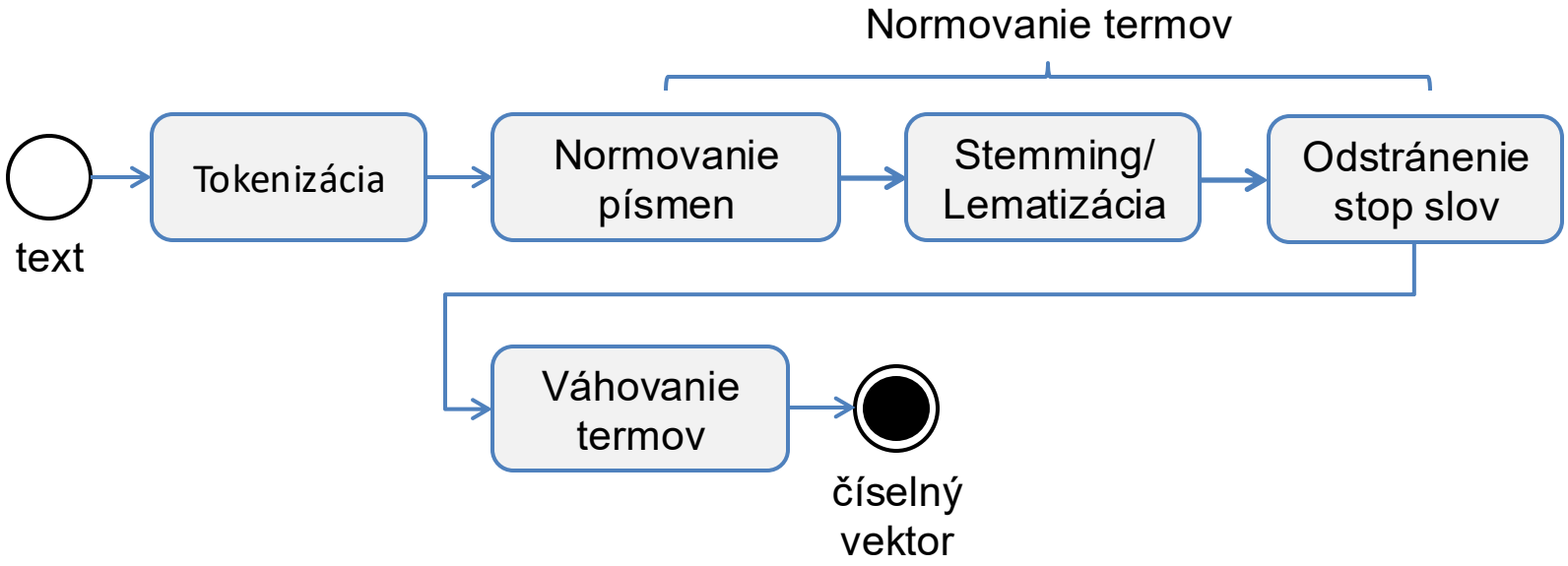
TEXT MINING 1

Objavovanie znalostí v textoch

Peter Bednár

Rozšírenie vektorovej reprezentácie dokumentov

Vektorová reprezentácia textov - indexovanie



Tokenizácia

- Rozdelenie textu na základné lexikálne jednotky - tokeny, ako napr. slová, interpunkčné znaky, číselné údaje, dátumy, internetové adresy a pod.
- Vstup: postupnosť znakov
- Výstup: postupnosť tokenov
- Pre európske jazyky sú slová oddelené medzerou alebo interpunkčnými znakmi - – ( ) , . ; : / „ “ ? ! …
  - Najčastejšie sa text rozdelí pomocou regulárnych výrazov
  - Výnimky sa následne ošetria podľa slovníka
    - Skratky, názvy, zložené výrazy
    - viac - menej → viac-menej

Tokenizácia - regulárne výrazy (1)

- Umožňujú kompaktne zapísať vzor pre postupnosť znakov
- Výskyt znaku

  . (bodka) - ľubovoľný znak

  [množina] - znak z množiny

  [^množina] - ľubovoľný znak okrem množiny

  \\ spätná lomka, \t tabulátor, \n nový riadok, \" úvodzovky
- Opakovanie znakov:

  ? - znak sa vyskytol 0 alebo raz

  * - znak sa vyskytol ľubovoľný počet krát (aj 0)

  + - znak sa vyskytol raz alebo viackrát

  {min, max} - znak sa vyskytol minimálne min a maximálne max krát, alebo {počet opakovaní}

Tokenizácia - regulárne výrazy (2)

- Kombinovanie výrazov

  výraz1 | výraz2 – výraz1 alebo výraz2

  (výraz) – zátvorky ohraničujú výraz pri kombinovaní

- Príklady:

  [a-záéíĺóŕúýčďľňšťžäô]+ – slovo v slovenskej abecede, všetky písmená malé, napr. vŕba

  [1-9][0-9]*(\.[0-9]*)? (€|EUR) – číselná hodnota v eurách, napr. 10.25 EUR

  [a-zA-Z0-9_-]+(\.[a-zA-Z0-9_-]+)*@([a-zA-Z0-9_-]+\.)+[a-z]{2,6} – emailová adresa, napr. peter.bednar@tuke.sk

Normovanie písmen

- Prevedieme znaky na malé písmená – výnimkou môžu byť názvy a skratky, kde sa môže striedaním veľkosti písmen meniť význam slova/termu
- Znormujeme znaky – napr. odstránime diakritiku, alebo v nemčine *umlaut* ß → ss, ae → ä → a, oe → ö → o, a pod.

Stemming

- Pre daný jazyk sa navrhnú pravidlá, ktoré prevedú rôzne tvary slova na ich spoločný koreň odstránením prípon a predpôn
- Relatívne presná metóda pre jazyky s jednoduchšou morfológiou, napr. pre angličtinu {clos-ing, clos-es, clos-er, clos-e} →clos+0
- Stemming nemusí byť jednoznačný, t.j. slová s rôznym významom môžu byť namapované na ten istý koreň
- Koreň už nemusí byť slovom (nemusí byť úplne zrozumiteľný)
- Pravidlá majú tvar [podmienka] S1 → S2: ak je splnená podmienka (napr. ak slovo končí na -s alebo obsahuje spoluhlásku), potom nahraď predponu/príponu S1 reťazcom S2
- Pre angličtinu sa najčastejšie používa Porterov stemmer

Porterov stemmer - príklady pravidiel

1a - Odstránenie prípon -s, -es

S →	cats → cat
IES → I	ponies → poni,
SSES → SS	caresses → caress

1b - Odstránenie prípon -d, -e, -ing

EED → EE	agreed → agree
ED →	plastered → plaster
ING →	plastering → plaster

1c - Zmena prípony y na i

Y → I	ponny → ponni

2 - Zmena prípon

ATIONAL → ATE	relational → relate

Lematizácia

- Každé slovo sa prevedie na základný tvar - lemu (neurčitok pre slovesá, 1. pád jednotného čísla pre podstatné mená, atď.)
- Napr.: peknej → pekný, prípadov → prípad
- Slovenčina má veľa prípon, nepravidelné slová, rôzne alternácie v základe slova (napr. otec, otcovi - bez e)
- Pre jazyky so zložitou morfológiou sa používa kombinácia modelov strojového učenia a morfologického slovníka, ktorý mapuje tvary slova na ich základný tvar
- Lematizácia môže byť nejednoznačná
  - Jeden tvar môže byť namapovaný na viacero lem (napr. mier → mier, alebo mieriť)

Odstránenie stop slov

- Ak zanedbáme poradie slov, najdôležitejšie pre reprezentáciu významu textu sú slovesá, podstatné mená a prídavné mená
- Odstránením neplnovýznamových slov sa zníži rozmer príznakového priestoru (neplnovýznamové slová spôsobujú pri modelovaní dátový šum)
- Odstránia sa slová zo slovníka tzv. stop slov
  - Spojky, predložky, častice, zámená, pomocné slovesá
  - Napr. pre slovenčinu: a, aby, aj, ako, ale, alebo, ani, áno, asi, bez, by, byť, cez, čo, či, dnes, do, další, ešte, ho, i, iba, ja, je, jeho, jej, k, kam, každý ...
- Môžu sa odstrániť aj čísla, dátumy, a pod., ako napr. 1 000, 26.10. 2015

Vektorová reprezentácia textov - príklad

Vstupný text:

Hodnota denného dovozu dosiahla iba 10 059 mil. barelov.

1. Tokenizácia:

Hodnota denného dovozu dosiahla iba 10 059 mil. barelov .

2. Normalizácia veľkosti písmen

hodnota denného dovozu dosiahla iba 10 059 mil. barelov .

3. Lematizácia

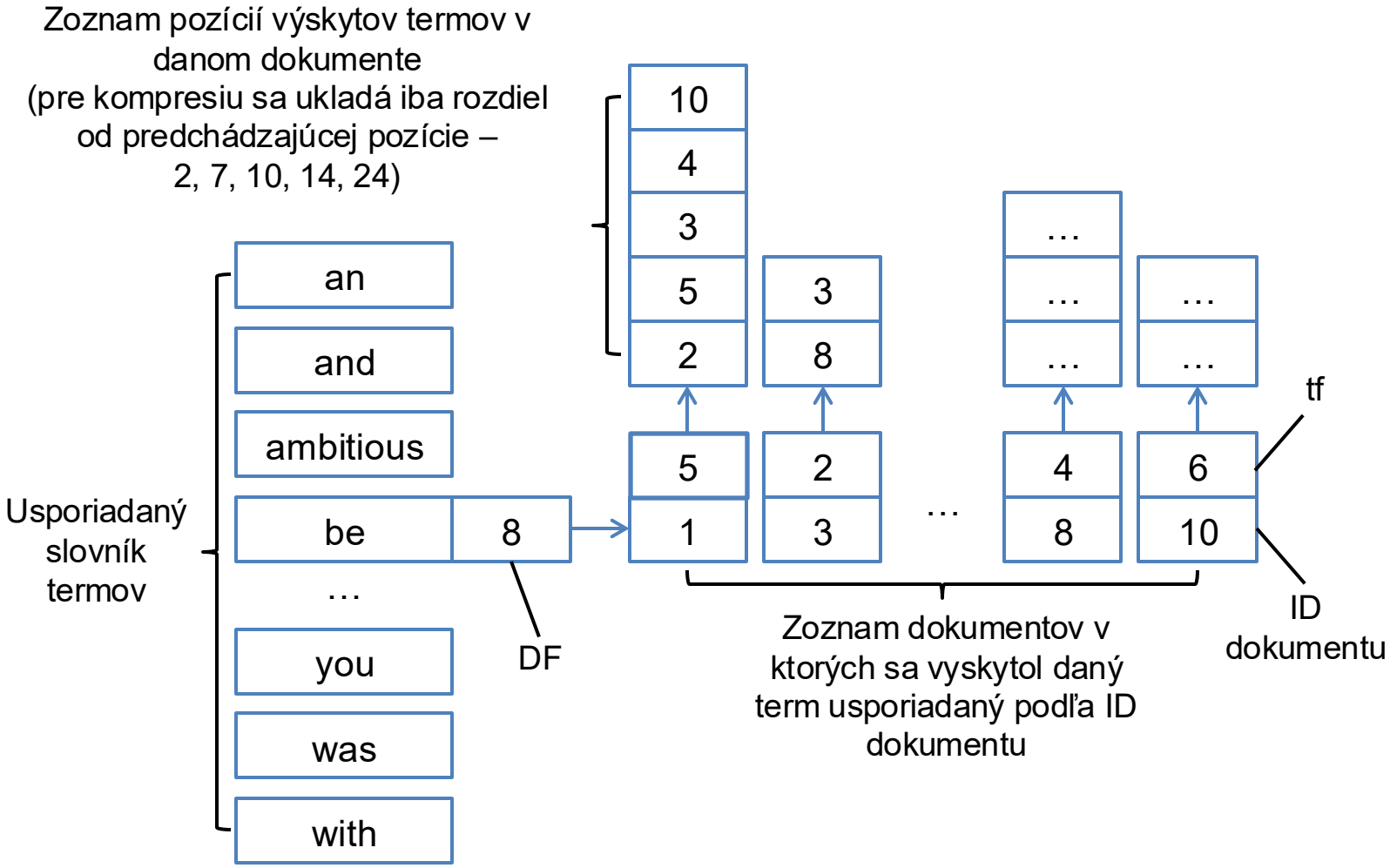
hodnota denný dovoz dosiahnuť iba 10 059 mil. barel .

4. Odstránenie stop slov

hodnota denný dovoz dosiahnuť barel

{barel, denný, dovoz, dosiahnuť, hodnota}

Rozšírený invertovaný index



Dopytovanie nad rozšíreným indexom

- Boolovské operátory
  - AND, OR, NOT
- Dokument môže byť rozdelený na sekcie (napr. nadpis, abstrakt, ...) a je možné ohraničiť v ktorej sekcii sa term vyskytuje/nevyskytuje
  - title:term
- Frázy a vyhľadávanie termov v okolí
  - "umelá inteligencia", "umelá inteligencia"~10
- Približné vyhľadávanie termov podľa editačnej vzdialenosti a regulárnych výrazov
  - te?t, te*t, text~

Výhody vektorovej reprezentácie textov

- Jednoduchý algebrický model reprezentácie textu
- Efektívny výpočet relevantnosti dokumentu voči dopytu
  - Vďaka indexovaniu pomocou invertovaných indexov
- Dobrá škálovateľnosť
  - Vyhľadávanie miliárd dokumentov / 10-150 ms na dopyt
  - Indexovanie viac než 100 000 dokumentov / 1s

Nevýhody vektorovej reprezentácie textov

- Zanedbáva sa poradie slov
  - tzv. nevieme reprezentovať veľkú časť významu textu
- Nejednoznačnosť termov
  - To isté slovo môže mať viacero významov a záleží na kontexte – každý význam by mal byť indexovaný ako samostatný term
  - Zhoršuje sa presnosť vyhľadávania
- Mnohotvárnosť pojmov
  - Jeden pojem môže byť reprezentovaný viacerými rôznymi slovami resp. slovnými spojeniami
  - Jedno slovo môže mať viacero tvarov – tokenizácia/lematizácia
  - Zhoršuje sa návratnosť vyhľadávania

Kontrolované slovníky

- Zoznam termov s jednoznačne definovaným významom používaných pri indexovaní
- Jeden pojem/koncept môže byť popísaný viacerými synonymickými termami (slovami, alebo slovnými spojeniami)
- Tezaurus – koncepty sú usporiadané v hierarchiách s reláciou všeobecnejší/špecifickejší
- Ontológie/znalostné grafy – medzi konceptami môžu byť aj ďalšie typy relácií

MeSH – Medical Subject Headings

- Tezaurus medicínskych pojmov https://meshb.nlm.nih.gov
- Približne 30 000 pojmov usporiadaných do viacnásobnej hierarchie (jeden pojem môže byť zaradený do viacerých stromových hierarchií), 177 000 termov
- Používa sa na indexovanie digitálnej knižnice MEDLINE/PubMed – 39 mil. citácií medicínskych textov z vedeckých časopisov a kníh, 2000-4000 záznamov denne

Wordnet

- Lexikálna ontológia https://en-word.net – databáza anglických slov ktoré sú zoskupené podľa významu a prepojené reláciami
- Synset – reprezentuje jednoznačný pojem, viac než 177 000 synsetov / 155 000 slov
- Jeden synset môže zoskupovať viac rôznych slov (synoným)
- Jedno slovo môže byť vo viacerých synsetoch (môže mať viacero významov)
- Synsety sú prepojené sémantickými reláciami
  - napr. všeobecnejší/špecifickejší, je časťou/má časť
- Na základnú anglickú verziu sa mapujú ostatné jazyky (napr. slovenský Wordnet http://korpus.juls.savba.sk/WordNet.html – takmer 25 000 synsetov)

Výhody kontrolovaných slovníkov

- Jednoznačne definovaný význam termov – zlepšuje sa presnosť vyhľadávania
- Vyhľadáva sa aj výskyt synoným
- Hierarchické usporiadanie umožňuje rozšíriť vyhľadávanie o špecifickejšie pojmy (expandovanie dopytov) – zlepšuje sa návratnosť vyhľadávania

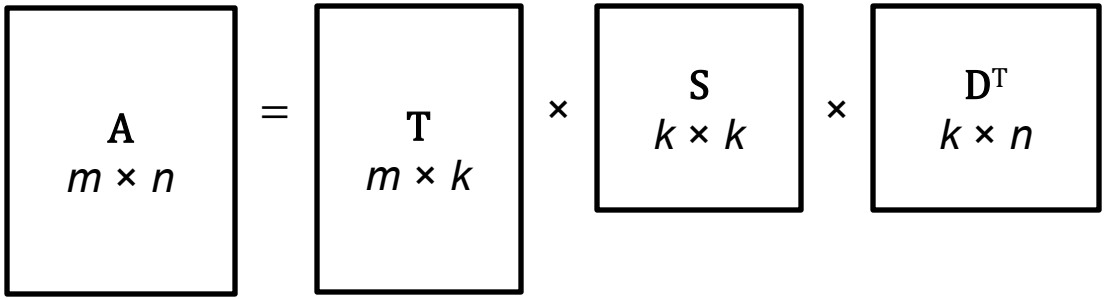
Nevýhody kontrolovaných slovníkov

- Je potrebné jednoznačne určiť výskyt termov pri indexovaní – pri manuálnom indexovaní veľká prácnosť
- Slovník je potrebné revidovať (rozširovať o novú terminológiu, odstrániť zastaralé/nepoužívané výrazy)
- Hierarchické usporiadanie termov nemusí zodpovedať informačným potrebám používateľa

Latentné Sémantické Indexovanie (LSI)

- Latentné Sémantické Indexovanie je založené na dekompozícii term-dokument matice na singulárne hodnoty a vektory

- - term-dokument matica o rozmeroch m - termov $\times$ n - dokumentov
- $\mathbf{T}$ - matica pravých singulárnych vektorov o rozmeroch
- - diagonálna matica kladných singulárnych hodnôt zoradených podľa veľkosti $s_1 \geq s_2 \geq \ldots \geq s_k \geq 0$ o rozmeroch
- - matica pravých singulárnych vektorov o rozmeroch $n \times k$



Interpretácia singulárnych hodnôt a vektorov (1)

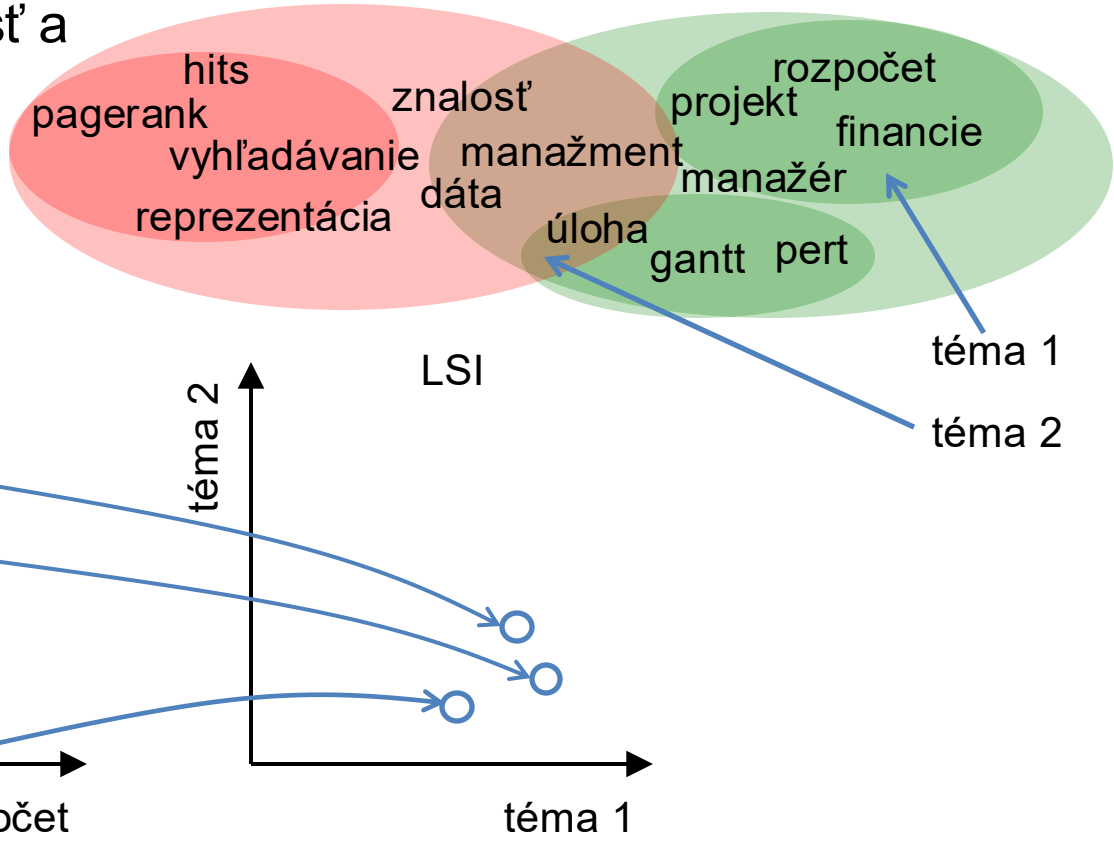
- Singulárne hodnoty a vektory rozkladajú maticu $\mathbf{A}$ na k komponentov, ktoré je možné prirovnať k „témam“ v dokumentoch
- Váha termu i pre dokument j sa rozloží podľa:

$$A_{i,j} = \underbrace{T_{i,1}s_1D_{j,1}}_{\text{z témy 1}} + \underbrace{T_{i,2}s_2D_{j,2}}_{\text{z témy 2}} + \cdots + \underbrace{T_{i,k}s_kD_{j,k}}_{\text{z témy } k}$$

- $T_{i,t}$ určuje, do akej miery term i vyjadruje tému t (jeden term môže vyjadrovať viacero tém a jedna téma môže byť vyjadrená viacerými termami)
- $D_{j,t}$ určuje, do akej miery dokument j obsahuje tému t (jeden dokument môže obsahovať viacero tém)
- $S_{t,t} = s_t$ určuje, ako je téma t dôležitá pre celú množinu dokumentov (maticu $\mathbf{A}$)

Interpretácia singulárnych hodnôt a vektorov (2)

- Pri LSI sa do spoločnej témy priradia termy, ktoré sa často v dokumentoch vyskytujú spoločne – vyhľadávanie v priestore LSI môže zlepšiť presnosť a návratnosť



Vyhľadávanie dokumentov v LSI

- Pri indexovaní nahradíme termy za témy, pričom zachováme iba r najdôležitejších tém, t.j. $(k - r)$ najmenších singulárnych hodnôt vynulujeme a vynecháme z reprezentácie

$$\boldsymbol{A} \approx \mathbf{T}_r \mathbf{S}_r \mathbf{D}_r^{\mathrm{T}}$$

- Transformované dokumenty sú uložené v riadkoch matice $\mathbf{D}_r$, ktorá má po redukcii r stĺpcov
- Vektor dopytu musíme najprv pretransformovať z TF-IDF reprezentácie na vektor tém:

$$\boldsymbol{q}_{\mathrm{LSI}} = \boldsymbol{q}_{\mathrm{TF-IDF}} \mathbf{T}_r \mathbf{S}_r^{-1}$$

- Podobnosť medzi transformovaným dopytom a dokumentami sa počíta rovnako ako pre TF-IDF váhovanie, napr. kosínusovou mierou podobnosti

Výhody LSI

- Môže zlepšiť návratnosť vyhľadávania – stačí ak dokument obsahuje súvisiace termy, ktoré sa vyskytujú často v kontexte termu ktorý je zadaný v dopite alebo jeho synonymá
- Môže zlepšiť presnosť vyhľadávania pretože polysémické slová s rôznym významom sú premietnuté podľa kontextu do iných latentných dimenzií
- Redukciou term-dokument matice sa potlačí šum spôsobený nepodstatnými termami

Nevýhody LSI

- Je potrebné vypočítať singulárnu dekompozíciu term-dokument matice, ktorú je zložité prepočítať pri indexovaní nových dokumentov
- Výsledná reprezentácia už nie je riedka matica, tzn. nie je možné použiť invertovaný index
  - Je potrebné vytvoriť zložitejší index nad vektormi pre rýchle vyhľadávanie k najbližších susedov
- Jednotlivé latentné dimenzie už nie je možné jednoducho interpretovať ako termy

Spracovanie prirodzeného jazyka

Problémy pri spracovaní prirodzeného jazyka

- Medzi hlavné problémy pri spracovaní prirodzeného jazyka patrí nejednoznačnosť a mnohotvárnosť
- Ďalším problémom je viacjazyčnosť.
  - V Európskej únii je 24 oficiálnych jazykov
  - Celkovo viac než 200, asi 60 je ohrozených
  - 51% dospelých občanov EU „rozumie“ angličtine, najpočetnejší materinský jazyk je nemčina (18%), slovenčinou hovorí 1%,
  - Ale napr. až 60% zákazníkov odradí online nakupovanie, ak nie je v ich rodnom jazyku
- Chyby, nespisovnosť

Nejednoznačnosť - úroveň znakov

- Nejednoznačnosť sa vyskytuje na všetkých úrovniach jazyka.
- Napr. už pri rozdelení textu - t.j. postupnosti znakov na slová
  - postupnosť Jean-Claude, hi-fi, MS-DOS označuje jedno slovo
  - ale napr. postupnosť Košice-Bratislava by mala byť rozdelená na Košice, pomlčka - a Bratislava
  - a naopak, postupnosť 1 000 000 € tvoria dve slová, číslovka 1 000 000 (hoci obsahuje medzery) a znak meny

Nejednoznačnosť - úroveň slov

- To isté slovo môže mať veľa významov
  - rys - zviera/výkres, čelo - časť tváre/hudobný nástroj (ide o tzv. homonymá, rovnako sa píšu, ale významovo nemajú nič spoločné)
  - oko - orgán zraku/reťaze/v polievke/na ponožke (ide o tzv. viacvýznamové slová odvodené na základe nejakej podobnosti popisovaného pojmu - napr. okrúhleho tvaru)

Nejednoznačnosť - úroveň viet

- Vo vete nie je jasné, aké sú vzájomné vzťahy medzi jednotlivými slovami
  - Firma Oracle odkúpila spoločnosť Sun po tom, čo sa zdvojnásobil jej obrat.
  - Nie je jasné, či sa zdvojnásobil obrat firme Oracle, alebo firme Sun

Nejednoznačnosť - prenesený význam

- Úmyselná nejednoznačnosť
  - Nejednoznačnosť sa často využíva napr. v reklamných sloganoch

- Slovné spojenia, ktorých doslovný význam sa nepoužíva
  - Prirovnania
- Slovné spojenia, pri ktorých autor zamýšľa iný než základný význam
  - Napr. použitie irónie, niektorých eufemizmov/vulgarizmov, slangových výrazov
  - Posunutie významu je kultúrne závislé – môže platiť iba pre určitú skupinu ľudí, napr. mládež

Mnohotvárnosť - úroveň slov (1)

- To isté slovo sa môže vyskytovať vo viacerých tvaroch
  - Hlavne pre jazyky s bohatou morfológiou, kam patrí aj slovenčina
  - Napr. pre prídavné mená máme 7 pádov x 3 osoby x 2 čísla = 42 možností, ktoré sa môžu odlišovať príponou – dobrá, dobrej, dobrému, ...
- Ten istý význam môžeme vyjadriť rôznymi slovami
  - získať/nadobudnúť - ide o tzv. synonymá
  - Zväčša nie sú univerzálne zameniteľné, v niektorých vetách vyjadrujú rovnaký resp. dostatočne podobný význam, a v niektorých sa už význam príliš posunie, alebo je lepšie uprednostniť jedno slovo kvôli štylistike (napr. nepoužívať slang)

Mnohotvárnosť - úroveň slov (2)

- Ten istý význam môžeme vyjadriť rôznymi symbolmi
  - Používanie skratiek
  - Emotikony

Mnohotvárnosť - úroveň viet (1)

- Voľný slovosled vo vete, kedy sa zmení poradie niektorých slov, ale nezmení sa význam
  - Firma Oracle odkúpila spoločnosť Sun po tom, čo sa zdvojnásobil jej obrat.
  - Firma Oracle odkúpila spoločnosť Sun po tom, čo sa jej obrat zdvojnásobil.
  - Po tom, čo sa jej obrat zdvojnásobil, bola firma Sun odkúpená firmou Oracle.
    - Ale tu už môže dôjsť k posunu významu, zdvojnásobil sa obrat spoločnosti Sun

Mnohotvárnosť - úroveň viet (2)

- Pre čo najefektívnejšiu komunikáciu používame čo najmenej slov – t.j. často sú slová vo vete vynechané a vetu je možné pochopiť iba podľa jej kontextu
  - Používanie zámen: Firma Oracle odkúpila Sun. Predtým sa však zdvojnásobil jej [znova môže byť priradenie nejednoznačné, aj keď najpravdepodobnejšie Oracle] obrat.
  - Nevyjadrený podmet: Firma Sun zdvojnásobila svoj obrat. Bola [firma Sun] však neskôr odkúpená firmou Oracle.
  - Vynechávanie známych faktov v rozhovore:
    - Kúpil si si iPhone? Áno - neodpovedáme celou vetou

Mnohotvárnosť - úroveň viet (3)

- A ďalšie spôsoby PARAFRÁZOVANIA
  - Koľkými spôsobmi viete vyjadriť spokojnosť/nespokojnosť s prednáškou?

Dôležitosť kontextu pre porozumenie textu

- Závisí na okolitom texte
  - na ostatných slovách vo vete, čo bolo uvedené v ostatných vetách, niekedy aj na netextových častiach – schémy, grafy, vzorce a pod.
- Niekedy je porozumenie závislé od toho, kto s kým hovorí
  - Napr. použitie slangu, nárečia, rozhovor rodiča s dieťaťom
- Niekedy závisí na situácii
  - Pomoc!
- A hlavne, človek používa uvažovanie na zjednoznačnenie významu a doplnenie chýbajúcich informácií
  - „Zdravý rozum“ – všeobecné znalosti
  - Ale aj veľmi špecifické znalosti pri odborných textoch

Jazyk a myslenie

- Práve súvislosť medzi myslením a pochopením jazyka robí počítačové spracovanie jazyka veľmi ťažkým

- **Ak dokáže stroj komunikovať v prirodzenom jazyku, ľudia ho budú považovať za inteligentného!**