Hurtownie danych i systemy Business Intelligence
Opis projektu

Autorzy: Jakub Kosterna, Patryk Wrona
Grupa ***Eurowizjanie***

Celem projektu jest zbudowanie hurtowni danych biorącej pod lupę Konkurs Piosenki Eurowizji. W najbliższych tygodniach zostanie przygotowany projekt ETL zaprojektowany w Visual Studio Integration Services. Będzie on przyjmował następujące źródła danych:

1. Tabela zawierająca informacje o wszystkich przyznanych punktach przez 45 lat trwania konkursu:
   https://data.world/datagraver/eurovision-song-contest-scores-1975-2019
   Dane będziemy pobierać **bezpośrednio z internetowego API *Data World***.

   Ramka zawiera ponad 50 000 obserwacji, gdzie każda odpowiada za ilość przyznanych punktów krajowi X od państwa Y w konkursie typu Z (pierwszy półfinał, drugi półfinał, finał), w sekcji punkty od jury / punkty od telewidzów.
   Poprawność danych została przetestowana z pomocą skryptu C++ - patrz *escVeryficationScript.cpp*, z użytkiem bazy przekształconej do odpowiedniego pliku .txt - *eurowizja.txt.* Rzeczywiście potwierdza on stan rzeczywisty, co zostało potwierdzone algorytmami wyznaczającymi sumaryczną liczbę zwycięstw poszczególnych państw oraz punkty finałowych zwycięskich występów.

   W pliku znajdują się pojedyncze błędy, takie jak parę literówek czy pojedyncze złe oznaczenie informacji, czy mamy do czynienia z wierszem punktów od państwa do tego samego państwa - ale wszystkie obserwacje źle wpisane zostały namierzone i będą łatwe do obsługi i transformacji w Visual Studio.

2. Ramka informacji o propozycjach na przełomie lat - zawierająca m. in. nazwę wykonawcy, tytuł, pozycje startowe w kolejnych pod-konkursach, a także tekst utworu
   https://github.com/Spijkervet/eurovision-dataset
   Jest to wolno **dostępny pojedynczy plik .csv**

3. … ? W planach znalezienie danych ekonomicznych i podobnych państw biorących udział w Eurowizji. Prawdopodobnie zostanie wzięty pod uwagę zbiór dotyczący wskaźnika HDI, najlepiej na przełomie lat. Potencjalne dodatkowe źródło danych:
   https://data.imf.org/?sk=388DFA60-1D26-4ADE-B505-A05A558D9A42&sId=1479329132316

4. (… ?)² Zastanawiamy się także nad połączeniem wykonawców i artystów z danymi o utworach ze *Spotify*, co przy sukcesywnym połączeniu dużej liczby utworów dałoby między innymi możliwość analizy gatunków (możliwe, że także i w hierarchii do nich przeznaczonej), tempa (uderzenia na minutę) czy specjalnych miar tworzonych i prezentowany przez szwedzki serwis - takich jak chociażby *danceability*.

W ramach projektu zostanie przygotowanych kilka hierarchii, w tym:

- geograficzna: państwo → część kontynentu → kontynent; np.
    - Polska → Europa Środkowa → Europa
    - Azerbejdżan → Azja Zachodnia → Azja
    - Australia → Australia → Australia

  Dane zostaną przygotowane ręcznie, ze względu na małą liczbę krajów biorących udział w konkursie (około 50), na podstawie informacji w Wikipedii.
- czasowa: rok → dekada → stulecie

W zamyśle hurtownia ma odpowiadać na pytania takie jak:

1) Czy dłuższe/krótsze piosenki (pod względem liczby słów w tekście) standardowo zyskują więcej aprobaty od jury? Hipoteza: bardziej rozbudowana treść powinna zwykle zdobywać więcej punktów u jurorów.
2) Czy dłuższe/krótsze teksty utworów są bardziej przekonujące dla telewidzów? Hipoteza: telewidzowie preferują prostsze piosenki, a co za tym idzie - krótsze pod względem przekazu lirycznego.
3) Czy jest zależność między pozycją startową a końcowym wynikiem? Hipoteza: utwory rozpoczynające konkurs kończą średnio na lepszej pozycji końcowej niż piosenki artystów, którzy prezentują swoje występy jako drudzy czy trzeci.
4) Czy bardziej rozwinięte / bogatsze kraje osiągają lepsze rezultaty w Konkursie? Albo: czy jest zależność między liczbą inwestycji państwa, a wynikiem? Hipoteza: tak.