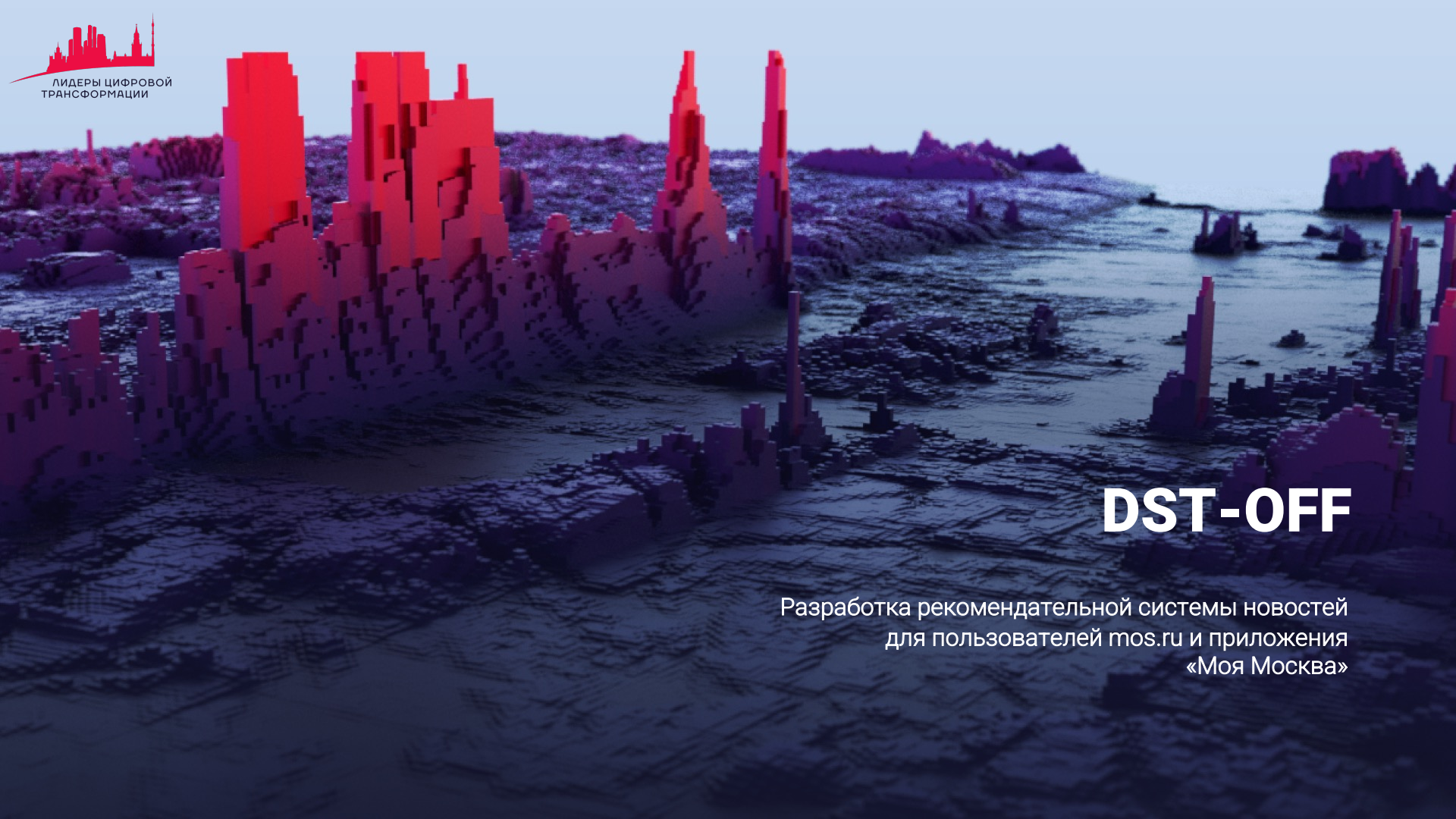
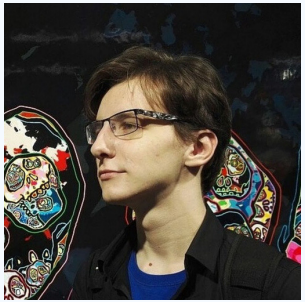


DST-OFF

Разработка рекомендательной системы новостей
для пользователей mos.ru и приложения
«Моя Москва»

Николай Ганибаев
Telegram: nganibaev
Капитан команды

- Python;
- Финансовое планирование;
- Методы моделирования процессов и программные средства для построения моделей;
- Основы программирования: типы и структуры данных;
- SQL;
- Архитектура приложений и базы данных



Александр Ганибаев
Telegram: AGanibaev
Аналитик

- Python;
- C/C++;
- Основы программирования: типы и структуры данных;
- Теория алгоритмов;
- R.

Васиф Фараджов
Telegram: valthazari
«Будущий data-scientist»

- Маркетинг и дизайн
- Python;
- Project Management;
- Теория машинного обучения.

DST-OFF

ЗАДАЧА И ЦЕЛЬ

Задача:
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.

Хакатон проводится в два этапа: основной и финальный. Данное решение является решением основного этапа хакатона.

Дано:
- исторические данные по активным пользователям сервиса mos.ru за август 2021 года за исключением последних 20 кликов;
- последние 20 кликов каждого пользователя являются контрольной выборкой - они будут использованы для оценки работы рекомендательной системы.

Цель:
- предсказать набор из 20 новостей для каждого пользователя.

ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

ВВОДНАЯ ИНФОРМАЦИЯ И ПОДХОД

Вводная информация для основного этапа конкурса:

- Имеется лог просмотра новостей авторизованными посетителями сервиса mos.ru с указанием id пользователя, id новости и даты/времени просмотра. 20 последних просмотров для каждого пользователя скрыты.
- Имеется датасет новостей с указанием id новости, заголовка, аннотации и текста, частично размеченный по тегам, сферам, темам, округам, районам, департаментам, персонам.
- Необходимо для каждого авторизованного пользователя определить 20 скрытых новостей.

Подход команды к решению задачи основного этапа (план):

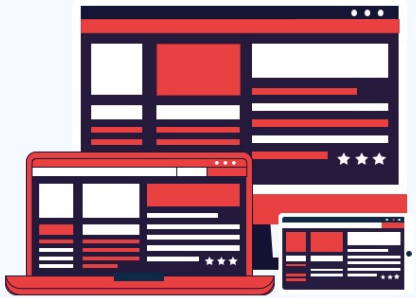
- Провести EDA логов
- Подготовить черновой вариант модели на основе информации из логов с помощью матричной факторизации, рассчитать точность работы модели.
- Провести EDA новостей
- Пересчитать модель с добавлением факторов, присутствующих в датасете новостей.
- Обогатить датасет новостей дополнительной разметкой с помощью NLP, добавить факторы к рекомендательной системе
- Сформировать датасет посетителей с определением предпочтений (времени посещения, специфики новостей, эмбеддинга), добавить факторы к рекомендательной системе
- Оформить решение в виде модели, выделив обучение и предсказание в отдельные скрипты
- Подготовить веб-демонстрацию с помощью streamlit
- Ход решения представлен ниже.

ХОД РЕШЕНИЯ

Общие соображения к реализации во время финального этапа:

Основная **идея** заключается **в более глубоком перекрестном анализе новостей и пользователей**. Мы собираем корпус текстов новостей mos.ru, на основе которого обучили нейросеть для решения NLP-задач, в т.ч. классификации и разметки (NER). Пользователей кластеризовали по профилям интересов на основе истории поведения. Отдельная модель определила профиль новых пользователей. Каждая новость получила вес для каждого профиля и при пересечении порога чувствительности будет предложена посетителям сайта.
Что на наш взгляд может заинтересовать посетителя сайта mos.ru и что можно будет включить в решение:

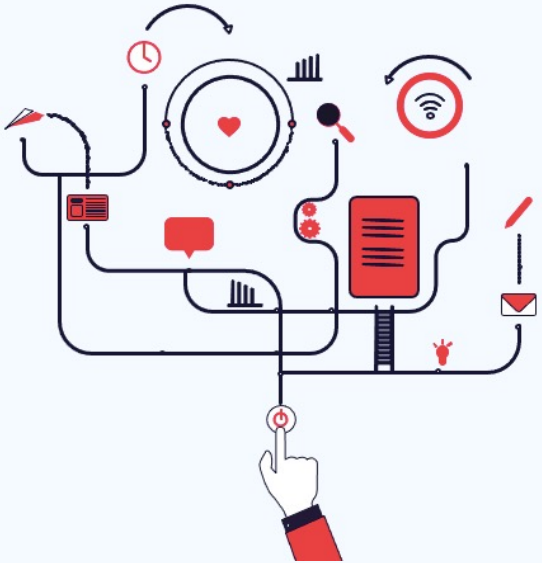
- Холодный старт для новых пользователей - предложить категории новостей на выбор для определения профиля;
- Для активных пользователей сделать предположения о сфере интересов: потенциально интересных темах, персоналиях, мероприятиях, районе;
- Предложить информацию о событиях в конкретном районе города;
- Предложить информировать об обсуждении проблем района, градостроительных -планов и т.п.
- Выбор ключевых слов, стоп-слов, департаментов, районов, чиновников и т.д.
- Городская афиша, события и мероприятия
- Обернуть в телеграм-бот



ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

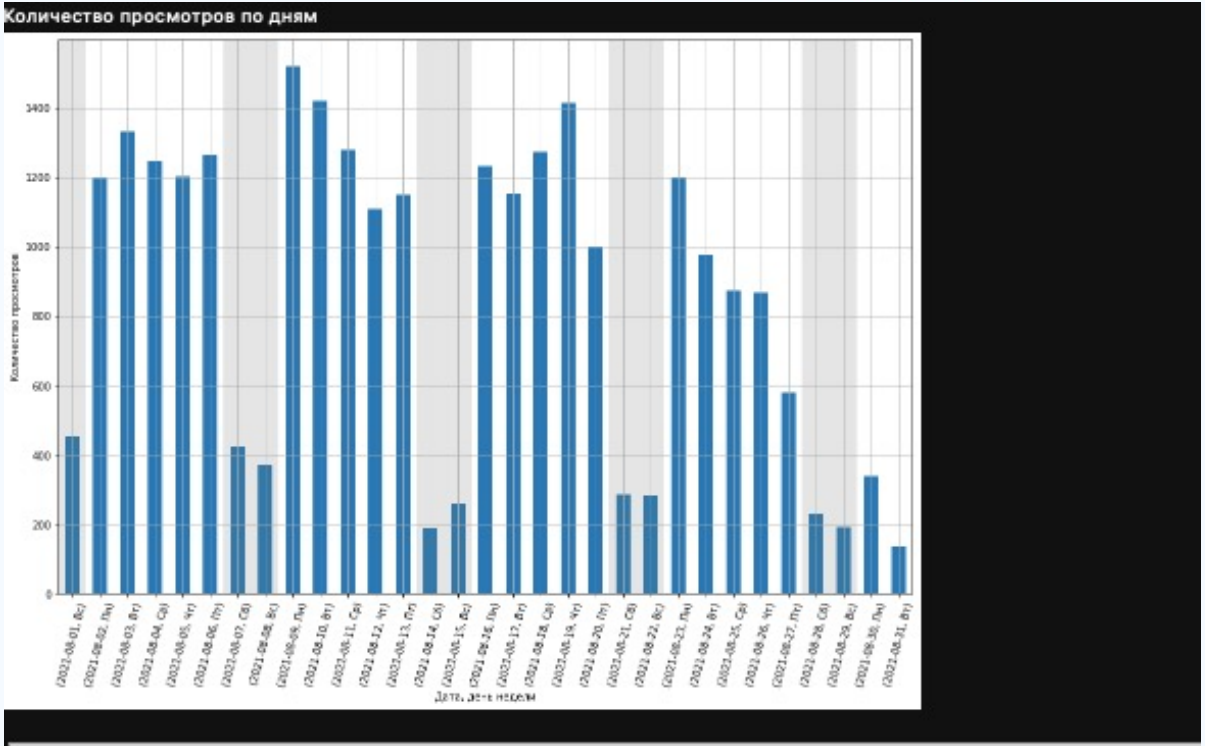
ЭТАПЫ РАБОТЫ

1. Импорт библиотек, настройки, служебные функции
2. Загрузка, очистка и обработка лог-файла, анализ
3. Подготовка датасета логов для поиска неявных закономерностей
4. Проведение разведывательного анализа данных датасета логов
5. Data Cleaning and Preprocessing function
6. Train Test Split function
7. Mean Average Precision at K function
8. Модель на основе лога просмотров без информации о новостях
9. Загрузка данных с информацией о новостях
10. Визуализация новостей с интерактивным отображением заголовка новости на диаграмме



ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

МАТЕРИАЛЫ К РЕШЕНИЮ

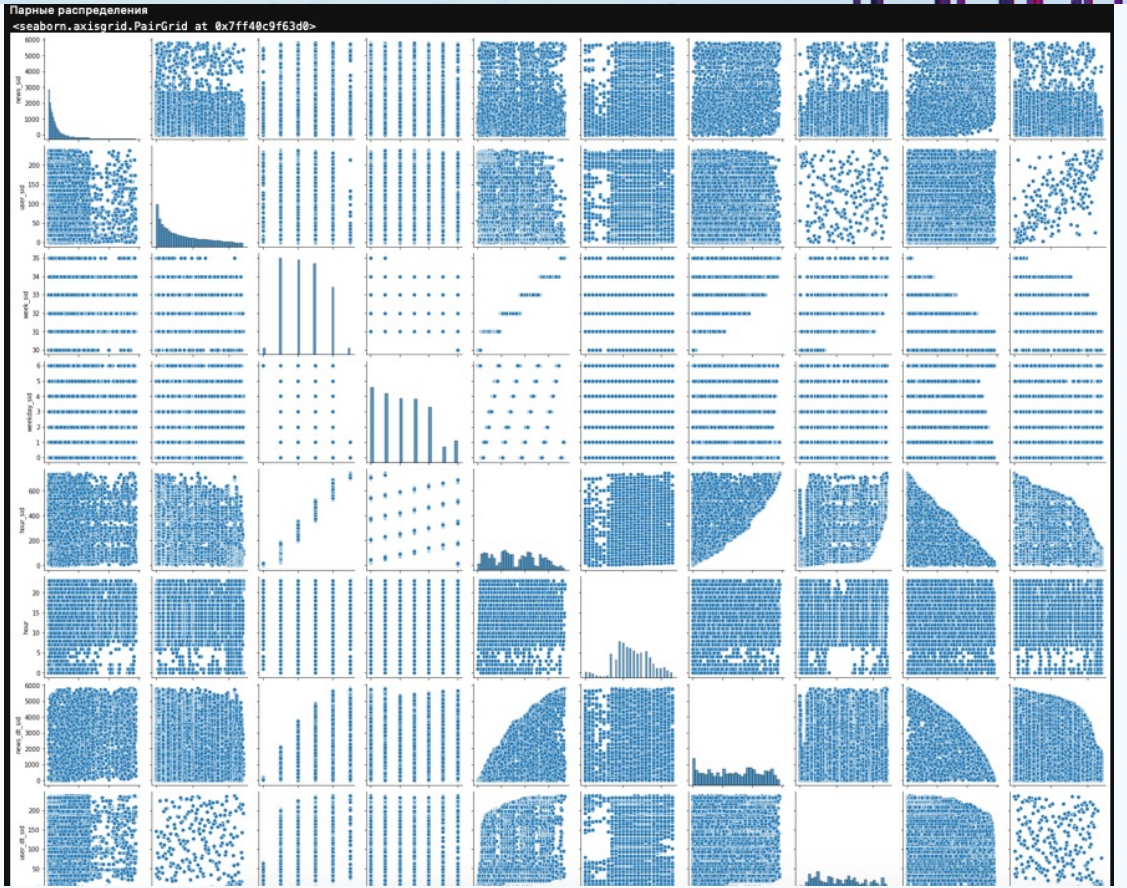


Из предоставленных данных видно, что на выходных в августе посещаемость сервиса кратно ниже будней.
Также наблюдается серьезное проседание посещаемости 24-27 и, особенно 30-31 августа.
Это говорит в пользу предположения, что именно на эти дни приходится большая часть скрытых контрольных просмотров.
Всего количество скрытых просмотров составляет 239 * 20 = 4780.

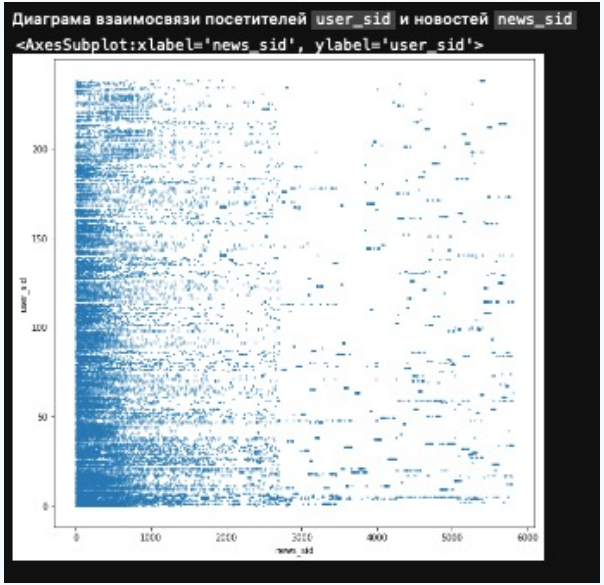
ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

МАТЕРИАЛЫ К РЕШЕНИЮ

Парные распределения

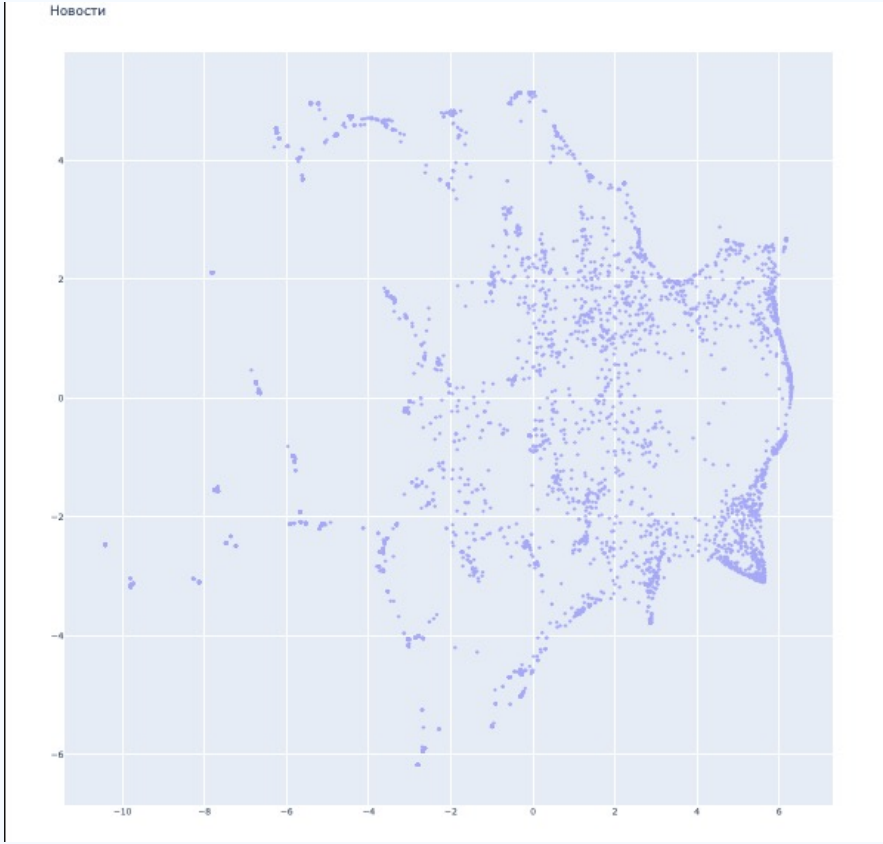


Диаграма взаимосвязи посетителей `user_sid` и новостей `news_sid`



ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

ВИЗУАЛИЗАЦИЯ НОВОСТЕЙ С ИНТЕРАКТИВНЫМ ОТОБРАЖЕНИЕМ ЗАГОЛОВКА НОВОСТИ НА ДИАГРАММЕ



Даже на глаз без дополнительной обработки сразу можно выделить несколько кластеров