Identifikácia previazania slov prostredníctvom ich rozloženia v dokumente

WIKT 2012

Tomáš Kučečka, Daniela Chudá, FIIT STU
kucecka@fiit.stuba.sk, chuda@fiit.stuba.sk

Problémová oblasť

Textové dokumenty, prirodzený jazyk

Zaujíma nás:

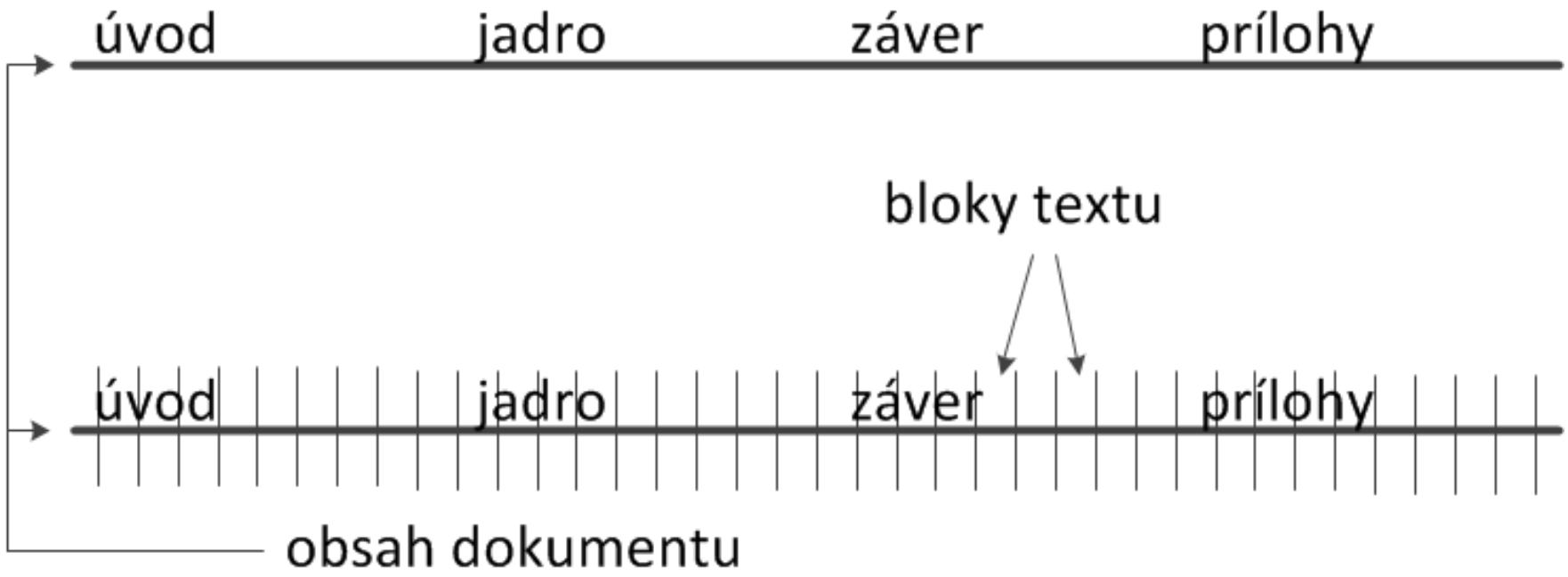
- téma dokumentu
- kľúčové slová, ktoré charakterizujú **časti** dokumentu

Problémy, ktoré adresujeme

- dokument je vnímaný ako celok, ktorý obsahuje len jednu tému
- prečo by sa všetky kľúčové slová mali týkať celého dokumentu

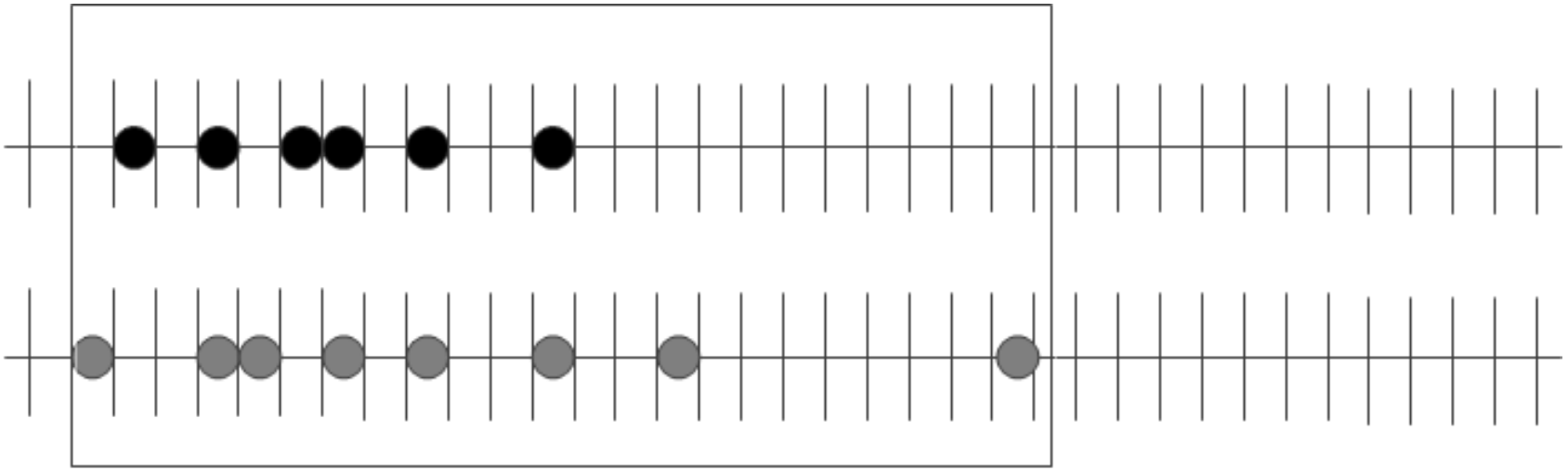
Návrh riešenia

1. Rozdelenie dokumentu na bloky textu pevnej dĺžky



Návrh riešenia

2. Sledovanie distribúcie kľúčových slov nad získanými blokmi



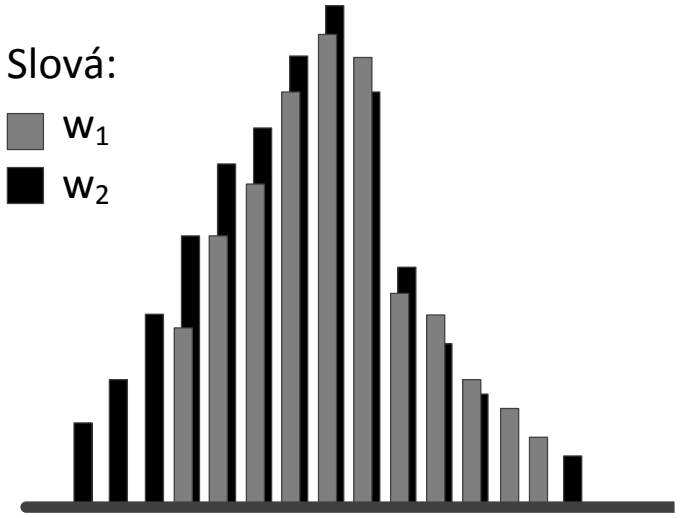
nakoľko sú si podobné z pohľadu distribúcie?

- výskyt kľúčového slova w_0
- výskyt kľúčového slova w_1

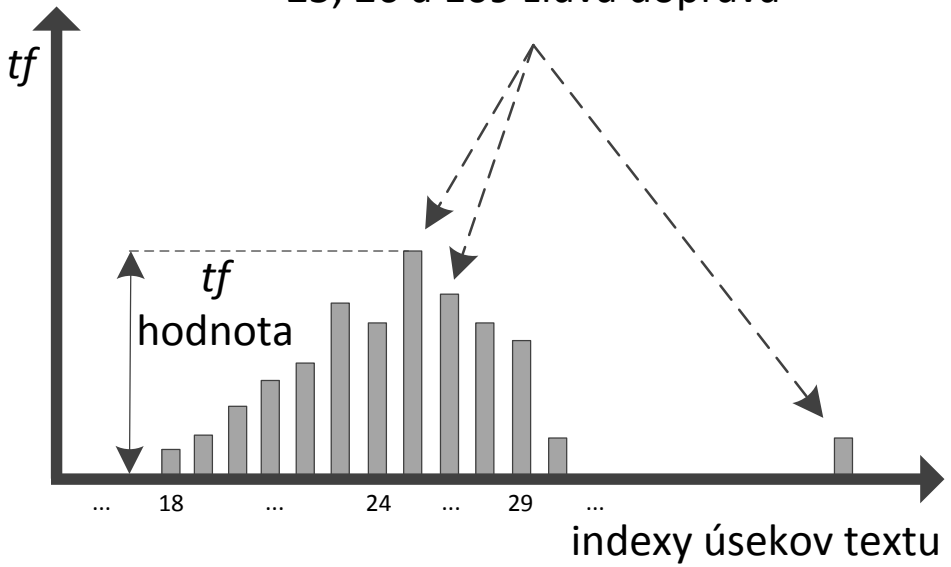
Porovnanie dvoch distribúcií

3. Porovnanie distribúcií kľúčových slov po dvojiciach Jensenovou–Shannonovou mierou podobnosti

a) multinomiálne rozdelenia slov nad úsekmi textu



b) slovo na úsekoch textu s indexmi 25, 26 a 109 zľava doprava



Očakávaný prínos

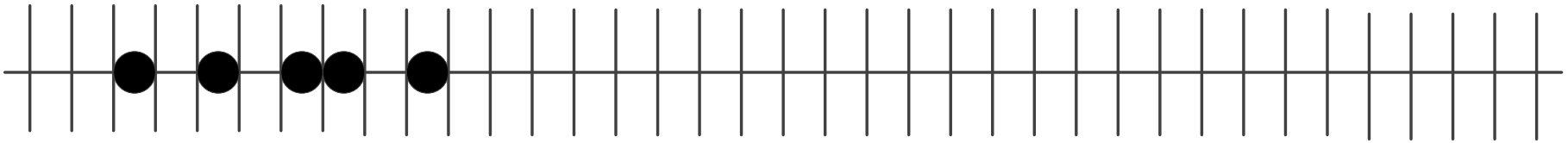
- Slová s podobným rozdelením:
  - tvoria segmenty
  - sú spolu vyskytujúce sa (často ich spolu nájdeme v texte) – previazanosť medzi dokumentmi
- Slová charakterizujú len tú časť dokumentu, kde sa vyskytujú
  - napr. iné slová v prílohe a iné v jadre dokumentu
  - slová charakterizujú celý dokument

Zhrnutie

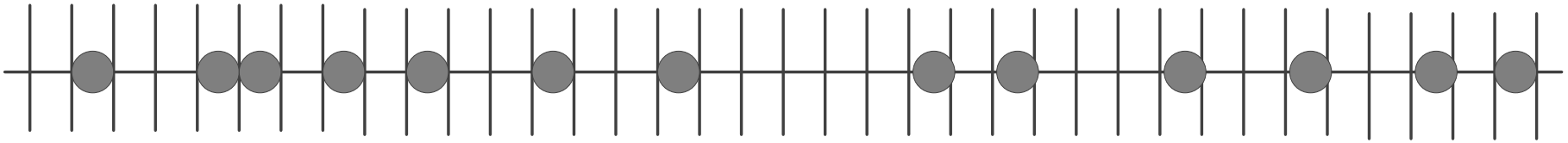
- Prístup k identifikácii previazanosti medzi slovami
- Prístup k segmentácii textu
- Prístup k identifikácii vzťahu slova k celkovému obsahu dokumentu

Aktuálny stav

- Dokončenie implementácie
- Experimentovanie
  - odhad parametrov - dĺžka bloku, hraničné hodnoty



- výskyt kľúčového slova w_0



- výskyt kľúčového slova w_1

segment