# Vyhľadávanie vzťahov v neštruktúrovanom texte pomocou sémantickej analýzy

Bc. Martin Jačala

Vedúci výskumného projektu: Mgr. Jozef Tvarožek

# Motivácia

- Veľké množstvo dostupných textov
  - Novinové články, blogy, ...
  - Opisujú osoby, miesta, názvy a pod.
- Veľká miera viacznačnosti

# Problém

- Výskyty entít v texte majú rôzny význam
  - Jaguar (zviera, automobil)
  - Michael Jordan (vedec, hráč NBA)
  - Big Blue (IBM, film)
- Bez kontexu nedokážeme vybrať správny význam

# Cieľ práce

- Navrhnúť a overiť metódu, ktorá rozlíši významy menných entít
- Rozlíšené významy sú identifikované vytvorením prepojenia do externého slovníka
- Využitie metód spracovania prirodzeného textu

# Cieľ práce

**Jaguar** (Panthera onca) is a big cat, a feline in the Panthera genus

http://en.wikipedia.org/wiki/Jaguar

http://en.wikipedia.org/wiki/Jaguar_Cars

Jaguar Cars Ltd., better known simply as **Jaguar**, is a British luxury car manufacturer

# Prehľad oblasti

- Porovnávanie dokumentov medzi sebou
  - Zhlukovanie rovnakých entít
  - Systém *Nominator*
- Špecifické skupiny problémov, slovníky
  - Rozlišovanie autorov vedeckých prác
  - Geografické lokality, *InfoXTract*

# Prehľad oblasti

- Dostupnosť veľkých textových korpusov
  - Wikipedia, dmoz.org
- Rozlišovanie na základe Wikipédie, 2006
  - Výpočet sémantickej podobnosti
  - Najlepšie metódy - presnosť 90%

# Formalizácia problému

- Problém môžme chápať ako hodnotenie vhodnosti jednotlivých významov

$$sim = \arg\max_d \ rank(q, d)$$

- pre každý z potencionálne možných významov

# Sémantická podobnosť

- Miera významovej podobnosti medzi termami (dokumentmi)
- Ako veľmi súvisí slovo A so slovom B
- Viaceré možnosti výpočtu
  - Graf, LSA, ESA, ...
- Podobnostné metriky

# Návrh riešenia

- Dva hlavné kroky
  - Predspracovanie a objavenie entít
  - Nájdenie vzťahu slova s jeho významom

# Určenie významu slova

- Nájdeme všetky vhodné významy z Wikipédie
  - Rozlišovacie stránky, presmerovania, hypertextové odkazy
- Určíme sémantickú podobnosť entity každým z významov (ESA)
- Vytvorenie vzťahu s najpodobnejším

# Explicitná sémantická analýza

- Východisko v LSA
- Pôvodne pre generovanie konceptov z textu
- Predpoklad - každý vstupný dokument reprezentuje koncept
- Výstup - odhad podobnosti dvoch slov vrámci sémantického priestoru

# Výpočet ESA

| | Jaguar Cars | Jaguar animal | … | OS X Jaguar |
|---|---|---|---|---|
| amazonia | 0 | 32945.5 | | 0 |
| wilderness | 120.43 | 11283.4 | | 0 |
| apple | 0 | 3220.8 | … | 44352 |
| industry | 1145.3 | 0 | | 3324.45 |
| car | 11120.11 | 395 | | 0 |
| … | | | | |
| music | 44 | 0 | … | 4356.5 |

# Výpočet podobnosti

- Pre každý dokument z množiny významov vypočítame podobnosť s vstupným textom

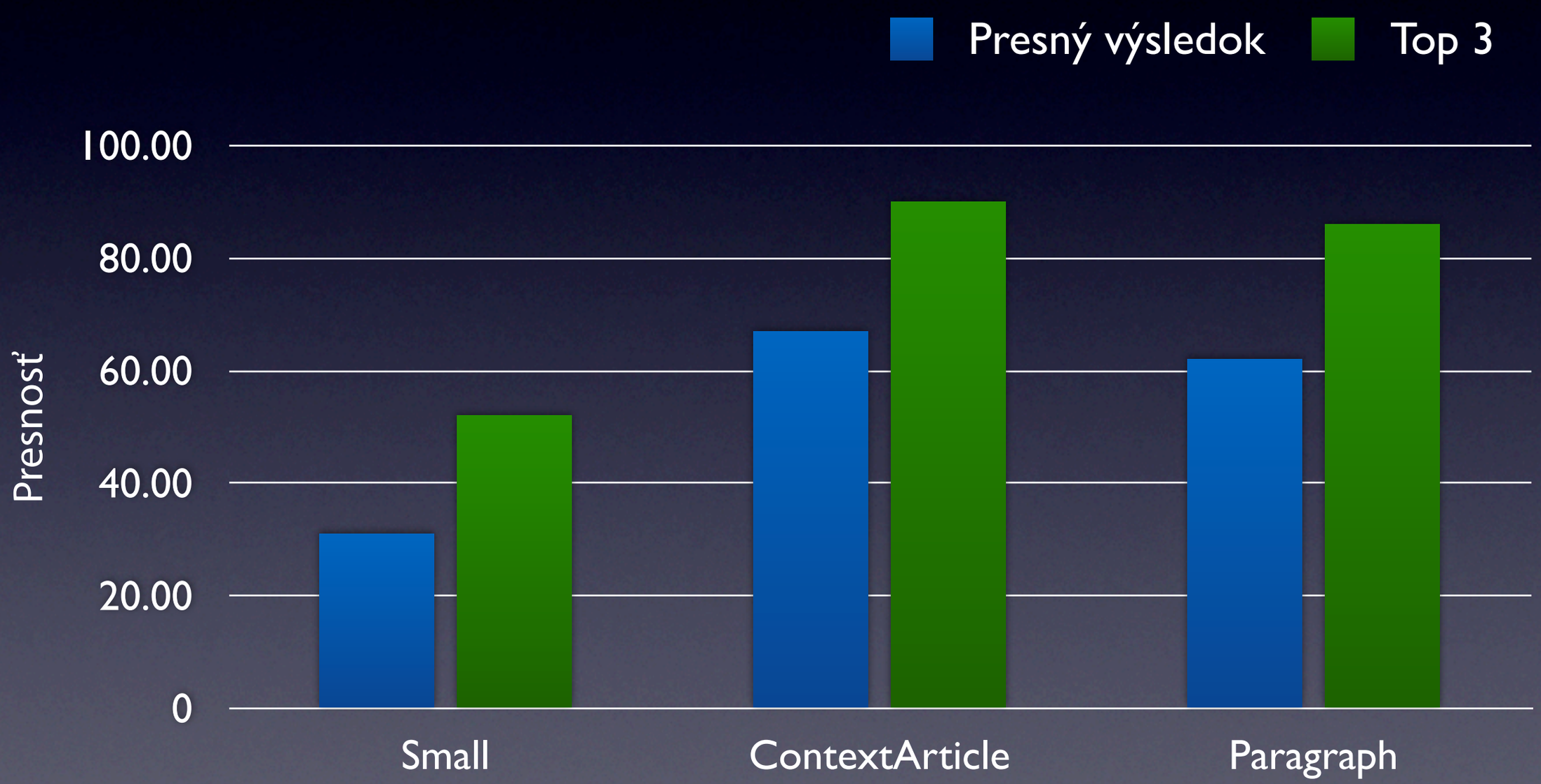
$$rank(q,d) = \frac{Q \cdot C_d}{\|Q\| \|C_d\|} = \frac{\sum_{i=1}^{n} Q_i \times C_{d_i}}{\sqrt{\sum_{i=1}^{n} (Q_i)^2} \times \sqrt{\sum_{i=1}^{n} (C_{d_i})^2}}$$

- Dokumenty reprezentujeme ako vektory
- Vypočítame kosínusovú podobnosť

# Overenie metódy

- Nie je štandardizovaná dátová vzorka pre overenie
- Používajú sa voľne dostupné texty
  - Niekoľko desiatok článkov
- Manuálne označkovať texty, porovnať s výsledkami metódy

# Dosiahnuté výsledky



# Ďalšia práca

- Porovnanie s metódou LSA
- Vplyv veľkosti sémantického priestoru
- Dodatočná metrika pre rozlíšenie najpravdepodobnejších kandidátov
- Implementácia demonštračného systému