


[github.com/](https://github.com/)

# natasha

— библиотека для извлечения  
структурированной информации из  
текстов на русском языке

Александр Кукушкин  
[lab.alexkuk.ru](http://lab.alexkuk.ru)

# Примеры / Разбор резюме



## Кукушкин Александр Викторович

Мужчина, 27 лет, родился 18 сентября 1990

+7 (916) 3668039 — предпочитаемый способ связи  
alexanderkuk@ya.ru

Проживает: Москва, м. Каширская  
Гражданство: Россия, есть разрешение на работу: Россия  
Не готов к переезду, не готов к командировкам

### Желаемая должность и зарплата

<b>Аналитик-разработчик</b>	<b>1 000 000</b>
Информационные технологии, интернет, телеком	руб.
<ul style="list-style-type: none"><li>• Программирование, Разработка</li><li>• Управление проектами</li><li>• Стартапы</li></ul>	
Занятость: полная занятость	
График работы: полный день	
Желательное время в пути до работы: не имеет значения	

### Опыт работы — 8 лет 7 месяцев

Май 2012 — **Видеок**

## Примеры / Разбор резюме

...

Опыт работы - 1 год 8 месяцев

Июль 2011 — Октябрь 2012

КФ ОАО Ростпечать

Семикаракорск

Бухгалтер-ревизор

Проведение ревизий, начисление зп,  
сдача отчета в ФСС, 2 НДФЛ, сдача  
бухгалтерской отчётности, поездки в  
ФНС

Май 2010 — Май 2011

ООО «Селикамск»

Махачкала

...

```
[
  {
    "period": {
      "start": {
        "year": 2011, "month": 7
      },
      "stop": {
        "year": 2012, "month": 10
      }
    },
    "position": "Бухгалтер-ревизор",
    "company": {
      "name": "КФ ОАО Ростпечать",
      "area": "Семикаракорск",
    }
  },
  ""
]
```

# Примеры / Контакты

О компании

Реклама на портале

Хостинг сайтов, домены

Информационная поддержка

Статистика посещаемости

Загрузка каналов Sakh.com

Контакты

Реквизиты

Команда

Вакансии

Промоматериалы

Информеры

Наши кнопки

История

10 лет

15 лет

## Контактная информация

Телефон:

- +7 4242 511090 — Билеты, Афиша
- +7 4242 511091 — Авто
- +7 4242 511092 — Недвижимость
- +7 4242 511105, 457099 — Новости
- +7 4242 511106, 457090 — Shoppy.ru
- +7 4242 511107 — Еда
- +7 4242 511109 — Факс
- +7 4242 511100, 457000 — Общий

Время работы:

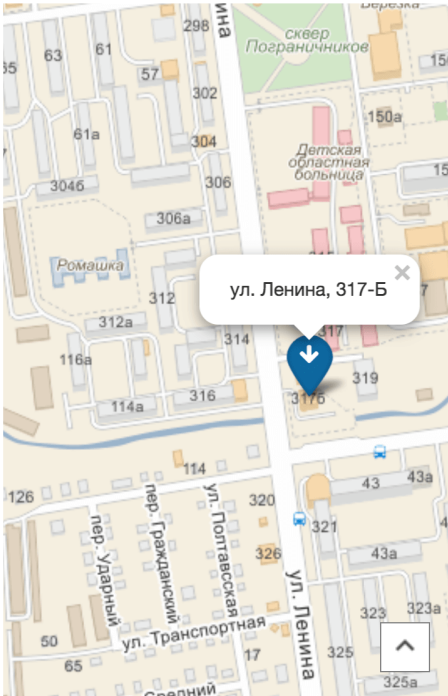
с 09:00 до 18:00  
сб, вс — выходной

Адрес: г. Южно-Сахалинск,  
ул. Ленина, 317-Б, 3-й этаж

Почтовый адрес: Россия, 693006,  
г. Южно-Сахалинск, ул. Ленина, 317-Б

Е-mail:

- reklama@sakh.com — реклама
- news@sakh.com — новости
- forum@sakh.com — администрация форума
- info@sakhalin.biz — справочник Сахалин.Бизнес
- support@sakh.com — техподдержка, хостинг
- auto@sakh.com — техподдержка Авто



ул. Ленина, 317-Б

## Примеры / Контакты

Луговой пр., д. 43-45, лит. Б

Магазин работает с 10.00 до 18.00  
Суббота, воскресенье – выходные

Телефон: 8 (812) 4954301

Факс: 8 812 3253523

shop@gold10.ru

Реквизиты: ООО «Голд 10 Маркет»  
ИНН: 7840038965; КПП: 784001001;

```
{
  "phones": ["+7 812 495-43-01"],
  "faxes": ["+7 812 325-35-23"],
  "emails": ["shop@gold10.ru"],
  "addresses": [
    {
      "parts": [
        {
          "name": "Луговой",
          "type": "проспект"
        },
        {
          "number": "43-45",
          "type": "дом"
        },
        {
          "value": "Б", "type": "литер"
        }
      ]
    }
  ],
  "inn": "7840038965",
  "kpp": "784001001"
}
```

# Примеры / Названия товаров

Купить игровую приставку, К

Alexander

← → ↻ ↺


Защищено

https://www.avito.ru/moskva/igry\_pristavki\_i\_programmy/igrovy\_e\_pristavki?user=1&view=gallery


☆ ⋮

Все Частные Компании


По умолчанию ▾




**Sony PlayStation 4 Slim. PRO. Xbox One S/X. PS3**  
18 890 р.  
м. Савеловская  
Сегодня 10:58




**Джойстики PS3, Геймпады PS3, Dualshock 3. Магазин**  
1 000 р.  
Возможна доставка  
Сегодня 10:30




**Джойстик PS2 PS3 PS4 Xbox 360**  
1 000 р.  
м. Братиславская  
Сегодня 09:43



**Геймпад с гироскопом для PS3 с пиксельным окрасом**  
1 000 р.  
м. Бабушкинская  
Сегодня 12:03




**Dualshock 4 белый**  
2 600 р.  
м. Академическая  
Сегодня 12:03




**Tom Clancy's : Ghost Recon - wildlands / StarWars**  
4 000 р.  
Возможна доставка  
Сегодня 12:01

VIP-объявления



**Продажа Playstation 3, игры**  
6 000 руб.  
Жулебино  
19 апреля 13:31



**Ps4**  
16 000 руб.

## Примеры / Названия товаров

Пульт PlayStation3

Продаю Ps4 pro

Сонька 3

Sony PlayStation 4 500Gb Slim

Новый жесткий кейс psp

Ps3 slim прошитая rebug 4.81 + fifa 18

```
{
  "accessory": {
    "type": "пульт"
  },
  "console": {
    "name": "PlayStation",
    "model": "3"
  }
}
{
  "console": {
    "name": "PlayStation",
    "model": "4",
    "version": "PRO"
  },
  "attributes": []
}
...
```

Библиотека «Наташа» — набор  
готовых правил для Yargy-парсера

[github.com/natasha/natasha](https://github.com/natasha/natasha)

CRF-  
теггер

[python-  
crfsuite](https://github.com/natasha/python-crfsuite)

Yargy-парсер — аналог  
яндексowego Томита-  
парсера для Python

[github.com/natasha/yargy](https://github.com/natasha/yargy)

Морфологический  
анализатор

[pymorphy2](https://github.com/natasha/pymorphy2)

Технология  
«Наташа»

[github.com/natasha](https://github.com/natasha)



# Принцип работы / Контекстно-свободные грамматики и словари

name → first

last

first last

last first

first middle last

last first middle

~~middle first last~~

abbr . abbr . last

last abbr . abbr .

Борис

Королёв

Николь Кидман

Анна Павловна Шерер

~~Иванович Иван Иванов~~

А.С.Пушкин

## Принцип работы / DSL

```
from yargy import rule, or_, Parser
from yargy.predicates import gram

FIRST = gram('Name')
LAST = gram('Surn')
MIDDLE = gram('Patr')
ABBR = gram('Abbr')

NAME = or_(
    rule(FIRST),
    rule(LAST),
    rule(FIRST, LAST),
    rule(LAST, FIRST),
    rule(FIRST, MIDDLE, LAST),
    rule(LAST, FIRST, MIDDLE),
    rule(ABBR, '.', ABBR, '.', LAST),
    rule(LAST, ABBR, '.', ABBR, '.'),
)
```

```
parser = Parser(NAME)
text = '''
Хочу поблагодарить учителей Бушуева
Вячеслава Владимировича и Бушуеву
Веру Константиновну.
'''

for match in parser.findall(text):
    start, stop = match.span
    print(text[start:stop])

>>> Бушуева Вячеслава Владимировича
>>> Бушуеву Веру Константиновну
```

## Принцип работы / Интерпретация и нормализация

```
from yargy.interpretation import fact
```

```
Name = fact(  
    'Name',  
    ['first', 'middle', 'last']  
)
```

```
FIRST = gram('Name').interpretation(  
    Name.first.inflected()  
)
```

```
LAST = gram('Surn').interpretation(  
    Name.last.inflected()  
)
```

```
MIDDLE = gram('Patr').interpretation(  
    Name.middle.inflected()  
)
```

```
NAME = or_(...)
```

```
parser = Parser(NAME)
```

```
text = '''
```

```
Хочу поблагодарить учителей Бушуева  
Вячеслава Владимировича и Бушуеву  
Веру Константиновну.  
'''
```

```
for match in parser.findall(text):  
    print(match.fact)
```

```
>>> Name(  
...     first='вячеслав',  
...     middle='владимирович',  
...     last='бушуев'  
... )
```

```
>>> Name(  
...     first='вера',  
...     middle='константиновна',  
...     last='бушуев'  
... )
```

## Принцип работы / Согласование

```
from yargy.relations import gnc_relation parser = Parser(NAME)
gnc = gnc_relation() text = '''
FIRST = gram('Name').interpretation( Хочу поблагодарить учителей Бушуева
Name.first.inflected() Вячеслава Владимировича и Бушуеву
).match(gnc) Веру Константиновну.
LAST = gram('Surn').interpretation(
Name.last.inflected()
).match(gnc)
MIDDLE = gram('Patr').interpretation(
Name.middle.inflected()
).match(gnc)
NAME = or_(...
```

```
for match in parser.findall(text):
    print(match.fact)
```

```
>>> Name(
...     first='вячеслав',
...     middle='владимирович',
...     last='бушуев'
... )
>>> Name(
...     first='вера',
...     middle='константиновна',
...     last='бушуева'
... )
```

## Принцип работы / Сборник готовых правил

```
from natasha import NamesExtractor

extractor = NamesExtractor()
text = '''
Хочу поблагодарить учителей Бушуева
Вячеслава Владимировича и Бушуеву
Веру Константиновну.
'''

for match in extractor(text):
    print(match.fact)
```

```
>>> Name(
...     first='вячеслав',
...     middle='владимирович',
...     last='бушуев',
...     nick=None
... )
>>> Name(
...     first='вера',
...     middle='константиновна',
...     last='бушуева',
...     nick=None
... )
```

Экстрактор	Пример	Аналоги
NamesExtractor	Отрицал существование <u>Иисуса</u> и пророка <u>Мухаммеда</u> , наделял <u>Иисуса Христа</u> качествами ожившего мертвеца	Томита-парсер
AddressExtractor	Офис работает с 14.00 до 17.00 по адресу <u>г.Красноярск ул.Парижской Коммуны,14 оф.14.</u>	pypostal*
DatesExtractor	Я посмотрел на инфляцию в России, взял период с <u>декабря 2002 года</u> по <u>декабрь 2015 года</u> Инфляция 246%.	dateparser
MoneyExtractor	В 1995 году стоимость <u>1 доллара</u> была около <u>800 рублей 50 копеек</u>	

...

Байкеры мотоклуба «Ночные волки» совместно с главой Чечни Рамзаном Кадыровым совершили мотопробег, посвященный дню рождения президента РФ Владимира Путина. Об этом сообщил в «ВКонтакте» сам Кадыров.

В составе "Эдмонта" голы забили Райан Джонс, Том Гилберт, а также Райан Смит, который поразил пустые ворота.

Следом за Свифт идут Кэти Перри (219 миллионов), Селена Гомес (205 миллионов) и Рианна (190 миллионов)

0.78 — f1-мера для имён на первой дорожке factRuEval-2016

0.95+ — у топовых решений

~0.90 — на глаз для русских имён

~1Кб — размер статьи на lenta.ru

~10 статей в секунду —  
производительность NamesExtractor

~20 статей в секунду — под PyPy

~200 статей в секунду — у Томи-  
парсера

Установка

```
pip install natasha
```

```
pip install yargy
```

Документация:

[natasha.readthedocs.io](https://natasha.readthedocs.io)

[yargy.readthedocs.io](https://yargy.readthedocs.io)

Примеры использования:

[github.com/natasha/natasha-examples](https://github.com/natasha/natasha-examples)

[github.com/natasha/yargy-examples](https://github.com/natasha/yargy-examples)