

Задача 3. Рекомендательная система для пользователей московских библиотек

Описание задачи

Сервис «Библиотеки Москвы» — это онлайн-витрина, где собраны все книги городских библиотек. Аудитория сервиса - это читатели московских библиотек (570 тысяч человек), а также люди, любящие читать, но не пользующиеся на данный момент библиотеками: родители, пенсионеры и тд. В библиотечном фонде Москвы содержатся 2,2 млн единиц данных, включающих в себя книги (художественная литература, научно-популярная и др.), периодические издания, аудио и видео материалы.

Авторизованным пользователям доступен Кабинет читателя (ЛК) на платформе «Библиотеки Москвы», где собраны исторические данные по заказам и выданным книгам, статусы и сроки взаимодействия с библиотеками (например, статусы готовности книги к получению в библиотеке, сроки возврата и т.д.), функционал «избранное», функционал поиска и бронирования книг онлайн.

Основная задача — создать рекомендательную систему на основе истории книг, взятых читателями в государственных библиотеках и просмотренными пользователями, а также данных о пользователях. Благодаря этим рекомендациям должно сокращаться время на поиск нужной или интересующей читателя литературы, а также расти количество книг, взятых в библиотеке. Как следствие, повышается удовлетворенность пользовательским опытом и лояльность к платформе и библиотекам Москвы.

Результатом работы должна быть рекомендательная система, интегрируемая в текущий сервис «Библиотеки Москвы» (<https://www.mos.ru/knigi/>, десктопная версия), рекомендации будут вынесены в результаты поиска и ЛК читателя для авторизованных пользователей, а также рекомендации для неавторизованных пользователей.

Расположение рекомендательной системы может быть комбинированным и остаётся на усмотрение команды. Потенциальные опции:

- Главная страница сервиса;
- Страница поиска;
- Страница книги;
- Процесс бронирования книги;
- Кабинет читателя (ЛК).

Требования к решению задачи

Обязательным условием решения является наличие сопроводительной документации к решению задачи. В ней необходимо описать: протестированные гипотезы и алгоритм работы решения, оценку работы модели / ансамбля моделей (в т.ч. используемые методы обработки данных, методы ML, оценку работы моделей/ ансамбля моделей с указанием итоговой оценки точности модели и анализа feature importance), дополнительные условия и ограничения, введенные командой для решения задачи.

Ограничения по работе модели:

1. Скорость работы модели < 1 сек.
2. Переобучение модели - не реже 1 раза в день, сокращение времени будет рассматриваться как дополнительное преимущество.

Формат передачи данных:

На вход модели подаются

- id пользователя.

Модель возвращает json со следующими полями:

- "recommendations" - список рекомендаций следующего формата:
 - "id" - книги,
 - "title" - заголовок книги,
 - "author" - автор или авторы книги.
- "history" - история заказов пользователя, взятая из dataset'a (нужно для быстрого визуального сравнения):
 - "id" - книги,
 - "title" - заголовок книги,
 - "author" - автор или авторы книги.

Пример:

```
{
  "recommendations": [
    {
      "id": 789,
      "title": "Красная шапочка",
      "author": "Перро"
    },
    {
      "id": 101112,
      "title": "Сказки",
      "author": "народ"
    },
    ...
  ],
  "history": [
    {
      "id": 123,
      "title": "Незнайка на Луне",
      "author": "Носов"
    },
    {
      "id": 456,
      "title": "Золотой ключик",
      "author": "Толстой"
    },
    ...
  ]
}
```

Также результаты должны быть направлены документом в соответствии с шаблоном result_task3.csv для проверки результатов модели.

Планируемое решение по интеграции - передача json через API.

Решение может быть опубликовано в публичном репозитории - в этом случае необходимо указать ссылку на хранилище.

Задача по созданию рекомендательной системы может решаться любым способом на усмотрение команды. Участники вправе использовать любые открытые библиотеки. Рекомендованный язык программирования - Python.

Комментарии к наборам данных (датасетам)

Для решения задачи будут предоставлены исторические данные (dataset_knigi_1) по активным пользователям сервиса с начала 2021 года за исключением последних 5 бронирований в формате: идентификатор пользователя, временная метка, url книги. Среди этих данных есть бронирования книг и добавления в избранное (последнее - оставлено в полном объёме). Подробная информация о структуре и содержании страницы доступна через json (books.json).

Также предоставлены данные по книговыдаче за последние 2 года, учитывающие общую выдачу книг в библиотеках (как заказы через сервис, так и выдачу напрямую в библиотеках). Датасеты по книговыдаче и описание файлов находятся в архиве datasets_biblioteki

Последние 5 бронирований мы поместили в контрольную выборку - они будут использованы для оценки работы рекомендательной системы. Цель: предсказать набор из 5 бронирований для каждого пользователя.

Допускается аргументированное использование наборов открытых данных города Москвы, федеральных наборов открытых данных и прочих открытых данных. При использовании данных за исключением предоставленных датасетов необходима аргументация и оценка их влияния на работу модели, а также описание сбора и обработки данных. Легкость сбора, интеграции и автоматической обработки внешних данных не обязательна, но будет рассматриваться как дополнительное преимущество.

В соответствии с ФЗ 152 "О персональных данных" от 27.07.2006 все данные должны обрабатываться и использоваться в деперсонализированном виде.

Требования к коду

1. Исходный код должен соответствовать сопроводительной документации;
2. Должна быть обеспечена возможность выполнения процедур сборки и запуска приведённого кода;
3. Сложные алгоритмические моменты в коде желательно сопроводить комментариями (будет расцениваться как дополнительное преимущество);
4. Участники могут использовать лицензионное ПО для своих целей при создании решения, но конечный продукт не должен зависеть от (включать в себя) ПО имеющего лицензию, в том числе с бесплатным периодом. Приветствуется использование открытых библиотек и открытого кода.

Требования к презентации

Требования к презентации указаны в шаблоне презентации.

Критерии, учитываемые при проведении предварительной экспертизы

1. Подход коллектива к решению задачи (дизайн решения: веб-интерфейс, логика вывода рекомендации, решение проблемы «холодного старта»);
2. Качество кода;
3. Соответствие решения выбранной коллективом задаче (полнота описания решения (сопроводительной документации, кода). При использовании внешних данных - прозрачность сбора и обработки данных, полнота описания);
4. Эффективность решения в рамках поставленной задачи (точность работы модели - Mean average precision at K).

Критерии, учитываемые при проведении финальной экспертизы

1. Подход коллектива к решению задачи;
2. Техническая проработка решения (полнота описания решения (сопроводительной документации, кода). При использовании внешних данных - прозрачность сбора и обработки данных, полнота описания);
3. Эффективность решения в рамках поставленной задачи (точность работы модели);
4. Выступление коллектива на питч-сессии.

Требования к сдаче решений на платформе:

1. Ссылка на сопроводительную документацию (.doc/.pdf);
2. Ссылка на репозиторий с кодом;
3. Ссылка на презентацию (требования указаны в шаблоне);
4. Ссылка на веб-интерфейс.