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

Vyhľadávanie informácií - Konzultácia 6

Katarína Juhásová

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

Finálny výstup spracovania dát pomocou MapReduce

Nižšie sú uvedené príklady riadkov z výstupných súborov obsahujúcich všetky potrebné vyparsované dáta. V týchto dátach sa nachádzajú aj autori bez článkov. Tých nebude možné žiadnym spôsobom zgrupiť, keďže nemajú informácie o študijných odboroch a spolu-autoroch, ktoré sú použité v podmienke pre zgrupovanie.

{"id": "1000347057", "name": "Frances Williams Browin", "paperCount": "1", "citationCount": "0", "papers": [{"id": "641188018", "title": "Coins have tales to tell : the story of American coins", "publicationDate": "1966-01-01", "referenceCount": "0", "citationCount": "0", "estimatedCitationCount": "0", "fieldsOfStudy": ["Art", "COinS"]}]}

{"id": "1000743255", "name": "Charlesworth Robert Koridon", "paperCount": "2", "citationCount": "9", "papers": [{"id": "1836499992", "title": "Bis-(aminotriazolyl-)-hydrocarbons", "publicationDate": "1953-12-04", "referenceCount": "1", "citationCount": "8", "estimatedCitationCount": "8", "fieldsOfStudy": ["Chemistry"], "coauthors": ["Shreve Randolph Norris"]}, {"id": "2280759494", "title": "Bis-(triazolyl)-azo dyes", "publicationDate": "1954-08-09", "referenceCount": "3", "citationCount": "1", "estimatedCitationCount": "1", "coauthors": ["Shreve Randolph Norris"]}]}

Tieto záznamy majú nasledujúcu štruktúru:

```
{
  “id”: “”,
  “name”: “”,
  “paperCount”: “”,
  “citationCount”: “”,
  “papers”: [
    {
      “id”: “”,
      “title”: “”,
      “publicationDate”: “”,
      “referenceCount”: “”,
      “citationCount”: “”,
      “estimatedCitationCount”: “”,
      “fieldsOfStudy”: [
        “”,
        “”
      ],
      “coauthors”: [
        “”
      ]
    }
```

```
    ]
}
```

Plán ďalšej práce

Do odovzdania projektu je ešte potrebné dokončiť nasledovné úlohy:

1. Importovanie dát do ElasticSearch
2. Zgrupenie autorov nad všetkými dátami
1. Exportovanie informácií ako počet zgrupených autorov, pôvodný celkový počet autorov vs. celkový počet autorov po zgrupovaní a pod.

Keďže autorov je veľa, pri zgrupovaní bude použitý ElasticSearch. Forcyklom budú postupne prejdení všetci autori. Pre každého z nich budú z vytvoreného indexu matchnutí tí autori, ktorí majú aspoň jedno meno zhodné s práve spracovávaným autorom. Títo autori budú porovnaní so spracovávaným autorom a v prípade, že nastane zhoda na základe mena a prienik množín študijných odborov pre týchto autorov bude neprázdna množina a podobne aj prienik množín spolu-autorov, v tom prípade budú títo autori zgrupení.