Schemat i transformacja danych

25 listopada 2017

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. Niestety ze względu na ograniczoną liczbę godzin ograniczymy się do zagadnień związanych z hurtowniami danych, wielowymiarowym modelowaniem, efektywną reprezentacją danych oraz prostymi zapytaniami analitycznymi.

2 Schemat danych △

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.

Dane są już w postaci, która pozwala na przetwarzanie prostych zapytań analitycznych. Jednak schemat ten można jeszcze poprawić. Zaproponuj modyfikację schematu tak, aby w efektywny sposób można było otrzymać odpowiedzi na następujące zapytania:

- Ranking popularności piosenek,
- 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.

3 Transformacja danych do nowego schematu *25p.*⋆

Treść

Należy przeprowadzić dane z bieżącej postaci do nowego schematu i umieścić je w systemie zarządzania bazą danych lub przechowywać je w inny sposób ułatwiający przetwarzanie danych. Po przeprowadzeniu transformacji wykonaj odpowiednie zapytania lub procedury, aby otrzymać odpowiedzi na zapytania z poprzedniego zadania.

W ramach zadania należy wyszczególnić i zaplanować wszystkie kroki transformacji, a następnie je wykonać używając dowolnych technik. Ocenie nie będzie podlegać efektywność wybranego rozwiązania. Jednak plan i wybrana technologia musi prowadzić do wykonania zadania w racjonalnym czasie.

Zadanie należy rozpocząć na zajęciach i dokończyć w domu. Wyniki należy zaprezentować w krótkim raporcie (maksymalnie 5 stron). Szablon raportu dostępny jest na stronie przedmiotu (`raport-2.tex`). Powinien on zawierać:

- schemat danych w formie początkowej oraz końcowej (czyli po transformacji),
- opis zaproponowanego procesu transformacji,
- czas przetwarzania oraz rozmiar danych po transformacji (dla poszczególnych relacji, indeksów i innych dodatkowych struktur danych, a także łącznie),
- postać wykonanych zapytań lub procedur razem z wynikiem i czasem wykonania (wyniki należy podać dla maksymalnie 15 pierwszych wierszy, dla każdego zapytania należy podać również liczbę wynikowych wierszy),
- porównanie oryginalnego i przetworzonego schematu pod względem wydajności i elastyczności wykorzystania.
- krótkie podsumowanie (np. zalety i wady wybranego rozwiązania, dyskusja na temat napotkanych trudności).