Конкурс Avito.ru:

Задача поиска контактной информации на изображениях и возможные подходы к её решению

Иван Гуз
к.ф.-м.н

Что такое Avito.ru?

- **#1 сайт объявлений в России***
  - 54млн посетителей / (с моб. версией)

- 150-200млн просмотров в день – **один из крупнейших сайтов объявлений в мире***

- 25 млн активных объявлений**: Недвижимость, Транспорт, Работа, Услуги, Товары для дома и т. д.
- Freemium бизнес модель
  - Продвижение объявлений с помощью платных сервисов
  - Реклама
  - Специальные услуги для SMB клиентов

Avito

* Сентябрь 2014. Статистика не включает данные с нативных приложений Android/iOS/Windows Phone
** 1 октября 2014

Ежемесячная аудитория Avito.ru

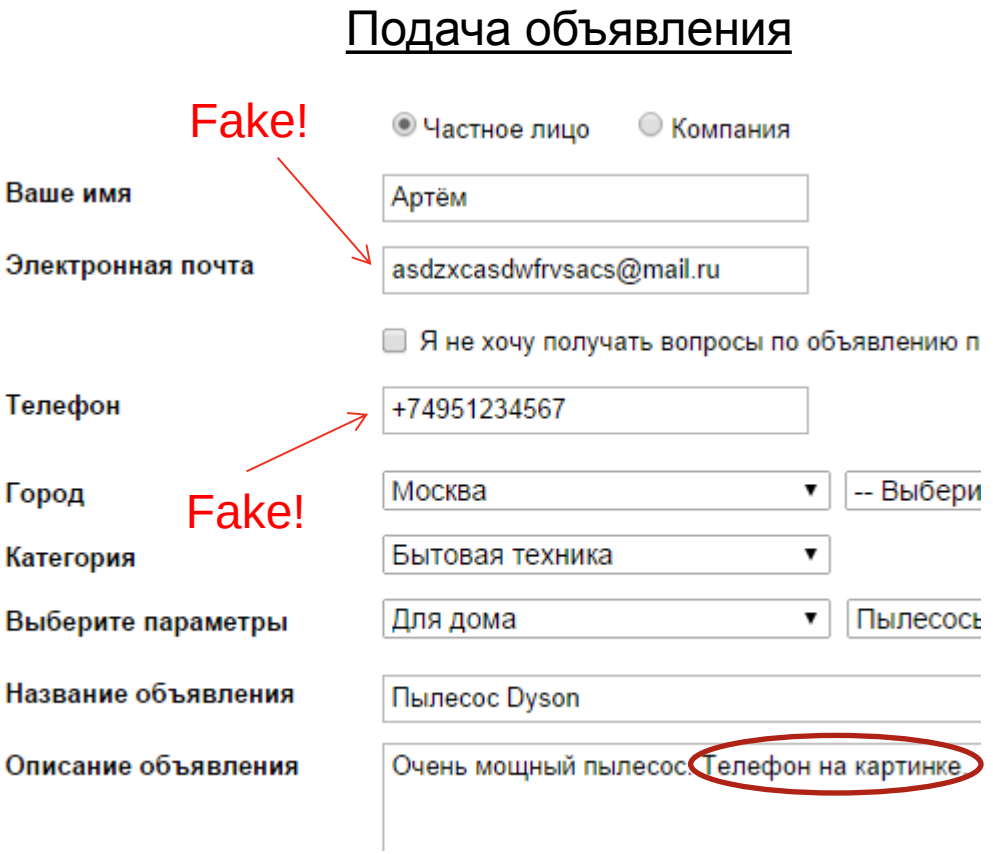
Уникальные посетители* avito.ru

* Статистика не включает данные с нативных приложений Android/iOS/Windows Phone
По данным LiveInternet

Новые объявления Gross vs Net

Почему контакты на фото запрещены?

- Для указания телефонов/email есть специальное поле. Спамеры заполняют их несуществующими или фиктивными значениями чтобы избежать блокировки по черным спискам
- Контакты в тексте объявления мы умеем эффективно находить. Их туда не спрятать
- Кроме телефонов на фото могут указывать прочие контакты (url, email, название компании), по которым также можно найти способ связи

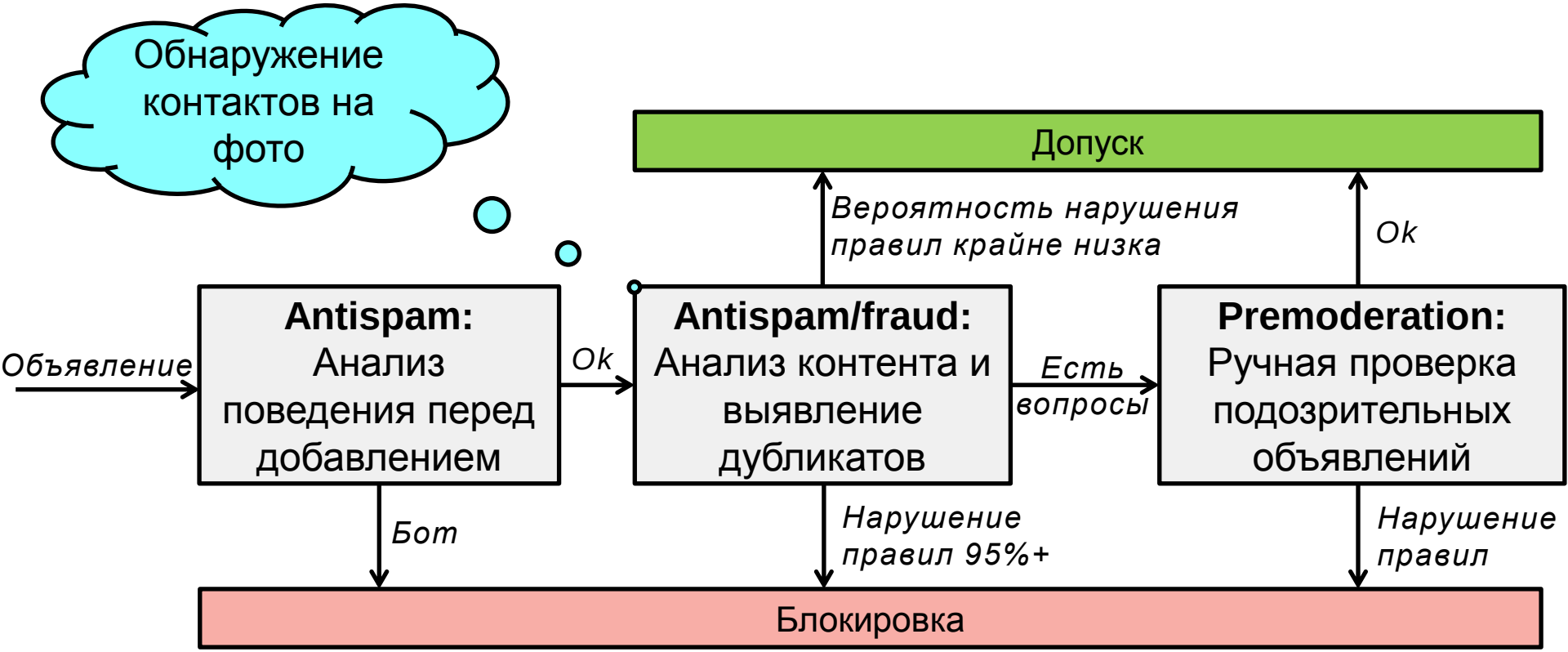


Примеры изображений

	Есть контакты	Нет контактов
Просто		 
Сложно	 	 

Где больше всего спама?
Там где стоимость товара самая большая – квартиры

Процесс модерации:
3 уровня защиты: Машины + Машины + Люди



Критерии качества моделей

Реальная жизнь:

- Команда модераторов, которая может проверить не более K объявлений
- 1 объявление –> 0,1,2,… изображений (фото)

Модель должна ранжировать объявления по вероятности наличия контактов. Чем лучше качество ранжирования в Top K – тем лучше модель

Доля с контактами среди первых i

$$\text{AP@K (Average Precision @ K)} = \sum_{i=1}^{k} P(i) * \left\{ \begin{matrix} 0, \text{если в } i \text{ нет контактов} \\ \frac{1}{\min(k,m)}, \text{если в } i \text{ есть контакты} \end{matrix} \right\}$$

Всего объявлений с контактами

Конкурс:

- Плоская выборка изображений
- Максимизировать AUC

Обучающие данные и этапы конкурса

1. За 3 месяца выгрузили все картинки объявлений, заблокированных по причине «контакты на фото»
2. За этот же период выгрузили случайную выборку картинок объявлений, которые были активны достаточно долго и не были заблокированы
3. Отдали 51 122 картинок (п.п 1, 2) на повторную ручную разметку

41 122 изображения – обучающая выборка (разметка дана)
10 000 изображений – тестовая выборка (разметки нет. На ней измеряется качество)

До 14 ноября* – обучение моделей, отправка (заморозка) моделей

10 000 изображений, аналогичная методика подготовки за более актуальный период

До 19 ноября – скоринг новых картинок с помощью готовых моделей

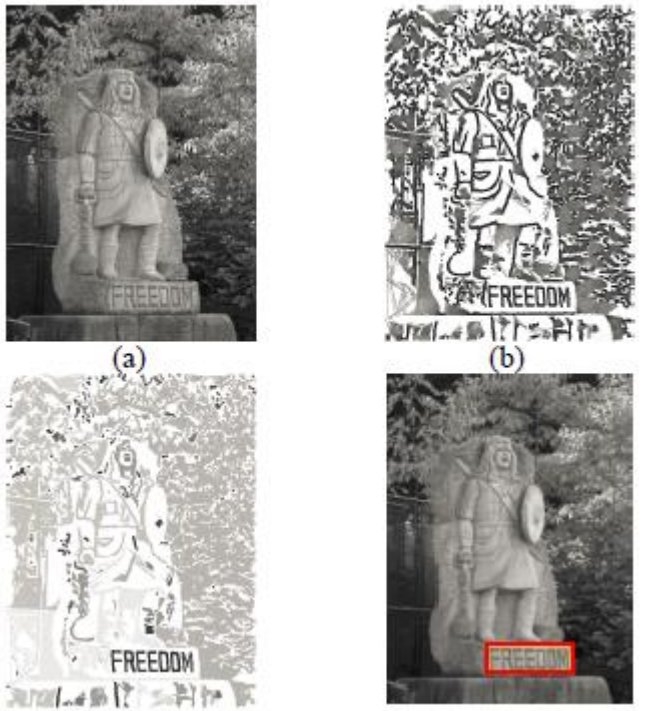
Определение победителей, проверка воспроизводимости – до 10 декабря

Важно! Использовать любую информацию кроме изображения бессмысленно! В контрольной выборке имена файлов будут случайными.

* До 5 ноября – регистрация новых участников

Подходы к решению задачи

- В идеале нужно решать задачу “поиска искусственно нанесенного на картинку текста” => как это делать – не ясно
- Плохо работает: бездумное использование OCR (пробовали Tesseract)
- Лучше работает: бинаризация текста и выделение связных компонент. По этому принципу работает алгоритм SWT (Stroke width transform). Какие компоненты называть текстом/контактами – это эвристика



Отбор победителей, правила и призы

- 1е место (100 000 руб. включая налоги) – максимальный AUC на контрольной выборке среди всех участников

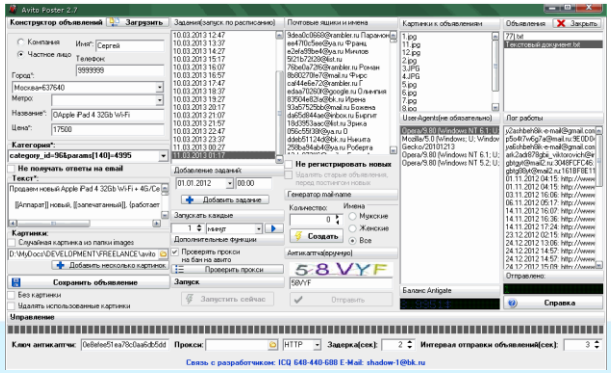
- 2е место (50 000 руб. включая налоги) – максимальный AUC на контрольной выборке среди оставшихся участников, **использующих ПО для некоммерческого использования**.

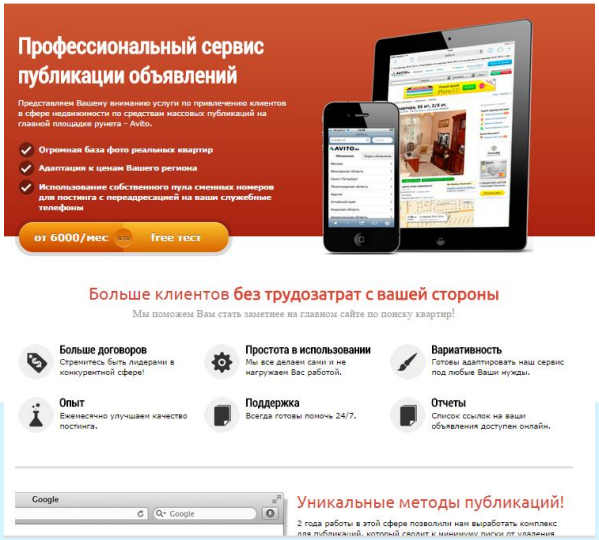
Участники: **только граждане РФ, достигшие 18 лет**

Победителю при необходимости нужно будет показать, как получается соответствующий результат с помощью отправленной модели

Чего мы хотим добиться?
Повысить порог входа в спамерский бизнес

	Любители (Физ. лица)	**Профессионалы (Юр. лица)**
Штат	1 чел	Разработчики, дизайнеры, продавцы, служба поддержки
Маркетинг	Форумы	Реклама, прямые продажи, Email
Объем спама	10-100 в день	100К+ в день
Бизнес модель	C2C	B2B
Обороты	Карманные деньги	MM+
Технологии	Толстый клиент	Облачные платформы
Скорость обновления	Неделя	Несколько раз в день





Спасибо за внимание!

machinelearning.ru

competition.avito.2014@forecsys.ru