Задача «Разработка рекомендательного алгоритма для читателей библиотеки»

Введение

В 2022 году ряд ведущих российских издательств, книжных ресурсов и библиотек отмечают рост спроса на электронные и бумажные книги. Фонды региональных и федеральных библиотек насчитывают миллионы экземпляров книг. Российская государственная библиотека (РГБ) — крупнейшая библиотека в нашей стране, предоставляющая современные цифровые сервисы, такие как электронная библиотека РГБ и Национальная электронная библиотека, благодаря которым читатели могут искать и просматривать книги, не выходя из дома.

Сейчас в фондах РГБ хранится более 47 миллионов книг и различных артефактов. Такой объем документов имеет огромную культурную, историческую и научную ценность, однако затрудняет процесс каталогизации. Поиск конкретных изданий и тематических подборок занимает время. Кроме непосредственного доступа к содержимому электронной книги, возникает потребность выполнять поиск по семантике книги и формировать ассоциативные связи между различными документами для того, чтобы в дальнейшем предлагать читателям более релевантный результат поиска, а также персонифицированные рекомендации.

Читатели библиотеки — это и любители художественных произведений, и специалисты, интересующиеся отраслевой литературой, ученые и соискатели, работающие в научных проектах. Разнообразие целевой аудитории усложняет разработку рекомендательной системы из-за высокой семантической сложности изданий.

В РГБ активно ведется процесс оцифровки документов, что делает возможным внедрение адаптивной системы поиска. В рамках чемпионата участникам предлагается разработать рекомендательный алгоритм для читателей библиотеки, который позволит осуществлять семантический поиск литературы и рекомендовать книги читателю на основе его персональных предпочтений. Такие подборки позволят посетителям библиотеки открыть для себя новые жанры, авторов и произведения, которые ранее им были неизвестны.

Условие задачи

Участникам необходимо для каждого из 16 753 пользователей сделать подборку из 20 рекомендаций. Порядок рекомендаций не учитывается, но очень важно, чтобы рекомендации основывались на интересе пользователя и были ему релевантны. Обратите внимание, что тестирующая система принимает только те решения, в которых содержится не более 20 рекомендаций для одного пользователя. Уникальных документов – 354 355.

Участники получают 3 таблицы: users.csv, items.csv, train_transactions.csv. **Users.csv** содержит информацию о читателях, где каждый читатель имеет свой уникальный номер читательского билета (chb). Таблица **items.csv**, содержит описание документов, которые доступны всем читателям, каждый документ имеет уникальный системный номер (sys_numb). Таблица **train_transactions.csv** устанавливает связь между users-items, показывает наличие взаимодействия читателя с документом.

Описание входных значений

users.csv:
chb – полный номер читательского билета
age – возраст читателя
gender – пол читателя
chit_type – тип читателя

items.csv:
sys_numb – системный номер документа
title – название документа
author – автор документа
izd – издательство
year_izd – год издания
bbk – ББК документа

train_transactions.csv:
chb - полный номер читательского билета
sys_numb – системный номер документа
date_1 – дата выдачи
is_real – был ли выдан заказ
type – тип книговыдачи (книговыдача/скачивание)
source – источник (один из трёх онлайн-просмотрщиков)
is_printed – печатный/электронный документ

Метрика

$$F1 = 2\,\frac{Precision * Recall}{Precision + Recall}$$

$$Precision = \frac{\text{кол} - \text{во релевантных рекомендаций}}{\text{кол} - \text{во рекомендаций}}$$

$$Recall = \frac{\text{кол} - \text{во релевантных рекомендаций}}{\text{кол} - \text{во реальных взаимодейсвий}}$$

*Кол-во взаимодействий в формуле Recall - это кол-во items с которыми пользователь взаимодействовал в реальности.

Правила чемпионата:

1. С момента открытия датасета до момента завершения приема решений репозиторий участника, в котором он ведет разработку по задаче текущего чемпионата, должен оставаться закрытым.

2. Участник обязан открыть доступ к репозиторию на чтение по ссылке (которая была прикреплена в ЛК в поле «Ссылка на код (гитхаб)») не позднее чем в течение 12 часов с момента окончания дедлайна отправки решений на региональном чемпионате.

3. Согласно п. 5.8 Положения в процессе верификации решений организаторы и технические эксперты, проверяющие решения участников, вправе назначить интервью с участниками чемпионата. Участник получит приглашение и ссылку на интервью не позднее, чем за 12 часов до публикации итогового лидерборда. Пропуск интервью участником является поводом для дисквалификации.

4. Организаторы вправе исключить участника из призовых позиций лидерборда за непредоставление одного из артефактов решения задачи: тизера, скринкаста, презентации, ссылки на репозиторий.

5. Организаторы вправе дисквалифицировать участника в случае выявление плагиата кода или несоблюдение Положения конкурса.

6. Участник, получивший 2 дисквалификации за сезон проекта, попадает в чёрный список с дальнейшим отстранением от участия в чемпионатах до конца сезона.