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

Zadanie z predmetu

Vyhľadávanie informácií

Bc. Michal Mészáros
Študijný odbor: IS
Vedúci: Ing. Ondrej Kaššák

Semester: tretí
Akademický rok: 2014/2015

Znenie zadania:

Sparsovanie osôb z wikipédie, vytvorenie jednoduchej služby 'mohli sa stretnúť?', ktorá po zadaní dvoch mien určí, či sa mohli dané osoby stretnúť (prekryv času ich života).
Všetky zdrojové kódy ukážkové dáta ako aj táto dokumentácia sú uložené na repozitári **https://github.com/MichalMeszaros/wikipedia.** Ak sa budem odkazovať na nejaký program alebo súbor jedná sa o tento repozitár.

Riešenie:

Dáta s ktorými som pracoval boli stiahnuté z voľne dostupného zdroja na wikipedii. Konkrétne som použil súbor enwiki-latest-pages-articles.xml.bz2[1]. Súbor mal po rozbalení niečo cez vyše 40 GB. Ako prvé bolo prefiltrovať dáta tak aby nám ostali len relevantné informácie potrebne na zodpovedanie otázku „Mohli sa stretnúť“. Po preštudovaní štruktúry dumpu z wikipédie zistil som, že informácie o osobách sú uložené v špecifickej štruktúre.:

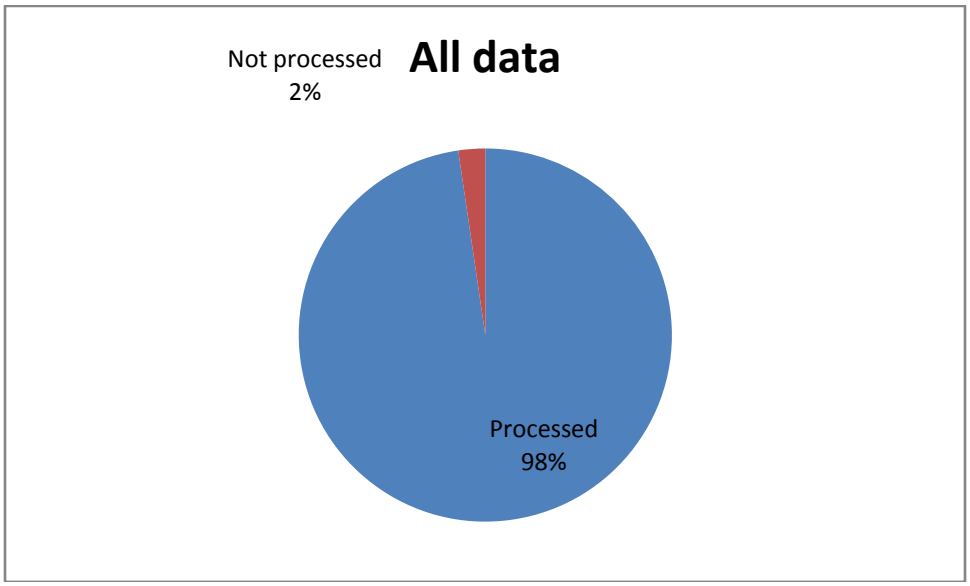
```
{{Persondata
| NAME              =Lincoln, Abraham
| ALTERNATIVE NAMES =
| SHORT DESCRIPTION =16th President of the United States
| DATE OF BIRTH     =February 12, 1809
| PLACE OF BIRTH    =Hardin County, Kentucky
| DATE OF DEATH     =April 15, 1865
| PLACE OF DEATH    =Washington, D.C.
}}
```

Pomocou príkazu grep v Unixovom prostredí som dáta prefiltroval tak aby mi ostali uz len relevantné informácie.
Príkaz vyzeral konkretne takto : **grep –A 7 “{{Persondata” enwiki-latest-pages-articles.xml > data.**
Následne sa nám vytvorí súbor data kde sú už uložené len prefiltrované dáta. Jeho veľkosť je približne 400 MB a obsahuje informácie o 1159110 osobách.
Tento súbor je následne vstupom pre program WikiParser ktorý ma za úlohu dať dátam konkrétny a jednotný tvar, s ktorým sa bude ďalej pracovať. Pre ukážkové vstupy a výstupy pozri zložku Data kde sú uložené všetky ukážkové vstupy aj výstupy.
Wikiparser pomocou regexov skonvertuje jednotlivé typy dátumov na mnou špecifický typ a konkrétne YYYY|MM|DD . Tento typ som si vybral preto aby sa mi to nemiešalo s ostatnými nakoľko tam nebol formát dátumu ktorý by obsahoval znak „|”.
Výstup tohto program je následne vstupom pre program CanTheyMeet, obsahuje logiku zisťovania či sa dve osoby mohli stretnúť podľa ich dátumov narodenia resp. úmrtia.
V programe je definovaná premenná lifelong ktorá nám pomáha určiť či sa dve osoby mohli stretnúť aj bez toho aby sme mali uvedené oba dátumy (narodenia aj úmrtia). Základná hodnota je 100 čiže predpokladáme že osoba žije sto rokov.

[1] Dostupné na http://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2

Štatistiky:

Ako som už vyššie spomenul celkový počet osôb ktoré boli vyparsované z wiki dumpu je 159110.
Z toho som spracoval 1132186 a teda nespracovaných bolo 26924. Tieto data boli nespracované pretože mali špecifický zápis dátumov a program WikiParser ich nevedel transformovať na daný formát.



Dalším zaújmavým údajom je počet osôb ktoré síce boli spracované ale nemali uvedené ani dátum narodenia a ani dátum úmrtia a teda nikdy nebolo možné zistiť informáciu či sa s niekým mohli stretnúť. Na nižšie uvedenom grafe je vidieť pomer dát.