Pattern matching

Metodický koncept k efektivní podpoře klíčových odborných kompetencí s využitím cizího jazyka ATCZ62 - CLIL jako výuková strategie na vysoké škole

Interreg
EVROPSKÁ UNIE
Rakousko-Česká republika
Evropský fond pro regionální rozvoj

Europäische Union
Evropská unie
Europäischer Fonds für regionale Entwicklung
Evropský fond pro regionální rozvoj

VŠTE

UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

Pattern matching – porovnávání vzorů

- Hledání jistého vzoru v dané sekvenci (hledání podřetězce v řetězci)
- Řetězce (strings)
  - Sekvence znaků
  - Abeceda – množina všech možných znaků
    - ASCII
    - Unicode
    - {0, 1}, {A, C, G, T}
  - *P* je řetězec délky *m*
    - Podřetězec *P[i..j]* řetězce *P*, skládající se ze znaků mezi *i* a *j*
    - Předpona (prefix)
    - Přípona (suffix)

Interreg
Rakousko-Česká republika
Evropský fond pro regionální rozvoj
Europäische Union
Evropská unie
Europäischer Fonds für regionale Entwicklung
Evropský fond pro regionální rozvoj
VŠTE
UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

Pattern matching – porovnávání vzorů

- Řetězec *T* (text) a *P* (pattern)
- Pattern matching – porovnávání vzorů – podřetězce *T,* který je roven (je stejný) jako *P*
- Aplikace
  - Textové editory
  - Vyhledávací nástroje
  - Biologický výzkum

Interreg
Rakousko-Česká republika
Evropský fond pro regionální rozvoj
Europäische Union
Evropská unie
Europäischer Fonds für regionale Entwicklung
Evropský fond pro regionální rozvoj
VŠTE
UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

Brute-Force (Hrubá síla) algoritmus

- Prochází text zleva doprava
- Porovnává vzor *P* s textem *T*, pro všechny možné pozice dokud
  - Není nalezena shoda
  - Nebyly vyzkoušeny všechny možné pozice
- Časová náročnost: *O(nm)*

Algorithm *BruteForceMatch*(*T*, *P*)
  Input text *T* délky *n* a vzor *P* délky *m*
  Output počáteční index podřetězce, nebo −1, pokud daný podřetězec neexistuje
  for $i \leftarrow 0$ **to** $n - m$
    { test shift *i* of the pattern }
    $j \leftarrow 0$
    while $j < m \wedge T[i + j] = P[j]$
      $j \leftarrow j + 1$
    if $j = m$
      return *i* { match at *i* }
    else
      return −1 { no match }

Interreg Rakousko-Česká republika
Evropský fond pro regionální rozvoj
Europäische Union
Evropská unie
Europäischer Fonds für regionale Entwicklung
Evropský fond pro regionální rozvoj
VŠTE
UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

Boyer-Moorův algoritmus

- Prohledáváme zprava doleva
- Index *i* ukazuje do textu *T*, index *j* ukazuje do *P*
- 4 případy při prohledávání
  - *T(i)* není v *P* vůbec, posuneme i dále o délku *P* (zarovnáme *P* na další písmeno v *T*, tedy *T(i+1))*
  - *T(i)* odpovídá *P(j)* - posuneme se v obou do leva a opakujeme (jako v brutální síle)
  - *T(i)* není *P(j),* ale *T(i)* je v *P* před indexem *j* -> zarovnáme *P* doprava tak, aby *T(i)* odpovídal jeho výskytu v *P* a opakujeme
  - *T(i)* není *P(j),* ale *T(i)* je v *P* za indexem *j* -> posuneme se doprava o 1 a opakujeme (nemůžeme se vracet, ale ani posunout o více dále)

Interreg Rakousko-Česká republika
Evropský fond pro regionální rozvoj
Europäische Union
Evropská unie
Europäischer Fonds für regionale Entwicklung
Evropský fond pro regionální rozvoj
VŠTE
UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

Boyer-Moorův algoritmus

- Vrátíme *i* - pokud nalezneme celý pattern
- Algoritmus je pro texty rychlejší, než brute force
- Přesto jeho složitost může být *O(mn + A),* kde A je velikost abecedy
- Rychlý pro velké abecedy
- Preprocesing – zjistí, na které pozici má které písmena (směrem zleva)
  - Je-li patern např.: „ABRAKADABRA„
  - A dostane index 10, B dostane index 8, K = 4, D = 6 a R = 9. Pro ostatní písmena přiřadíme -1.
  - Naimplementujeme tedy funkci Last(char input), která podle písmene vrátí tento index, a pak pro případy 3 a 4 porovnáváme Last(*T(i)*) a *j*, a tím pak víme, jestli i posunout na Last(*T(i)*) (pokud je Last(*T(i)*) < *j*) nebo pouze *i++*

Interreg Rakousko-Česká republika
Evropský fond pro regionální rozvoj
Europäische Union
Evropská unie
Europäischer Fonds für regionale Entwicklung
Evropský fond pro regionální rozvoj
VŠTE
UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

Knuth-Morris-Pratt (KMP) algoritmus

- Prohledává text zleva doprava
- Nedělá všechna porovnání
- Pokud narazíme na neshodu, posune se o více než o jedno písmeno.
  - O maximální prefix *P(0 .. j-1)*, který je suffixem *P(1 .. j-1)*
  - Suffix a prefix se nesmí překrývat.
- Najdeme-li kus *P* (od začátku, tedy prefix), znaky tohoto prefixu odpovídají textu, není třeba je kontrolovat znovu
- Konec nalezeného podřetězce může být také obsažen v začátku tohoto podřetězce. Takovou shodou je samozřejmě celá nalezená část P, proto hledáme od P+1. Takže jdeme od konce nalezeného kusu P zleva a zprava, a ve chvíli, kdy nenalezneme shodu, víme, o kolik se můžeme posunout.
- Toto lze předpočítat do tabulky – pak je vše O(1)

Interreg Rakousko-Česká republika
Evropský fond pro regionální rozvoj
Europäische Union
Evropská unie
Europäischer Fonds für regionale Entwicklung
Evropský fond pro regionální rozvoj
VŠTE
UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

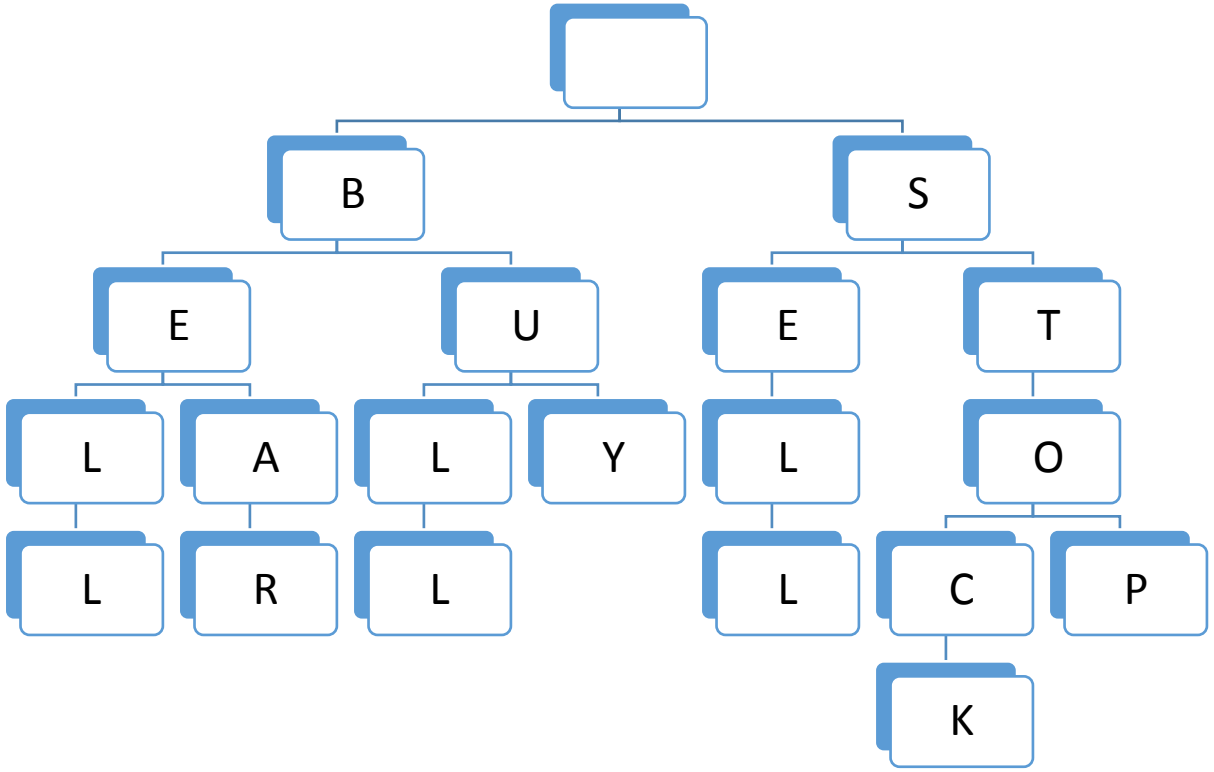
Trie

- Struktura pro předzpracování textu
- V každém uzlu je jednopísmeno
- Délka cesty z uzlu k vrcholu = pořadí písmene ve slovu
- Vyhledávání: *O(dm)*, kde *d* je velikost abecedy, *m* délka slova
- Do Trie je možné ukládat i celý text, v každém uzlu je pak jedno slovo

Interreg
Rakousko-Česká republika
Evropský fond pro regionální rozvoj
Europäische Union
Evropská unie
Europäischer Fonds für regionale Entwicklung
Evropský fond pro regionální rozvoj
VŠTE
UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

Trie

S={BELL, BEAR, BULL, BUY, SELL, STOCK, STOP}

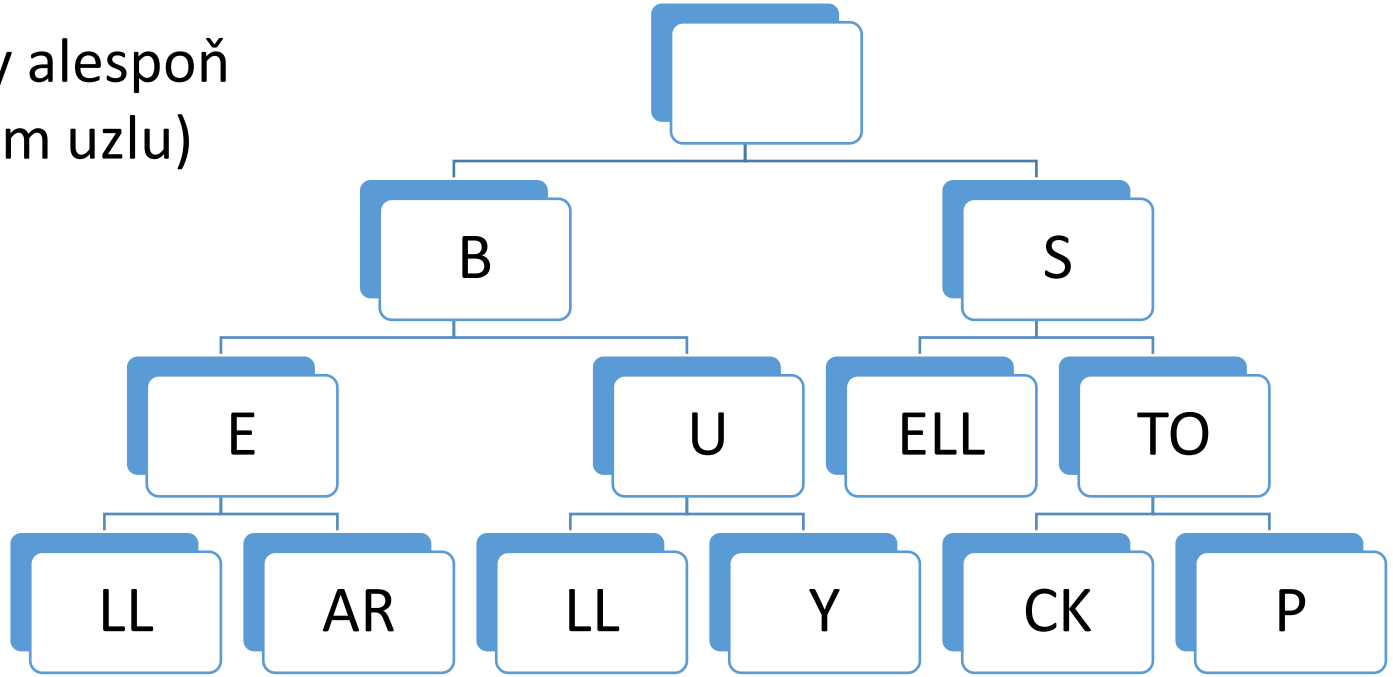


Interreg
Rakousko-Česká republika
Evropský fond pro regionální rozvoj
Europäische Union
Evropská unie
Europäischer Fonds für regionale Entwicklung
Evropský fond pro regionální rozvoj
VŠTE
UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

Komprimovaný trie

S={BELL, BEAR, BULL, BUY, SELL, STOCK, STOP}

Komprimovaný trie obsahuje uzly alespoň stupně dva (dvě písmena v jednom uzlu)



Interreg
Rakousko-Česká republika
Evropský fond pro regionální rozvoj
Europäische Union
Evropská unie
Europäischer Fonds für regionale Entwicklung
Evropský fond pro regionální rozvoj
VŠTE
UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA