANANA:

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Gradnja avtomata za vzorec BANANA:

$P =$ B A N A N A

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Gradnja avtomata za vzorec BANANA:

	$P =$	B	A	N	A	N	A
$\Sigma:$		0	1	2	3	4	5
A							
B							
N							

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Gradnja avtomata za vzorec BANANA:

	$P =$	B	A	N	A	N	A
$\Sigma:$		0	1	2	3	4	5
A		0					
B		1					
N		0					

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Gradnja avtomata za vzorec BANANA:

	$P =$	B	A	N	A	N	A
$\Sigma:$		0	1	2	3	4	5
A		0	2				
B		1	1				
N		0	0				

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Gradnja avtomata za vzorec BANANA:

	$P =$	B	A	N	A	N	A
$\Sigma:$		0	1	2	3	4	5
A		0	2	0			
B		1	1	1			
N		0	0	3			

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Gradnja avtomata za vzorec BANANA:

	$P =$	B	A	N	A	N	A
$\Sigma:$		0	1	2	3	4	5
A		0	2	0	4		
B		1	1	1	1		
N		0	0	3	0		

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Gradnja avtomata za vzorec BANANA:

	$P =$	B	A	N	A	N	A
$\Sigma:$		0	1	2	3	4	5
A		0	2	0	4	0	
B		1	1	1	1	1	
N		0	0	3	0	5	

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Gradnja avtomata za vzorec BANANA:

	$P =$	B	A	N	A	N	A
$\Sigma:$		0	1	2	3	4	5
A		0	2	0	4	0	6
B		1	1	1	1	1	1
N		0	0	3	0	5	0

Delo z nizi

Matevž
Jekovec

Imejmo $T = \text{AAABAAABAAABAAAB}$ in $P = \text{AAAA}$.

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Imejmo $T = AAABAAABAAABAAAB$ in $P = AAAA$.

Premislimo: Prva zgrešitev se pojavi, ko je okno na poziciji $i = 0$ in znaku $j = 3$, če štejemo od 0 dalje.

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Imejmo $T = \text{AAABAAABAAABAAAB}$ in $P = \text{AAAA}$.

Premislimo: Prva zgrešitev se pojavi, ko je okno na poziciji $i = 0$ in znaku $j = 3$, če štejemo od 0 dalje.

Ideja: Namesto, da premaknemo okno za 1 znak naprej, ga premaknemo za 4 znake naprej, saj ne bo ujemanja!

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Imejmo $T = AAABAAABAAABAAAB$ in $P = AAAA$.

Premislimo: Prva zgrešitev se pojavi, ko je okno na poziciji $i = 0$ in znaku $j = 3$, če štejemo od 0 dalje.

Ideja: Namesto, da premaknemo okno za 1 znak naprej, ga premaknemo za 4 znake naprej, saj ne bo ujemanja!

Za koliko pa lahko okno premaknemo, če se znak nahaja v vzorcu, a se ne ujema?

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Imejmo $T = AAABAAABAAABAAAB$ in $P = AAAA$.

Premislimo: Prva zgrešitev se pojavi, ko je okno na poziciji $i = 0$ in znaku $j = 3$, če štejemo od 0 dalje.

Ideja: Namesto, da premaknemo okno za 1 znak naprej, ga premaknemo za 4 znake naprej, saj ne bo ujemanja!

Za koliko pa lahko okno premaknemo, če se znak nahaja v vzorcu, a se ne ujema?

Primeri:

- ▶ Vzorec AAAA: če ni ujemanja, lahko vedno premaknemo okno za 4 mesta naprej, saj je znak *tuj* vzorcu.

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Imejmo $T = AAABAAABAAABAAAB$ in $P = AAAA$.

Premislimo: Prva zgrešitev se pojavi, ko je okno na poziciji $i = 0$ in znaku $j = 3$, če štejemo od 0 dalje.

Ideja: Namesto, da premaknemo okno za 1 znak naprej, ga premaknemo za 4 znake naprej, saj ne bo ujemanja!

Za koliko pa lahko okno premaknemo, če se znak nahaja v vzorcu, a se ne ujema?

Primeri:

- ▶ Vzorec AAAA: če ni ujemanja, lahko vedno premaknemo okno za 4 mesta naprej, saj je znak *tuj* vzorcu.
- ▶ Vzorec ABAB: če se pojavi *B* na vhodu in pričakujemo *A*, potem preskočimo le en znak oz. obratno. Za tuje znake isto kot zgoraj premaknemo okno za 4 mesta naprej.

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Okno lahko premaknemo naprej za toliko, da se neujemajoč znak besedila pokriva z zadnjo pojavitvijo znaka v vzorcu.

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Okno lahko premaknemo naprej za toliko, da se neujemajoč znak besedila pokriva z zadnjo pojavitvijo znaka v vzorcu.
Algoritem: Zgradimo preskočno tabelo, ki hrani indekse najbolj desnih pojavitev vseh možnih znakov abecede Σ .

$$P = \quad B \quad A \quad N \quad A \quad N \quad A$$

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Okno lahko premaknemo naprej za toliko, da se neujemajoč znak besedila pokriva z zadnjo pojavitvijo znaka v vzorcu.
 Algoritem: Zgradimo preskočno tabelo, ki hrani indekse najbolj desnih pojavitev vseh možnih znakov abecede Σ .

$P =$	B	A	N	A	N	A
	0	1	2	3	4	5
A						
B						
N						

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Okno lahko premaknemo naprej za toliko, da se neujemajoč znak besedila pokriva z zadnjo pojavitvijo znaka v vzorcu.

Algoritem: Zgradimo preskočno tabelo, ki hrani indekse najbolj desnih pojavitev vseh možnih znakov abecede Σ .

$P =$	B	A	N	A	N	A
	0	1	2	3	4	5
A	-1	1	1	3	3	5
B	0	0	0	0	0	0
N	-1	-1	2	2	4	4

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Okno lahko premaknemo naprej za toliko, da se neujemajoč znak besedila pokriva z zadnjo pojavitvijo znaka v vzorcu.
 Algoritem: Zgradimo preskočno tabelo, ki hrani indekse najbolj desnih pojavitev vseh možnih znakov abecede Σ .

$P =$	B	A	N	A	N	A	
	0	1	2	3	4	5	right[c]
A	-1	1	1	3	3	5	5
B	0	0	0	0	0	0	0
N	-1	-1	2	2	4	4	4

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Iskanje: Okno postavimo na začetek $i = 0$ in preverjamo znake od desne proti levi $j = |P| - 1$:

- ▶ Če se znak c v besedilu ujema z znakom v oknu, nadaljujemo s preverjanjem naslednjega znaka v oknu in besedilu.

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Iskanje: Okno postavimo na začetek $i = 0$ in preverjamo znake od desne proti levi $j = |P| - 1$:

- ▶ Če se znak c v besedilu ujema z znakom v oknu, nadaljujemo s preverjanjem naslednjega znaka v oknu in besedilu.
- ▶ Če se c ne ujema in je $\text{right}[c] = -1$, potem premakni okno za $|P|$ znakov naprej.

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Iskanje: Okno postavimo na začetek $i = 0$ in preverjamo znake od desne proti levi $j = |P| - 1$:

- ▶ Če se znak c v besedilu ujema z znakom v oknu, nadaljujemo s preverjanjem naslednjega znaka v oknu in besedilu.
- ▶ Če se c ne ujema in je $\text{right}[c] = -1$, potem premakni okno za $|P|$ znakov naprej.
- ▶ Če se c ne ujema in $\text{right}[c] \neq -1$, potem premaknemo okno za $j - \text{right}[c]$ znakov.

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Iskanje: Okno postavimo na začetek $i = 0$ in preverjamo znake od desne proti levi $j = |P| - 1$:

- ▶ Če se znak c v besedilu ujema z znakom v oknu, nadaljujemo s preverjanjem naslednjega znaka v oknu in besedilu.
- ▶ Če se c ne ujema in je $\text{right}[c] = -1$, potem premakni okno za $|P|$ znakov naprej.
- ▶ Če se c ne ujema in $\text{right}[c] \neq -1$, potem premaknemo okno za $j - \text{right}[c]$ znakov.
 - ▶ Izjema: V primeru, da bi okno premaknili nazaj, ga premaknemo za 1 znak naprej.

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Iskanje: Okno postavimo na začetek $i = 0$ in preverjamo znake od desne proti levi $j = |P| - 1$:

- ▶ Če se znak c v besedilu ujema z znakom v oknu, nadaljujemo s preverjanjem naslednjega znaka v oknu in besedilu.
- ▶ Če se c ne ujema in je $\text{right}[c] = -1$, potem premakni okno za $|P|$ znakov naprej.
- ▶ Če se c ne ujema in $\text{right}[c] \neq -1$, potem premaknemo okno za $j - \text{right}[c]$ znakov.
 - ▶ Izjema: V primeru, da bi okno premaknili nazaj, ga premaknemo za 1 znak naprej.

T: NA DREVESU RASTEJO BANANE.

P: BANANA

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Iskanje: Okno postavimo na začetek $i = 0$ in preverjamo znake od desne proti levi $j = |P| - 1$:

- ▶ Če se znak c v besedilu ujema z znakom v oknu, nadaljujemo s preverjanjem naslednjega znaka v oknu in besedilu.
- ▶ Če se c ne ujema in je $\text{right}[c] = -1$, potem premakni okno za $|P|$ znakov naprej.
- ▶ Če se c ne ujema in $\text{right}[c] \neq -1$, potem premaknemo okno za $j - \text{right}[c]$ znakov.
 - ▶ Izjema: V primeru, da bi okno premaknili nazaj, ga premaknemo za 1 znak naprej.

T: NA DREVESU RASTEJO BANANE.

P: BANANA

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Iskanje: Okno postavimo na začetek $i = 0$ in preverjamo znake od desne proti levi $j = |P| - 1$:

- ▶ Če se znak c v besedilu ujema z znakom v oknu, nadaljujemo s preverjanjem naslednjega znaka v oknu in besedilu.
- ▶ Če se c ne ujema in je $\text{right}[c] = -1$, potem premakni okno za $|P|$ znakov naprej.
- ▶ Če se c ne ujema in $\text{right}[c] \neq -1$, potem premaknemo okno za $j - \text{right}[c]$ znakov.
 - ▶ Izjema: V primeru, da bi okno premaknili nazaj, ga premaknemo za 1 znak naprej.

T: NA DREVESU RASTEJO BANANE.

P: BANANA

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Iskanje: Okno postavimo na začetek $i = 0$ in preverjamo znake od desne proti levi $j = |P| - 1$:

- ▶ Če se znak c v besedilu ujema z znakom v oknu, nadaljujemo s preverjanjem naslednjega znaka v oknu in besedilu.
- ▶ Če se c ne ujema in je $\text{right}[c] = -1$, potem premakni okno za $|P|$ znakov naprej.
- ▶ Če se c ne ujema in $\text{right}[c] \neq -1$, potem premaknemo okno za $j - \text{right}[c]$ znakov.
 - ▶ Izjema: V primeru, da bi okno premaknili nazaj, ga premaknemo za 1 znak naprej.

T: NA DREVESU RASTEJO BANANE.

P:

BANANA

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Iskanje: Okno postavimo na začetek $i = 0$ in preverjamo znake od desne proti levi $j = |P| - 1$:

- ▶ Če se znak c v besedilu ujema z znakom v oknu, nadaljujemo s preverjanjem naslednjega znaka v oknu in besedilu.
- ▶ Če se c ne ujema in je $\text{right}[c] = -1$, potem premakni okno za $|P|$ znakov naprej.
- ▶ Če se c ne ujema in $\text{right}[c] \neq -1$, potem premaknemo okno za $j - \text{right}[c]$ znakov.
 - ▶ Izjema: V primeru, da bi okno premaknili nazaj, ga premaknemo za 1 znak naprej.

T: NA DREVESU RASTEJO BANANE.

P: BANANA

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Uporabimo znanje iz zgoščevalnih tabel:

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Uporabimo znanje iz zgoščevalnih tabel:

1. Najprej izračunamo zgoščeno vrednost vzorca H_p .

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Uporabimo znanje iz zgoščevalnih tabel:

1. Najprej izračunamo zgoščeno vrednost vzorca H_P .
2. Postavimo okno na začetek $i = 0$ in izračunamo zgoščeno vrednost znakov pod oknom H_T .

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Uporabimo znanje iz zgoščevalnih tabel:

1. Najprej izračunamo zgoščeno vrednost vzorca H_P .
2. Postavimo okno na začetek $i = 0$ in izračunamo zgoščeno vrednost znakov pod oknom H_T .
3. Če $H_T = H_P$, preverimo vse znake pod oknom, da se prepričamo o enakosti, in vrnemo mesto pojavitve.

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Uporabimo znanje iz zgoščevalnih tabel:

1. Najprej izračunamo zgoščeno vrednost vzorca H_P .
2. Postavimo okno na začetek $i = 0$ in izračunamo zgoščeno vrednost znakov pod oknom H_T .
3. Če $H_T = H_P$, preverimo vse znake pod oknom, da se prepričamo o enakosti, in vrnemo mesto pojavitve.
4. Če $H_T \neq H_P$, se premaknemo za en znak naprej, izračunamo novo zgoščeno vrednost znakov pod oknom H_T in gremo na korak 3.

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Ideja: Za izračun nove zgoščene vrednosti okna nam ni treba prebrati ponovno vseh znakov v njem, ampak le primnožimo a , prištejemo vrednost novega znaka in odstranimo zgoščeno vrednost najstarejšega znaka. V konstantnem času tako premaknemo okno za en znak naprej.

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Ideja: Za izračun nove zgoščene vrednosti okna nam ni treba prebrati ponovno vseh znakov v njem, ampak le primnožimo a , prištejemo vrednost novega znaka in odstranimo zgoščeno vrednost najstarejšega znaka. V konstantnem času tako premaknemo okno za en znak naprej.

Robin-Karpova zgoščevalna funkcija:

$$H = c_1 a^{|P|-1} + c_2 a^{|P|-2} + \dots + c_{|P|} a^0$$

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Ideja: Za izračun nove zgoščene vrednosti okna nam ni treba prebrati ponovno vseh znakov v njem, ampak le primnožimo a , prištejemo vrednost novega znaka in odstranimo zgoščeno vrednost najstarejšega znaka. V konstantnem času tako premaknemo okno za en znak naprej.

Robin-Karpova zgoščevalna funkcija:

$$H = c_1 a^{|P|-1} + c_2 a^{|P|-2} + \dots + c_{|P|} a^0$$

Ko pride nov znak c' , odstranimo najstarejši znak c_{old} in izračunamo nov

$$H' = H \cdot a - c_{old} \cdot a^{|P|} + c'$$

Delo z nizi

Matevž
Jekovec

Številsko drevo omogoča iskanje po začetkih besed. Kaj pa po korenu ali priponi?

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Delo z nizi

Matevž
Jekovec

Številsko drevo omogoča iskanje po začetkih besed. Kaj pa po korenu ali priponi?

Uvod

Hramba
besedila

Priponsko drevo je stisnjeno številsko drevo, ki hrani vse predpone besedila (ali besed).

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Številsko drevo omogoča iskanje po začetkih besed. Kaj pa po korenu ali priponi?

Priponsko drevo je stisnjeno številsko drevo, ki hrani vse predpone besedila (ali besed).

0 1 2 3 4 5 6 7 8 9 10 11
T=ABRAKADABRA\$

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Številsko drevo omogoča iskanje po začetkih besed. Kaj pa po korenu ali priponi?

Priponsko drevo je stisnjeno številsko drevo, ki hrani vse predpone besedila (ali besed).

0 1 2 3 4 5 6 7 8 9 10 11
T=ABRAKADABRA\$

0 ABRAKADABRA\$
1 BRAKADABRA\$
2 RAKADABRA\$
3 AKADABRA\$
4 KADABRA
5 ADABRA\$
6 DABRA\$
7 ABRA\$
8 BRA\$
9 RA\$
10 A\$
11 \$

Priponsko drevo (ang. *Suffix tree*)

Delo z nizi

Matevž
Jekovec

Številsko drevo omogoča iskanje po začetkih besed. Kaj pa po korenu ali priponi?

Priponsko drevo je stisnjeno številsko drevo, ki hrani vse predpone besedila (ali besed).

Uvod

Hramba
besedila

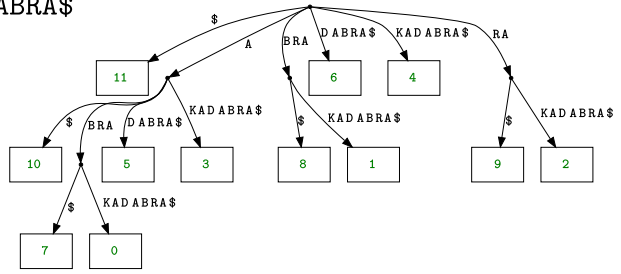
Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

0 1 2 3 4 5 6 7 8 9 10 11
T=ABRAKADABRA\$

- 0 ABRAKADABRA\$
- 1 BRAKADABRA\$
- 2 RAKADABRA\$
- 3 AKADABRA\$
- 4 KADABRA
- 5 ADABRA\$
- 6 DABRA\$
- 7 ABRA\$
- 8 BRA\$
- 9 RA\$
- 10 A\$
- 11 \$



Delo z nizi

Matevž
Jekovec

Priponsko polje je varianta priponskega drevesa, zapisana v polju.

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Priponsko polje je varianta priponskega drevesa, zapisana v polju.

Iskanje: podobno kot v urejem polju običajnih števil — z bisekcijo!

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Priponsko polje je varianta priponskega drevesa, zapisana v polju.

Iskanje: podobno kot v urejem polju običajnih števil — z bisekcijo!

0 1 2 3 4 5 6 7 8 9 10 11
T=ABRAKADABRA\$

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Priponsko polje je varianta priponskega drevesa, zapisana v polju.

Iskanje: podobno kot v urejem polju običajnih števil — z bisekcijo!

0 1 2 3 4 5 6 7 8 9 10 11
T=ABRAKADABRA\$

0 ABRAKADABRA\$
1 BRAKADABRA\$
2 RAKADABRA\$
3 AKADABRA\$
4 KADABRA
5 ADABRA\$
6 DABRA\$
7 ABRA\$
8 BRA\$
9 RA\$
10 A\$
11 \$

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Priponsko polje je varianta priponskega drevesa, zapisana v polju.

Iskanje: podobno kot v urejem polju običajnih števil — z bisekcijo!

0 1 2 3 4 5 6 7 8 9 10 11
T=ABRAKADABRA\$

11 \$
10 A\$
7 ABRA\$
0 ABRAKADABRA\$
5 ADABRA\$
3 AKADABRA\$
8 BRA\$
1 BRAKADABRA\$
6 DABRA\$
4 KADABRA
9 RA\$
12 RAKADABRA\$

Priponsko polje (ang. *Suffix array*)

Delo z nizi

Matevž
Jekovec

Priponsko polje je varianta priponskega drevesa, zapisana v polju.

Iskanje: podobno kot v urejem polju običajnih števil — z bisekcijo!

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

0 1 2 3 4 5 6 7 8 9 10 11
T=ABRAKADABRA\$
SA
11 \$
10 A\$
7 ABRA\$
0 ABRAKADABRA\$
5 ADABRA\$
3 AKADABRA\$
8 BRA\$
1 BRAKADABRA\$
6 DABRA\$
4 KADABRA
9 RA\$
12 RAKADABRA\$

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Priponsko polje je varianta priponskega drevesa, zapisana v polju.

Iskanje: podobno kot v urejem polju običajnih števil — z bisekcijo!

0 1 2 3 4 5 6 7 8 9 10 11
T=ABRAKADABRA\$

LCP SA

/	11	\$
0	10	A\$
1	7	ABRA\$
4	0	ABRAKADABRA\$
1	5	ADABRA\$
1	3	AKADABRA\$
0	8	BRA\$
3	1	BRAKADABRA\$
0	6	DABRA\$
0	4	KADABRA
0	9	RA\$
2	12	RAKADABRA\$

Delo z nizi

Matevž
Jekovec

Enostavna (sicer ne prav učinkovita) gradnja — uporabimo splošen algoritem za urejanje in sprogramiramo svoj comparator.

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Delo z nizi

 Matevž
 Jekovec

Enostavna (sicer ne prav učinkovita) gradnja — uporabimo splošen algoritem za urejanje in sprogramiramo svoj comparator.

Uvod

 Hramba
 besedila

 Iskanje vzorca
 v besedilu

 Podobnost
 med nizi

Vaje in naloge

```

1  const char *text="abcabc";
2  const int N=6;
3
4  bool suffixCompare( int a, int b ) {
5      int i;
6      for (i=0; (i<N - max(a,b)) &&
7          (text[a+i])==(text[b+i]); i++);
8
9      if ( i == N - max(a,b) ) { return a>b; }
10     else { return (text[a+i] < text[b+i]); }
11 }
12
13 int main() {
14     vector<int> sa = {0,1,2,3,4,5};
15     std::sort(sa.begin(), sa.end(), suffixCompare);
16 }
```

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Levenshteinova razdalja ali *urejevalna razdalja* δ med dvema nizoma T_1 in T_2 je najmanjše število vstavljanj, brisanj ali zamenjav posameznih znakov, da spremenimo T_1 v T_2 .

Delo z nizi

Formula za izračun razdalje med nizi $T_1[0..i]$ in $T_2[0..j]$:

Matevž
Jekovec

$$\delta(i, j) = \min \begin{cases} \delta(i-1, j) + 1 \\ \delta(i, j-1) + 1 \\ \delta(i-1, j-1) + 1 & \text{če } T_1[i] \neq T_2[j] \\ \delta(i-1, j-1) & \text{če } T_1[i] = T_2[j] \end{cases}$$

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Delo z nizi

Formula za izračun razdalje med nizi $T_1[0..i]$ in $T_2[0..j]$:

$$\delta(i, j) = \min \begin{cases} \delta(i-1, j) + 1 \\ \delta(i, j-1) + 1 \\ \delta(i-1, j-1) + 1 & \text{če } T_1[i] \neq T_2[j] \\ \delta(i-1, j-1) & \text{če } T_1[i] = T_2[j] \end{cases}$$

Izračun razdalje se naredi z Levenshteinovo matriko:

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Delo z nizi

 Formula za izračun razdalje med nizi $T_1[0..i]$ in $T_2[0..j]$:

$$\delta(i, j) = \min \begin{cases} \delta(i-1, j) + 1 \\ \delta(i, j-1) + 1 \\ \delta(i-1, j-1) + 1 & \text{če } T_1[i] \neq T_2[j] \\ \delta(i-1, j-1) & \text{če } T_1[i] = T_2[j] \end{cases}$$

Uvod

 Hramba
 besedila

 Iskanje vzorca
 v besedilu

 Podobnost
 med nizi

Vaje in naloge

Izračun razdalje se naredi z Levenshteinovo matriko:

B A N A N A

Delo z nizi

 Formula za izračun razdalje med nizi $T_1[0..i]$ in $T_2[0..j]$:

$$\delta(i, j) = \min \begin{cases} \delta(i-1, j) + 1 \\ \delta(i, j-1) + 1 \\ \delta(i-1, j-1) + 1 & \text{če } T_1[i] \neq T_2[j] \\ \delta(i-1, j-1) & \text{če } T_1[i] = T_2[j] \end{cases}$$

 Matevž
 Jekovec

Uvod

 Hramba
 besedila

 Iskanje vzorca
 v besedilu

 Podobnost
 med nizi

Vaje in naloge

Izračun razdalje se naredi z Levenshteinovo matriko:

	B	A	N	A	N	A
0	1	2	3	4	5	6
A	1					
N	2					
A	3					
N	4					
A	5					
S	6					

Delo z nizi

 Formula za izračun razdalje med nizi $T_1[0..i]$ in $T_2[0..j]$:

$$\delta(i, j) = \min \begin{cases} \delta(i-1, j) + 1 \\ \delta(i, j-1) + 1 \\ \delta(i-1, j-1) + 1 & \text{če } T_1[i] \neq T_2[j] \\ \delta(i-1, j-1) & \text{če } T_1[i] = T_2[j] \end{cases}$$

Uvod

 Hramba
 besedila

 Iskanje vzorca
 v besedilu

 Podobnost
 med nizi

Vaje in naloge

Izračun razdalje se naredi z Levenshteinovo matriko:

		B	A	N	A	N	A
	0	1	2	3	4	5	6
A	1	1	1	2	3	4	5
N	2						
A	3						
N	4						
A	5						
S	6						

Delo z nizi

Formula za izračun razdalje med nizi $T_1[0..i]$ in $T_2[0..j]$:

$$\delta(i, j) = \min \begin{cases} \delta(i-1, j) + 1 \\ \delta(i, j-1) + 1 \\ \delta(i-1, j-1) + 1 & \text{če } T_1[i] \neq T_2[j] \\ \delta(i-1, j-1) & \text{če } T_1[i] = T_2[j] \end{cases}$$

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Izračun razdalje se naredi z Levenshteinovo matriko:

		B	A	N	A	N	A
	0	1	2	3	4	5	6
A	1	1	1	2	3	4	5
N	2	2	2	1	2	3	4
A	3						
N	4						
A	5						
S	6						

Delo z nizi

Formula za izračun razdalje med nizi $T_1[0..i]$ in $T_2[0..j]$:

$$\delta(i, j) = \min \begin{cases} \delta(i-1, j) + 1 \\ \delta(i, j-1) + 1 \\ \delta(i-1, j-1) + 1 & \text{če } T_1[i] \neq T_2[j] \\ \delta(i-1, j-1) & \text{če } T_1[i] = T_2[j] \end{cases}$$

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Izračun razdalje se naredi z Levenshteinovo matriko:

		B	A	N	A	N	A
	0	1	2	3	4	5	6
A	1	1	1	2	3	4	5
N	2	2	2	1	2	3	4
A	3	3	2	2	1	2	3
N	4						
A	5						
S	6						

Delo z nizi

Formula za izračun razdalje med nizi $T_1[0..i]$ in $T_2[0..j]$:

$$\delta(i, j) = \min \begin{cases} \delta(i-1, j) + 1 \\ \delta(i, j-1) + 1 \\ \delta(i-1, j-1) + 1 & \text{če } T_1[i] \neq T_2[j] \\ \delta(i-1, j-1) & \text{če } T_1[i] = T_2[j] \end{cases}$$

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Izračun razdalje se naredi z Levenshteinovo matriko:

		B	A	N	A	N	A
	0	1	2	3	4	5	6
A	1	1	1	2	3	4	5
N	2	2	2	1	2	3	4
A	3	3	2	2	1	2	3
N	4	4	3	2	2	1	2
A	5						
S	6						

Delo z nizi

Formula za izračun razdalje med nizi $T_1[0..i]$ in $T_2[0..j]$:

$$\delta(i, j) = \min \begin{cases} \delta(i-1, j) + 1 \\ \delta(i, j-1) + 1 \\ \delta(i-1, j-1) + 1 & \text{če } T_1[i] \neq T_2[j] \\ \delta(i-1, j-1) & \text{če } T_1[i] = T_2[j] \end{cases}$$

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Izračun razdalje se naredi z Levenshteinovo matriko:

		B	A	N	A	N	A
	0	1	2	3	4	5	6
A	1	1	1	2	3	4	5
N	2	2	2	1	2	3	4
A	3	3	2	2	1	2	3
N	4	4	3	2	2	1	2
A	5	5	4	3	2	2	1
S	6						

Delo z nizi

 Formula za izračun razdalje med nizi $T_1[0..i]$ in $T_2[0..j]$:

$$\delta(i, j) = \min \begin{cases} \delta(i-1, j) + 1 \\ \delta(i, j-1) + 1 \\ \delta(i-1, j-1) + 1 & \text{če } T_1[i] \neq T_2[j] \\ \delta(i-1, j-1) & \text{če } T_1[i] = T_2[j] \end{cases}$$

Uvod

 Hramba
 besedila

 Iskanje vzorca
 v besedilu

 Podobnost
 med nizi

Vaje in naloge

Izračun razdalje se naredi z Levenshteinovo matriko:

		B	A	N	A	N	A
	0	1	2	3	4	5	6
A	1	1	1	2	3	4	5
N	2	2	2	1	2	3	4
A	3	3	2	2	1	2	3
N	4	4	3	2	2	1	2
A	5	5	4	3	2	2	1
S	6	6	5	4	3	3	2

Delo z nizi

Formula za izračun razdalje med nizi $T_1[0..i]$ in $T_2[0..j]$:

$$\delta(i, j) = \min \begin{cases} \delta(i-1, j) + 1 \\ \delta(i, j-1) + 1 \\ \delta(i-1, j-1) + 1 & \text{če } T_1[i] \neq T_2[j] \\ \delta(i-1, j-1) & \text{če } T_1[i] = T_2[j] \end{cases}$$

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Izračun razdalje se naredi z Levenshteinovo matriko:

		B	A	N	A	N	A
	0	1	2	3	4	5	6
A	1	1	1	2	3	4	5
N	2	2	2	1	2	3	4
A	3	3	2	2	1	2	3
N	4	4	3	2	2	1	2
A	5	5	4	3	2	2	1
S	6	6	5	4	3	3	2

Levenshteinova razdalja se nahaja spodnjem desnem kotu $\delta(|T_1|, |T_2|)$.

Delo z nizi

Formula za izračun razdalje med nizi $T_1[0..i]$ in $T_2[0..j]$:

$$\delta(i, j) = \min \begin{cases} \delta(i-1, j) + 1 \\ \delta(i, j-1) + 1 \\ \delta(i-1, j-1) + 1 & \text{če } T_1[i] \neq T_2[j] \\ \delta(i-1, j-1) & \text{če } T_1[i] = T_2[j] \end{cases}$$

Uvod

Hramba besedila

Iskanje vzorca v besedilu

Podobnost med nizi

Vaje in naloge

Izračun razdalje se naredi z Levenshteinovo matriko:

		B	A	N	A	N	A
	0	1	2	3	4	5	6
A	1	1	1	2	3	4	5
N	2	2	2	1	2	3	4
A	3	3	2	2	1	2	3
N	4	4	3	2	2	1	2
A	5	5	4	3	2	2	1
S	6	6	5	4	3	3	2

Levenshteinova razdalja se nahaja spodnjem desnem kotu $\delta(|T_1|, |T_2|)$. Pot nam pove, katere ukaze je potrebno izvesti.

Delo z nizi

Matevž
Jekovec

Za izračun najdaljšega skupnega podniza lahko uporabimo priponsko drevo obeh nizov in pogledamo najgloblje skupno vozlišče.

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Delo z nizi

 Matevž
 Jekovec

Uvod

 Hramba
 besedila

 Iskanje vzorca
 v besedilu

 Podobnost
 med nizi

Vaje in naloge

Za izračun najdaljšega skupnega podniza lahko uporabimo priponsko drevo obeh nizov in pogledamo najgloblje skupno vozlišče.

Lahko pa uporabimo matriko, podobno Levenshteinovi, le da spremenimo pogoj:

$$\begin{aligned}
 &LCSuff(T_1[1..i], T_2[1..j]) = \\
 &\begin{cases} LCSuff(T_1[1..i-1], T_2[1..j-1]) + 1 & \text{če } T_1[i] = T_2[j] \\ 0 & \text{sicer} \end{cases}
 \end{aligned}$$

Delo z nizi

 Matevž
 Jekovec

Uvod

 Hramba
 besedila

 Iskanje vzorca
 v besedilu

 Podobnost
 med nizi

Vaje in naloge

Za izračun najdaljšega skupnega podniza lahko uporabimo priponsko drevo obeh nizov in pogledamo najgloblje skupno vozlišče.

Lahko pa uporabimo matriko, podobno Levenshteinovi, le da spremenimo pogoj:

$$\begin{aligned}
 & LCSuff(T_1[1..i], T_2[1..j]) = \\
 & \begin{cases} LCSuff(T_1[1..i-1], T_2[1..j-1]) + 1 & \text{če } T_1[i] = T_2[j] \\ 0 & \text{sicer} \end{cases}
 \end{aligned}$$

		B	A	N	A	N	A
	0	0	0	0	0	0	0
A	0	0	1	0	1	0	1
N	0	0	0	2	0	2	0
A	0	0	1	0	3	0	3
N	0	0	0	2	0	4	0
A	0	0	0	0	3	0	5
S	0	0	0	0	0	0	0

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Najdaljše skupno podzaporedje (ang. *longest common subsequence*) je podobno najdaljšemu skupnemu podnizu, le da se lahko med posameznimi znaki, ki so prisotni v obeh nizih, vrinjene tudi tuji znaki.

Delo z nizi

 Matevž
 Jakovec

Uvod

 Hramba
 besedila

 Iskanje vzorca
 v besedilu

 Podobnost
 med nizi

Vaje in naloge

Najdaljše skupno podzaporedje (ang. *longest common subsequence*) je podobno najdaljšemu skupnemu podnizu, le da se lahko med posameznimi znaki, ki so prisotni v obeh nizih, vrinjene tudi tuji znaki.

$$LCS(T_1[1..i], T_2[1..j]) =$$

$$\begin{cases} LCS(T_1[1..i-1], T_2[1..j-1]) + 1 & \text{če } T_1[i] = T_2[j] \\ \max(LCS(T_1[1..i], T_2[1..j-1]), LCS(T_1[1..i-1], T_2[1..j])) & \\ \text{če } T_1[i] \neq T_2[j] \end{cases}$$

Delo z nizi

 Matevž
 Jekovec

Uvod

 Hramba
 besedila

 Iskanje vzorca
 v besedilu

 Podobnost
 med nizi

Vaje in naloge

Najdaljše skupno podzaporedje (ang. *longest common subsequence*) je podobno najdaljšemu skupnemu podnizu, le da se lahko med posameznimi znaki, ki so prisotni v obeh nizih, vrinjene tudi tuji znaki.

$$LCS(T_1[1..i], T_2[1..j]) =$$

$$\begin{cases} LCS(T_1[1..i-1], T_2[1..j-1]) + 1 & \text{če } T_1[i] = T_2[j] \\ \max(LCS(T_1[1..i], T_2[1..j-1]), LCS(T_1[1..i-1], T_2[1..j])) & \text{če } T_1[i] \neq T_2[j] \end{cases}$$

		B	A	N	A	N	A
	0	0	0	0	0	0	0
A	0	0	1	1	1	1	1
N	0	0	1	2	2	2	2
A	0	0	1	2	3	3	3
N	0	0	1	2	3	4	4
A	0	0	1	2	3	4	5
S	0	0	1	2	3	4	5

Delo z nizi

Matevž
Jekovec

Burkhard-Kellerjevo drevo se uporablja pri črkovalnikih in je uporabno za iskanje podobnih besed. Vozlišča predstavljajo besede, povezave med vozlišči pa njihovo urejevalno razdaljo.

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Burkhard-Kellerjevo drevo se uporablja pri črkovalnikih in je uporabno za iskanje podobnih besed. Vozlišča predstavljajo besede, povezave med vozlišči pa njihovo urejevalno razdaljo. Gradnja: Začnemo s poljubnim pojmom. Nato dodamo novo geslo tako, da izračunamo urejevalno razdaljo D med trenutnim vozliščem in novim geslom. Če povezava v smeri razdalje ne obstaja, novo geslo pripnemo obstoječemu vozlišču. Sicer sledimo povezavi in izračunamo urejevalno razdaljo med novim vozliščem in našim geslom.

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Burkhard-Kellerjevo drevo se uporablja pri črkovalnikih in je uporabno za iskanje podobnih besed. Vozlišča predstavljajo besede, povezave med vozlišči pa njihovo urejevalno razdaljo. Gradnja: Začnemo s poljubnim pojmom. Nato dodamo novo geslo tako, da izračunamo urejevalno razdaljo D med trenutnim vozliščem in novim geslom. Če povezava v smeri razdalje ne obstaja, novo geslo pripnemo obstoječemu vozlišču. Sicer sledimo povezavi in izračunamo urejevalno razdaljo med novim vozliščem in našim geslom.

Iskanje najbolj podobnih besed z urejevalno razdaljo δ :
 Začnemo pri korenu in izračunamo urejevalno razdaljo D med iskanim geslom in trenutnim vozliščem. Če je $D \leq \delta$, potem izpišemo trenutno vozlišče. Preiskujemo tiste otroke, ki imajo razdaljo $\geq D - \delta$ ali $\leq D + \delta$.

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

gesla

ura

urar

urnik

avto

pes

pesem

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

gesla

ura

ura

urar

urnik

avto

pes

pesem

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

gesla

ura

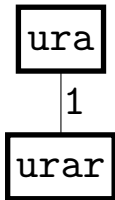
urar

urnik

avto

pes

pesem



Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

gesla

ura

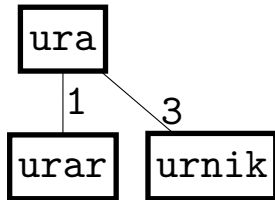
urar

urnik

avto

pes

pesem



Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

gesla

ura

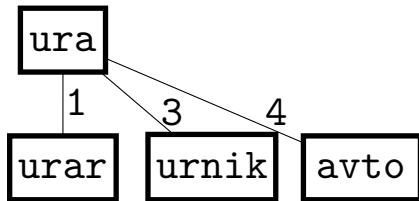
urar

urnik

avto

pes

pesem



Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

gesla

ura

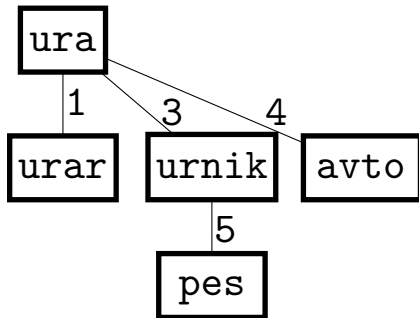
urar

urnik

avto

pes

pesem



Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

gesla

ura

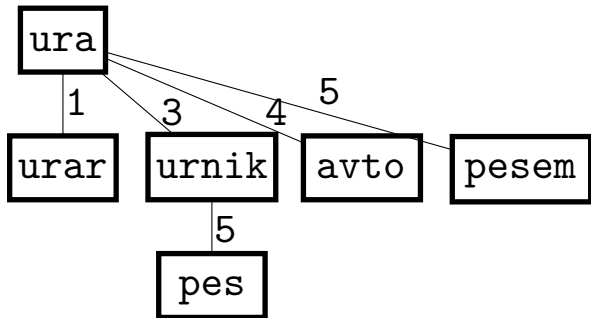
urar

urnik

avto

pes

pesem



Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

gesla

ura

urar

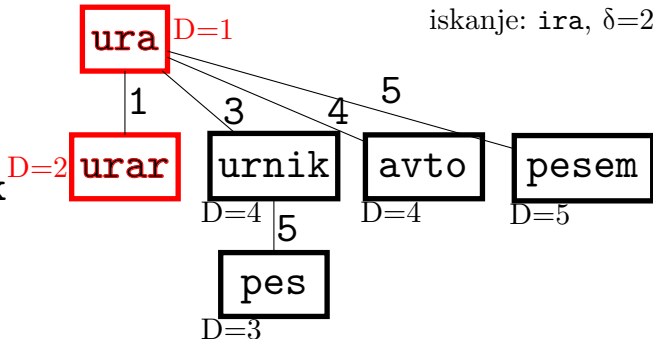
urnik

avto

pes

pesem

iskanje: ira, $\delta=2$



Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Vaje:

- ▶ S pomočjo urejanja (npr. quicksort) uredi vnešene nize po abecedi.
- ▶ Sprogramiraj številsko drevo, ki podpira operaciji vstavljanja in iskanja.
- ▶ Sprogramiraj gradnjo in iskanje po priponskem polju.
- ▶ Sprogramiraj BK-drevo, ki podpira operaciji vstavljanja in iskanja.
- ▶ Sprogramiraj Edit Distance, Longest Common Substring in Longest Common Subsequence s pomočjo dinamičnega programiranja.

Delo z nizi

Matevž
Jekovec

Uvod

Hramba
besedila

Iskanje vzorca
v besedilu

Podobnost
med nizi

Vaje in naloge

Vaje:

- ▶ S pomočjo urejanja (npr. quicksort) uredi vnešene nize po abecedi.
- ▶ Sprogramiraj številsko drevo, ki podpira operaciji vstavljanja in iskanja.
- ▶ Sprogramiraj gradnjo in iskanje po priponskem polju.
- ▶ Sprogramiraj BK-drevo, ki podpira operaciji vstavljanja in iskanja.
- ▶ Sprogramiraj Edit Distance, Longest Common Substring in Longest Common Subsequence s pomočjo dinamičnega programiranja.

UVa naloge:

- 499 What's The Frequency, Kenneth?
- 454 Anagrams
- 164 String Computer - Edit distance
- 290 Palindroms ↔ smordnilaP
- 335 Processing MX Records