|  |  |
| --- | --- |
| Projekt – sieci neuronowe | Data złożenia projektu: X.Y.Z |
| Numer grupy projektowej: 1 | Imię i nazwisko I: Weronika Bednarz  Imię i nazwisko II: Izabela Bereżnicka |

System Rekomendacji Filmów

# Opis problemu i danych

**Ogólny opis dziedziny, jakiej dotyczą dane.**

Dane są wykorzystywane do tworzenia bardziej wydajnych systemów. Systemy rekomendacji są rodzajem systemów filtrowania informacji, ponieważ poprawiają jakość wyników wyszukiwania i dostarczają elementy, które są bardziej istotne dla wyszukiwanej pozycji lub są związane z historią wyszukiwania użytkownika.

Są one wykorzystywane do przewidywania oceny lub preferencji, które użytkownik nadałby danemu elementowi. Prawie każda duża firma technologiczna zastosowała je w takiej czy innej formie: Amazon używa ich do sugerowania produktów klientom, YouTube używa ich do decydowania, który film ma być odtwarzany jako następny, a Facebook używa ich do polecania stron do polubienia i osób do obserwowania.

W naszym projekcie będziemy budować podstawowy system rekomendacji filmów przy użyciu [TMDB 5000 Movie Dataset](https://www.kaggle.com/tmdb/tmdb-movie-metadata).

**Ilość rekordów, nazwy i ilość zmiennych wejściowych z podziałem na jakościowe i ilościowe.**

Pierwszy zbiór danych zawiera następujące cechy:

* movie\_id - unikalny identyfikator każdego filmu,
* obsada - nazwiska głównych i drugoplanowych aktorów,
* crew - nazwisko reżysera, montażysty, kompozytora, scenarzysty itp.

Drugi zbiór danych ma następujące cechy:

* budżet - budżet, w którym film został nakręcony,
* gatunek - gatunek filmu, akcja, komedia, thriller itp.,
* strona główna - link do strony głównej filmu,
* id - Jest to w rzeczywistości identyfikator filmu, tak jak w pierwszym zestawie danych,
* keywords - słowa kluczowe lub tagi związane z filmem,
* original\_language - język, w którym film został nakręcony,
* original\_title - tytuł filmu przed tłumaczeniem lub adaptacją,
* overview - krótki opis filmu,
* popularity - liczba określająca popularność filmu,
* production\_companies - Dom produkcyjny filmu,
* production\_countries - kraj, w którym film został wyprodukowany,
* release\_date - data premiery filmu,
* revenue - światowy przychód wygenerowany przez film,
* runtime - czas trwania filmu w minutach,
* status - "Released" lub "Rumored",
* tagline - slogan filmu,
* title - Tytuł filmu,
* vote\_average - średnia ocen otrzymanych przez film,
* vote\_count - liczba otrzymanych głosów.

**Podstawowe statystyki (min, max, średnia, odchylenie standardowe).**

**Histogram dla zmiennej wyjściowej. Jaki problem jest rozwiązywany (regresja, klasyfikacja, segmentacja)?**

ll

# Obróbka danych

Ilość brakujących rekordów i sposób rozwiązania tego problemu, normalizacja/standaryzacja danych wejściowych, transformaty wykonane na danych (zależnie od problemu, np. dla szeregów czasowych średnia z ostatnich 10 rekordów, dla rozpoznawania obrazów przetworzenie przez filtry graficzne). Sposób podziału na zbiór uczący oraz zbiór testujący.

# Opis zastosowanych sieci neuronowych

Typ i architektura sieci, ilość epok uczenia, sposób uczenia sieci. Informacja o własnej implementacji/ użytym frameworku. Co najmniej 5 różnych architektur sieci.

# Dyskusja wyników oraz wnioski

Zebrane w tabeli wyniki sieci z podziałem na zbiór uczący i testujący. Co najmniej 2 miary, wraz z komentarzem, dlaczego zostały wybrane takie, a nie inne. Określenie (subiektywne), czy wyniki są satysfakcjonujące wraz z uzasadnieniem. Wnioski, dalsze propozycje rozwoju projektu.