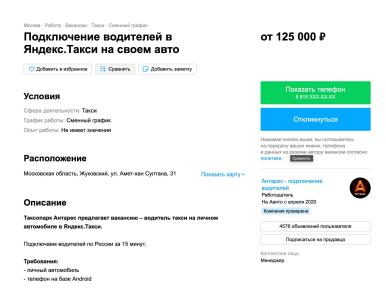


Опорный

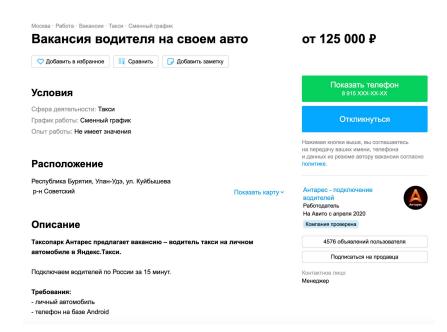




Кандидаты



Опорный



Кандидат

avito.tech

Объявление считается повторным, если:

- Совпадает с другим не полностью, но в них есть что-то одинаковое (например, фото или описание).
- Фото и описания разные, но по контексту ясно, что речь об одном и том и тоже товаре или услуге.
- Вы создали общее объявление для разных вещей, а потом сделали по отдельному на каждую их них.

- Такое же размещено с другого профиля.
- Вы удалили или досрочно сняли с публикации первое объявление. Тогда разместить похожее тоже не выйдет, пока у первого не закончится срок действия (30 дней).
- Исключение: вы можете публиковать одинаковые объявления в разных городах, если оно относится к категории «Услуги», «Готовый бизнес» или «Работа».

Стандартный пайплайн. Какие проблемы?

- Собрать датасет
- Разметить датасет
- Преобразовать в фичи
- Обучить модельку
- ?????
- profit

Сбор датасета

Холодный старт

- Как искать пары?
- Как обеспечить разнообразие?
- Как размечать?

Сбор датасета

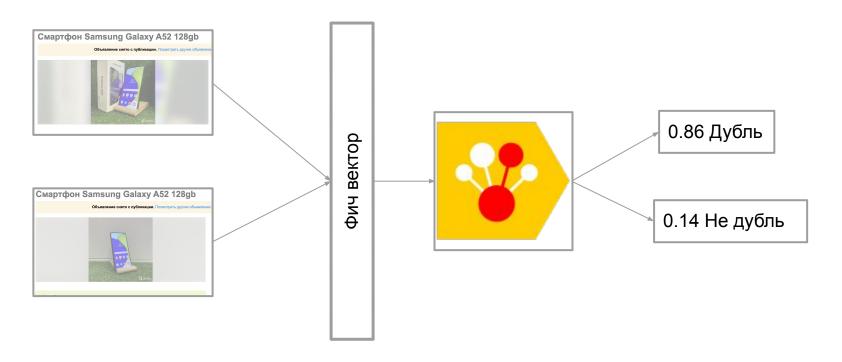
Холодный старт

- Как искать пары?
- Как обеспечить разнообразие?
- Как размечать?

Уже есть модель

- Правила не устарели?
- Ошибка модели?
- Ошибка модератора?
- Как увеличить полноту?

Обучение модели



Модель готова

Что дальше?

Пора катить?

Модель готова

Что дальше?

Пора катить?

Есть проблема

Как искать кандидатов?

Как искать кандидатов?

- Нужно искать быстро
- Нужно искать среди тысяч-сотен тысяч самые подходящие
- Нужно находить небольшое количество

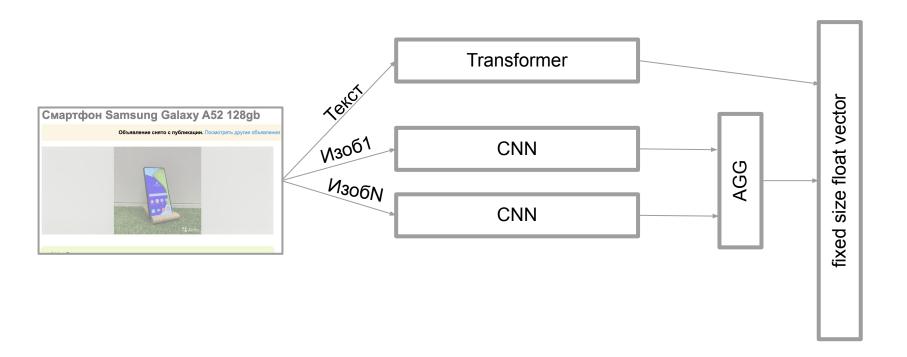
Как искать кандидатов?

- По полному совпадению атрибутов
- По частичному совпадению атрибутов
- По семантической близости

Семантическая близость

- **Текст**
- Изображения

Семантическая близость



Как искать в векторном пространстве?

Как искать в векторном пространстве?

- Postgres
- Faiss
- Sphinx

Пайплайн

- Находим N кандидатов
- Прогоняем через бустинг N пар опорный - кандидат
- Для предсказаний выше порога автодействия применяем действие
- Для предсказаний выше порога отправки модератору отправляем на ручную проверку
- Иначе пропускаем на сайт