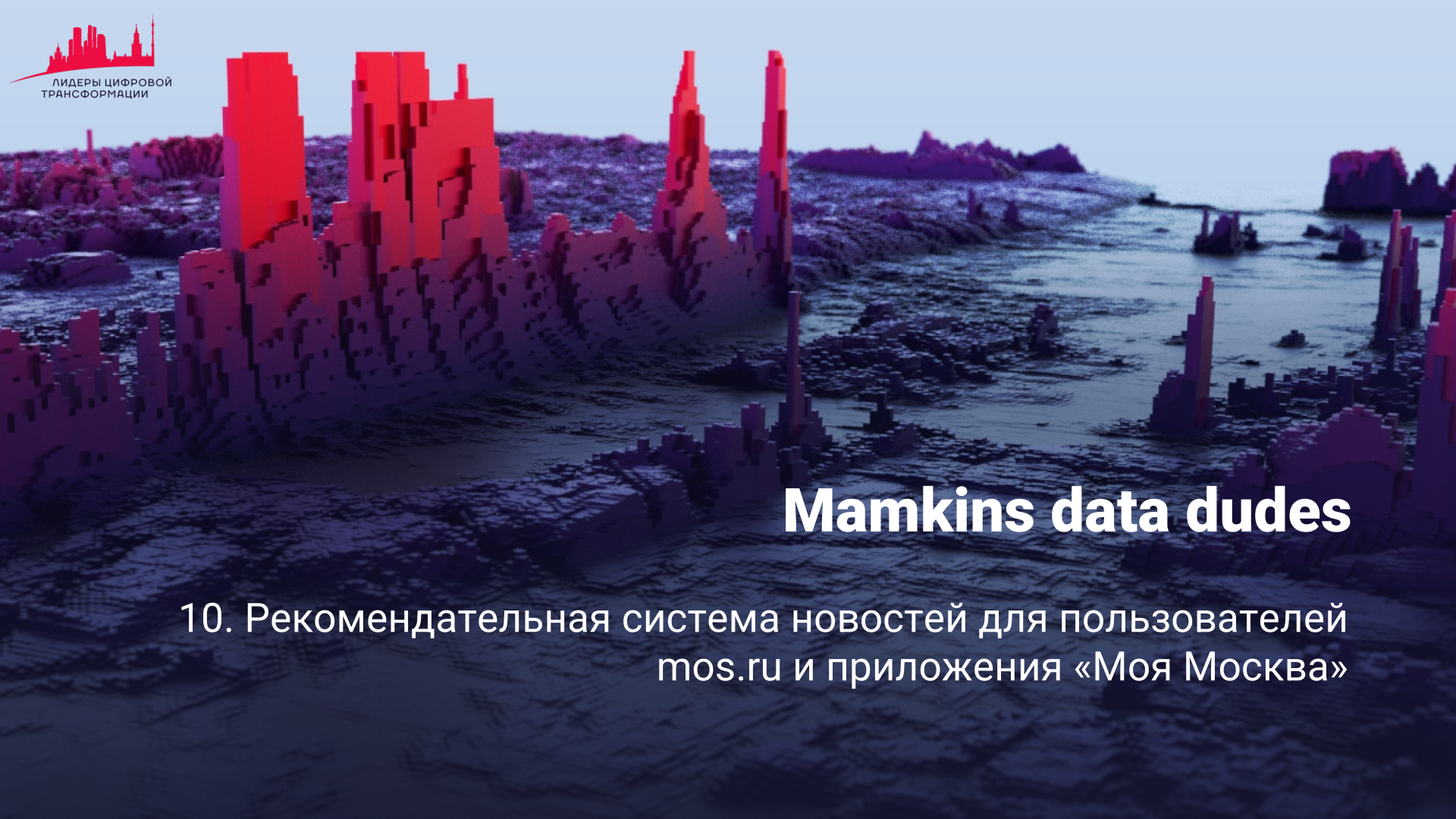
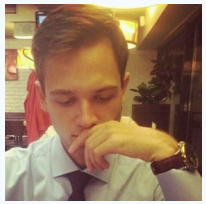


Mamkins data dudes

10. Рекомендательная система новостей для пользователей mos.ru и приложения «Моя Москва»



Стегура Никита
+7 909 920 99-94
@nstegura
Developer

Андрианова Маргарита
+7 916 661 94-65
@MargaritaAndrianova
Team Lead

Джалагония Индико
+7 999 989 26-85
@NikaParen
Data Scientist

Волчугин Денис
+7 985 965 11-87
@volchugin
Developer

Mamkins data dudes

Рекомендательная система

ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

Основные параметры

Актуальность

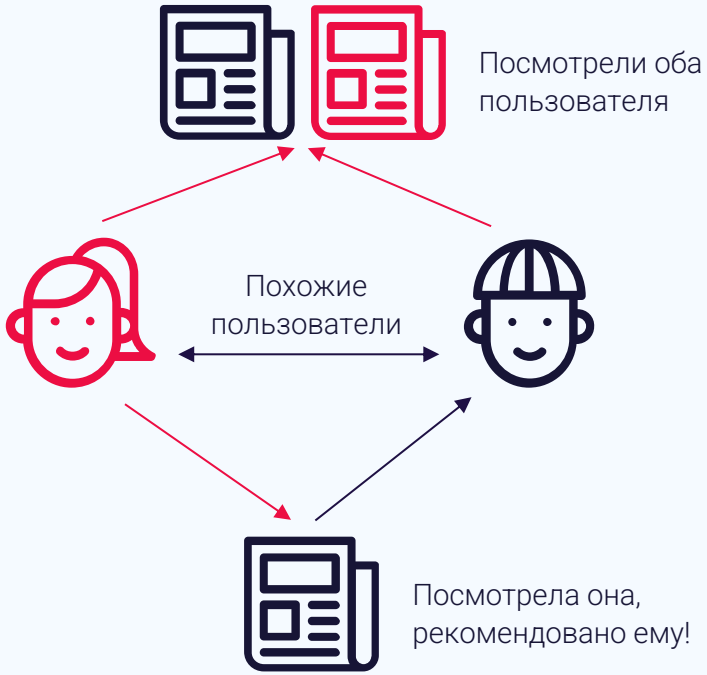
Специфика новостей в том, что новизна является главным фактором в выборе контента. Поэтому в рекомендациях в первую очередь мы старались учесть актуальность новостей.

Для сравнения свежести новостей мы опирались на количество дней с даты публикации материала.

Популярность

Чем выше популярность новости, тем она более востребована. Это наиболее явный фактор для рекомендаций.

Ранжирование новостей = количество просмотров / количество дней с даты публикации



Модель

Специфика задачи:
- быстрое устаревание
- небольшое количество актуальных материалов (от 40 до 65 в будни и от 10 до 20 в выходные)

Модель рекомендаций ItemItemRecommender позволяет достаточно быстро подобрать набор новостей на основе метода ближайших соседей с учетом значения актуальности.

60%

пользователей получили хотя бы 1 точную рекомендацию

9 из 20

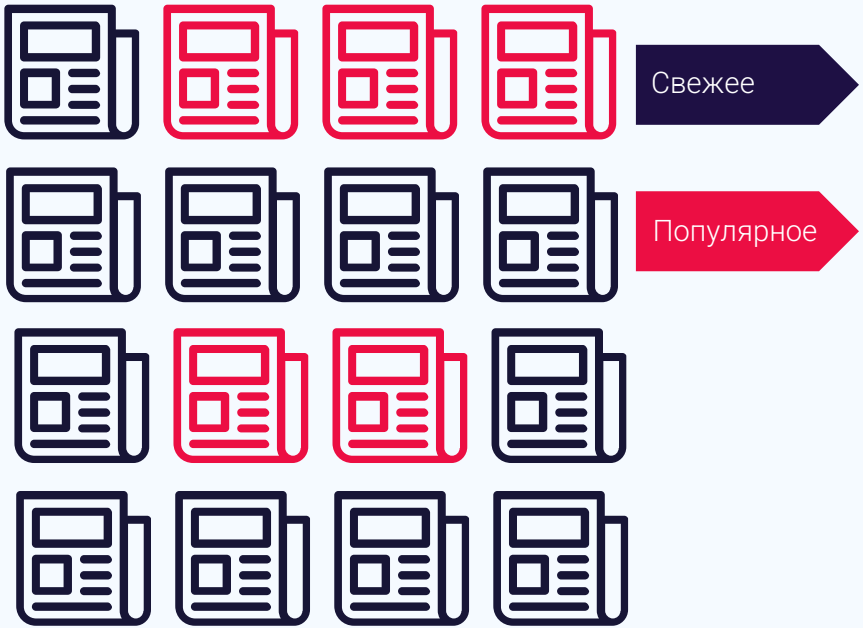
максимальное значение предсказанных рекомендаций на 1 пользователя

Показатели основаны на тестовых результатах для модели.

360ms – время на обучение

получение рекомендаций для 1 пользователя ~1-3ms

“

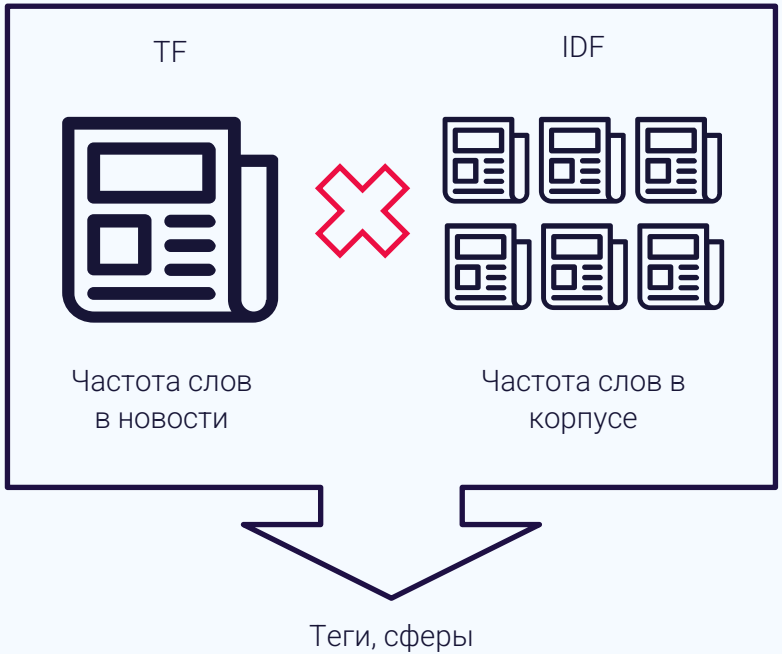


Новые пользователи

Ранжирование новостей для новых пользователей также основано на актуальности. Чем старше новость, тем более популярной она должна быть, чтобы попасть в рекомендации.

Если и пользователь новый и новость новая, то на старте мы искусственно повышаем ранк новости, позволяя выйти в рекомендации, чтобы определить реальную заинтересованность к новости.

Автоматическая разметка материалов



Логика

Используемый алгоритм на базе метрики TF-IDF (TF — term frequency, IDF — inverse document frequency).
Он выбирает наиболее весомые слова на основе частоты употребления в документе в сравнении с полным корпусом.

Полный корпус составляется на основе всех новостей, их тегов и сфер, исключая стоп-слова.

Результатом алгоритма является набор тегов и сфер для переданного текста новости.

Выводы

ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

Применение и развитие

Применение

Предложенная модель рекомендаций наиболее применима для дополнительного блока новостей на странице самой новости.

Авторазметку материалов можно использовать для помощи новостной редакции при создании новых материалов.

Это позволит собрать больше данных и проверить работоспособность моделей.

Развитие

При наличии больших данных можно использовать гибридный тип коллаборативной фильтрации, который позволит улучшить качество рекомендательной системы и распространить ее применение на разводящие страницы (главная страница и раздел новости).

ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

Ссылки

Репозиторий проекта
Демо-стенд
Сопроводительная документация