Slovenská technická univerzita
Fakulta informatiky a informačných technológií
Ilkovičova 2, 842 16 Bratislava

# Vyhľadávanie informácií - Konzultácia 3

Katarína Juhásová

Cvičiaci: Mgr. Martin Šeleng, PhD.
Študijný odbor: Inteligentné softvérové systémy
Ročník: 1. ročník, Ing. štúdium
Akademický rok: 2020/2021

# Analýza autorov

Zo všetkých autorov som si vybrala mená prvých 10,000,000 a následne som ich zoradila abecedne. Pri analáze týchto mien som zistila nasledovné:

- nekonzistentné kódovanie, napr. "Zuzana Obertov&#x00E1"
- výskyt nezmyselných mien ako "Åsa Össbo", čo však môže byť spôsobené iným kódovaním ako UTF-8
- výskyt mien ako "yellowlabel", "wpadmin", "woodshed production", čo zrejme nie sú osoby, ale nie je možné s tým nič urobiť a teda aj k týmto menám budem ďalej pristupovať ako ku klasickým menám
- výskyt niekoľkých opakujúcich sa mien, napr. "paul keppel"
- výskyt rôznych mien, ktoré majú byť zrejme jedna osoba: "yayah Chanafiah" a "yayah chanafiah"

Na identifikáciu rôznych mien patriacich jednej osobe plánujem skúsiť nasledovné:

- zjednotiť kódovanie
- zjednotiť tie, ktoré sú pri lowercase zhodné
- zohľadniť výskyt/absenciu titulu a prípadne zjednotiť vhodné skupiny
- pri menách pozostávajúce z viacerých častí (viac ako 2) skúsiť permutácie a na základe prípadnej zhody zvážiť zjednotenie
- zjednotiť mená, ktoré majú vo formátoch “meno priezvisko” a “priezvisko, meno” zhodné obe mená

Zjednotenie autorov na základe vyššie uvedených podmienok bude samozrejme prevedené iba za podmienky, že články pre autorov budú mať aspoň 1 spoločný študijný odbor.

# Zmeny v porovnaní s návrhom riešenia

Počas práce na projekte som dospela k záveru, že bude vhodnejšie najskôr priradiť študijné odbory k článkom a následne články k autorom bez akéhokoľvek zjednocovania alebo rozdeľovania. Následne sa pokúsim nájsť autorov, ktorí by mohli byť jedna osoba a na základe podobnosti študijných odborov pre články určiť, či majú byť niektorí autori zjednotení alebo nie. Tento spôsob je jednoduchší ako zbytočne zjednocovať autorov a následne ich rozdeľovať ako to bolo opísaná v originálnom návrhu riešenia. Upravená architektúra riešenia sa nachádza na ďalšej strane.

# Architektúra riešenia