Modelowanie wielowymiarowe i transformacja danych

2 marca 2019

Opis pliku z zadaniami

Wszystkie zadania na zajęciach będą przekazywane w postaci plików `.pdf`, sformatowanych podobnie do tego dokumentu. Zadania będą różnego rodzaju. Za każdym razem będą one odpowiednio oznaczone:

- Zadania do wykonania na zajęciach oznaczone są symbolem $\triangle$ – nie są one punktowane, ale należy je wykonać w czasie zajęć.
- Punktowane zadania do wykonania na zajęciach oznaczone są symbolem $\diamond$ – należy je wykonać na zajęciach i zaprezentować prowadzącemu, w wypadku nie wykonania zadania w czasie zajęć lub nieobecności, zadania staje się zadaniem do wykonania w domu ($\star$).
- Zadania do wykonania w domu oznaczone są symbolem $\star$ – są one punktowane, należy je dostarczyć w sposób podany przez prowadzącego i w wyznaczonym terminie (zwykle przed kolejnymi zajęciami).
- Zadania programistyczne można wykonywać w dowolnym języku programowania, używając jedynie biblioteki standardowej dostępnej dla tego języka.

1 Studium przypadku △

Treść

Podczas zajęć laboratoryjnych będziemy używać danych związanych z konkursem *Million Song Dataset Challenge*. Dotyczy on stworzenia systemu rekomendującego piosenki dla użytkowników pewnego serwisu. Dokładny opis danych można znaleźć na stronach:

- http://www.kaggle.com/c/msdchallenge
- http://labrosa.ee.columbia.edu/millionsong/

W naszych zadaniach zrobimy pierwsze kroki w kierunku stworzenia systemu rekomendacyjnego.

Na zajęciach będziemy wykorzystywać okrojony i zmodyfikowany zbiór danych Million Song Dataset (MSD). Należy pobrać dwa pliki ze strony przedmiotu z następującymi informacjami:

- `unique_tracks.txt` – zawiera informacje takie jak identyfikator ścieżki, identyfikator utworu, nazwę artysty oraz tytuł utworu,
- `triplets_sample_20p.txt` – zawiera identyfikator użytkownika, identyfikator utworu oraz datę odsłuchania.

W ostatnim ćwiczeniu celem będzie obliczenie odpowiedzi na następujące zapytania:

- Ranking popularności utworów,
- Ranking użytkowników ze względu na największą liczbę odsłuchanych unikalnych utworów,
- Artysta z największą liczbą odsłuchań,
- Sumaryczna liczba odsłuchań w podziale na poszczególne miesiące,
- Wszyscy użytkownicy, którzy odsłuchali wszystkie trzy najbardziej popularne piosenki zespołu Queen.

Jednak wcześniej zadaniem będzie obliczenie podstawowych statystyk dla tych danych, identyfikacja głównych mankamentów tych danych oraz zaproponowanie sposobu na ich rozwiązanie. Następnie zadaniem będzie zaimplementowanie transformacji danych do postaci, która ułatwi przetwarzanie tych danych oraz będzie pozbawiona głównych problemów. Wszystkie obliczenia należy zrealizować wykorzystując narzędzia powłoki systemowej (np., `bash`), takie jak: `head`, `cat`, `paste`, `sort`, `uniq`, `grep`, `sed`, `wc`, `cut`, `comm`, `echo`, `date`, `join`, `nl`, `tr`, `date`, `uniq`, `awk` (wspierający również tabele mieszające i przetwarzanie wielu plików). Do pomiaru czasu trwania poszczególnych operacji można wykorzystać program `time`.

Użytkownicy systemu `MacOS` powinni zwrócić uwagę na różnice w implementacji narzędzi powłoki systemowej. W celu ich zniwelowanie można zainstalować wersję `GNU` narzędzi.

2 Obliczanie podstawowych statystyk *5p.*◇

Treść

Policz następujące statystyki dla obu wejściowych plików danych:

1. Liczba wierszy i unikalna liczba wierszy
2. Liczba unikalnych wartości w każdej kolumnie
3. Przedział dat, dla których zebrane są dane (**uwaga**: znacznik czasowych jest wartością wygenerowaną losowo)

Przed obliczeniem wszelkich statystyk warto zajrzeć jak wyglądają przykładowe wiersze w obu plikach:

```
head unique_tracks.txt
head triplets_sample_20p.txt
```

Jednym z problemów w liczeniu statystyk będzie podział obu plików na kolumny ze względu na wykorzystanie specyficznego separatora. Jak możemy rozwiązać ten problem? Spróbuj wykorzystać takie narzędzia jak `fold`, `sort` oraz `tr`. W celu zamiany separatora skorzystaj z narzędzia `sed`.

W celu policzenia liczby wierszy i unikalnej liczby wierszy dla pierwszego zbioru danych wystarczy zapisać następujące komendy:

```
wc -l unique_tracks.txt
sort unique_tracks.txt | uniq | wc -l
```

Policzenie unikalnej liczby wartości w pierwszej kolumnie dla pierwszego zbioru danych może zostać obliczony przynajmniej na dwa sposoby:

```
cat unique_tracks-transformed.txt | cut -d "," -f1 | sort | uniq
    | wc -l

cat unique_tracks-transformed.txt | awk -FS "," 'BEGIN{c = 0} {
    if(!t[$1]) c++} END{print c}'

sort unique_tracks.txt | uniq | wc -l
```

3 Poprawiony schemat bazy danych △

Treść

Przeanalizuj oryginalny schemat danych dla problemu Million Song Dataset (MSD). Zastanów się, jakie widzisz problemy. Na podstawie analizy oraz wyników z poprzedniego zadania zaproponuj nowy schemat.

Zastanów się, jakie elementy można poprawić w oryginalnym schemacie, tak aby zapewnić jego elastyczność (np. nowe informacje na temat użytkowników), dokładniejszą obsługę informacji o dacie, mniejszą złożoność pamięciową oraz lepszą wydajność czasową przy rozrastających się informacjach o poszczególnych obiektach w tym problemie.

Przy projektowaniu nowego schematu przypomnij sobie informacje dotyczące schematu gwiazdy przedstawione na wykładzie.

4 Transformacja danych z linii komend *10p.*◇

Treść

Zrealizuj zadanie transformacji danych do przygotowanego schematu przy użyciu komend unixowej powłoki systemowej (np., bash). Głównymi zadaniami jest generacja sztucznych kluczy głównych oraz odpowiednie rozdzielenie kolumn do wynikowych tabel. Do tego problemu można podejść na dwa sposoby. W pierwszym wykorzystywane są tabele mieszające. Dlatego warto wykorzystać program `awk` (program ten potrafi obsługiwać na wejściu więcej niż jeden plik).

Drugi sposób polega na sortowaniu. W jego realizacji warto wykorzystać takie narzędzia jak `sort`, `join` oraz `cut` i `nl`. Poniżej przedstawiony jest kod, który tworzy wymiar utworu.

```
time join -t "," -1 2 -2 2 -o 1.1,2.1,2.2,2.3,2.4 \
  <(cat unique_tracks-transformed.txt | cut -d ',' -f2 | sort -u
      | nl -s "," -w1) \
  <(cat unique_tracks-transformed.txt | sort -t "," -u -k2,2) >
     songs-s
```

Zadanie to ma dwa cele. Pierwszym jest doskonalenie znajomości narzędzi powłoki systemowej. Drugim jest opracowanie i realizacja planu wykonania procedury transformacji danych wykorzystując elementarne operacje haszowania oraz sortowania.

5 Zapytania z linii komend *10p.*◇

Treść

Używając powłoki systemowej można także zadać zapytania bazodanowe. Rozpatrzmy pierwsze zapytanie wspomniane na początku tego dokumentu. Jego celem jest wyświetlenie rankingu popularności utworów. Poniżej przedstawione są dwa sposoby zapisania tego zapytania w powłoce systemowej.

Pierwszy sposób wykorzystuje tablice mieszające w programie `awk`.

```
time gawk '
BEGIN{FS=","; OFS=","}
{count[$3]++}
END{for (s in count){print s,count[s]};}
' facts | sort -t ',' -k2nr,2 | head | gawk '
BEGIN{FS=","; OFS=","}
FILENAME=="-"{songid[$1]=$2}
FILENAME=="songs"{if ($1 in songid) print $1,$5,songid[$1]}
' - songs | sort -t ',' -k3nr,3
```

Drugi sposób bazuje na sortowaniu.

```
join -t ',' -1 2 -2 1 -o 1.2,2.4,2.5,1.1 \
  <(cat facts-s | cut -d ',' -f3 | sort | uniq -c | sed 's/^
      *//;s/ /,/' | sort -t ',' -k1rn,1 | head | sort -t ',' -k2
      ,2) \
  <(cat songs-s | sort -t ',' -k1,1) | sort -t ',' -k4nr,4
```

Korzystając z powyższych przykładów wykonaj pozostałe zapytania.

Poniżej przedstawione są wyniki dla wszystkich pięciu zapytań (dla wszystkich zapytań wyświetl maksymalnie 10 wierszy, w ostatnim zapytaniu wyświetl dodatkowo liczbę wszystkich użytkowników).

- Zapytanie 1:

```
You're The One Dwight Yoakam 145267
Undo Bjork 129778
Revelry Kings Of Leon 105162
Sehr kosmisch Harmonia 84981
Horn Concerto No. 4 in E flat K495: II. Romance (Andante
    cantabile) Barry Tuckwell/Academy of St Martin-in-the-
    Fields/Sir Neville Marriner 77632
Dog Days Are Over (Radio Edit) Florence + The Machine 71300
Secrets OneRepublic 58472
Canada Five Iron Frenzy 54655
Invalid Tub Ring 53494
Ain't Misbehavin Sam Cooke 49073
```

- Zapytanie 2:

```
ec6dfcf19485cb011e0b22637075037aae34cf26 1040
119b7c88d58d0c6eb051365c103da5caf817bea6 641
b7c24f770be6b802805ac0e2106624a517643c17 637
4e73d9e058d2b1f2dba9c1fe4a8f416f9f58364f 592
d7d2d888ae04d16e994d6964214a1de81392ee04 586
6d625c6557df84b60d90426c0116138b617b9449 584
113255a012b2affeab62607563d03fbdf31b08e7 561
c1255748c06ee3f6440c51c439446886c7807095 547
db6a78c78c9239aba33861dae7611a6893fb27d5 529
99ac3d883681e21ea68071019dba828ce76fe94d 499
```

- Zapytanie 3:

```
Coldplay 201081
```

- Zapytanie 4:

```
01 2353423
02 2142793
03 2354696
04 2280733
05 2358146
06 2277770
07 2353362
08 2354811
09 2281371
10 2355043
11 2278369
12 2338840
```

- Zapytanie 5:

```
$Number of users: 34
00832bf55ed890afeb2b163024fbcfaf58803098
01cb7e60ba11f9b96e9899dd8da74c277145066e
0ac20218b5168c10b8075f1f8d4aff2746a2da39
1084d826f98b307256723cc5e9a3590600b87399
11abd6aaa54a50ed5575e8af9a485db752b97b42
28daf225834bae38f86555c8a03bca3bbf0e535d
429303f0cacab81f0c03ddfd7c2d472c8373e130
476a5902414891326ebcd8f6d9b5849f462704fa
4cd2428f7bfcff1e2423bbdfc1437a1572c23700
5283f472d868bfac68805acb83f35fd7142e3afd
```