Rafał Wasilewski, 160384

Diagnostyka Systemów

Sprawozdanie

Temat projektu: System rekomendacji filmów lub książek na podstawie dotychczasowych ocen.

Cel projektu

Celem projektu było stworzenie systemu rekomendacji filmów lub książek na podstawie dotychczasowych ocen.

Zbiór danych

Zdecydowano się na realizację systemu rekomendacji filmów ze względu na łatwość w pozyskaniu danych zawierających oceny poszczególnych produkcji. W tym celu posłużono się zbiorem ze strony MovieLens (movielens.org), która docelowo zajmuje się rekomendacją filmów dla użytkowników. Dataset zawiera ponad sto tysięcy ocen, 3686 tagów przeznaczonych dla 9742 filmów. Zbiór danych pochodzi z 2018 roku, jednak do realizacji zadania taki zbiór zdecydowanie wystarczy. Użytkownicy zostali wybrani losowo, każdy z nich ocenił co najmniej 20 filmów.

Możliwe metody realizacji systemu rekomendującego

Jednym z możliwych rozwiązań jest filtrowanie na podstawie treści (ang. Content-Based Filtering). Ta strategia opiera się na danych dostarczonych na temat rzeczy, w tym przypadku filmów. Algorytm rekomenduje produkty, które są podobne do tych, które użytkownik polubił w przeszłości. Podobieństwo jest zazwyczaj obliczane na podstawie danych, które posiadamy o produktach, a także na podstawie wcześniejszych preferencji użytkownika. Przykładowo, jeżeli użytkownik słuchał głównie utworów podobnego typu, to zostaną mu polecone utwory należące do tej samej kategorii.

Innym sposobem jest filtrowanie kolaboratyne (ang. Collaborative Filtering). Filtrowanie to opiera się na kombinacji zachowań użytkownika oraz porównywaniu i zestawianiu ich z zachowaniami innych użytkowników w bazie danych. Główna różnica pomiędzy tymi dwoma filtrami polega na tym, iż na filtrowanie kolaboratywne wpływa interakcja wszystkich użytkowników z elementami, natomiast w przypadku filtrowania opartego na treści brane są pod uwagę tylko dane danego użytkownika. Technika ta dzieli się na dwa typy: oparta na użytkownikach oraz oparta na elementach. W filtrowaniu grupowym opartym na użytkownikach ustalamy wynik podobieństwa między dwoma użytkownikami. Na podstawie podobieństwa polecane są przedmioty lubiane/kupione przez jednego użytkownika drugiemu użytkownikowi, zakładając, że na podstawie podobieństwa obu użytkowników zaprezentowane rzeczy mogą się spodobać drugiej osobie. Przykładem firmy stosującej to rozwiązanie jest Netflix. W przypadku filtrowania opartego na elementach (ang. Item-Based Collaborative Filtering) obliczane jest podobieństwo elementu do istniejącego elementu, który jest używany/obejrzany przez istniejących użytkowników. Następnie na

podstawie stopnia podobieństwa można powiedzieć, że jeśli użytkownik X lubi pozycję A, a nowa pozycja Y jest najbardziej podobna do pozycji A, to polecenie pozycji Y użytkownikowi X jest bardzo sensowne.

Opis rozwiązania

Ze względu na posiadane dane, wymagania projektu oraz wady filtrowania opartego na użytkownikach (ludzie są z reguły kapryśni, gusta zmieniają się od czasu do czasu, dodatkowo macierze zawierające większą ilość użytkowników są trudne w realizacji) zdecydowano się na implementację metody filtrowania opartego o elementy(przedmioty). Opisany zbiór danych pochodzi ze strony Kaggle (https://www.kaggle.com/shubhammehta21/movie-lens-small-latest-dataset). Na wstępie zredukowano ilość danych, aby usprawnić działanie systemu. W zbiorze zostali jedynie użytkownicy, którzy ocenili co najmniej 50 filmów, a filmy muszą posiadać co najmniej 10 ocen. Następnie wartości NaN oznaczające brak oceny zamieniono na bardziej intuicyjne zera. W programie wykorzystano algorytm k najbliższych sąsiadów w celu zidentyfikowania najbardziej pasujących filmów na podstawie ocen użytkowników. Do obliczania odległości, a właściwie do obliczania podobieństwa poszczególnych wektorów (filmów z ocenami jako elementy wektora) wykorzystano podobieństwo cosinusowe wyrażone następującym wzorem:

$$Cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \ \|\vec{b}\|}$$

Przy czym a oraz b to kolejne wektory odpowiadające filmom wraz z ocenami. Odległość euklidesowa nie jest zalecana ze względu na ilość danych i podobieństwo w otrzymywanych wynikach. Podobieństwo jest obliczanie dla każdego wektora, co zajmuje trochę czasu. Na podstawie podobieństwa wybierane są filmy, które mają najwyższy współczynnik podobieństwa. W ten sposób podając tytuł filmu można otrzymać listę rekomendowanych filmów, gdzie liczba filmów polecanych jest równa zmiennej k.

Podsumowanie

Algorytm w sposób zadowalający wskazuje rekomendowane filmy na podstawie ocen użytkowników. Jest to ocena czysto subiektywna, jednak obliczone współczynniki podobieństwa, a także podobieństwo otrzymanych gatunków to potwierdza.