

Relevancia dôležitých slov z článkov v digitálnej knižnici

Bc. Máté Vangel

Vedúci práce: Ing. Róbert Móro
Diplomová práca
10.6.2015

Motivácia a ciele práce

- Kľúčové slová sú dôležité pre široké spektrum úloh
 - ➔ Modelovanie domény
 - ➔ Modelovanie používateľa
 - ➔ Sumarizácia textu
 - ➔ Navigácia, vyhľadávanie

Extrakcia kľúčových slov v digitálnych knižniciach

- Text
- Metadata
- Abstrakt
- Autorom uvedené kľúčové slová
- Značky (Tagy)
- Citácie

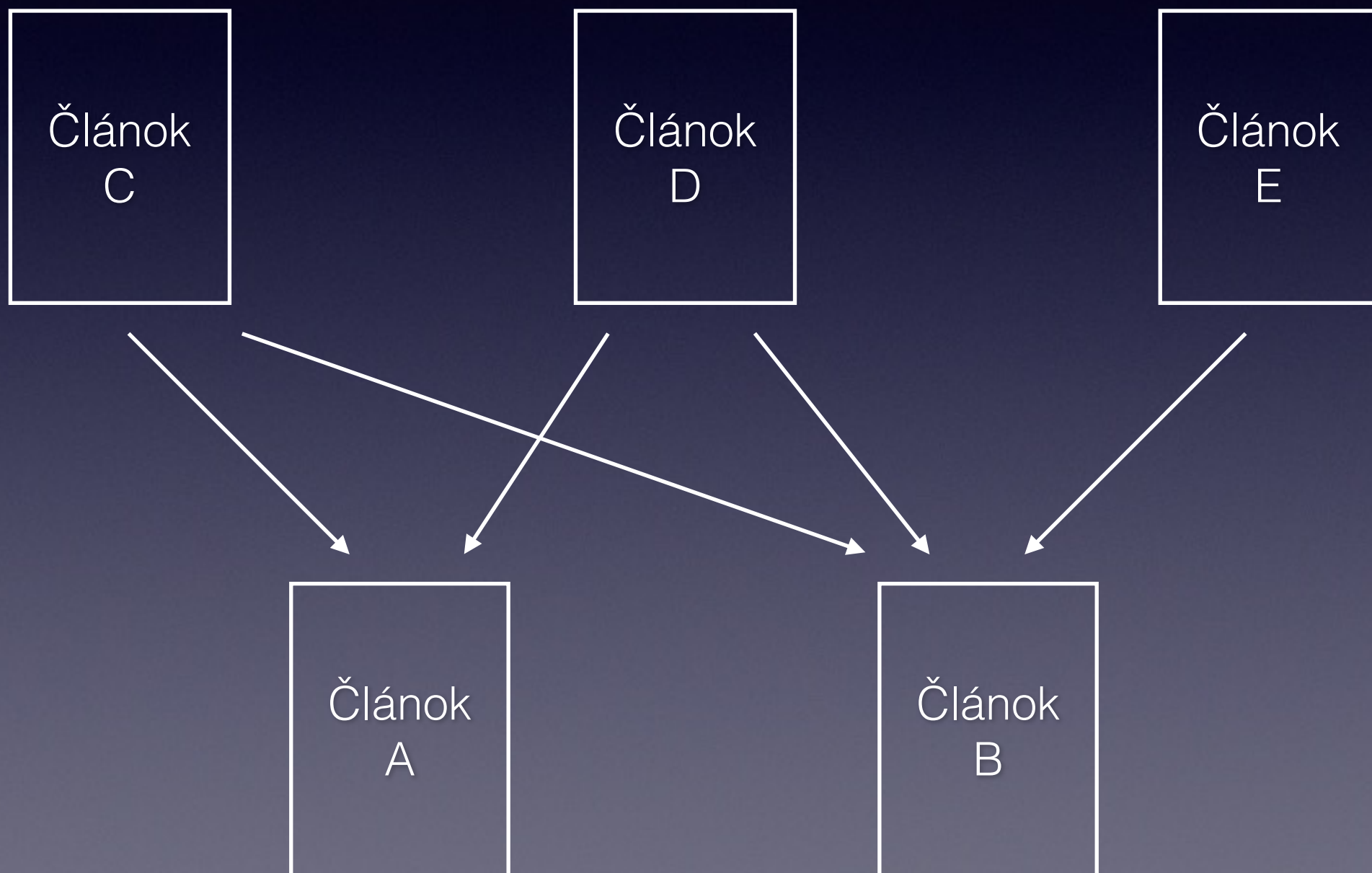
Bibliografické odkazy - Citácie

- Referencie iných autorov na daný článok
- Vyzdvihujú konkrétny aspekt článku, ktorý autor citácie považuje za relevantný (*Elkiss et al. 2008*)
- Majú charakter sumarizácie (*Abu-Jbara & Radev 2011*)
- Ak pri extrakcii kľúčových slov používame akýkoľvek nenulový citačný kontext, potom extrahované slová budú relevantnejšie (*Qazvinian et al. 2010; Ritchie et al. 2008*)
- Počet a rozmanitosť extrahovaných informácií (aspektov) z citačných viet postupne začne konvergovať, aj keď neustále analyzujeme aj nové citácie (*Elkiss et al. 2008*)

Bibliografické odkazy - Citácie

- Relevantné vlastnosti
 - ➔ Kontext citácií
 - ➔ Umiestnenie
 - ➔ Okolie
 - ➔ Ko-citácie

Ko-citácie



Ko-citácie

- Podobnosť dokumentov
 - Váha ko-citácie (CCI)
 - Index ko-citácie (CCPI)

Hlavná hypotéza

Zohľadnenie ko-citácií prinesie zlepšenie presnosti pri extrakcii kľúčových slov oproti tomu, keď extrakčná metóda využíva len citácie, v prípade ak článok má malý počet citácií.

Metóda na extrakciu a určenie relevancie kľúčových slov pomocou citácií a ko-citácií

1. Identifikácia citačných prepojení
2. Identifikácia citačných kontextov
3. Identifikácia ko-citácií
 - Výpočet CCI a filtrovanie článkov

Metóda na extrakciu a určenie relevancie kľúčových slov pomocou citácií a ko-citácií

4. Vytvoríme 3 zdroje textu (text článku, citačné kontexty, ko-citácie)
5. Odstránenia stop slov a lematizácia
6. Extrakcia kľúčových slov zvlášť pre každý zdroj textu (TF-IDF)
7. Normalizácia
8. $w = w(text) + w(kontext) + w(ko-citácia)$

Overenie metódy

- Dátová sada ACL Anthology Network
 - 26 000 vedeckých článkov
 - Strojové učenie
 - NLP
- Annota —> kvantitatívne vyhodnotenie na základe explicitnej spätnej väzby od používateľov (expertov)

Overenie - Použité metódy na extrakciu kľúčových slov

M1 —> Extrakcia z textu

M2 —> (M1) + citačné kontexty

M3 —> (M2) + ko-citácie

Overenie - Použité metódy na extrakciu kľúčových slov

M1 —> Extrakcia z textu

M2 —> (M1) + citačné kontexty

M3 —> (M2) + ko-citácie

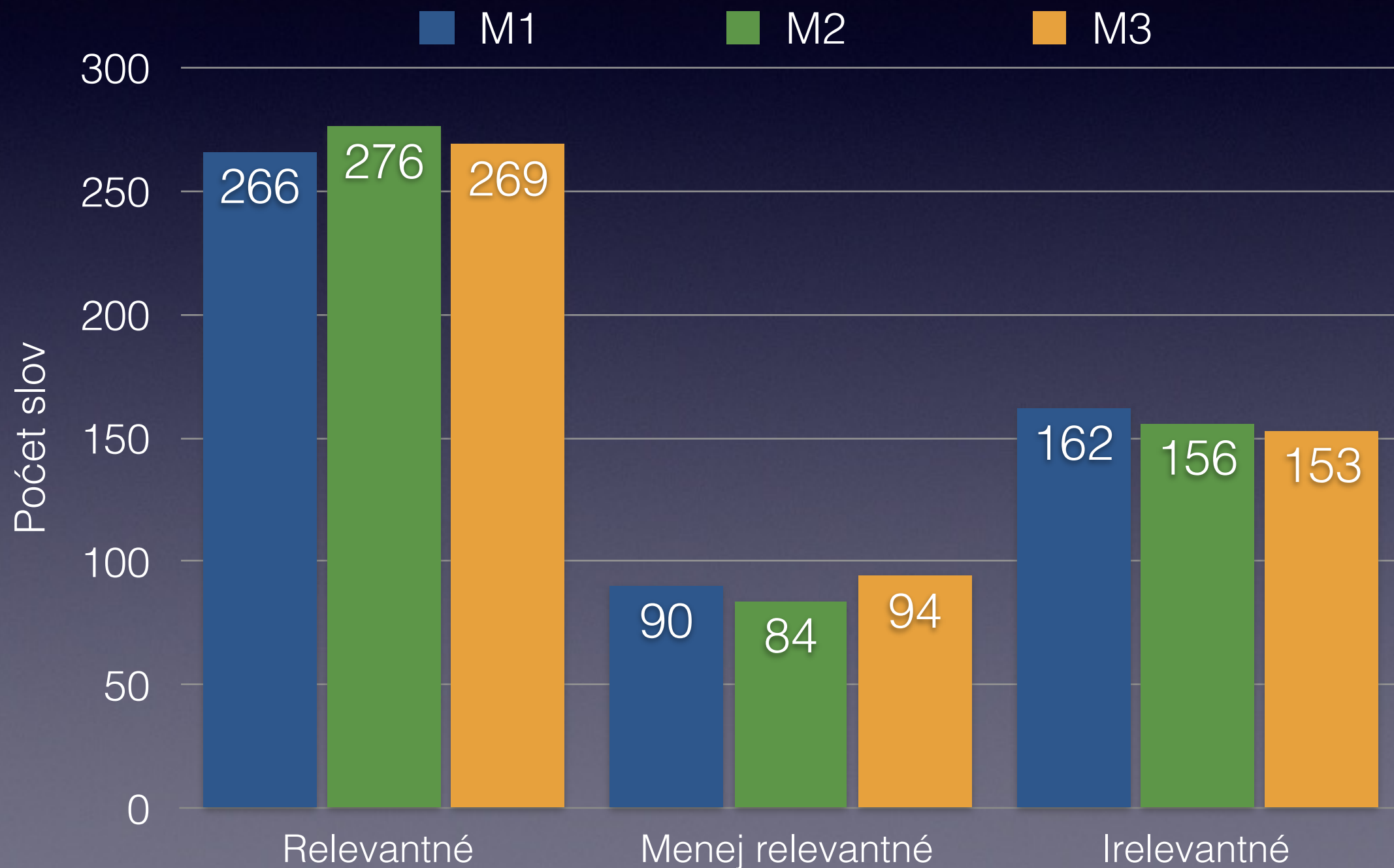
$$\underline{P(1) < P(2) < P(3)}$$

Ukážka

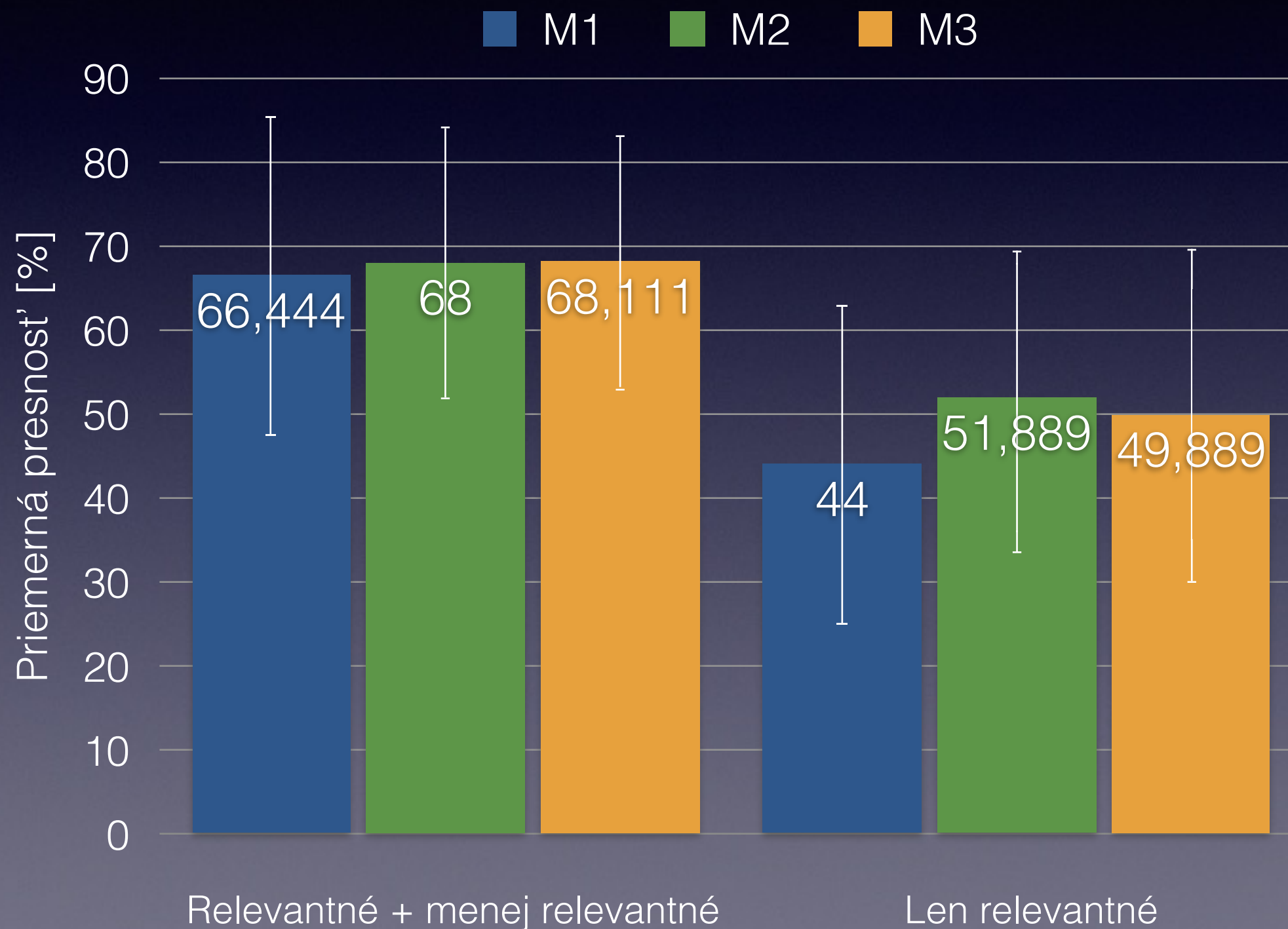
Výsledky overenia

- 8 účastníkov
- 844 ohodnotených slov
- 45 rôznych článkov

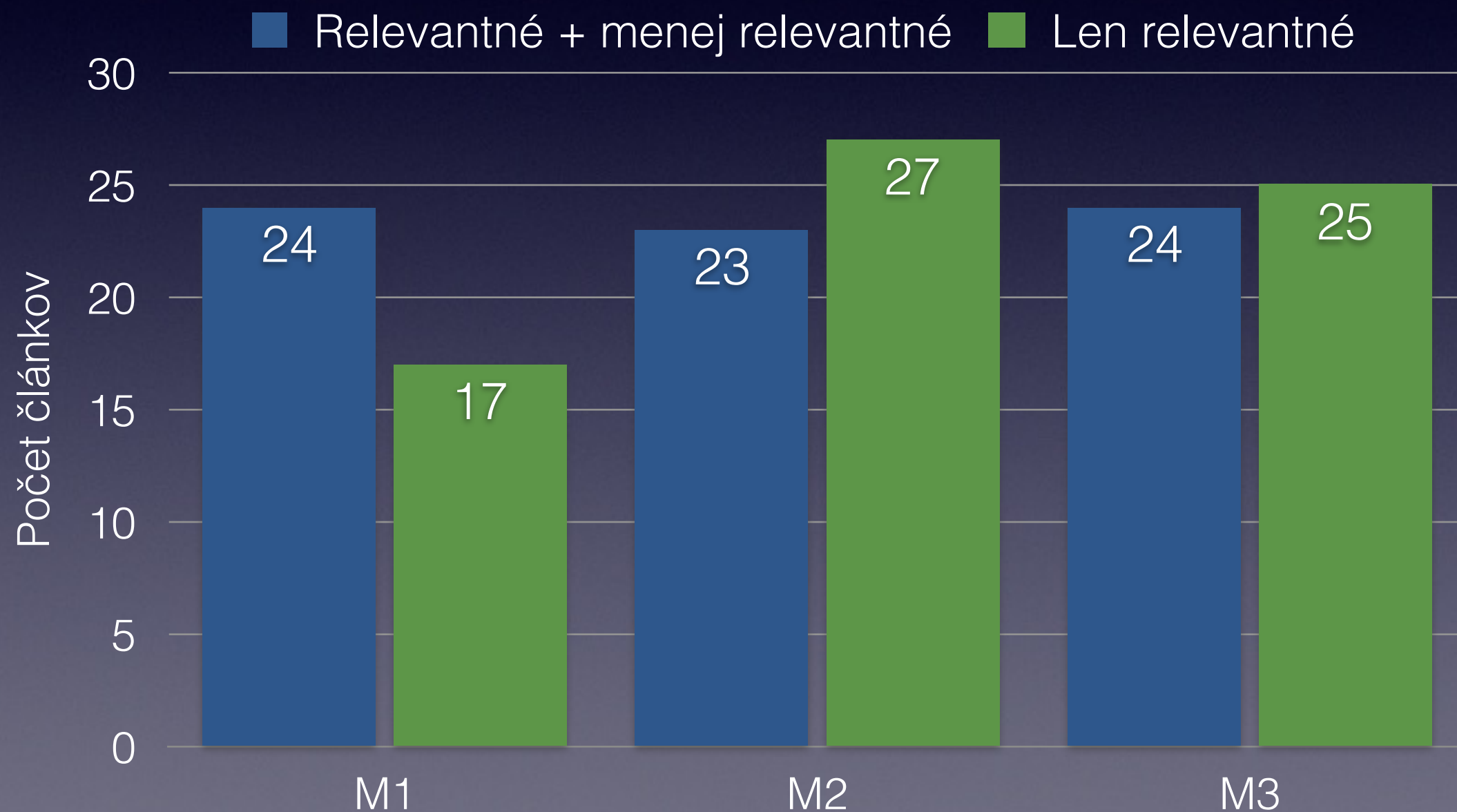
Globálny prehľad z pohľadu jednotlivých slov a metód



Priemerná presnosť z pohľadu metód



Úspešnosť jednotlivých metód z pohľadu článkov



Podrobnejšia analýza získaných výsledkov

- Rozdelili sme články na dve skupiny na základe počtu ich citácií (*hranica 20*)
- (1.) M2 —> 39%, M3 —> 26%
- (2.) M2 —> 8%, M3 —> 17%

Nové slová extrahované pomocou ko-citácií

- 119 z 844
 - 39 relevantné
 - 26 menej relevantné
 - 54 irelevantné

Možnosti d'alšej práce

- Vyskúšať inú extrakčnú metódu
- Iné hodnoty a kombinácie parametrov CCI a CCPI
- Rozšírenie kolekcie stop slov
- Rozsiahlejší experiment

Zhrnutie a záver

- Extrakcia kľúčových slov a určenie ich relevancie pomocou citácií a ko-citácií
- Modulárna metóda
- Kvantitatívne overenie na základe explicitnej spätnej väzby od expertov
- Čiastočne sa nám podarilo overiť našu hypotézu a poukázali sme na to, že analýza ko-citácií môže byť vhodná pri extrakcii kľúčových slov

- Ritchie, A., Robertson, S., Teufel, S.: Comparing citation contexts for information retrieval. In: CIKM '08: Proc. of the 17th ACM Conference on Information and Knowledge Management, ACM Press, (2008), pp. 213-222.
- Abu-jbara, A. & Radev, D., 2011. Coherent Citation-Based Summarization of Scientific Papers. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, Oregon, pp. 500–509.
- Qazvinian, V., Radev, D., Özgür, A.: Citation summarization through keyphrase extraction. In: COLING '10: Proc. of the 23rd International Conference on Computational Linguistics, (2010), pp. 895–903.
- Elkiss, A. et al., 2008. Blind Men and Elephants : What Do Citation Summaries Tell Us About a Research Article ? Journal of the American Society for Information Science and Technology, 59(2003), pp.51–62.

Aká je znovupoužiteľnosť metódy?

- Modulárna implementácia
- Použiteľná aj pre iné dátové sady, ktoré obsahujú texty článkov a citačné prepojenia

Akého charakteru sú extrahované kľúčové slová po obsahovej stránke? Sú medzi extrahovanými kľúčovými slovami „typickí predstavitelia“ správnych resp. nesprávnych?

- Relevantné a menej relevantné:
 - question - Entity based Q&A Retrieval
 - keywords - Text Analysis for Automatic Image Annotation
 - topic - Summarizing Definition from Wikipedia
 - bayes - Multi-Document Summarization using Sentence-based Topic Models
- Irelevantné:
 - skip - Metadata-Aware Measures for Answer Summarization in Community Question Answering
 - Nanba - Automatic Classification Of Citation Function
 - email - Context-Enhanced Citation Sentiment Detection

Sú v ko-citáciách naozaj cenné kľúčové
slová, ktoré sa inde nevyskytujú? Nie je
problém v tom, že sa k nim nevieme dostať?

Ako by sa aspoň časť tej „hlušiny“ mohla odstrániť?

- Efektívnejšie filtrovať mená autorov
- Filtrovať skratky a špecifické slová
- Rozšíriť kolekciu stop slov a skonkretizovať na danú doménu