Zestaw 1 – Netflix Prize Data

Pochodzenie danych to `https://www.kaggle.com/netflix-inc/netflix-prize-data`

Dane zawierają oceny filmów wystawione przez użytkowników platformy Netflix.

Zbiór danych

Wykorzystywane są dwa zbiory danych.

Pierwszy `netflix-prize-data.zip` to główny, "strumieniowy" zbiór plików mających format csv i następujące pola:

- `date` – data wystawienia oceny w formacie `YYYY-MM-DD`
- `film_id` – identyfikator filmu
- `user_id` – identyfikator użytkownika
- `rate` – ocena filmu

Drugi zbiór statyczny `movie_titles.csv` zawiera następujące pola:

- `ID` – identyfikator filmu
- `Year` – rok produkcji
- `Title` – tytuł filmu

ETL – obraz czasu rzeczywistego

Agregacja na poziomie filmów (id i tytuł filmu) w poszczególnych miesiącach.

Wartości agregatów to:

- liczba ocen
- suma ocen
- liczba (unikalnych) osób, która dokonała oceny (zakładamy możliwość wielokrotnej oceny pojedynczego filmu przez poszczególne osoby.

Wykrywanie "anomalii"

Wykrywanie "anomalii" ma polegać na wykrywaniu wysokiego zainteresowania danym filmem w danym okresie czasu. Program ma być parametryzowany przez:

- D – długość okresu czasu wyrażoną w dniach
- L – liczbę ocen (minimalna)
- O – średnią ocenę (minimalna)

Wykrywanie anomalii ma być dokonywane każdego dnia.

Przykładowo, dla parametrów D=30, L=100, O=4 program każdego dnia będzie raportował te filmy, które w ciągu ostatnich 30 dni uzyskały co najmniej 100 ocen dających średnią ocenę co najmniej 4.

Raportowane dane mają zawierać

- analizowany okres - okno (start i stop)
- tytuł filmu
- liczbę ocen
- średnią ocenę

Załóż, że dane mogą być nieuporządkowane – mogą być opóźnione o jeden dzień.