Start-UP ХАБАРОВСК

devDV

ION DV

СТАРТАПУ ИИ:

NLP – автоматическая обработка текста на естественном языки

Обсуждаем: https://t.me/devdvAI, https://t.me/devdvStartup

Репозиторий: https://github.com/akumidv/startup-khv-ai-study

ЗАЧЕМ START-UP ХАБАРОВСК В СФЕРЕ ИТ?

ОБОРОТ ОТРАСЛИ

⇧ С **0,7** МЛРД.
ДО **3-5** МЛРД.РУБ

ЗАНЯТОСТЬ

⇧ С **500** ДО
1.5-2 ТЫС.
СПЕЦИАЛИСТОВ

ЗП

⇧ УРОВЕНЬ
КВАЛИФИЦИРОВАННЫХ
ЗП С.ПЕТЕРБУРГА И
МОСКВЫ

НАЗНАЧЕНИЕ ЦИКЛА

ПОМОЧЬ ФОРМИРОВАНИЮ СООБЩЕСТВА #AI И #MASHINELEARNING

ПО ВОЗМОЖНОСТИ СФОРМИРОВАТЬ КОМАНДЫ

В ИДЕАЛЕ СФОРМИРОВАТЬ ПРОДУКТЫ ДЛЯ СТАРТАПОВ

ПОМОЧЬ С ТРУДОУСТРОЙСТВОМ ИЛИ РАБОТОЙ НА ФРИЛАНС

КОГДА НУЖНА ОБРАБОТКА ТЕКСТОВ

- Категоризация текстов, например писем
- Классификация новостей
- Сопоставление вопроса и ответа
- Антиспам
- Определение языка

НОРМАЛИЗАЦИЯ ПРЕДЛОЖЕНИЙ И СЛОВ

МИНИМАЛЬНАЯ ПОДГОТОВКА

- Тримминг
- Нижний регистр
- Удаление пунктуации и специальных символов
- Удаление дублирующихся слов

ДОПОЛНИТЕЛЬНАЯ ПОДГОТОВКА

- Лемматизация => к словарной форме
- Стемминг => основа слов
- Удаление «лишних»(стоп) слов

ТОКЕНИЗАЦИЯ

РАЗБИЕНИЕ ТЕКСТА НА МАССИВ СЛОВ

'Если вы не можете объяснить это своей бабушке, вы сами этого не понимаете’

```
[
  'Если',    'вы’,   'не',      'можете’,  'объяснить', 'это',
  'своей',   'бабушке’, 'вы',     'сами’, 'этого',    'не’,
  'понимаете'
]
```

СТЕММИНГ

СЛОВА ПРИВЕДЕННЫЕ К ОСНОВЕ

['Если’, 'вы’, 'не’, 'можете’, 'объяснить’, 'это',
'своей', 'бабушке’, 'вы', 'сами’, 'этого', 'не’,
'понимаете']

['есл', 'вы’, 'не', 'может’, 'объясн’, 'эт’,
'сво', 'бабушк’, 'вы', 'сам’, 'эт', 'не’,
'понима’]

МЕШОК СЛОВ

ЧАСТОТА С КОТОРОЙ ВСТРЕЧАЕТСЯ СЛОВО

greet = ['Привет', 'Здравствуй', 'Добрый день’, 'Добрый вечер', 'Здравствуйте', 'Приветствую’, 'Здорова', 'Доброе утро']

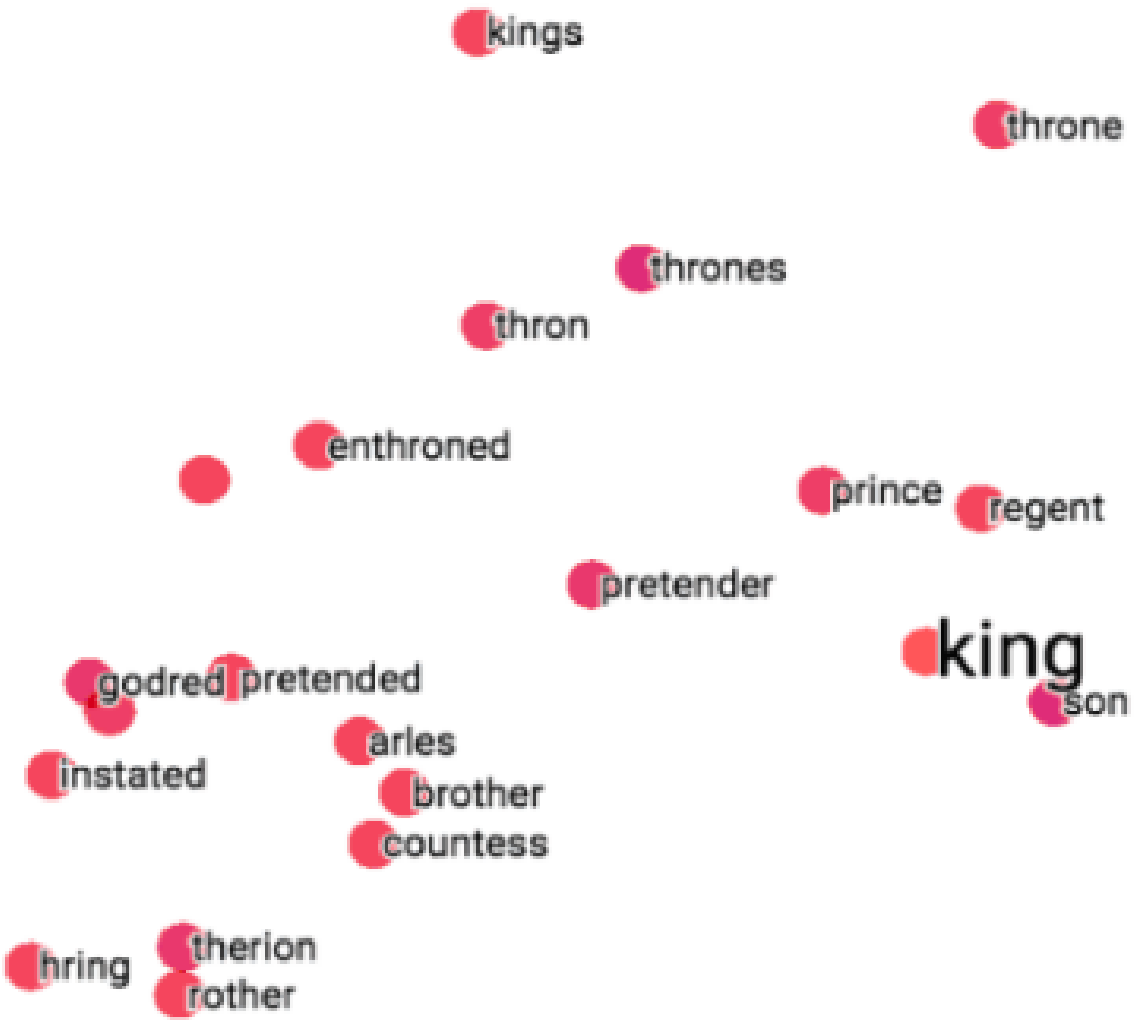
bye = ['Пока', 'До встречи', 'До свидания’, 'Прощай', 'Еще увидимся', 'скоро увидимся', 'до новых встречь’, 'Дотвиданья’]

```
{ 'привет': 1,
  'здравств': 2,
  'добр': 3,
  'ден': 1,
  'вечер': 1,
  'приветств': 1,
  'здоров': 1,
  'утр': 1,
  'пок': 1,
  'до': 2,
  'встреч': 2,
  'свидан': 1,
  'проща': 1,
  'ещ': 1,
  'увид': 2,
  'скор': 1,
  'нов': 1,
  'дотвидан’: 1 }
```

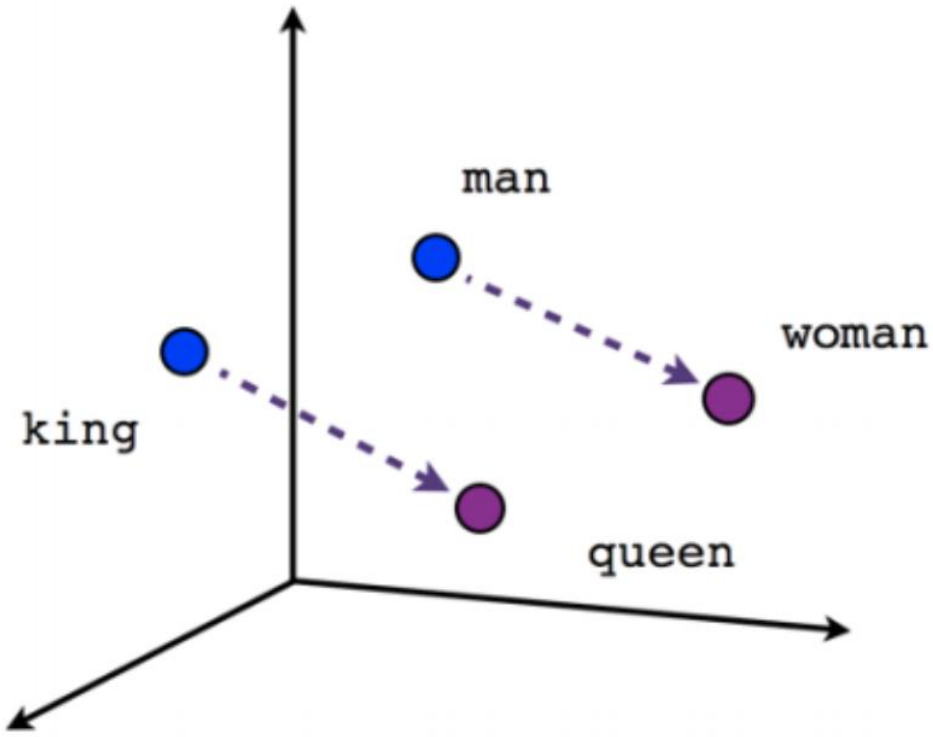
КЛАССИФИКАЦИЯ БАЙЕСА(ВЕРОЯТНОСТЬ)

	ПРИВЕТСТВИЕ	ПРОЩАНИЕ
'Добрых суток'	0.235	0.058
'Покедова'	0.529	0.529
'Увидимся'	0.176	0.058

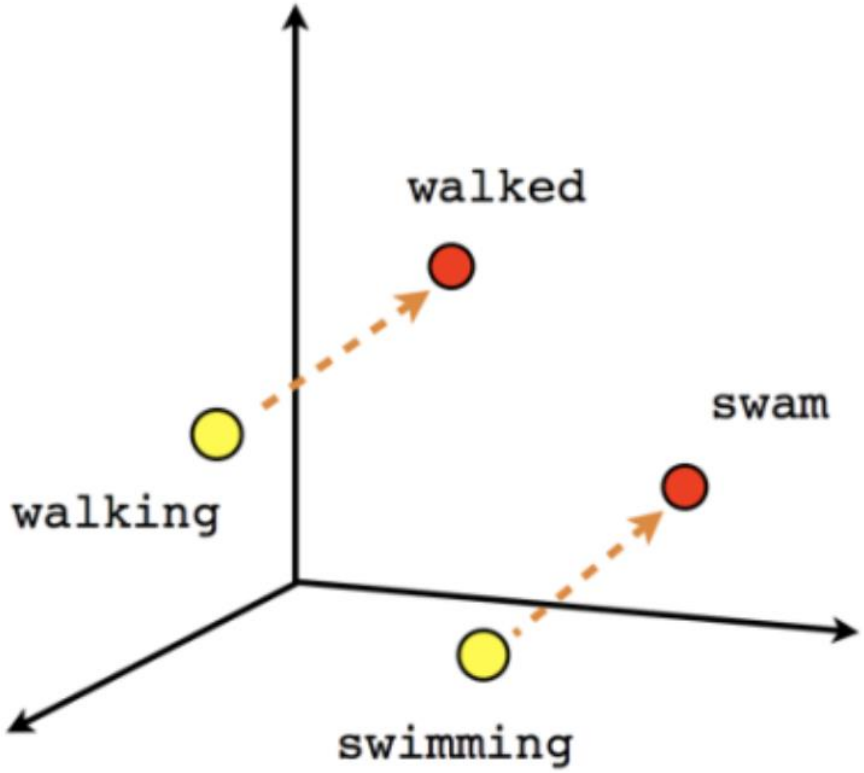
КАК ВЫГЛЯДИТ ОБЛАКО СЛОВ



БЛИЗОСТЬ СЛОВ В ОБЛАКЕ



Male-Female



Verb tense

ВЕКТОР СЛОВ

In comparison to dogs, cats have not undergone major changes during the domestication process.

СЛОВАРЬ

```
in:0
 comparison:1
 to:2
 dogs:3
 cats:4
 have:5
 not:6
 undergone:7
 major:8
 changes:9
 during:10
 the:11
... 240 more items
```

ВЕКТОР

```
[
  1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
  1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
  0, 0, 0, 0,
  ... 152 more items
]
```

КОСИНУСНАЯ БЛИЗОСТЬ ПРЕДЛОЖЕНИЙ

	ДИСТАНЦИЯ
• In comparison to dogs, cats have not undergone major changes during the domestication process.	0.0000
• As cat simply catenates streams of bytes, it can be also used to concatenate binary files, where it will just concatenate sequence of bytes.	0.95275
• A common interactive use of cat for a single file is to output the content of a file to standard output.	0.8644
• Domestic cats are similar in size to the other members of the genus Felis, typically weighing between 4 and 5 kg (8.8 and 11.0 lb).	0.7327

N-GRAM-Ы СЛОВ

This is Big Data AI Book

Uni-Gram	This	Is	Big	Data	AI	Book
Bi-Gram	This is	Is Big	Big Data	Data AI	AI Book	
Tri-Gram	This is Big	Is Big Data	Big Data AI	Data AI Book		

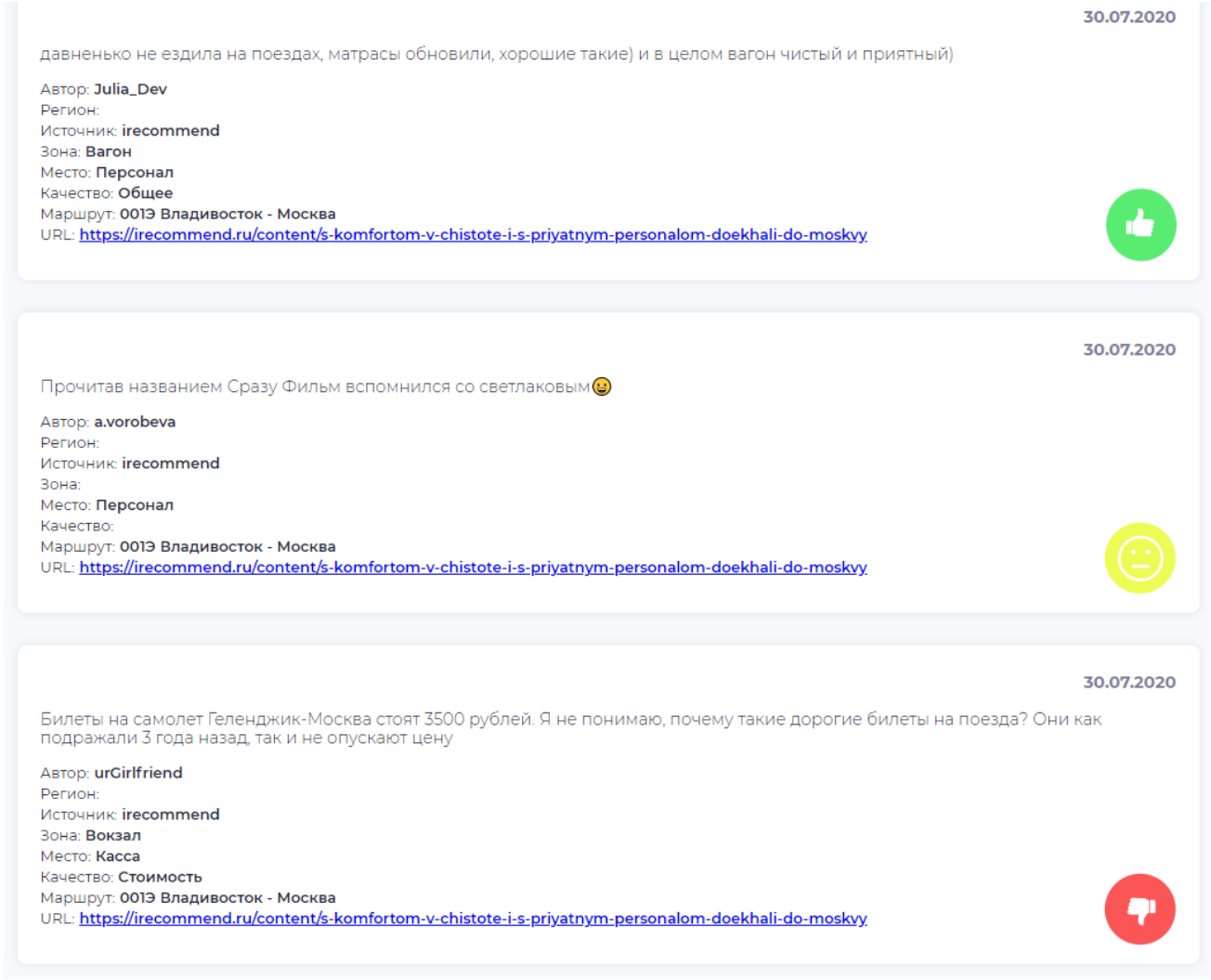
N-GRAM-Ы СИМВЛОВ


fastText
Library for efficient text classification and representation learning

fasttext.cc

ПРИМЕР FASTEXT ДЛЯ КЛАССИФИКАЦИИ

https://rzhd-reputation.iondv.ru/portal



30.07.2020

давненько не ездила на поездах, матрасы обновили, хорошие такие) и в целом вагон чистый и приятный)

Автор: **Julia_Dev**
Регион:
Источник: **irecommend**
Зона: **Вагон**
Место: **Персонал**
Качество: **Общее**
Маршрут: **001Э Владивосток - Москва**
URL: https://irecommend.ru/content/s-komfortom-v-chistote-i-s-priyatnym-personalom-doekhali-do-moskvy

30.07.2020

Прочитав названием Сразу Фильм вспомнился со светлаковым

Автор: **a.vorobeva**
Регион:
Источник: **irecommend**
Зона:
Место: **Персонал**
Качество:
Маршрут: **001Э Владивосток - Москва**
URL: https://irecommend.ru/content/s-komfortom-v-chistote-i-s-priyatnym-personalom-doekhali-do-moskvy

30.07.2020

Билеты на самолет Геленджик-Москва стоят 3500 рублей. Я не понимаю, почему такие дорогие билеты на поезда? Они как подражали 3 года назад, так и не опускают цену

Автор: **urGirlfriend**
Регион:
Источник: **irecommend**
Зона: **Вокзал**
Место: **Касса**
Качество: **Стоимость**
Маршрут: **001Э Владивосток - Москва**
URL: https://irecommend.ru/content/s-komfortom-v-chistote-i-s-priyatnym-personalom-doekhali-do-moskvy

СТРУКТУРА НЕЙРОСЕТИ FASTTEXT

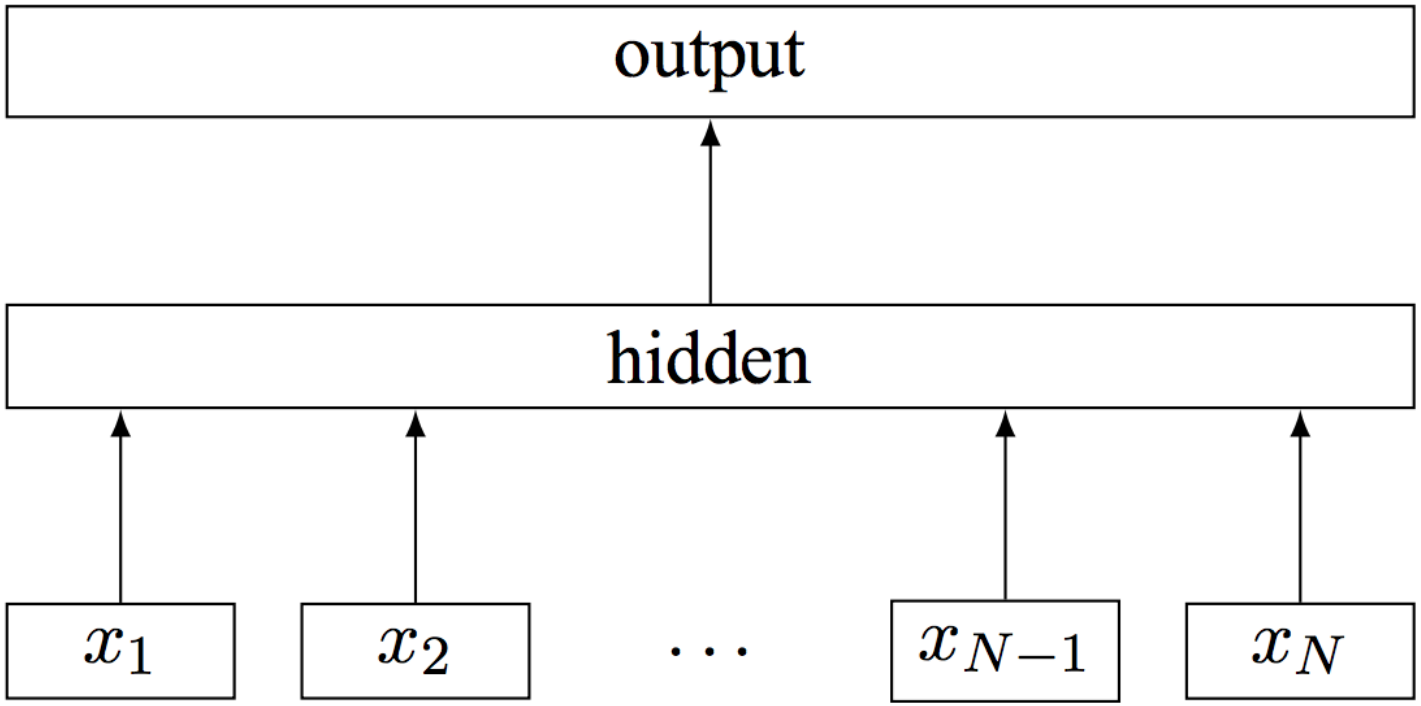


Figure 1: Model architecture of `fastText` for a sentence with N ngram features $x_1, \ldots, x_N$. The features are embedded and averaged to form the hidden variable.

РАЗМЕТКА

__label__greet Привет
__label__greet Здравствуй
__label__greet Добрый день
__label__greet Добрый вечер
__label__greet Здравствуйте
__label__greet Приветствую
__label__greet Здорова
__label__greet Доброе утро

__label__bye Пока
__label__bye До встречи
__label__bye До свидания
__label__bye Прощай
__label__bye Еще увидимся
__label__bye скоро увидимся
__label__bye до новых встречь
__label__bye Дотвиданья

СМОТРИМ ПРИМЕРЫ

ЧТО МОЖНО СДЕЛАТЬ ДЛЯ УЛУЧШЕНИЯ МОДЕЛЕЙ?

ДОРОЖНАЯ КАРТА МЕРОПРИЯТИЙ

СКРАПИНГ

ТЕКСТ

- Векторная оценка схожести предложений
- Классификация текста fasstext
- **Предсказание текста, нейронные сети**

ВРЕМЕННЫЕ РЯДЫ

- Прогнозирование на нейронных сетях
- AutoML (hyperopt)

ИЗОБРАЖЕНИЯ

- Классификация на нейронных сетях
- Детекция CV и нейронные сети

UPWORK ИЛИ СТАРТАП

КОНТАКТЫ

ОБСУЖДАЕМ

https://t.me/devdvAI

https://t.me/devdvStartup

РЕПОЗИТОРИЙ

https://github.com/akumidv/startup-khv-ai-study

АНДРЕЙ КУМИНОВ

+7 914 770 5846

https://facebook.com/akuminov

https://vk.com/akumidv