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

Vyhľadávanie informácií - Konzultácia 5

Katarína Juhásová

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

Spracovanie všetkých dát pomocou Hadoop MapReduce

Krok 1: Vyparsovanie autorov

Vyparsovanie autorov do .jsonl.bz2 súboru. Príklady výsledných riadkov súboru:

{"id": "1000002831", "name": "Daniel W. Klohs", "paperCount": "1", "citationCount": "1"}
{"id": "1000004387", "name": "Anna Konieczko", "paperCount": "2", "citationCount": "0"}
{"id": "1000005149", "name": "Etienne Louguet", "paperCount": "1", "citationCount": "0"}

Krok 2: Vyparsovanie článkov

Vyparsovanie článkov do .jsonl.bz2 súboru. Príklady výsledných riadkov súboru:

{"id": "1000000068", "title": "Telephone interface controller for unattended operation", "publicationDate": "1981-01-20", "referenceCount": "2", "citationCount": "4", "estimatedCitationCount": "4"}

{"id": "1000000379", "title": "Processing alcholic beverage distillates", "publicationDate": "1946-12-21", "referenceCount": "6", "citationCount": "10", "estimatedCitationCount": "10"}

{"id": "1000000679", "title": "Dry admixture for hydraulic binders", "publicationDate": "2005-06-29", "referenceCount": "4", "citationCount": "0", "estimatedCitationCount": "0"}

Krok 3: Vyparsovanie študijných odborov

Z doterajšej práce s dátami bolo zistené, že práca s hierarchiou študijných odborov nie je podstatná a neprináša žiadne zlepšenie pokiaľ ide o zgrupovanie autorov. Z toho dôvodu boli vyparsované študijné odbory bez záznamov o ich rodičovských odboroch. Príklady riadkov z .jsonl.bz2 súboru:

{"id": "100053769", "name": "Siemens"}
{"id": "100065037", "name": "Lexical chain"}
{"id": "100068826", "name": "Hemolymph"}

Krok 4: Prepojenie id článkov a študijných odborov

Príklady riadkov z .jsonl.bz2 súboru:

{"paperId": "2463659692", "fieldId": "100001284", "fieldName": "Public economics"}
{"paperId": "2463632037", "fieldId": "100001284", "fieldName": "Public economics"}
{"paperId": "2464075420", "fieldId": "100001284", "fieldName": "Public economics"}

Krok 5: Prepojenie článkov a študijných odborov

Príklady riadkov z .jsonl.bz2 súboru:

{"id": "1000001125", "title": "Molluscs of the river Muchawka and selected habitat factors", "publicationDate": "2001-01-01", "referenceCount": "0", "citationCount": "0", "estimatedCitationCount": "0", "fieldsOfStudy": ["Fishery", "Ecology", "Habitat"]}

{"id": "1000000679", "title": "Dry admixture for hydraulic binders", "publicationDate": "2005-06-29", "referenceCount": "4", "citationCount": "0", "estimatedCitationCount": "0", "fieldsOfStudy": ["Materials science", "Cement", "Portland cement", "Composite material"]}

{"id": "1000000668", "title": "Horan, Wall & Walker - Guide to Australian Business RecordsArchival Resources", "publicationDate": "1999-11-22", "referenceCount": "0", "citationCount": "0", "estimatedCitationCount": "0", "fieldsOfStudy": ["Biography", "Engineering"]}

Krok 6: Prepojenie id článkov a autorov

Príklady riadkov z .jsonl.bz2 súboru:

{"paperId": "954575375", "authorId": "1000002831", "name": "Daniel W. Klohs", "paperCount": "1", "citationCount": "1"}

{"paperId": "958545020", "authorId": "1000004387", "name": "Anna Konieczko", "paperCount": "2", "citationCount": "0"}

{"paperId": "2608047901", "authorId": "1000004387", "name": "Anna Konieczko", "paperCount": "2", "citationCount": "0"}

Krok 7: Prepojenie článkov a autorov

Príklady riadkov z .jsonl.bz2 súboru:

{"id": "1000000679", "title": "Dry admixture for hydraulic binders", "publicationDate": "2005-06-29", "referenceCount": "4", "citationCount": "0", "estimatedCitationCount": "0", "fieldsOfStudy": ["Materials science", "Cement", "Portland cement", "Composite material"], "authors": [{"id": "1000007118", "name": "JingLUO", "paperCount": "1", "citationCount": "0"}]}

{"id": "1000001125", "title": "Molluscs of the river Muchawka and selected habitat factors", "publicationDate": "2001-01-01", "referenceCount": "0", "citationCount": "0", "estimatedCitationCount": "0", "fieldsOfStudy": ["Fishery", "Ecology", "Habitat"], "authors": [{"id": "1000007204", "name": "Thisa Maitê Furlanetto Bordignon", "paperCount": "1", "citationCount": "0"}]}

{"id": "100000002", "title": "Electron Spin Resonance Investigations of Oxygen-Centered Free Radicals in Biological Systems", "publicationDate": "1988-01-01", "referenceCount": "23", "citationCount": "7", "estimatedCitationCount": "7", "fieldsOfStudy": ["Radical", "Superoxide", "Nuclear magnetic resonance", "Xanthine oxidase", "Hydroxyl radical", "Chemistry", "Electron paramagnetic resonance", "Hyperfine structure", "Oxygen", "Photochemistry", "Autoxidation"], "authors": [{"id": "1000004387", "name": "Anna Konieczko", "paperCount": "2", "citationCount": "0"}, {"id": "1000000829", "name": "BianYanjie", "paperCount": "1", "citationCount": "0"}]}

Krok 8 (prebieha spracovanie): Vytvorenie samostatných záznamov pre jednotlivých autorov

Po poslednom spracovaní by mal byť pre všetkých autorov vytvorený záznam s nasledujúcou štruktúrou:

```
{
   “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. upravenie funkcie na zgrupovanie autorov nie len na základe podobnosti mena a spoločných študijných odborov ale aj podľa spoločných spoluautorov podieľajúcich sa na článkoch
2. zgrupenie týchto autorov nad všetkými dátami
3. vytvorenie indexu
4. exportovanie informácií ako počet zgrupených autorov, pôvodný celkový počet autorov vs. celkový počet autorov op zgrupovaní a pod.