КУРСОВ ПРОЕКТ

ПО СИСТЕМИ, ОСНОВАНИ НА ЗНАНИЯ

Тема:

Система за генериране на препоръка за закупуване на книги (recommender system)

Студенти:

Иван Иванов Петров, 3MI0700035

Петя Ангелова Личева, 3MI0700022

София, януари 2025 г.

Съдържание

[1. Формулировка на задачата 1](#_Toc18459)

[2. Използвани данни 2](#_Toc15295)

[3. Използвани алгоритми 3](#_Toc6408)

[4. Описание на програмната реализация 3](#_Toc23803)

[5. Примери, илюстриращи работата на програмната система 3](#_Toc12397)

[6. Литература 3](#_Toc43)

### Формулировка на задачата

Система за препоръки на книги, като препоръките се базират на рейтинг от читателите/потребителите. На данните би могло да се приложи ре-филтрация по жанрове и по възраст на читателите (евентуално). Основен проблем, който би следвало да се реши преди ре-филтрацията е, че 40% от читателите не са посочили възрастта си, което води до нужда от поправки/зачистване в данните преди ре-филтрацията.

### Използвани данни

Описаните по-долу **CSV таблици** са взети от **dataset-а в** <https://www.kaggle.com/datasets/arashnic/book-recommendation-dataset,> а този ресурс е предоставен и в **графа Литература** на документацията **под наименованието Сетове с данни**.

#### Users

Таблица, която съдържа информация за потребителите, наричани още читатели в рамките на нашата система. Информацията, която се съхранява за читателите е техните уникални идентификационни номера (User-ID), локациите им (Location) и възрастта им. Потребителските идентификатори са анонимизирани и съпоставими с цели числа. Данните за местоположението на читателите и тяхната възраст се предоставят, ако има такива. В противен случай тези полета приемат NULL стойности. Това предполага, че ако те бъдат включвани в методите за категоризация и препоръки на книги на други потребители, трябва да се подсигурим за коректността на данните и евентуално да ги почистим от NULL стойности, тъй като те не внасят никаква информация и пречат на процеса по препоръка на книги.

#### Books

Таблица, която предоставя информация за книгите, от които имаме в наличност и ще можем да препоръчваме на нашите читатели. Книгите са уникално идентифицируеми от техните ISBN номера като трябва предварително да се погрижим, че всички ISBN номера са уникални и валидни. Таблицата също така предоставя информация за съдържанието на книгата като загавие на книгата (Book-Title), автор на книгата (Book-Author), година на публикация на книгата (Year-Of-Publication) и издателство на книгата

(Publisher). **Тази информация е предоставена от Amazon Web Services.** Трябва да се отбележи също така, че при наличие на съавторство над някоя книга, само имената на първият автор са предоставени. В таблицата също така се съхранява информация за линка към корицата на съответната книга (линк към уеб сайта на Amazon). Тези линкове може да се показват под различни форми (Image-URL-S - къси, Image-URL-M - средно дълги, Image-URL-L - дълги).

#### Ratings

Таблица, която описва оценките на книгите. В нея се съдържа информация кой потребител (User-ID) коя книг (ISBN) а е оценил и каква е била самата оценка (число в интервала от 1 до 10, вписана в полето Book-Rating). Ако някоя книга все още не е била оценена от даден потребител, оценката, която автоматично се вписва в полето за оценка (Book-Rating) е 0.

### Използвани алгоритми

За решаването на задачата се използват корелация по рейтинга на книгите, за да се определи най-препоръчвната и k-NN алгоритъма, за да се открият други книги с подобни характеристики, които да бъдат препоръчвани на читателите.

### Описание на програмната реализация

Тук се включва описание на структурата на създадената програмна система и спецификация на отделните компоненти (процедури, функции, дефиниции на класове и др.), написани от студентите.

### Примери, илюстриращи работата на програмната система

Тук се включват примери (в това число снимки на екрани и др.), които илюстрират основните случаи на работа/използване на създадената програмна система, съответните данни и получените резултати.

### Литература

* **Сетове с данни:** https://www.kaggle.com/datasets/arashnic/book-recommendation-dataset
* **Ръководство за създаване на система за препоръки в 1 час (част 1):** https://towardsdatascience.com/how-did-we-build-book-recommender-systems-in-an-hour-the-fundamentals-dfee054f978e
* **Ръководство за създаване на система за препоръки в 1 час (част 2):** https://towardsdatascience.com/how-did-we-build-book-recommender-systems-in-an-hour-part-2-k-nearest-neighbors-and-matrix-c04b3c2ef55c
* **Kaggle notebook за по-добро разбиране на данните:** https://www.kaggle.com/code/arashnic/recom-i-data-understanding-and-simple-recomm
* **Еволюция на алгоритмите за системи, отправящи препоръки (част 1):**  https://medium.com/@anicomanesh/evolution-of-recommendation-algorithms-part-i-fundamentals-and-classical-recommendation-bb1c0bce78a9