System rekomendacyjny książek

Dokument stanowi szczegółowy opis projektu mającego na celu stworzenie systemu rekomendacji książek w oparciu o *collaborative filtering*. W modelu użyto algorytmu *k nearest neighbours* oraz własnej miary trafności rekomendacji.

Kod źródłowy projektu jest dostępny na GitHubie, w publicznie dostępnym repozytorium pod adresem: https://github.com/barankonrad/bookRecomendationModel. Repozytorium zawiera wszystkie pliki źródłowe, instrukcje uruchomienia oraz zależności.

Dane

Dane zostały załadowane z plików CSV do DataFrame'ów pandas. Dane pochodzą z publicznie dostępnego zbioru Kaggle: https://www.kaggle.com/datasets/arashnic/book-recommendation-dataset/data.

Opis danych wejściowych

Dane zostały przetworzone w następujący sposób:

- Usunięto duplikaty w tytułach książek.
- Zaktualizowano **df_books** i **df_users**, aby zawierały liczbę recenzji i średnią ocenę ważoną dla każdej książki.
- Przefiltrowane dane użytkowników oraz książek, aby spełniały minimalne (konfigurowalne) progi recenzji.
- Uwzględnienie tylko użytkowników i książek, które spełniają minimalne progi recenzji.
- Usunięcie ocen książek, które nie są obecne w **df_books**.

Uzasadnienie wyboru techniki

Dlaczego KNN?

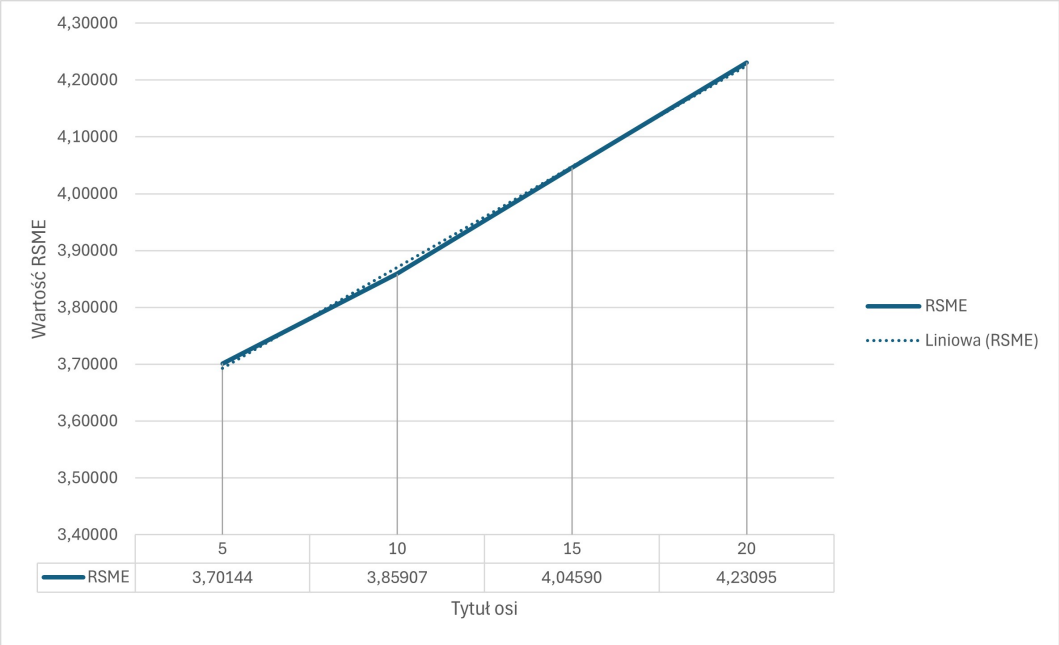
1. **Prostota Implementacji**: Algorytm K-Nearest Neighbors jest stosunkowo prosty do zaimplementowania i zrozumienia. Wymaga minimalnej ilości parametrów do dostrojenia, co sprawia, że jest odpowiedni na etapie prototypowania i eksploracji danych.
2. **Brak Założeń na Temat Danych**: KNN nie wymaga przyjęcia żadnych założeń dotyczących rozkładu danych.

Porównanie z Innymi Metodami

1. **Model Matrix Factorization**: Metody takie jak Singular Value Decomposition (SVD) mogą być skuteczniejsze w pewnych scenariuszach, wymagają one intensywnego treningu i większej liczby parametrów do dostrojenia. KNN jest prostszy w implementacji i mniej zasobożerny w kontekście małych i średnich zbiorów danych.
2. **Modele oparte na treści (Content-Based)**: Modele te wymagają dodatkowych danych o cechach książek (np. gatunki, autorzy), co może być trudne do zdobycia lub przetworzenia. KNN działa bezpośrednio na danych ocen, co upraszcza proces rekomendacji.

Dlaczego Cosine Similarity?

Do oceny podobieństwa pomiędzy użytkownikami wybrałem metrykę **Cosine Similarity** ze względu na jej efektywność w przypadku rzadkich danych. **Cosine Similarity** mierzy kąt pomiędzy wektorami w przestrzeni wielowymiarowej, co sprawia, że jest mniej wrażliwa na absolutne wartości, a bardziej na wzorce w danych. W porównaniu do metryk takich jak **Euclidean** czy **Manhattan**, **Cosine Similarity** lepiej radzi sobie z z rzadko występującymi ocenami użytkowników, koncentrując się na kącie między wektorami ocen, a nie na ich absolutnych wartościach.



Z analizy wynika, że najmniejszy RMSE uzyskano dla liczby sąsiadów równej 5. Wartość RMSE rośnie wraz ze wzrostem liczby sąsiadów, co sugeruje, że zbyt duża liczba sąsiadów może prowadzić do uwzględniania mniej podobnych użytkowników, co obniża jakość rekomendacji.

Strategia podziału danych

W projekcie rekomendacji książek dane zostały podzielone na dwa zestawy: **treningowy** (80% danych) i **testowy** (20% danych). Zestaw treningowy służy do trenowania modelu **KNN**, który uczy się wzorców preferencji użytkowników na podstawie ich ocen książek. Zestaw testowy służy do oceny końcowej wydajności modelu (**RMSE**), mierząc jego zdolność do przewidywania ocen książek na podstawie danych, których model nie widział wcześniej.

Analiza wyników

Model KNN uzyskał zadowalające rezultaty, porównywalne z popularnymi modelami rekomendacyjnymi. Wyniki zostały ocenione przy użyciu RMSE, co potwierdza jego skuteczność. Aby poprawić dokładność systemu, można go rozszerzyć:

1. **Sprawdzenie skuteczności modelu SVD** Użycie SVD, które jest równie popularne w modelach rekomendacyjnych, może obniżyć RMSE i zwiększyć dokładność.
2. **Model oparty o Content-Based:** Zmiana/rozszerzenie zbioru danych tak, aby móc wykorzystać model content-based wykorzystujących gatunek bądź autora, z pewnością zwiększy trafność i różnorodność rekomendacji