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

Vyhľadávanie informácií - Projekt

Katarína Juhásová

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

Téma: Microsoft Academic Graph, zjednotenie používateľov vystupujúcich pod inými menami na základe článkov s rovnakej oblasti, štatistiky o počtoch autorov, publikácii

Cieľom projektu spracovať dostupné dáta o článkoch a ich autoroch takým spôsobom, aby bolo možné určiť, ktoré články boli napísané rovnakým autorom, pričom formát mena nemusí byť zhodný vo všetkých článkoch (napr. uvedené celé meno vs. skratky, rôzne poradia mien...). Zároveň by malo byť možné identifikovať opačné prípady, kde sa síce autor podľa mena javí rovnaký pre viaceré články, avšak rozličné odborné zameranie článkov vylučuje možnosť, že by išlo o tú istú osobu.

Riešenie bude pozostávať z 3 častí. V prvej časti budú spracované mená a následne vytvorené zhluky na základe rovnakých mien. Tu bude potrebné vyriešiť problém rôznych formátov mena jedného autora, pričom bude treba dbať na správnosť fungovania riešenia aj pre netradične dlhé mená. V nasledujúcej časti budú títo autori prerozdelení na základe odborného zamerania článkov. Po tejto fáze by mali byť zjednotené všetky rôzne mená, pod akými sú jednotlivú autori podpísaní v článkoch.

V poslednej časti budú spracované dáta použité na získanie rôznych štatistík, či už ide o výber top autorov s najväčším počtom publikácií alebo zoznam najpopulárnejších študijných odborov. Prípadne by mohlo byť možné si vyžiadať zoznam článkov pre konkrétne meno autora a pod.

Na implementáciu riešenia bude použitý jazyk Python spolu s framework-om Hadoop na distribuované spracovanie dát.

Existujúce riešenia

V súčastnosti je tento problém obsiahnutý aj v projekte Microsoft Academic Graph (MAG) (https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/?from=http%3A%2F%2Fresearch.microsoft.com%2Fmag), kde sú okrem iného zachytené vzťahy medzi publikáciami, autormi, inštitúciami či študijnými odbormi. Podobným projektom je Open Academic Graph (https://www.openacademic.ai/oag/) , ktorý prepojil MAG a AMiner.

Dáta

Bohužiaľ stále sa mi nepodarilo zohnať raw data pre Microsoft Academic Graph.