Slovenská technická univerzita
Fakulta informatiky a informačných technológií
Ilkovičova 3, 842 16 Bratislava 4

Vyhľadávanie informácii

Záznamník práce na projekte

2022/23, ZS

Radovan Cyprich

Dátum: 16.12.2022

ZÁZNAM KONZULTÁCII

Konzultácia č. 1: 27. 9. 2022 - vybraná téma O42 - Crawlovanie viacerých stránok s mačkami, následne merge-ovanie týchto datasetov o mačkách do jedného záznamu (Python)

Konzultácia č. 2: 4. 10. 2022 - Mačky. Definuje VINF cez teamový projekt. Pseudokód bude.

Konzultácia Prezentácia č. 3: 8. 11. 2022 - Tri krát robí scraper z podstránky, ktoré spája do jedného datasetu. Veľa času zabralo čistenie dát. Používa regex pri tvorbe datasetu. Do nasledujúcej konzultácie bude pracovať na väčšom sub-datasete. Pridá zlepšenie vyhľadávania.

Konzultácia č. 4: 30. 11. 2022 - Pridáva paralelné spracovanie. Opätovné pridanie vyhľadávania.

Konzultácia č. 5 - 13. 12. 2022 - Konečné odovzdanie projektu.

NÁPLŇ ZADANIA

- Crawlovanie stránok o mačkách z rôznych zdrojov
- Spojenie datasetov na základe spoločných kritérii
- Parsovanie záznamov o mačkach
- Transformácia dát a čistenie datasetu
- Indexovanie sparsovanych záznamov
- Vytvorenie efektívneho vyhľadavania nad viacerými stĺpcami
- Efektívne filtrovanie dát (filtre na viaceré atriúty nie len na meno)
- Vytvorenie konzolového menu pre používateľa

RIEŠENIE

Naše riešenie pozostáva z piatich hlavných častí.

- crawlovanie databázy (knižnica ***re*** na prácu s regulárnymi výrazmi)
- transformovanie atribútov a parsovanie dát (nástroj pySpark)
- mergovanie datasetov (knižnica pandas)
- indexovanie (nástroj pyLucene)
- ivyhľadávanie (nástroj pyLucene)

Implementáciu sme vykonávali v jazyku python s použitím pomocných knižníc a nástrojov spomenutých vyššie.

CRAWLER

- Crawler na nemeckú stránku s databázou mačiek. Postupným prechádzaním idčiek v url vyťahujem informácie z jednotlivých stránok. Informácie spracuvávam a vytiahnuté dáta čiastočne transformujem vzhľadom na navhnutú schému databázy ešte pred zapisovaním do súboru.

- Pri crawlovaní používame regex na odstráňovanie tagov, nepotrebných znakov a slov zo stiahnutého html dokumentu.

- Nemecká databáza obsahujúca 200 000 záznamov sa nám podarila stiahnuť celá. Poľskú databázu vzhľadom na timeouty a neustále rušenie http requestov zo strany servera sme nestihli stiahnuť a následne upraviť celú, ale len okolo 50 000 záznamov za mesiac a pol.

POSTUP

- Prechádzanie jednotlivých stránok postupne po ID v URL.
- Otvorenie stránok (dokopy troch 3) s informáciami o mačke, jej oceneniach a rodičoch
- Vytiahnutie celého html danej stránky
- Filtrovanie štrukturovaného textu v tagoch pomocou regulárnych výrazov
- Dáta, ktoré sme vytiahli vložíme do slovníka, ktorý je štrukturovaný presne ako databázova schéma, ktorú sme navrhli v tímovom projekte
- Dáta transformujeme do formátu s nami určenymi oddelovačmi a koncom záznamu
- Riadok/záznam o mačke zapíšeme do textového súboru

NEMECKÁ STRÁNKA O MAČKÁCH

- https://www.felidae-ev.de/stammdaten.php?id=2&type=sd
  Stránka s prvou mačkou
- https://www.felidae-ev.de/stammdaten.php?id=2&type=sb
  Stránka s inf. o jej rodičoch
- https://www.felidae-ev.de/stammdaten.php?l=0&id=2&type=a
  Stránka obsahujúca inf. o jej úspechoch

- Pre získanie všetkých informácií o jednej mačke, ktoré potrebujeme sme museli získať dáta z troch rôznych stránok (teda 3 rôzne URL) a pre každú sme písali vlastné regulárne výrazy na získanie dát, ktoré potrebujeme.

TRANSFORMÁCIA DÁT

Po čiastocnej transformácii dát pri samotnom crawlovaní a následnom parsovani dát, bolo potrebné urobiť viacero transformácii vzhľadom na zmeny našej databázovej schémy, ale aj nevhodné a nekvalitné dáta, ktoré sme si všimli neskôr. Jednalo sa hlavne o tieto transformácie:

- Transformácia celých názvov plemien na kód rasy
- Transformácia formátu dátumov z eg. 01.01.2000 na 2000-01-01
- Transformácia kolóniek reprezentujúcich mená rodičov bolo potrebné odstránenie hodnôt *Unknown* a zmena formátu hodnôt *Foundation*
- Transformácia ID rodičov, ktoré sme potrebovali mať naviazané na naše IDčka, nie na IDčka v nemeckej databáze

Implementoval som transformáciu klasickym spracovaním v pythone. Následne som rovnaké transformácie spravil aj pomocou nástroja pySpark, distribuovane. Pri spracovaní som si dáta

ukladal vo formáte RDD (Resilient Distributed Dataset), čo je základná dátová štruktúra nástroja spark. Sú to nemenné distribuované kolekcie objektov akéhokoľvek typu. Ako už názov napovedá, ide o záznamy údajov odolné voči chybám, ktoré sa nachádzajú na viacerých uzloch. Využil som funkciu map, v ktorej som paralélne upravoval až 6 stĺpcov. Následne som tieto dáta rovnakým spôsobom ako pri obyčajnom spracovaní sparsoval do súboru v žiadanom formáte.

Zdroje: https://spark.apache.org/docs/latest/

MERGE DATABÁZ

V rámci tímového projektu sme museli vytvoriť merge script, ktorý bude mergovať viacero súborov do jedného. Tento merge script mal na starosti môj kolega z tímového projektu Matej Delinčák. Tento skript som si upravil pre potreby tohto predmetu a pomocou neho som spojil 4 databázy. Tieto databázy sme si rovnomerne rozdelili aby sme za čo najkratší čas stiahli čo najviac mačiek. Ja sťahujem ešte poľskú a perzskú databázu avšak scrapovanie týchto databáz vzhľadom na zabezpečenie stránok bude trvať dlhšie.

Po transformovaní nemeckej databázy uloženej v súbore catsSPARK.txt sme si túto databázu uložili do dátoveho súboru ako cats_german.csv, keďže náš script je prispôsobený na mergovanie csv súborov. Následne som vzal tri ďalšei databázy od kolegov a spustil merge skript.

- Nemecká databáza = 200 000 záznamov, stiahol Radovan Cyprich
- link: https://www.felidae-ev.de/
- Švedská databáza = 507 000 záznamov, stiahol Matej Delinčák
- link: https://www.sverak.se/
- Nórska databáza = 197 000 záznamov, stiahol Matej Kuráň
- link: https://katt.nrr.no/Katter/kissat
- Fínska databáza = 292 000 záznamov, stiahol Matej Kuráň
- link: http://kissat.kissaliitto.fi/kissat

- Výsledná databáza v rámci tohto projektu obsahuje 1 190 000 záznamov a súbor má veľkosť 181 MB.

Z týchto štyroch databáz bolo najťažšie stiahnuť práve tú nemeckú keďže sme museli prehľadávať až tri stránky, kdeto pri ostatných len jeden html element div. Nemecká databáza taktiež obsahovala najviac informácii o mačkách oproti ostatným, preto aj čistenie a transformácie stĺpcov boli oveľa náročnejšie. V rámci tímového projektu sme stiahli taktiež databázy pawpeds a dve ruské, ktoré mal na starosti kolega z druhého cvičenia.

VYHĽADÁVANIE

V poslednej časti projektu sme implementovali fulltextové vyhľadávanie nad stĺpcami NAME, BREED a kombinovane nad oboma. Najskôr sme vytvorili posting list čisto pomocou jazyka python (Konzultacia 4) a nevyužívali sme, žiadne nástroje. Tento spôsob fullltextového vyhľadávania bol veľmi pomalý, ale funkčný. V neskoršej fáze projektu sme využili sme knižnicu pylucene, ktorej funkcionalitu sme zabezpečili pomocou imagu v Dockeri. Na začiatku sme si pomocou IndexWriter z pyLucene vytvorili index. Zaindexovali sme stĺpce ID,NAME a BREED. Stĺpec ID sme potrebovali preto, aby sme vedeli zistiť pozíciu mačky v datasete, aby sme ju mohli vypísať. Následne pomocou searchera vyhľadávame nad týmito stĺpcami. Používateľ ma k dispozícii interaktívne menu, kde si môže zvoliť, čo chce vyhľadávať. Proces vyhľadávania podrobnejšie vysvetľujeme v časti používateľskej príručky.

Link na pylucene image: https://hub.docker.com/r/coady/pylucene

MODULY A DÁTA

Tento projekt som štrukturoval do viacerých modulov. Každý modul (.py súbor) je zodpovedný za niečo iné. Každý z týchto modulov sa spúšťa samostatne a každý má špecifický dátový alebo konzolový výstup.

- **regex_crawler.py** - modul zodpovedný za crawlovanie nemeckej databázy
- **transformations.py** - modul, v ktorom sa nachádzajú transformovacie funkcie a taktiež sa tu vykonáva obyčajná transformácia dát
- **pyspark.py** - modul, v ktorom sa transformujú dáta pomocou nástroja pyspark volaním funkcii z modulu transformations vo funkcii map nad vytvoreným RDD
- **merge.py** - modul, ktorý načíta všetky datasety v zložke ***data*** a následne ich spojí do jedného výsledného datasetu
- **pylucene.py** - modul, ktorý po spustení a vytvorení indexu nad niektorými stĺpcami umožňuje vyhľadávanie nad jedným alebo viacerými stĺpcami, je to taktiež výsledný modul aj s používateľským rozhraním

Dáta na mergovanie sa nachádzajú v priečinku ***data***, kde sú databázy vo formáte .csv

- Dátový výstup modulu regex_crawler je **catsCRAWLED.txt**
- Dátový výstup modulu transformations je **catsTRANSFORMED.txt**
- Dátový výstup modulu pyspark je **catsSPARK.txt**
- Dátový výstup modulu merge je **allcats.txt**

So súborom allcats.txt pracujeme v module pylucene.py a teda je to náš finálny dátovy súbor, nad ktorým vyhľadávame.

POUŽÍVATEĽSKÉ ROZHRANIE

Pri vyhľadávaní som vytvoril jednoduchú konzolovú aplikáciu, ktorá po naindexovaní stĺpcov zobrazí používateľovi menu. V menu máme možnosť výberu stĺpca, nad ktorým chceme

vyhľadávať. Pri vyhľadávaní a vytváraní indexu zaznamenávame čas, ktorý vypisujeme do konzoly.

```
CONNECTED
Index time taken: 36.226125717163086 seconds
INDEXES CREATED

------------------------------------------------------------
Options:
1 - Search -> FIELD NAME
2 - Search -> FIELD BREEDS
3 - Search -> FIELDS NAME & BREEDS
4 - Search -> SAMPLE TEST 1 (Searching all cats names 'Victoria')
5 - Search -> SAMPLE TEST 1 (Searching all cats names 'Vyctoria')
6 - Search -> SAMPLE TEST 2 (Searching all cats with breed 'SEY' which states for Seychellos breed)
7 - Search -> SAMPLE TEST 3 (Searching all cats with breed 'RAG' meaning ragdols named 'Oswald')
x - EXIT SEARCH
```

Po stlačení tlačidiel napíšeme meno resp. rasu mačky, ktorú chceme hľadať.

```
------------------------------------------------------------
Options:
1 - Search -> FIELD NAME
2 - Search -> FIELD BREEDS
3 - Search -> FIELDS NAME & BREEDS
x - EXIT SEARCH
3
Enter name to find:
Oswald
Enter breed to find:
RAG
Search time over NAME taken: 0.20371580123901367 seconds
Search time over BREED taken: 0.6898245811462402 seconds
5 total matches.

188919|Oswald|Finland|190815||FI SK LO 1718244|FI|FI|||RAG|ruskeanaamio colourpoint| n|2017-08-22|M|978101081710958|||
526999|Clara Oswald|Norway|35659||(NO) NRR LO 182829|NO|NO|||RAG|sjokoladetabbymasket bicolour| b 03 21|2016-09-13|F||
93997|Oswald|Finland|93999||FIN SRK LO 69276|FIN|FIN|||RAG|sininaamio colourpoint| a|2003-05-26|M|||||FI*Hobbit-Dolls|
484852|Oswald a Camelot Treasure|German Database|192408|||||||RAG||o 33|2018-08-25|M||||||485066|485069|Dollnouveau Co
999752|FIN*Hobbit-Dolls Oswald|SVERAK||(FI)SRK LO 69276|(FI)SRK LO 69276|FI|FI|||RAG||a||M||||||1000822|1000823||||;


------------------------------------------------------------
```

INŠTALÁCIA

Pre úspešné spustenie jednotlivých modulov je potrebné mať nainštalovaný python3. Ďalej musíme mať nainštalovaný nástroj Spark a správne nastavené cesty k systémovým premenným. Posledná vec je vytvorenie nového python interpretera pomocou docker-pull metódy, ktorá nám automaticky stiahne a naštartuje pylucene kontajner. Okrem základných balíčkov jazyka python je potrebné si taktiež importovať aj tieto:

- Pandas
- Pathlib
- Pyspark

- Pylucene
- Paths
- Re
- Requests

TESTOVANIE

Testovanie sme vykonávali na štyroch vzorkách.

Search → SAMPLE TEST 1a (Searching all cats names 'Victoria')
Predpokladali sme, že budú existovať nejaké mačky s menom Victoria. Takýchto mačiek sa v našej databáze nachádzalo 855.

```
4
Search time over NAME taken: 0.2676219940185547 seconds
Search succesful 855 total matches.
```

Search → SAMPLE TEST 1b (Searching all cats names 'Vyctoria')
Predpokladali sme, že na druhej strane nebudú žiadne mačky s menom Vyctoria. Tento test sa potvrdil nenašli sa žiadne výsledky.

```
5
Search time over NAME taken: 0.03773236274719238 seconds
Search unsuccesful no records found!
```

Search → SAMPLE TEST 2 (Searching all cats with breed 'SER' which states for Serval breed)
Tretí test sme chceli zistiť koľko servalov máme v databáze a pretože toto plemeno sme našli len v nemeckej databáze. Test sa nám potvrdil našlo sa 60 mačiek z tohto plemena a len z nemeckej databázy.

```
6
Search time over BREED taken: 0.06681561470031738 seconds
Search succesful 60 total matches.
```

Search → SAMPLE TEST 3 (Searching all cats with breed 'RAG' meaning ragdols named 'Oswald')
Poseldný test sme chceli zistiť, koľko ragdolov s menom Oswald máme v databáze. Takáto kombinácia mena a rasy v našej databáze existuje. Našlo sa 5 záznamov.

```
7
Search time over NAME taken: 0.08867835998535156 seconds
Search time over BREED taken: 0.27196526527404785 seconds
Search succesful 5 total matches.

484852|Oswald a Camelot Treasure|German Database|192408||||||RAG||o 33|2018-08-25|M||||||485066|485069|Dollnouveau Cowgirl Cind
93997|Oswald|Finland|93999||FIN SRK LO 69276|FIN|FIN|||RAG|sininaamio colourpoint| a|2003-05-26|M|||||FI*Hobbit-Dolls|88539|9833
526999|Clara Oswald|Norway|35659||(NO) NRR LO 182829|NO|NO|||RAG|sjokoladetabbymasket bicolour| b 03 21|2016-09-13|F|||||NO*Glas
999752|FIN*Hobbit-Dolls Oswald|SVERAK||(FI)SRK LO 69276|(FI)SRK LO 69276|FI|FI|||RAG||a||M||||||1000822|1000823||||;
188919|Oswald|Finland|190815||FI SK LO 1718244|FI|FI|||RAG|ruskeanaamio colourpoint| n|2017-08-22|M|978101081710958||||FI*Ragtai
```

ZHODNOTENIE

V tomto projekte som po implementovaní crawleru transformoval dáta do dátového súboru podľa našej databázovej schémy. Zvyšné dáta sme prevzali od kolegov z tímového projektu, kedže nemecká databáza obsahovala len 200 tisíc záznamov.Naučili sme sa pracovať s nástrojmi PySpark a PyLucene. Vytvorili sme efektívne vyhľadávanie celých slov avšak vyhľadávanie nam podstringami sme neimplementovali pomocou PyLucene. Funkčnosť vyhľadávania sme overili niekoľkými testami kde sme si overili naše riešenie. Tým, že sme zlúčili viacero rôznych databáz niekedy nastala situácia, že jedna mačka bola vo výsledku vyhľadávania viac krát. Naše riešenie nepodporuje vyradzovanie duplicít, preto by sme túto časť mohli implementovať v budúcnosti.