# System rekomendacji filmów

Marcin Starzak, Michał Cyran

Celem projektu było opracowanie systemu rekomendacji filmów, który umożliwia przewidywanie ocen filmów na podstawie preferencji użytkowników. System opiera się na danych z bazy Movie Lens Small Latest Dataset (kaggle.com/datasets/...), która zawiera informacje o ocenach filmów, ich tagach oraz gatunkach. W projekcie wykorzystano kombinację metod znajdowania najbliższego wektora cech filmów oraz dekompozycji macierzy metodą Singular Value Decomposition (SVD), aby stworzyć efektywny i jak najbardziej dokładny model rekomendacji.

Model operuje na liście ostatnich *n* ocenionych przez użytkownika filmów jako wejściu. Na tej podstawie, model generuje teoretyczny profil filmu oparty na uśrednionych cechach tagów, a następnie wybiera film, który ma najbliższy wektor cech (o najmniejszej normie) do tego teoretycznego profilu. Kolejnym krokiem jest przekazanie *k* najbardziej zbliżonych filmów do algorytmu SVD, który dokonuje predykcji ocen dla każdego z tych filmów. Proces ten ma na celu ocenę, czy proponowany film pasuje do dotychczasowych preferencji ocenianych przez użytkownika oraz innych użytkowników o podobnych gustach.

## Uzasadnienie Wyboru Techniki/Modelu

Wybraliśmy kombinację metod znajdowania najbliższego wektora cech filmów oraz SVD ze względu na ich komplementarne zalety:

- **Znajdowanie najbliższego wektora cech filmów:**
  - **Bezpośredniość**: Metoda ta pozwala bezpośrednio porównywać filmy na podstawie ich cech, takich jak tagi i gatunki.
  - **Skalowalność**: Algorytm jest skalowalny i może być łatwo dostosowany do dużych zbiorów danych.
  - **Szybkość obliczeń**: Metoda ta jest szybka w obliczeniach, co jest kluczowe dla systemów rekomendacji działających w czasie rzeczywistym.

- **SVD (Singular Value Decomposition):**
  - **Redukcja wymiarowości:** SVD pozwala na zredukowanie wymiarowości danych, co zmniejsza złożoność obliczeniową i zwiększa szybkość modelu.
  - **Wysoka dokładność:** SVD jest znane z wysokiej dokładności w przewidywaniu brakujących wartości w macierzach użytkownik-film.
  - **Elastyczność:** Technika ta dobrze radzi sobie z danymi o różnej gęstości i jest w stanie efektywnie przewidywać oceny nawet dla użytkowników z niewielką ilością ocen.
  - Dodatkowo, wzorowaliśmy się pracą Application of Dimensionality Reduction in Recommender System -- A Case Study

## Opis Danych Wejściowych

Dane wejściowe pochodzą z bazy MovieLens, zawierającej trzy główne pliki:

- **ratings.csv**: Zawiera oceny filmów przez użytkowników.
- **movies.csv**: Zawiera informacje o filmach, takie jak tytuły i gatunki.
- **tags.csv**: Zawiera tagi przypisane filmom przez użytkowników.

Przed przetwarzaniem danych dokonano następujących operacji:

- **Czyszczenie danych**: Usunięto zbędne kolumny (timestamp i userId) oraz tagi pojawiające się tylko raz.
- **Przygotowanie tagów**: Każdemu tagowi przypisano unikalny ID i utworzono macierz tagów dla każdego filmu.
- **Normalizacja ocen**: Oceny zostały znormalizowane przez zastąpienie brakujących ocen średnią oceną filmu oraz odjęcie średniej oceny użytkownika.

## Testowanie modelu

Niestety, z powodu braku odpowiednich danych, nie przeprowadzono testów modelu na dedykowanym zestawie testowym. Wymagałoby to odpowiedniej grupy testerów, którzy ocenialiby przedstawione propozycje. W efekcie, nie posiadamy wartości RMSE ani innych miar oceny jakości modelu.

Przeprowadzaliśmy jednak praktyczne testy, które skutkowały propozycjami filmów o dobrych ocenach i gatunkowo zbliżonych do filmów wejściowych.

## Analiza Wyników

Model oparty na kombinacji metod znajdowania najbliższego wektora cech filmów oraz SVD ma potencjał osiągnięcia zadowalających wyników, jednak konieczne jest przeprowadzenie testów na dedykowanym zestawie testowym w celu oceny jego dokładności i skuteczności. Dalsze badania mogą również obejmować optymalizację hiperparametrów oraz eksplorację innych zaawansowanych technik uczenia maszynowego.

## Propozycje Dalszych Kroków

- **Przeprowadzenie testów**: Konieczne jest przeprowadzenie testów modelu na dedykowanym zestawie testowym w celu dokładnej oceny jego wydajności.
- **Optymalizacja hiperparametrów**: Eksperymentowanie z różnymi wartościami parametrów modelu może pomóc w poprawie jego dokładności i skuteczności.
- **Eksploracja zaawansowanych technik**: Wprowadzenie zaawansowanych metod uczenia maszynowego, takich jak głębokie sieci neuronowe, może dalsze zwiększyć efektywność modelu rekomendacji filmów.