Start-UP ХАБАРОВСК

devDV

ION DV

СТАРТАПУ ИИ:
Извлечение данных в веб и программные роботы

Ч.1 Основы

Start-UP
ХАБАРОВСК

Инноваторы, предприниматели

Медиа среда и сообщества

Крупный бизнес Инвесторы

ИТ-Парк, ресурсы и поддержка

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

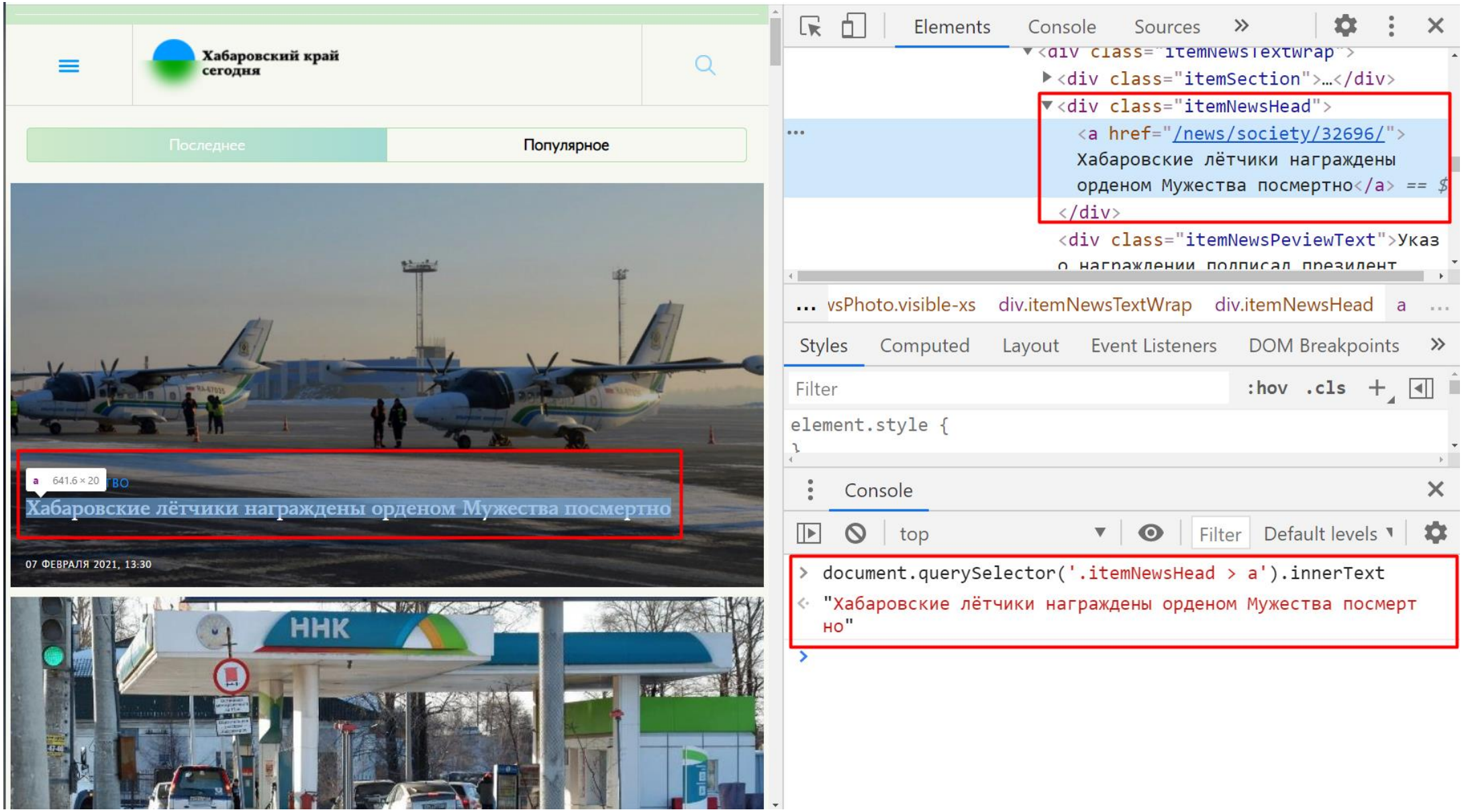
ПОМОЧЬ ФОРМИРОВАНИЮ СООБЩЕСТВА #AI И #MASHINELEARNING

ПО ВОЗМОЖНОСТИ СФОРМИРОВАТЬ КОМАНДЫ

В ИДЕАЛЕ СФОРМИРОВАТЬ ПРОДУКТЫ ДЛЯ СТАРТАПОВ

ПОМОЧЬ С ТРУДОУСТРОЙСТВОМ ИЛИ РАБОТОЙ НА ФРИЛАНС

БАЗА ДЛЯ СКРАПИНГА: HTML



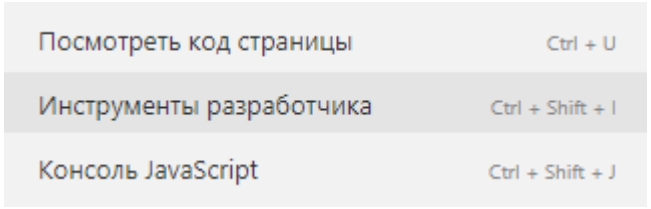
СЕЛЕКТОРЫ

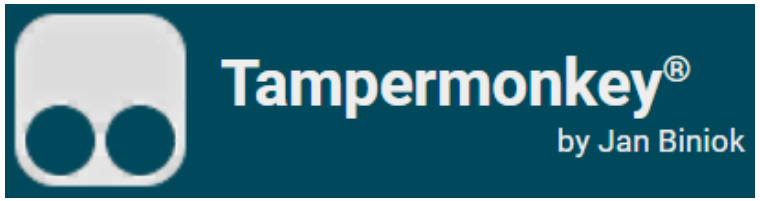
```
document.querySelector ('*’)          // Любые элементы

document.querySelector ('div’)        // Элементы с таким тегом

document.querySelector ('#login’)     // Элементы с данным id

document.querySelector ('.news’)      // Элементы с таким классом

document.querySelector ('[name="article"]') // Селекторы на атрибут

document.querySelector (':visited’)   // «Псевдоклассы»
```

ОТНОШЕНИЯ И СОЧЕТАНИЯ

```
$(‘.imgWrap.marker3’)                  // С двумя классами

$('.itemNewsHead > a ‘)                // a непосредственный потомок .itemNewsHead

$('.itemNewsWrap  .itemNewsHead') //  .itemNewsHead непосредственный
                                       //  потомок .itemNewsWrap

$(‘.itemNewsWrap:nth-child(2)’)        // 2й потомок itemNewsWrap

$(‘.itemNewsHead:nth-of-type(2)’)  // 2й потомок itemNewsWrap с тем же тегом
```

ПРОСТЕЙШИЙ ПАРСИНГ В БРАУЗЕРЕ






```
const newsHeads = document.querySelectorAll('.itemNewsWrap .itemNewsHead > a')
const firstHeadText = newsHeads[1].innerText;

console.log('Всего заголовков:', newsHeads.length)
console.log('Заголовок первого элемента:', firstHeadText)

// Сохраняем данные в файл
const saveLink = document.createElement('a');
saveLink.href = 'data:text,Всего заголовков ' + newsHeads.length +
                '\nЗаголовок первого элемента ' + firstHeadText;
saveLink.download = "data.txt";
saveLink.click();
```

ДВА ПОДХОДА

КАКОЙ ПОБЕДИТ?

ПОЛУЧИТЬ СТРАНИЦУ И ПАРСИТЬ HTML

VS

ОТКРЫТЬ БРАЗУЕР И ПАРСИТЬ В НЕМ

ПОЛУЧЕНИЕ СТРАНИЦЫ И ПАРСИНГ


```
const home = await superagent
  .get('https://todaykhv.ru/')
const htmlHome = home.res.text

let $ = cheerio.load(htmlHome);

const nH = $('.itemNewsWrap .itemNewsHead > a’)


console.log(nH.first().text())
```

```
home = requests.get('https://todaykhv.ru/’)

htmlHome = home.text

soup = BeautifulSoup(htmlHome, 'html.parser’)

newsHeads = soup.select('.itemNewsWrap
.itemNewsHead > a')

print(newsHeads[0].text)
```

ДЛЯ ЧЕГО НУЖНО УПРАВЛЯТЬ БРАУЗЕРОМ?

АВТОРИЗАЦИЯ

ПЕРЕХОДЫ ПО СТРАНИЦАМ ПОСЛЕ ЗАПОЛНЕНИЯ ДАННЫХ

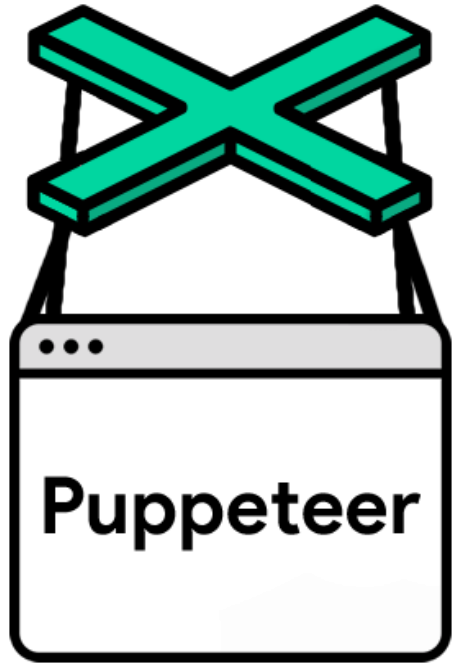
REACT: МОДЕЛЬ ПОЛУЧЕНИЯ ДАННЫХ И РЕНДЕРИНГА ПРЕДСТАВЛЕНИЯ В БРАУЗЕРЕ

СКРОЛЛИНГ И ПОДГРУЗКА ДАННЫХ

БЛОКИРОВКИ ПО ПРОФИЛЯМ ЗАПРОСОВ НЕ ПОХОЖИХ НА ЧЕЛОВЕКА

УПРАВЛЕНИЕ БРАУЗЕРОМ

vs



ОСОБЕННОСТИ SELENIUM

- РАЗНЫЕ БРАУЗЕРЫ
- СЕРВЕР И GRID
- ОСНОВНЫЕ ЯЗЫКИ

- ТЯЖЕЛЫЙ ДЛЯ ТОНКОГО УПРАВЛЕНИЯ
- НИЗКАЯ СКОРОСТЬ

ОТКРЫТИЕ СТРАНИЦЫ И ПАРСИНГ

JS

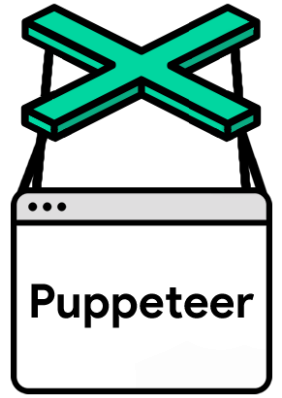
python

Selenium

```
let driver = await new
Builder().forBrowser('chrome').build();

await driver.get('https://todaykhv.ru/’);



const newsHeads = await
driver.findElements(By.css('.itemNewsWrap
.itemNewsHead > a’))

const firstHeadText = await newsHeads[1].getText();

console.log(`Главная заголовков: ${newsHeads.length}`);
console.log(`Заголовок первого элемента:
${firstHeadText}\n`);
```

```
driver = webdriver.Chrome()

driver.get('https://todaykhv.ru')

WebDriverWait(driver,
10).until(EC.presence_of_element_located((By.CSS_SELEC
TOR, '.itemNewsWrap .itemNewsHead > a')))

    newsHeads =
driver.find_elements_by_css_selector('.itemNewsWrap
.itemNewsHead > a')
    firstHeadText = newsHeads[0].text

    print('Главная, заголовков:', len(newsHeads))
    print('Заголовок первого элемента:', firstHeadText)
```

ОСОБЕННОСТИ PUPPETEER



УПРАВЛЕНИЕ БРАУЗЕРОМ НА НИЗКОМ УРОВНЕ

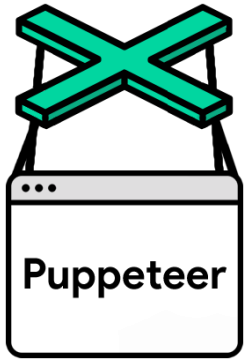
ВСТРОЕННЫЕ ФУНКЦИИ РАБОТЫ С ОБЪЕКТАМИ

СКОРОСТЬ И HEADLESS

НЕТ ВНЕШНЕГО СЕРВЕРА С GRID

JS, ОСТАЛЬНЫЕ ЯЗЫКИ НЕ ПРИЖИЛИСЬ

ОТКРЫТИЕ СТРАНИЦЫ И ПАРСИНГ



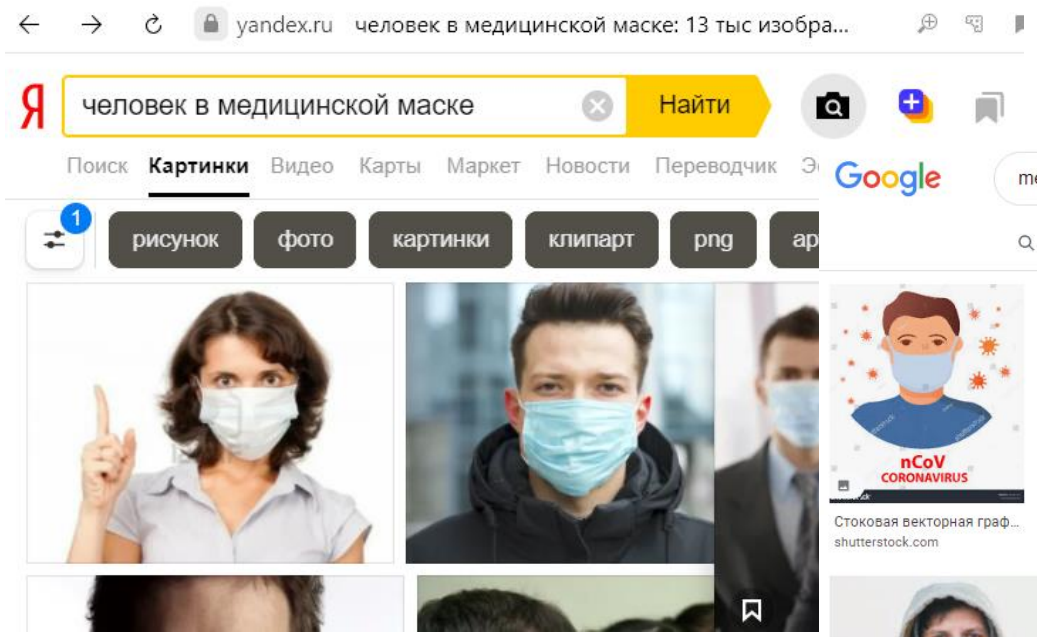

```
const browser = await puppeteer.launch({headless: false});
let page = await browser.newPage();

await page.goto('https://todaykhv.ru/', {waitUntil: 'load'});

await page.watiForSelector('.itemNewsWrap .itemNewsHead
> a', {timeout: 10000})

const newsHeads = await page.$$('.itemNewsWrap
.itemNewsHead > a')

const firstHeadText = await newsHeads[0].evaluate(el =>
el.innerText)

console.log(`Всего, заголовков: ${newsHeads.length}`)
console.log(`Заголовок первого элемента: ${firstHeadText}`)
```

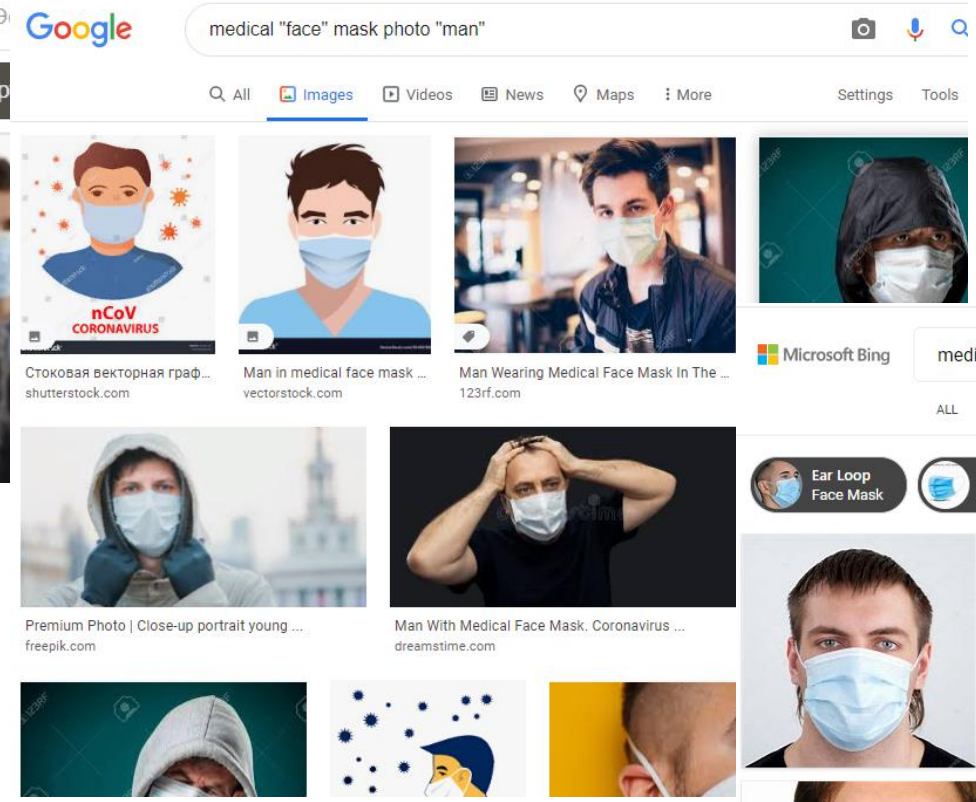
```
browser = await launch(headless=False)
page = await browser.newPage()

await page.goto('https://todaykhv.ru/')

await page.waitForSelector('.itemNewsWrap .itemNewsHead > a',
timeout=10000)

newsHeads = await page.JJ('.itemNewsWrap .itemNewsHead > a')

firstHeadText = await page.evaluate('el => el.innerText', newsHeads[0])

print('Главная, заголовков:', len(newsHeads))
print('Заголовок первого элемента:', firstHeadText)
```

ЗАЧЕМ ЕЩЁ НУЖНО?

ГДЕ БОЛЬШЕ ВСЕГО РАЗМЕЧЕННЫХ ИЗОБРАЖЕНИЙ?

В ПОИСКОВИКАХ ЕСТЬ РАЗМЕТКА





ПРОГРАММНЫЕ РОБОТЫ

ДЕЛАЕМ В СИСТЕМЕ ВМЕСТО ЧЕЛОВЕКА

Ввод данных

Перенос данных между двумя системами

Выполнение рутинных проверок и действий

ПРАКТИКА

ПАРСИМ НОВОСТИ

https://meduza.ru

- ЗАГОЛОВКИ
- ТЕКСТЫ
- ДАТЫ
- ССЫЛКИ

(*) ПАРСИМ ПОИСК

ЗАДАНИЕ НА ДОМ

ПАРСИМ ПО НАЗВАНИЮ ЛЮБОЙ ОРГАНИЗАЦИИ

https://news.yandex.ru

ЗАГОЛОВКИ
ТЕКСТЫ
ДАТЫ
ССЫЛКИ

ОРГАНИЗАЦИОННЫЕ ВОПРОСЫ

ВРЕМЯ И ДЕНЬ НЕДЕЛИ?

НУЖНО ЕЩЕ ПРАКТИЧЕСКОЕ ЗАНЯТИЕ ПО СКРАПИНГУ ДАННЫХ?

* работа с файлами и управление интерфейсом
* управление в виндовс

НУЖНО ПО ИНТЕГРАЦИОННОМУ ТЕСТИРОВАНИЮ (НА JS)?

МОЖНО КОНТРИБУТИТЬ IONDV

ПОСЛЕ СКРАПИНГА

ТЕКСТ

- Нормализация, токенизация
- Векторная оценка схожести предложений
- Классификация текста fasstext, нейронные сети

ВРЕМЕННЫЕ РЯДЫ

- Прогнозирование на нейронных сетях
- AutoML (hyperopt)

ИЗОБРАЖЕНИЯ

- Классификация на нейронных сетях
- Детекция CV и нейронные сети

КОНТАКТЫ

ОБСУЖДАЕМ

https://t.me/devdvAI

АНДРЕЙ КУМИНОВ

+7 914 770 5846

https://facebook.com/akuminov

https://vk.com/akumidv