System rekomendacyjny wypożyczalni filmów z rozmytością

Raport końcowy

Piotr Iśtok
Piotr Jacak

Czerwiec 2025

Spis treści

1 Dane

Dane do omawianego systemu rekomendacyjnego zostały pobrane ze strony https://grouplens.org/datasets/movielens/
Zbiór danych zawiera trzy pliki:

- *movies.csv* - tabela z kolumnami: identyfikator filmu, tytuł filmu, gatunek filmu (gatunki wypisane obok siebie, oddzielone znakiem '|').
- *tags.csv* - tabela z kolumnami: identyfikator użytkownika dodającego tag, identyfikator filmu, tag, znacznik czasowy dodania tagu.
- *ratings.csv* - tabela z kolumnami: identyfikator użytkownika, identyfikator filmu, ocena (w skali od 0 do 5), znacznik czasowy dodania oceny.

2 Eksploracyjna Analiza Danych - EDA

2.1 Cel i zakres analizy

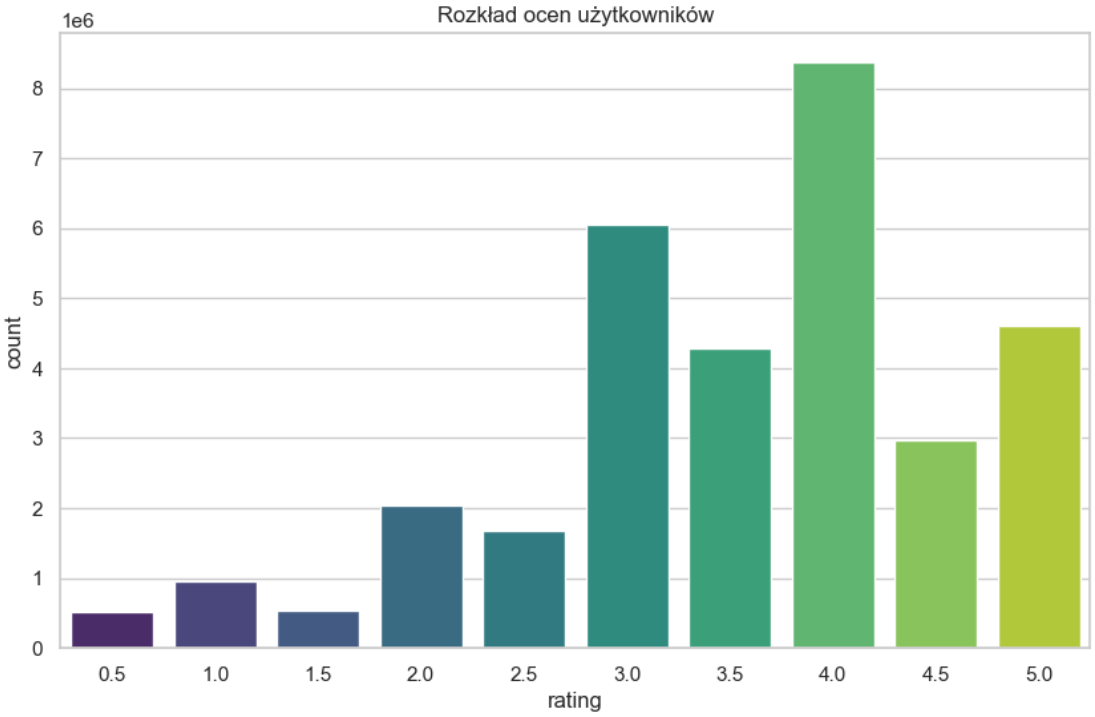
Celem eksploracyjnej analizy danych było lepsze zrozumienie struktury, rozkładu i jakości danych wejściowych oraz identyfikacja potencjalnych problemów, które mogłyby wpłynąć na skuteczność systemu rekomendacyjnego. Analiza objęła trzy podstawowe źródła danych: oceny użytkowników (*ratings.csv*), informacje o filmach (*movies.csv*) oraz przypisane tagi (*tags.csv*).

2.2 Struktura zbioru danych

- Liczba ocen: 32000204
- Liczba użytkowników: 200948
- Liczba unikalnych filmów: 87585
- Liczba tagów: 140979

2.3 Rozkład ocen

Rozkład ocen użytkowników jest asymetryczny, z przewagą ocen pozytywnych – najwięcej ocen znajduje się w przedziale 3.0–4.0. Średnia ocena wynosi 3.54, a mediana to 3.5, co sugeruje umiarkowanie pozytywne nastawienie użytkowników do ocenianych filmów.

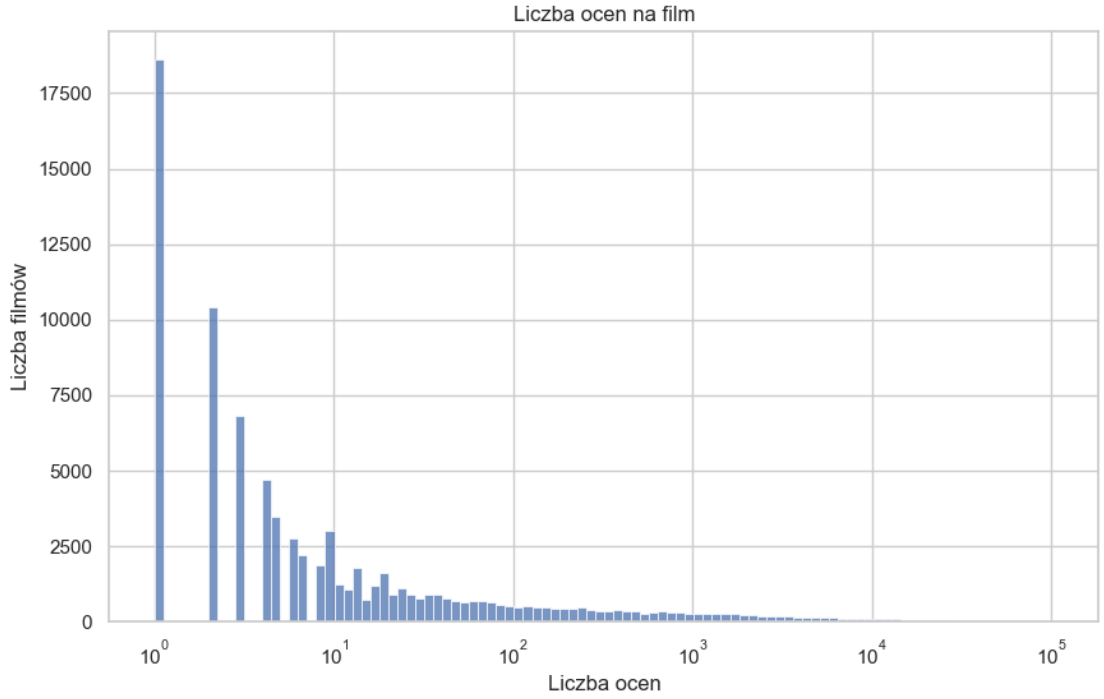


Rysunek 1: Rozkład ocen użytkowników

2.4 Aktywność użytkowników i popularność filmów

- Średnia liczba ocen na użytkownika: 159.25
- Mediana liczby ocen na użytkownika: 73
- Najwięcej ocen użytkownika: 33332
- Najmniej ocen użytkownika: 20

Następnie został jeszcze zbadany rozkład liczby ocen na film. Wyniki są pokazane poniżej.

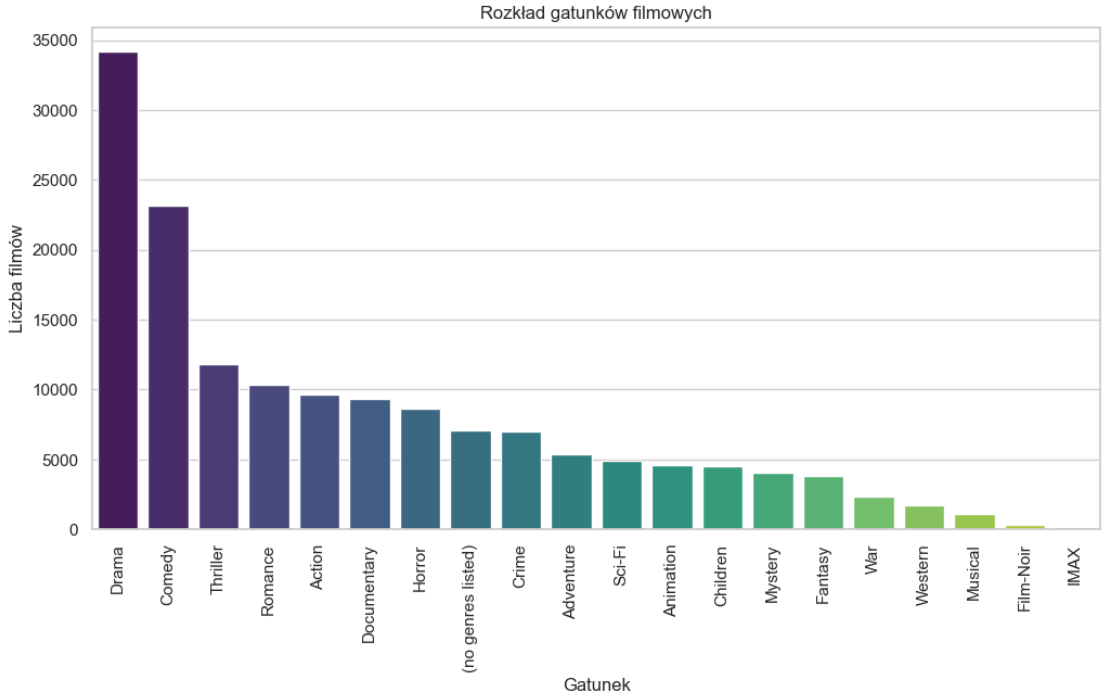


Rysunek 2: Rozkład liczby ocen filmów

Rozkład liczby ocen na użytkownika i na film ukazuje typową sytuację dla systemów rekomendacyjnych – nieliczni użytkownicy są bardzo aktywni, a nieliczne filmy bardzo popularne.

2.5 Gatunki filmów

W analizie uwzględniono zarówno surowe liczenie przynależności filmu do gatunku (bez wag), jak i podejście z uwzględnieniem proporcji, gdy film należy do wielu kategorii. Najczęściej występujące gatunki to: Drama, Comedy i Thriller. Poniżej pokazano wykres rozkładu liczby filmów o danym gatunku.

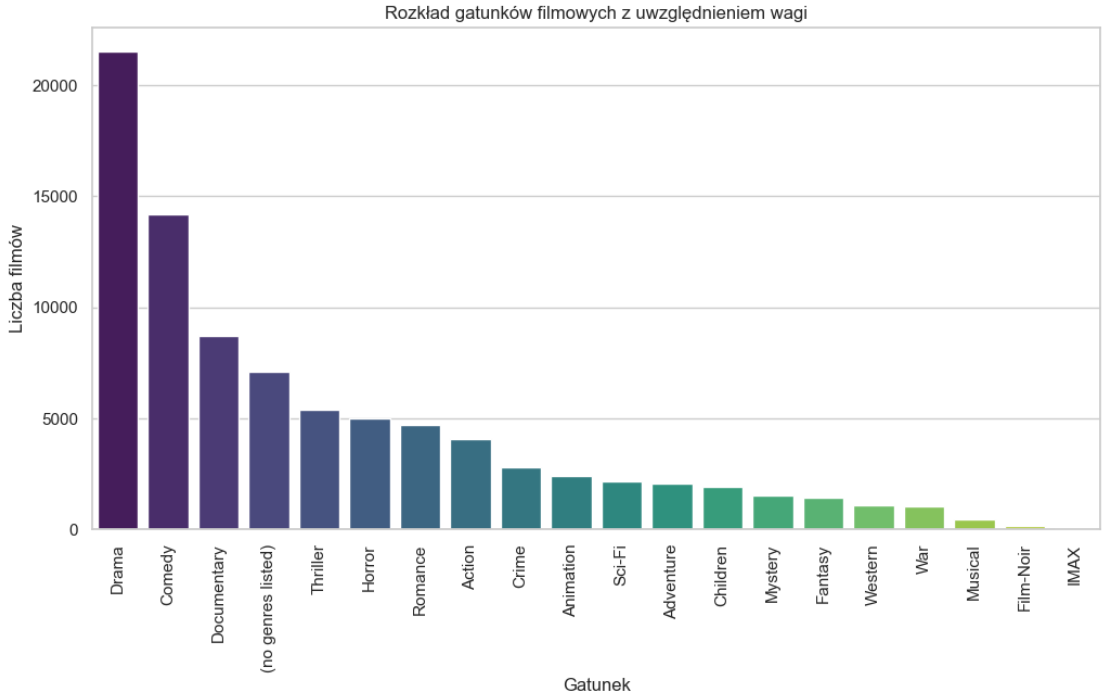


Rysunek 3: Rozkład gatunków filmów

Została następnie wyznaczona mediana, wartość maksymalna oraz obliczona średnia liczba gatunków na film.

- Mediana liczby gatunków na film: 1
- Maksymalna liczba gatunków filmu: 10
- Średnia liczba gatunków na film: 1.76

W drugim wariancie przyjęliśmy, że przynależność filmu do gatunku filmowego będzie dzielona przez ilość wszystkich gatunków, do których należy. W takim przypadku oczywiście nie ma sensu badania średniej, wartości maksymalnej oraz mediany ilości gatunków filmu, bo wynoszą one 1.

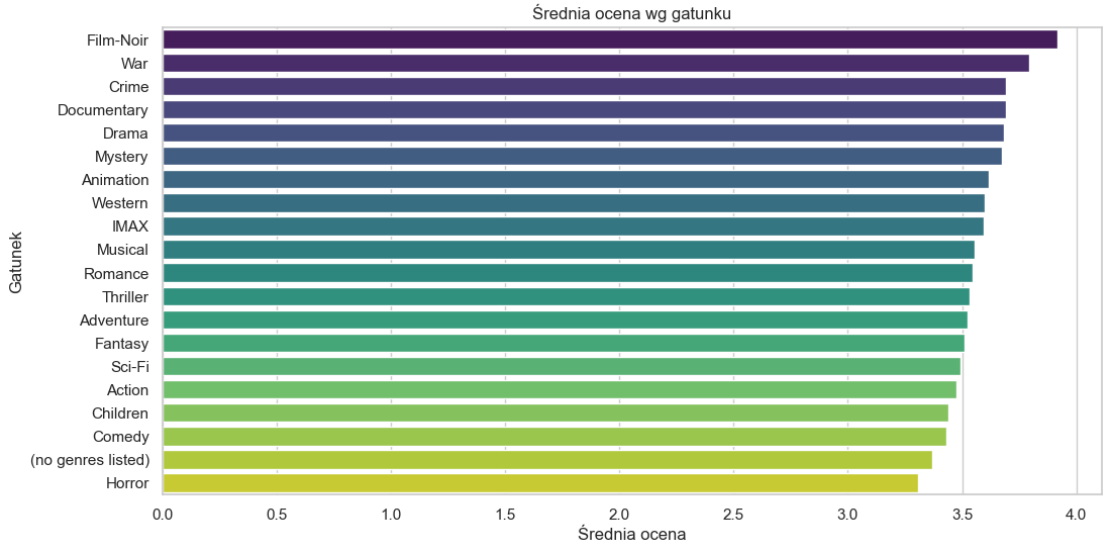


Rysunek 4: Rozkład gatunków filmów wyważona w zależności od ilości gatunków, do których należy

Można zauważyć, że znaczenie gatunku Thrille spada i znajduję się już tylko na 5 miejscu i jego 3 miejsce zajmuje gatunek Documentary.

2.6 Analiza jakościowa filmów

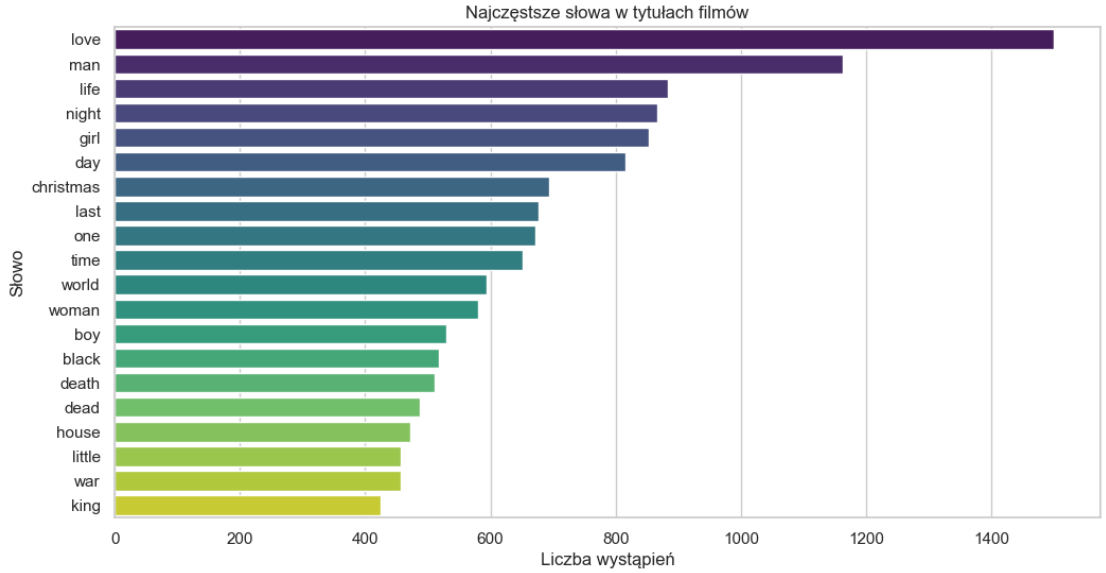
Dla każdego gatunku obliczono średnią ocenę. Najwyżej oceniane były m.in. filmy z gatunków takich jak: Film-Noir i Documentary. Taka analiza wspiera dobór cech do części content-based modelu.



Rysunek 5: Średnia ocena według gatunku, do którego należy film

2.7 Analiza tytułów filmów

Z tytułów wydobyto lata produkcji oraz najczęstsze słowa – zarówno przed, jak i po oczyszczeniu (usunięcie stopwords i lematyzacja). Celem tego etapu było stworzenie efektywnych cech tekstowych do modelu content-based filtering.



Rysunek 6: Najczęstsze słowa w tytułach filmów

2.8 Wnioski

Eksploracja danych pozwoliła:

- potwierdzić przydatność danych wejściowych do dalszego modelowania,
- wskazać potencjalne ograniczenia,
- przygotować odpowiednie cechy dla obu komponentów systemu rekomendacyjnego.

3 Content-based filtering

Poniżej przedstawiono sposób implementacji metody content-based filtering w omawianym systemie rekomendacyjnym.

1. Za pomocą biblioteki *pandas* wczytano tabele z plików *csv*.
2. Pogrupowano tabelę *tags* i połączono wszystkie tagi dla danego filmu w jedną komórkę tabeli.
3. Połączono tabele *tags* oraz *movies* na podstawie identyfikatora filmu. Stworzono połączony wektor zawierający wszystkie tagi oraz gatunki dla danego filmu, oddzielone spacją.
4. Z biblioteki *scikit-learn* wykorzystano funkcję *TfidVectorizer*. Funkcja tworzy macierz, gdzie wiersze to poszczególne filmy, natomiast kolumny to słowa występujące w wektorach utworzonych wyżej (funkcja bierze pod uwagę tylko te słowa, które wystąpiły co najmniej dwa razy i pomija spójniki w języku angielskim). Wartości w macierzy to wartości $TF - IDF$. Wartość $TF - IDF$ to iloczyn wartości TF i IDF, gdzie:
   - TF:
   $$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$
   gdzie $n_{i,j}$ jest liczbą wystąpień słowa dla danego filmu, a mianownik jest sumą liczby wystąpień wszystkich słów dla danego filmu.
   - IDF:
   $$idf_i = log\frac{|D|}{|\{d : t_i \in d\}|} \tag{2}$$
   gdzie $|D|$ to liczba filmów, a $|\{d : t_i \in d\}|$ - liczba filmów zawierających przynajmniej jedno wystąpienie danego słowa.
5. Rekomendacja dla danego użytkownika
   - Zbudowanie profilu użytkownika: z macierzy $TF - IDF$ stworzenie wektora filmów, które użytkownik ocenił powyżej 4 - wartości w takim wektorze to średnia wartości w macierzy dla danego filmu.
   - Obliczenie podobieństwa cosinusowego między wektorem użytkownika a całą macierzą $TF-IDF$. Podobieństwo jest w skali $[0, 1]$ co przekłada się na rozmytość - **film może tylko w pewnym stopniu przynależeć do zbioru polecanych filmów**
   - Wybranie filmów nieobejrzanych przez użytkownika i tych z najwyższą wartością podobieństwa.

4 Collaborative filtering

Poniżej przedstawiono sposób implementacji metody collaborative item-item filtering w omawianym systemie rekomendacyjnym.

1. Za pomocą biblioteki *pandas* wczytano tabele z plików *csv*.
2. Przy pomocy biblioteki *scipy* stworzony macierz rzadką *użytkownik-film*.
3. Przeskalowano oceny, tak aby różnicę między ocenami były większe.
4. Wykorzystanie funkcji Okapi BM25 w celu wyważenia wpływu bardzo popularnych filmów, które mogłyby dominować podobieństwa.
5. Przy pomocy biblioteki *implicit* wyznaczenie macierzy podobieństwa cosinusowego przy zapamiętaniu tylko k-najbliższych sąsiadów (w naszym wypadku 500).
6. Rekomednacje:
   (a) Dla istniejącego użytkownika w modelu: *model.recommend()*.
   (b) Dla pojedynczego filmu: *model.similar_items()*.
   (c) Dla nowego użytkownika: stworzenie ręcznego wektora ocen, użycie macierzy podobieństwa *model.similarity* i samodzielne obliczenia

5 System hybrydowy

Poniżej przedstawiono implementację finalnego systemu hybrydowego, w omawianym systemie eksperckim:

1. Zarekomendowanie 2000 najbardziej wskazanych filmów dla danego użytkownika przy użyciu metody CBF oraz 2000 filmów przy użyciu metody CF.
2. Wartości podobieństwa ($[0, 1]$) uzyskane za pomocą metody CF zostały pomnożone przez $0,6$, natomiast wartości podobieństwa uzyskane za pomocą metody CBF przez $0,4$ - średnia ważona uwzględniająca lepsze wyniki otrzymywane w ogólności przez CF.
3. Posortowanie otrzymanych wartości podobieństwa i zwrócenie filmów, dla których końcowe wartości podobieństwa są najwyższe.

6 Przykładowa rekomendacja

Poniżej na rysunku 7 znajduje się przykładowa rekomendacja dziesięciu filmów przy użyciu systemu hybrydowego dla użytkownika o ID 5.

```
display(recommend_hybrid(user_id=5))
```

```
1
```

	movieId	title	genres	score
0	589	Terminator 2: Judgment Day (1991)	Action|Sci-Fi	0.811341
1	648	Mission: Impossible (1996)	Action|Adventure|Mystery|Thriller	0.655648
2	350	Client, The (1994)	Drama|Mystery|Thriller	0.653102
3	527	Schindler's List (1993)	Drama|War	0.618531
4	1036	Die Hard (1988)	Action|Crime|Thriller	0.617678
5	1198	Raiders of the Lost Ark (Indiana Jones and the...	Action|Adventure	0.610672
6	508	Philadelphia (1993)	Drama	0.602836
7	377	Speed (1994)	Action|Romance|Thriller	0.600000
8	485	Last Action Hero (1993)	Action|Adventure|Comedy|Fantasy	0.570667
9	204	Under Siege 2: Dark Territory (1995)	Action	0.561166

Rysunek 7: Rekomendacje filmów dla użytkownika o ID 5

7 Wnioski

W przedstawionym projekcie stworzono i przeanalizowano system rekomendacyjny dla wypożyczalni filmów. W ramach prac wykonano pełną eksploracyjną analizę danych, zaprojektowano dwa niezależne modele rekomendacyjne – content-based filtering (CBF) oraz collaborative filtering (CF), a następnie zintegrowano je w systemie hybrydowym.

Na podstawie przeprowadzonych eksperymentów i analiz sformułowano następujące wnioski:

- Dane wejściowe (oceny, tagi, gatunki) cechują się odpowiednią jakością i ilością, umożliwiającą efektywne tworzenie modeli rekomendacyjnych.
- Content-based filtering pozwala na precyzyjne profilowanie użytkownika w oparciu o jego preferencje gatunkowe i tematyczne, jednak jego skuteczność może być ograniczona w przypadku nowych użytkowników lub filmów bez wystarczających metadanych.
- Collaborative filtering, dzięki wykorzystaniu informacji o zachowaniach innych użytkowników, cechuje się wysoką skutecznością, lecz cierpi na tzw. problem zimnego startu.
- Wprowadzenie rozmytości w interpretacji podobieństwa (np. poprzez podobieństwo cosinusowe w skali $[0, 1]$) pozwala modelować niepewność rekomendacji i lepiej dopasować wyniki do rzeczywistych preferencji użytkownika.
- System hybrydowy, łączący zalety obu podejść i odpowiednio je ważony, osiąga lepsze rezultaty niż pojedyncze komponenty, oferując bardziej trafne i stabilne rekomendacje.

Zrealizowany projekt potwierdza, że połączenie klasycznych metod rekomendacyjnych z podejściem rozmytym oraz analizą treści może prowadzić do skutecznego i elastycznego systemu rekomendacyjnego, gotowego do dalszego rozwoju i adaptacji.