Porovnávanie dokumentov s využitím znalostí na pozadí

Bc. Martin Kováčik

Vedúci projektu: Mgr. György Frivolt

Motivácia

- Veľké množstvo informácií (stále rastie)
- Hlavný problém:
  - efektívne vyhľadávanie
  - filtrovanie
  - integrácia dát
  - kategorizácia dát a klastrovanie
- Jeden z potrebných nástrojov:
  Určenie podobnosti
- Cieľom je určovanie a ohodnocovanie podobnosti dokumentov

Motivácia

- Jeden z pohľadov na podobnosť:

  Tematický prekryv informácií prezentovaných v porovnávaných dokumentoch

- Iná formulácia:

  Dokument o autach je podobnejší dokumentu o motorkách, ako dokumentu o programovaní

- Manuálne ohodnocovanie podobnosti je príliš náročné a nedeterministické

- Cieľ: Aproximácia ľudského spôsobu vyhodnocovania podobnosti

Problémy

- Prirodzený jazyk
  - Prirodzený jazyk je viacznačný
  - Viaceré spôsoby vyjadrenia rovnakej informácie
  - Veľa nadbytočných prvkov a tvarov
- Interpretácia prirodzeného jazyka z pohľadu podobnosti
- Model dokumentu a vyhodnotenie podobnosti

Interpretácia informácií v dokumente

- Dôraz na vyzdvihnutie obsahovej hodnoty dokumentu

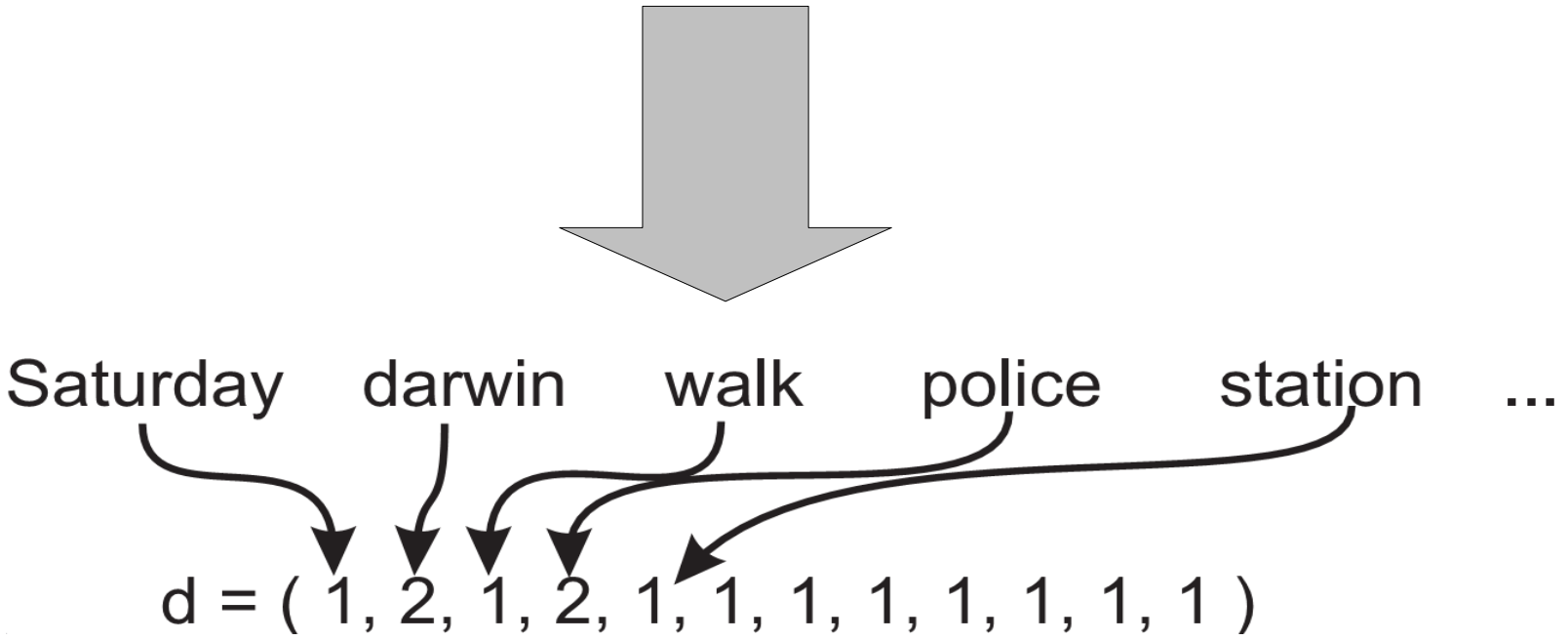
- Množina prepojených konceptov
  - Vytvára graf
  - Vhodná pre získavanie znalostí

- **Zastúpenie konceptuálnych prvkov**
  - Dokument ako zastúpenie znakov - Termov
  - Bag-Of-Words
  - Predpoklad:

    Rovnaké témy sú vyjadrené prostredníctvom rovnakých slov

BOW – Vektorová reprezentácia

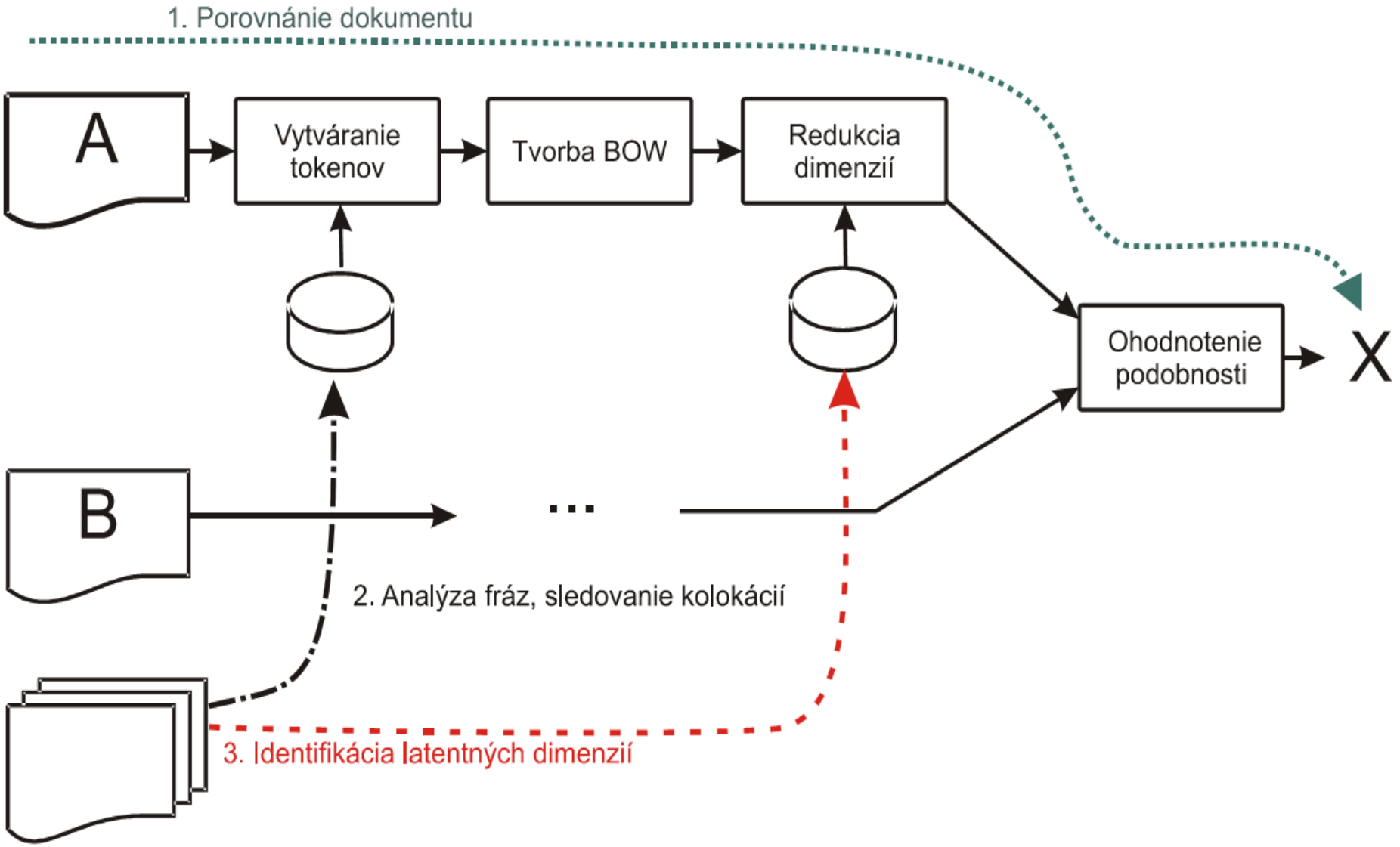
...But on Saturday Darwin walked in to a police station in central London. Darwin told police that he did not remember where he had been for last five years...



BOW – Vektorová reprezentácia

- Problémy
  - Veľká dimenzionalita
  - Vektory sú riedke
- Viacero metód vychádzajúcich z BOW reprezentácie dokumentu
  - Redukcia dimenzií
  - Kompaktnejšie vektory
- Výsledkom analýzy metód je zovšeobecnený proces zostavenia BOW a jeho použitia na porovnanie podobnosti

Proces porovnávania



Vytváranie tokenov

- Token = Slovo
  - *Stemming / Lematizácia* – Použitý Porter Stem (SnowBall)
  - *Stop Words* – Majú zanedbateľný vplyv na vyhodnotenie podobnosti
    - Jazykovo špecifické

      a, an, and, are, as, the, then, there, these, they, this

      (Apache Lucene)
    - Doménovo špecifické

      affected, affecting, again, chem, importance, usefullness

      (MEDLINE Bibliography Database)
  - Databázy zostavované expertne a na základe štatistických vlastností
    - Overovanie štatistického vplyvu na ohodnocovanie podobnosti

Vytváranie tokenov

- Token = N-gram (postupnosť n-slov)
  - Čiastočne zohľadnený kontext
  - Problém: náhodná kolokácia vs. viacslovné pomenovanie
  - Niektoré slová sa vyskytujú častejšie spolu so špecifickým slovami

    Los Angeles, San Francisco

  - Ako s inými slovami

    The post, The computer

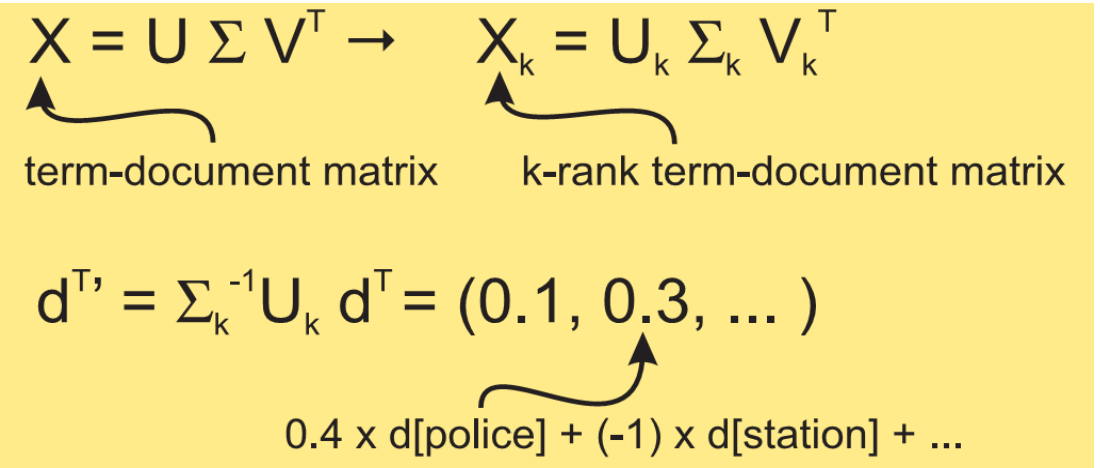
  - Využívaný korpus ***WortSchatz Universität Leipzig***
    - Štatistické informácie o kolokáciách

Tvorba BOW

- Sledovaný výskyt termov
- Použité metriky
  - Term Frequency
  - Binárne ohodnotenie prítomnosti
- Možné priame porovnanie
- Veľký počet dimenzií a dimenzie s malým vplyvom na identifikáciu tém

Transformácia BOW

- Cieľ: Vytvorenie vektora s menším počtom dimenzií
- Vytváranie syntetických (latentných) dimenzií
  - Zhlukovanie synoným a príbuzných slov
  - Znižovanie vplyvu častých slov
- Aplikovaná metóda LSA (Latent Semantic Analysis)
  - Využíva sa SVD



Transformácia BOW

- Problémy
  - Veľká náročnosť na pamäť a výpočtové prostriedky
  - Výpočtová náročnosť:

$$O(min(mn^2, nm^2))$$

- Prínosy
  - Identifikácia synonym
  - Riešenie problému s viacznačnými slovami

Určenie podobnosti

- Vzdialenostné metriky
  - COS – Preferovaná v IR
  - Euler, Manhatan, ...

- Metrika založená na teórií informácie

$$ItSim(A,B)=\frac{I(common(A,B))}{I(desc(A,B))}=\frac{2\cdot\sum_t min\{p_{A,t},pB,t\}\log\pi(t)}{\sum_t p_{A,t}\log\pi(t)+\sum_t p_{B,t}\log\pi(t)}$$

Vyhodnotenie výkonnosti

- Realizované nad množinou n-dokumentov
- Každá dvojica je ohodnotená podobnosťou
- p_k – dvojica dokumentov s k-tou najväčšou podobnosťou

$$precision(p_k) = \frac{\#(\forall\ p_j : j < k \wedge ITP(p_j))}{k} = \frac{\#\,Relevantné\ výsledky}{\#\,Všetky\ výsledky}$$

- ITP ~ binárna podobnosť
  - Dva dokumenty sú binárne podobné ak majú aspoň jeden spoločný znak (kľúčové slovo, kategória, ...)

Vyhodnotenie výkonnosti

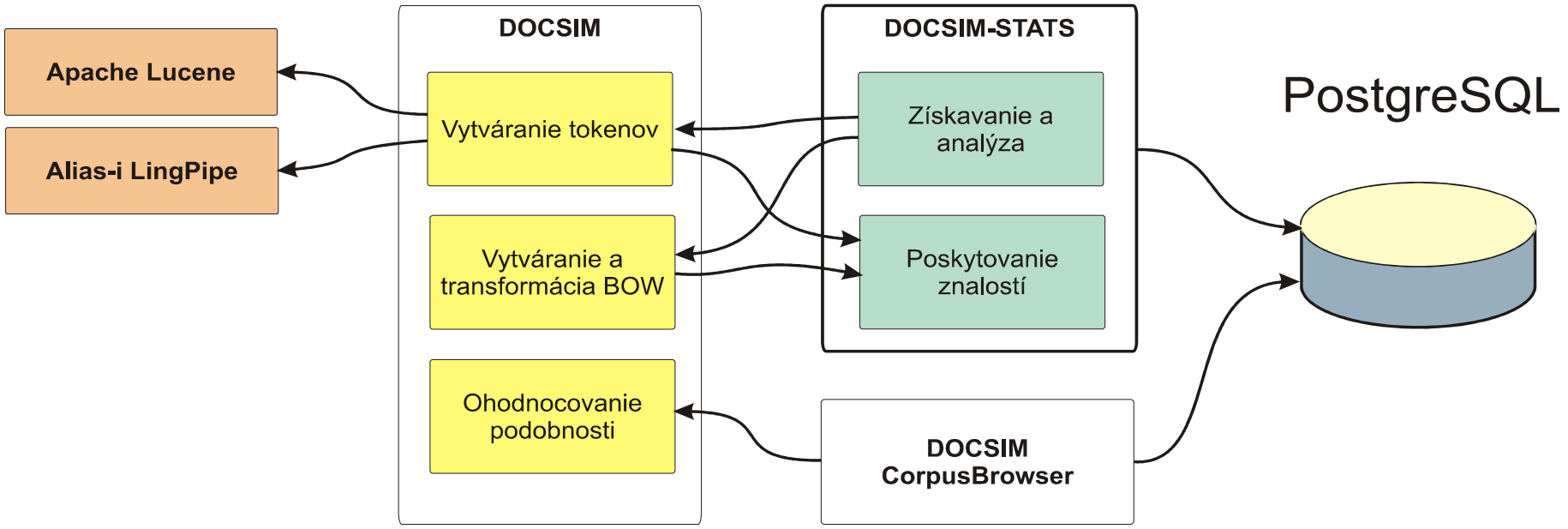
- Celkové ohodnotenie s využitím Priemernej presnosti

$$AvP = \frac{\sum_{k=1}^{\|D\|} ITP(p_k)\, precision(p_k)}{\|D\|}$$

- Vrátil mi k-ty pár relevantný výsledok?
- Ohodnotenie viacerých metód umožňuje relatívne porovnanie výkonnosti
- Distribúcia ITP určuje maximálnu možnú AvP
  - vzájomna podobnosť dokumentov v množine

Architektúra riešenia

- DOCSIM – Porovnávanie dokumentov
- DOCSIM-STATS – Analýza korpusu a získavanie znalostí
- CorpusBrowser – Vyhodnocovanie výkonu a správa konfigurácií



Implementácia

- DOCSIM
  - Java
  - Proces tokenizácie využíva rámec Apache Lucene, i-Aliasi Ling Pipe
- DOCSIM-STATS
  - Java, GNU Octave, PostgreSQL, PL/pgSQL
- Corpus Browser
  - Java, Swing, PostgreSQL, PL/pgSQL

Prezentácia programu Corpus Browser

Porovnávanie dokumentov s využitím znalostí na pozadí

Bc. Martin Kováčik

Vedúci projektu: Mgr. György Frivolt

Testovanie vybraných metód

Porovnávanie dokumentov s využitím znalostí na pozadí

Bc. Martin Kováčik

Vedúci projektu: Mgr. György Frivolt

Testovacie údaje

- Testovacie údaje
  - Databáza publikácií ACM (Projekt Mapekus)
    - ITP vyhodnocované na základe kľúčových slov a kategórií
    - Reálne vedecké publikácie
  - Reuters-21578
    - ITP vyhodnocované na základe priradených tém
    - Expertne pripravené články
- Výber dokumentov náhodným vzorkovaním

Referenčná metóda

- PostgreSQL PG_TRGM Similarity
  - Využitie trigramov (na úrovní písmen)
  - Dostupná priamo ako funkcia v databáze PostgreSQL
- Výkon porovnávačov je porovnávaný s touto metódou.

Konfigurácia testu

Označenie	Postup tokenizácie	Vyhodnotenie podobnosti
cos1	1. Tokenizácia na základe gramatiky – rozdelenie na slová, odstránenie interpunkcie a označenie typu tokenu podľa preddefinovaných regulárnych výrazov. 2. Všetky termy sú konvertované na malé písmená	kos. vzdialenosť
itsim1	1. Tokenizácia na základe gramatiky – rozdelenie na slová, odstránenie interpunkcie a označenie typu tokenu podľa preddefinovaných regulárnych výrazov. 2. Všetky termy sú konvertované na malé písmená	ItSim
cos2	1. Tokenizácia na základe gramatiky – rozdelenie na vety, odstránenie interpunkcie. 2. Vytváranie tokenov z viet 3. Filtrovanie tokenov na základe typu (filtrované sú čísla a akronymy) 4. Všetky termy sú konvertované na malé písmená 5. Sú odstránené zastavovacie slová (Slovnik z rámca Lucene) 6. Stemming (Snowball)	kos. vzdialenosť

Výsledky testov

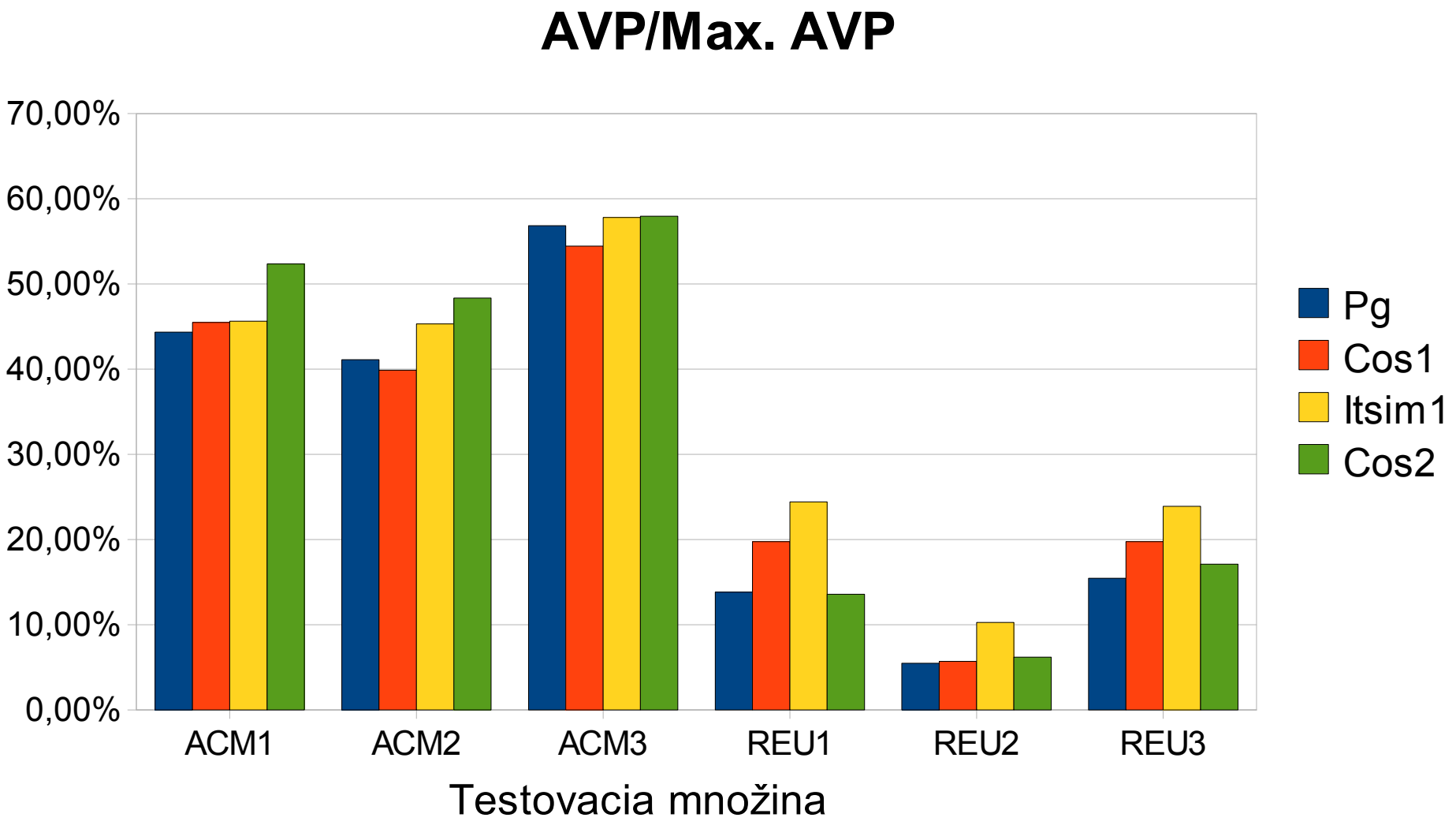
Charakteristika údajov

	ACM1	ACM2	ACM3	REU1	REU2	REU3
ITP=0	793	786	700	1028	1092	993
ITP=1	432	439	525	197	133	232
Max. AVP	0,35	0,36	0,43	0,16	0,11	0,19

Výsledky

	ACM1	ACM2	ACM3	REU1	REU2	REU3
Pg	0,16	0,15	0,2	0,05	0,02	0,05
Cos1	0,16	0,14	0,19	0,07	0,02	0,07
Itsim1	0,16	0,16	0,2	0,09	0,04	0,08
Cos2	0,18	0,17	0,2	0,05	0,02	0,06

Výsledky testov



Výsledky testov

- Počet unikátnych termov v testovacej množine

Metóda	ACM1	ACM2	ACM3	REU1	REU2	REU3
Cos1, IT	23174	19112	19779	2348	2540	2134
Cos2	14488	11506	12645	1731	1949	1506
Cos2 / ITSim	**62,52%**	**60,20%**	**63,93%**	**73,72%**	**76,73%**	**70,57%**

- Kompaktnejšia reprezentácia zvyšuje presnosť COS2 oproti COS1

Interpretácia výsledkov

- PG metóda má nižšiu presnosť
  - Nezohľadňuje slová (trigramy) – stráca sa významové prepojenie dokumentov
- COS2 metóda realizuje filtrovanie termov
  - Výhoda pri ACM – veľký počet
  - Nevýhoda pri Reuters (Malé texty)
- ITSIM1 výkonnejšia ako COS1
  - Prekryv vektorov v tomto prípade lepšie koreluje s podobnosťou

Zhodnotenie

- Porovnávanie BOW reprezentácií – zovšeobecnený proces
- Použitie porovnávača Out-Of-The-Box v aplikácií + možnosť vyhodnotiť výkon
- Vytvorené stavebné bloky pre:
  - Vytvorenie, Filtrovanie
  - Transformácia
  - Ohodnotenie
- Znalosti:
  - štatistické vlastnosti textu získané analýzou korpusu
  - StopWords, Lematizačné pravidlá, Transformácie
- Článok na IIT.SRC 2008

Q & A

Porovnávanie dokumentov s využitím znalostí na pozadí

Bc. Martin Kováčik

Vedúci projektu: Mgr. György Frivolt