Analiza bazy danych top 1000 filmów z IMDb

Dokumentacja projektu

Łukasz Kiecka

Dokumentacja:

Program służy do analizy bazy danych top 1000 filmów z IMDb. Wszystkie opcje przeznaczone do prowadzenia analizy są dostępne z poziomu stworzonego GUI.

Do analizy korzystamy z bibliotek takich jak:

* PyQt5
* Pandas
* Numpy
* Plotly

Na początku operujemy na kolumnach wczytanej bazy danych, aby odpowiednią ją przeformatować.

Poszczególne elementy GUI umieszczone są dzięki QGridLayout, gdzie:

U góry znajdziemy zdjęcia

Po lewej stronie GUI znajdziemy przyciski generujące wykresy kolumnowe dotyczące odpowiednych danych gdzie oś Y to ilość wystąpień a oś X to domyślnie (bez zmiany wartości na umieszonym na dole GUI QSpinBoxie) 3 odpowiednie wartości (nazwa przycisku określa jakie). W komendach, gdzie odpowiednia funkcja jest podpięta do przycisku korzystamy z „lambda”, aby bezproblemowo móc użyć wewnątrz funkcji, aby docelowo dla pustego QSpinBoxa wyświetlane były 3 wartości na osi X. Dla przykładu po wpisaniu 12 w QSpinBox i kliknięciu przycisku „reżyser” w przeglądarce (biblioteka plotly) otwiera nam się taki wykres kolumnowy z odpowiednio ustawionymi kolorami pasującymi do motywu strony IMDb:

Po lewej stronie znajduje się również przycisk „Tabela kolumnowa - ilość filmów w latach 1920-2020”, który niezależnie od liczby w spinboxie prezentuje wykres ilości filmów w tych latach.

Wszystkie użyteczności po lewej stronie używają plotly.express i odpowiednio „bar” oraz „histogram”. One pobierają odpowiednio posortowane dane z data\_frame oraz wyświtlają te dane z odpowiednio ustwionym layoutem.

Tło w GUI ustawione jest dzięki QImage, QPalette, QBrush, a 3 inne obrazy dzięki QPixmap. Każde z przycisków są to QPushButton w tym te obsługujące filtracje(po lewej stronie GUI) mają dodatkowo przypisany setToolTip w celu jaśniejszego przedstawienia możliwości programu.

Po środku znajduję się przycisk generujący graf połączeń wykorzystujący bibliotekę plotly (plotly.graph\_objects). Na osiach X – rok wydania, Y – Wynik Metascore wyświtla on odpowiednio filmy jako kropki oraz łączy, te które posiadają tego samego reżysera. Operujemy tutaj sporo na słownikach i listach oraz dodajemy odpowiednie wartości do nodes oraz egdes, aby zostały one potem przedstawione na grafie.

Po prawej stronie GUI znajdziemy przyciski aktywujące konkretne funkcje filtrujące dane. Odpowiednie wartości wykorzystywane w filtracji wpisujemy w QLineEdit oraz po kliknięciu przycisku otrzymujemy informację zwrotną (odpowiedź lub prośba o wpisanie odpowiedniej watrości w pole) dzięki osobnemu okienku. Szczególowy opis powyższych funkcji do filtracji znajduję się poniżej:

Funkcje:

Funkcja filtr\_aktor pobiera od użytkownika imię i nazwisko aktora, a następnie w pętli while, która wykonuje się tylko kiedy warunek jest prawdziwy za pomocą pętli for porównuje ciąg znaków wpisany przez użytkownika z 4 wektorami zawierającym dane dotyczące aktorów.Jeśli w którymkolwiek wektorze w konkretnym sprawdzanym w danym momencie i-tym wierszu, zapisany jest ten sam ciąg znaków, co podał użytkownik to do wektora o nazwie tytul jest dopisywany nowy element odpowiadający tytułowi filmu, który występuje w sprawdzanym w danym momencie i-tym wierszu wektora series\_title, za pomocą funkcji append oraz zmienna maks, która początkowo ma wartość 0 zwiększa się o jeden. Pętla ta wykonywana jest n razy, gdzie n jest długością wektora star1 uzyskaną za pomocą funkcji len, która zwraca liczbę jego elementów. Pętla zaczyna się od i = 0 a w momencie kiedy i osiągnie wartość n pętla zakończy swoje działanie i dla i=n nie będą wykonywały się żadne instrukcje znajdujące się w pętli. Po zakończeniu działania pętli for sprawdzany jest warunek logiczny. Jeśli zmienna maks po wykonaniu całej pętli ma wartość 0 to oznacza, że w bazie danych nie znajduje się podanym przez użytkownika aktor i program wyświetla taki komunikat w nowym oknie. W przeciwnym wypadku zmienna again zmienia wartość na 0 dzięki czemu kończy się działanie pętli while, a następnie funkcja zwraca tytuły filmów, w których grał podany aktor.

Funkcja czas\_trwania po pobraniu od użytkownika tytułu filmu wykonuje w pętli while, która wykonuje się tylko kiedy warunek jest prawdziwy, pętle for. Pętla for wykonuje się od i=0 do a w momencie kiedy i=n pętla kończy swoje działanie ( n liczba elementów wektora series\_title uzyskana za pomocą funkcji len). Podczas działania pętli sprawdzany jest warunek logiczny czy w wektorze series\_title w i-tym wierszu znajduje się taki sam ciąg znaków co podany przez użytkownika. Jeśli warunek ten jest prawdziwy to zmienna czas\_trwania przyjmuje wartość odpowiadającą elementowi znajdującemu się w i-tym wierszu wektora runtime, czyli długość trwania konkretnego filmu. Po zakończeniu działania pętli for sprawdzany jest warunek logiczny czy zmienna czas\_trwania jest równa 0. Jeśli tak jest to oznacza, że w bazie danych nie ma informacji o podanym przez użytkownika filmie i użytkownik dostaje taką informację. Natomiast w przeciwnym razie pętla while kończy swoje działanie i funkcja zwraca informację o długości trwania danego filmu.

Funkcja najbardziej\_kasowy\_film\_z\_calej\_bazy zwraca tytuł oraz wynik filmu, który jest najbardziej rentowny z całej bazy danych. W pętli for, która wykonuje się n razy, gdzie n odpowiada liczbie elementów wektora gross, program sprawdza warunek logiczny czy i-ty element wektora gross jest większy od zmiennej maks, która wyjściowo przyjmuje wartość 0. Jeśli warunek jest prawdziwy to zmienna maks przyjmuje wartość odpowiadającą elementowi z i-tego wiersza wektora gross oraz zmienna tytul przyjmuje wartość odpowiadającą i-temu elementowi wektora series\_title. Po zakończeniu działania pętli funkcja zwraca tytuł oraz wynik filmu.

Funkcja najbardziej\_kasowy\_film jest podobna do funkcji powyżej, ale zwraca najbardziej kasowy film dla podanego przez użytkownika roku. Funkcja po pobraniu od użytkownika roku wykonuje w pętli while, która wykonuje się tylko kiedy warunek jest prawdziwy, pętle for. Pętla for wykonuje się od i=0 do a w momencie kiedy i=n pętla kończy swoje działanie ( n liczba elementów wektora gross uzyskana za pomocą funkcji len). Podczas działania pętli sprawdzany jest warunek logiczny czy w wektorze released\_year i-ty element ma tą samą wartość co rok podany przez użytkownika. Jeśli warunek ten jest prawdziwy to zmienna maks(wyjściowa wartość 0) przyjmuje wartość odpowiadającą i-temu elementowi wektora gross, a zmienna tytul przyjmuje wartość i-tego elementu wektora series\_title. Po zakończeniu działania pętli for sprawdzany jest warunek logiczny czy zmienna maks jest równa 0. Jeśli tak jest to oznacza, że w bazie danych nie ma informacji o podanym przez użytkownika roku lub podał on błędne dane i użytkownik dostaje taką informację. Natomiast w przeciwnym razie pętla while kończy swoje działanie i funkcja zwraca informację o tytule filmu i jego wyniku.

Funkcja opis\_filmu zwraca opis filmu do podanego przez użytkownika filmu. Funkcja po pobraniu od użytkownika tytułu filmu wykonuje w pętli while, która wykonuje się tylko kiedy warunek jest prawdziwy, pętle for. Pętla for wykonuje się od i=0 do a w momencie kiedy i=n pętla kończy swoje działanie ( n liczba elementów wektora series\_title uzyskana za pomocą funkcji len). Podczas działania pętli sprawdzany jest warunek logiczny czy w wektorze series\_title i-ty element ma tą samą wartość co tytuł podany przez użytkownika. Jeśli warunek ten jest prawdziwy to zmienna opis(wyjściowa wartość- pusty ciąg znaków) przyjmuje wartość odpowiadającą i-temu elementowi wektora overview. Po zakończeniu działania pętli for sprawdzany jest warunek logiczny czy zmienna opis przechowuje pusty ciąg znaków. Jeśli tak jest to oznacza, że w bazie danych nie ma informacji o podanym przez użytkownika filmie lub podał on błędne dane i użytkownik dostaje taką informację. Natomiast w przeciwnym razie pętla while kończy swoje działanie i funkcja zwraca opis filmu.

Funkcja najlepiej\_oceniane\_filmy zwraca 5 najwyżej ocenianych filmów z podanego przez użytkownika gatunku. Funkcja po pobraniu od użytkownika gatunku wykonuje w pętli while, która wykonuje się tylko kiedy warunek jest prawdziwy, pętle for. Pętla for wykonuje się od i=0 do a w momencie kiedy i=n pętla kończy swoje działanie ( n liczba elementów wektora genre uzyskana za pomocą funkcji len). Podczas działania pętli sprawdzany jest warunek logiczny czy w wektorze genre i-ty element ma tą samą wartość co gatunek podany przez użytkownika. Jeśli warunek ten jest prawdziwy to do macierzy ocena dodawany jest nowy element, który w pierwszej kolumnie posiada ocenę i-tego filmu a w drugiej tytuł i-tego filmu. Po zakończeniu działania pętli for sprawdzany jest warunek logiczny czy liczba elementów macierzy ocena jest równa 0. Jeśli tak jest to oznacza, że w bazie danych nie ma informacji o podanym przez użytkownika gatunku lub podał on błędne dane i użytkownik dostaje taką informację. Natomiast w przeciwnym razie macierz ocena jest sortowana rosnąco za pomocą funkcji sort oraz pętla while kończy swoje działanie, a funkcja zwraca tytuły 5 najlepiej ocenianych filmów, wyświetlając 5 ostatnich wartości z drugiej kolumny macierzy ocena.