Fakulta Informatiky a Informačných Technológií
Slovenská Technická Univerzita v Bratislave

Lukáš Grejták

Zachytávanie údajov o profesionálnych hráčoch z hry Counter - Strike

Vyhľadávanie informácií

Študijný program: Inteligentné Softvérové Systémy

Cvičiaci: Ing. Igor Stupavský

December 2023

Obsah

Úvod

Téma

Zachytávanie údajov o profesionálnych hráčoch z hry Counter-Strike

Stránka

liquipedia.com

GitHub

https://github.com/grejty/VINFproject

Zámer projektu

V tomto projekte je našim cieľom crawlovať dáta o profesionálnych Counter-Strike hráčoch zo stránky liquipedia. Dáta musíme vyparsovať z HTML-iek a následne rozumne uložiť.

V druhej časti naše dáta doplníme o populačné dáta k jednotlivým rodným krajinám hráčov za pomoci distribuovaného spracovania v Sparku. Zdrojom týchto dát bude dostupný wikipedia dump o veľkosťi 80GB.

Nakoniec pre naše obohatené dáta vytvoríme indexer za pomoci Lucene a následne aj search, kde pomocou zadaných dopytov budeme získavať informácie o hráčoch.

Ako informácie sme sa rozhodli v logike programu počítať, či dvaja zadaní hráči spolu počas svojej kariéry mohli hrať. Ako doplňujúcu informáciu vypočítame aj pomer profesionálnych hráčov v jednotlivých krajinách - práve na základe dát spracovaných z wiki dumpov.

Nakoniec vytvoríme aj unit testy, ktoré naše vyhľadávacie dopyty skontrolujú.

Pseudokód

1. Skontrolovanie robots.txt súboru a kontrola, či môzeme adresu crawlovať
2. Nastavenie time-outu na 30s (aby sme zabránili prípadnému IP banu)
3. Získanie URL adries na jednotlivé regióny hráčov za pomoci regexu
4. Crawlovanie cez URL regiónov a zachytávanie URL jednotlivých hráčov
5. Sťahovanie HTML kódov stránok s profilmi hráčov
6. Parsovanie a ukladanie dát do .csv súboru
7. Prehľadanie wiki dumpu za pomoci Sparku
8. Obohatenie predtým zachytených dát o nové data nájdené vo wiki dumpe
9. Vytvorenie indexera, ktorý pre každú hodnotu zapíše jej polohu v súbore
10. Search funkcia, ktorá pomocou indexera odpovie na zadaný dopyt.

Dáta

Útržok vyparsovaných dát z liquipedie:

Header: Dáta

Nick: stikle-

Overview: Klesti “stikle-” Kola (born July 5, 1998) is an Albanian professional Counter-Strike: Global Offensive coach.

Name: Klesti Kola

Nationality: Albania

Born: July 5, 1998 (age 25)

Status: Active

Status Years Active (Player):

Status Years Active (Coach): 2020 – Present

Role: Coach

Team: Sangal Esports

Approx. Total Winnings: $1,006

Games: Global Offensive

Útržok dát, ktoré vrátil Spark:

country	population
Afghanistan	38,346,720
Albania	2,793,592
Algeria	45,400,000
Argentina	46,621,847
Austria	9,027,999
Azerbaijan	10,353,296
…	…

Miery úspešnosti po spracovaní 80GB wiki dumpu:

Podarilo sa nám nájsť zhodu až vyše 97% - v dumpe sa nenachádzali len 2 krajiny z pôvodných dát:

- Amount of nationalities in parsed data (not enriched): **82**
- Found **80** nationalities in the dump. (Success rate**: 97.56%**)

Boli sme schopní obohatiť až 98% záznamov:

- Players with Population data: **1971**
- Players without Population data: **39**
- Success Rate: **98.06%**

Obohatené dáta:

K pôvodným dátam, ktorých útržok sme opísali vyššie, sme pridali tieto dva ďalšie stĺpce. Jeden reprezentuje populáciu danej krajiny a druhý počet hráčov z danej krajiny:

- **Population**: 2,793,592
- **Nationality_Count**: 9

Výsledky unit testov:

```
Enter search query: (t - test cases, q - exit)

t

Query: Freelance peacemaker

Nick: peacemaker, Years Active(Player): 2002 – 2015

Number of residents per professional player in Portugal: 209,347

Nick: Freelance, Years Active(Player): 2012 – 2019

Number of residents per professional player in Switzerland: 989,145

Result:

The two players could have played together.

Expected: The two players could have played together.

Correct.
```

Query: karl flex0r

Nick: *karl, Years Active(Player): 2009 – 2016, 2016 – 2017, 2020*

Number of residents per professional player in China: *16,415,698*

Nick: *flex0r, Years Active(Player): 2004 – 2019, 2021 – 2022*

Number of residents per professional player in France: *1,000,626*

Result:

The two players could have played together.

Expected: *The two players could have played together.*

Correct.

Query: RobbaN dukiiii

Nick: *dukiiii, Years Active(Player): 2013 – Present*

Number of residents per professional player in Austria: *752,333*

Nick: RobbaN, Years Active(Player): *2003 – 2012*

Number of residents per professional player in Sweden: *93,282*

Result:

The two players could not have played together.

Expected: *The two players could not have played together.*

Correct.

Query: Jee tomsku

Nick: *Jee, Years Active(Player): 2021 – Present*

Number of residents per professional player in China: *16,415,698*

Nick: *tomsku, Years Active(Player): 2010 – 2018*

Number of residents per professional player in Finland: *124,768*

```
Result:

The two players could not have played together.

Expected: The two players could not have played together.

Correct.
```

Konzultácie

Konzultácia č. 2

19.10.2023 - Crawluje dáta, používa regex, parsuje údaje. Do budúcej konzultácie urobí indexáciu.

Konzultácia č. 3

3.11.2023 - Funkčná vlastná indexácia a search funkcia. Do budúcej konzultácie doplní dáta s wikipédiou.

Konzultácia č. 4 - Prezentácia

16.11.2023 - Prezentácia OK. Do budúcej konzultácie paralelné spracovanie údajov.

Konzultácia č. 5

2.12.2023 - Konečné odovzdanie a prezentácia OK.