

Задача 10. Рекомендательная система новостей для пользователей mos.ru и приложения “Моя Москва”

Описание задачи

Mos.ru - официальный портал Мэра и Правительства Москвы, рассчитанный на различные сегменты горожан. Сайтом пользуются как авторизованные пользователи, которым доступно оформление услуг, так и неавторизованные пользователи, которые ищут полезную информацию о жизни города в новостях, афише, различных сервисах и услугах.

Основная задача - помочь жителям Москвы ориентироваться в новостях и дать возможность выбирать то, что интересно именно им. Благодаря этим рекомендациям пользователям должен показываться контент, более релевантный к просмотру, а также расти количество просмотренных страниц и глубина просмотра. Как следствие, улучшается удовлетворенность пользовательским опытом у пользователей и повышается лояльность к portalу.

Командам нужно будет изучить сценарии потребления новостей на mos.ru и разработать рекомендательную систему, предлагающую новости для авторизованных и неавторизованных пользователей. В решении также нужно предусмотреть автоматическую разметку новостей по органам исполнительной власти и их руководителям, тематикам, тегам и др. Рекомендации будут вынесены в текущий сервис “Новости” (<https://www.mos.ru/news/>, десктопная версия).

Новости - тот продукт, который, как правило, живёт недолго - свежие новости должны чаще попадать в рекомендации. Мы рекомендуем учесть это в решении.

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 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" - новости,



- "title" - заголовок новости,
- "date" - дата публикации новости.
- "history" - история заказов пользователя, взятая из dataset'a (нужно для быстрого визуального сравнения):
 - "id" - новости,
 - "title" - заголовок новости,
 - "date" - дата публикации новости.

Пример:

```
{
  "recommendations": [
    {
      "id": 789,
      "title": "Прием экзаменов в ВУЗ",
      "date": "2021-10-08 12:00"
    },
    {
      "id": 101112,
      "title": "ЕГЭ",
      "date": "2021-10-09 12:01"
    },
    ...
  ],
  "history": [
    {
      "id": 123,
      "title": "Открылся садик",
      "date": "2021-10-10 12:02"
    },
    {
      "id": 456,
      "title": "Школа online",
      "date": "2021-10-11 12:03"
    },
    ...
  ]
}
```

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

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



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).

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_news_1) по активным пользователям сервиса за месяц за исключением последних 20 кликов в формате: идентификатор пользователя, временная метка, url страницы с новостями. Последние 20 кликов мы поместили в контрольную выборку - они будут использованы для оценки работы рекомендательной системы. Цель: предсказать набор из 20 новостей для каждого пользователя. Подробная информация о структуре и содержании новостей доступна через json (news.json).

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 презентации

Требования к презентации указаны в шаблоне презентации.

Критерии, учитываемые при проведении предварительной экспертизы

1. Подход коллектива к решению задачи
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 (обоснованность выбранных методов для разработки модели, алгоритма, сервиса);
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.

