

*Презентация проекта*

---

iPavlov

Евгений Егоров  
Илья Игашов  
Арсений Крохалев  
Евгений Рубаненко

---

---

# Проблема пользователя

---

- Из-за отсутствия инструментария для работы с русским текстом разработчики и исследователи тратят много времени и сил на создание собственных моделей и алгоритмов работы с русским текстом: **высокий порог вхождения в область разработки диалоговых систем**
- Как следствие — **отсутствие качественных продуктов в области обработки русского текста и диалоговых систем (чат-ботов, рекомендательных систем и т.д.)**

---

# Целевая аудитория

---

- Конечные пользователи
- Разработчики и исследователи
  - Лаборатории
  - Стартапы
  - Большой и средний бизнес



---

# Количественные исследования

---

- Около 50% опрошенных пользуются чат-ботами
- Чат-боты и голосовые помощники используются для выполнения примитивных команд
- Пользователи считают, что на данный момент уровень развития русскоязычных диалоговых систем низок
- Разработчики сталкиваются с нехваткой данных на русском языке

---

# Наше решение

---

- Датасеты на русском языке
- Средства их сборки
- Алгоритмы анализа текстов, работающие «из коробки»
- Готовые продукты для отдельных бизнес-проектов

---

# Наше решение: Data Scientist

---

## **Проблема:**

нет библиотеки, которая одновременно предоставляет необходимые данные, а также удобные инструменты для работы с ними

## **Предлагаемое решение:**

библиотека, в которой будут собраны данные, модели и средства для работы с ними

## **Сроки:**

планируется закончить разработку к концу 2018 года



---

# Наше решение: Data Scientist

---

## Рыночные факторы:

### Спрос:

- все респонденты, которым в исследованиях необходимо работать с данными на русском языке (57% от всех опрошенных), считают, что такая библиотека необходима
- ученые, работающие с более распространенными языками (напр., английский) (32% от всех опрошенных), не видят особой нужды в такой библиотеке, но будут рады новым удобным функциям

### Существующие предложения:

Open-source библиотеки: Deeplearning4j, TensorFlow, Keras, Theano, nltk и др.

---

# Наше решение: Стартап

---

## Проблема:

за короткий период времени необходимо сделать рабочий продукт или его прототип

## Предлагаемое решение:

обеспечить разработчиков всеми необходимыми инструментами для быстрого развертывания / разработки своего продукта (удобные функции для обработки данных, возможность легко и быстро сделать чат-бота, использующего ML & DA и др.)

## Сроки:

планируется начать разработку в ближайшее время (как будет разработан MVP (см п.1) и получено дополнительное финансирование);  
на реализацию закладывается 1 год



---

# Наше решение: Стартап

---

## Рыночные факторы:

### Спрос:

все 100% опрошенных считают, что данные инструменты необходимы, а те, которые есть, — написаны недостаточно качественно; респондентам не хватает существующего функционал

### Существующие предложения:

под каждую из задач есть разные решения, но все они состоят из множества трудоемких шагов: PyTorch, Heroku, Dialogflow

---

# Наше решение: Большой и средний бизнес

---

## Проблема:

неэффективность существующей системы поддержки клиентов

## Предлагаемое решение:

бот-помощник / сайт / десктопное приложение / фреймворк, который автоматически подсказывает человеку из команды поддержки варианты ответа на вопрос пользователя

## Сроки:

данный инструмент находится в финальной стадии разработки  
оценочное время завершения разработки — лето 2018

---

# Наше решение: Большой и средний бизнес

---

## Рыночные факторы:

Основным фактором стал “заказ” Сбербанка

## Спрос:

- многие из опрошенных (63% от всех опрошенных) высказались, что используют ботов-помощников
- так же было замечено, что компании работают в этом направлении (например, Тинькофф).

## Существующие предложения:

- сейчас все разработки в данной области являются результатом работы отдельной команды разработчиков; многие до сих пор пользуются текстовыми скриптами для решения данной проблемы (например, call-центры);
- какого-то инструмента для создания такой системы еще нет



# Существующие решения

Framework	Text Classification	Text Generation	Text Summarization	NER	POS tagging	Word embeddings	Dependency Parsing	SRL	Language Modeling	Machine Translation	Speech Recognition
Deep learning 4j	✓				✓	✓	✓				
Keras	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓
Tesnor Flow	✓	✓	✓	✓	✓	✓			✓	✓	✓
Theano	✓	✓	✓	✓	✓	✓		✓		✓	✓

---

# Существующие решения

## Вывод

---

- Превзойти конкурентов, сделав лишь универсальное средство разработки, не получится: уже есть библиотеки, которые дают возможность выполнять многие известные задачи, связанные с обработкой и анализом текста
- Необходимо добиться того, что наше средство будет давать лучшие результаты в рассматриваемых задачах
- Больше высокоуровневых решений для низкоквалифицированных разработчиков

# Метрики

Что измеряем	Как измеряем
Качество работы продукта и его составных частей	Всевозможные метрики качества моделей машинного обучения
«Успешность» продукта	Количество пользователей, обратная связь
	Появление на рынке новых продуктов, использующих наши решения
Ошибки, баги и т.д.	Взаимодействие с разработчиками через github



---

# NER: Сбор Данных

---

- Данные с юридических форумов:

- legal-forum.ru
- yurist-forum.ru
- zonazakona.ru

*<...> Соседи тещи в деревне установили на своем доме камеру видеонаблюдения за ее участком, заявив, что она, якобы, ворует у них плоды и овощи и теперь они имеют возможность за нею наблюдать. <...>  
Подпадает ли это под ст 152,2 и если да, то как грамотно поступить - вызывать милицию или прямо обращаться в суд?*

- Данные из социальных сетей

- объявления в ВК
- продажа квартир в ВК

*Продам шаль 3 000₽ за одну (голубая, сереневая, красная). Возможен торг. г. Екатеринбург. По всем вопросам обращаться в личку.*

# NER: Яндекс.Толока

9:17 0,01\$ Разметка текста 2

Инструкция



Я, К\*\*\* (паспорт \*\*\*), зарегистрированная и проживающая по адресу : г. Сургут, ул. 30 лет Победы, \*\*\*; согласно Конституции Российской Федерации, в частности статьи 24, пункта 2 о предоставлении возможности ознакомления с документами и материалами, непосредственно затрагивающими мои права и свободы

При этом суд, по мнению судебной коллегии не учел, что несовершеннолетний К. А., несмотря на регистрацию в комнате размером 18, 6 кв. м по адресу : <...>, фактически в нее вселен не был, проживал с родителями на другой жилой площади, в связи с чем не приобрел право пользования спорной жилой площадью.

Минздравсоцразвития РФ письмом 30.09.2005 года № 4677-ВС (л.д.)

(л.д.)

Статью написала Елена Чудинова (С) для деловой газеты "Взгляд".

начните добровольную ликвидацию с переходом на упрощенную процедуру по Статья 224.

, сконцентрировались в местной курилке ... И уже себя, млин, обществом считают ... И еще какие - то глобальные выводы делают вселенской же глупости, прям как Шариков в "Собачем сердце".

1) Андрей С.

Условия : зарплата по результатам собеседования (обсуждается индивидуально), офис станция метро ул. 1905г.



# NER: Active Learning

