Przetwarzanie masywnych danych

12 grudnia 2018

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.
- 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).

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.

2 Przetwarzanie masywnych danych *10p.*◇

Treść

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 utworu, identyfikator wykonania, nazwę artysty oraz tytuł utworu,
- `triplets_sample_20p.txt` – zawiera identyfikator użytkownika, identyfikator utworu oraz datę odsłuchania.

Wykorzystując dowolną technologię oblicz 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.

Podczas rozwiązywania zadania zwróć uwagę na czas przygotowania odpowiednich struktur danych i ich rozmiaru oraz czas wykonywania zapytań. Na następnych zajęciach każde rozwiązanie będzie dokładnie omówione z punktu widzenia wykorzystanej technologii, czasu wykonywania i zajętości pamięci. Zaproponowane rozwiązanie może być dopasowane do tego konkretnego zadania. Podejście cechujące się ogólnością będzie rozważane na kolejnych zajęciach. Zwróć także uwagę na jakość danych (tzn. występujące anomalie w danych).

Rozwiązanie powinno działać zgodnie z poniższą specyfikacją techniczną:

1. Wykonanie odpowiedniego skryptu z rozwiązaniem powinno zostać przedstawione podczas zajęć lub jako zrzut wyjścia konsoli systemowej.
2. Rozwiązania dla wszystkich podpunktów należy wypisać na standardowe wyjście jedno po drugim.
3. Dla zapytania 1. wypisz na standardowe wyjście 10 najpopularniejszych piosenek posortowanych w kolejności malejącej, każdą piosenkę w osobnej linii w formacie:

```
<tytuł utworu 1> <nazwa wykonawcy 1> <ilość odsłuchań 1>
<tytuł utworu 2> <nazwa wykonawcy 2> <ilość odsłuchań 2>
...
<tytuł utworu 10> <nazwa wykonawcy 10>  <ilość odsłuchań 10>
```

4. Dla zapytania 2. wypisz na standardowe wyjście 10 użytkowników z największą liczbą odsłuchanych unikalnych piosenek posortowanych w kolejności malejącej, każdego użytkownika w osobnej linii w formacie:

```
<id użytkownika 1> <ilość odsłuchanych unikatowych utworów 1>
<id użytkownika 2> <ilość odsłuchanych unikatowych piosenek 2>
...
<id użytkownika 10> <ilość odsłuchanych unikatowych piosenek 10>
```

5. Dla zapytania 3. wypisz na standardowe wyjście nazwę najpopularniejszego wykonawcy w formacie:

```
<nazwa wykonawcy> <sumaryczna ilość odsłuchań jego utworów>
```

6. Dla zapytania 4. wypisz na standardowe wyjście miesiące, każdy w osobnej linii w formacie:

```
1 <sumaryczna ilość odsłuchań w miesiącu 1>
2 <sumaryczna ilość odsłuchań w miesiącu 2>
3 <sumaryczna ilość odsłuchań w miesiącu 3>
...
12 <sumaryczna ilość odsłuchań w miesiącu 12>
```

   lub

```
01 <sumaryczna ilość odsłuchań w miesiącu 1>
02 <sumaryczna ilość odsłuchań w miesiącu 2>
03 <sumaryczna ilość odsłuchań w miesiącu 3>
...
12 <sumaryczna ilość odsłuchań w miesiącu 12>
```

7. Dla podpunktu 5. wypisz na standardowe wyjście 10 pierwszych id użytkowników w osobnych liniach spełniających warunek zgodnie z porządkiem alfabetycznym (posortowanych rosnąco po id) oraz liczbę wszystkich takich użytkowników w formacie:

```
<id użytkownika 1>
<id użytkownika 2>
...
<id użytkownika 10>
Number of users: <liczba użytkowników spełniających warunek>
```