

Работа со СПИСКОМ СОТРУДНИКОВ

Команда проекта:
Камаев Кирилл
Боков Федора



Описание проекта





Описание проекта

1

Проверка на
валидность

2

Исправление
опечаток при
возможности

3

Генерация адреса
почты и пароли

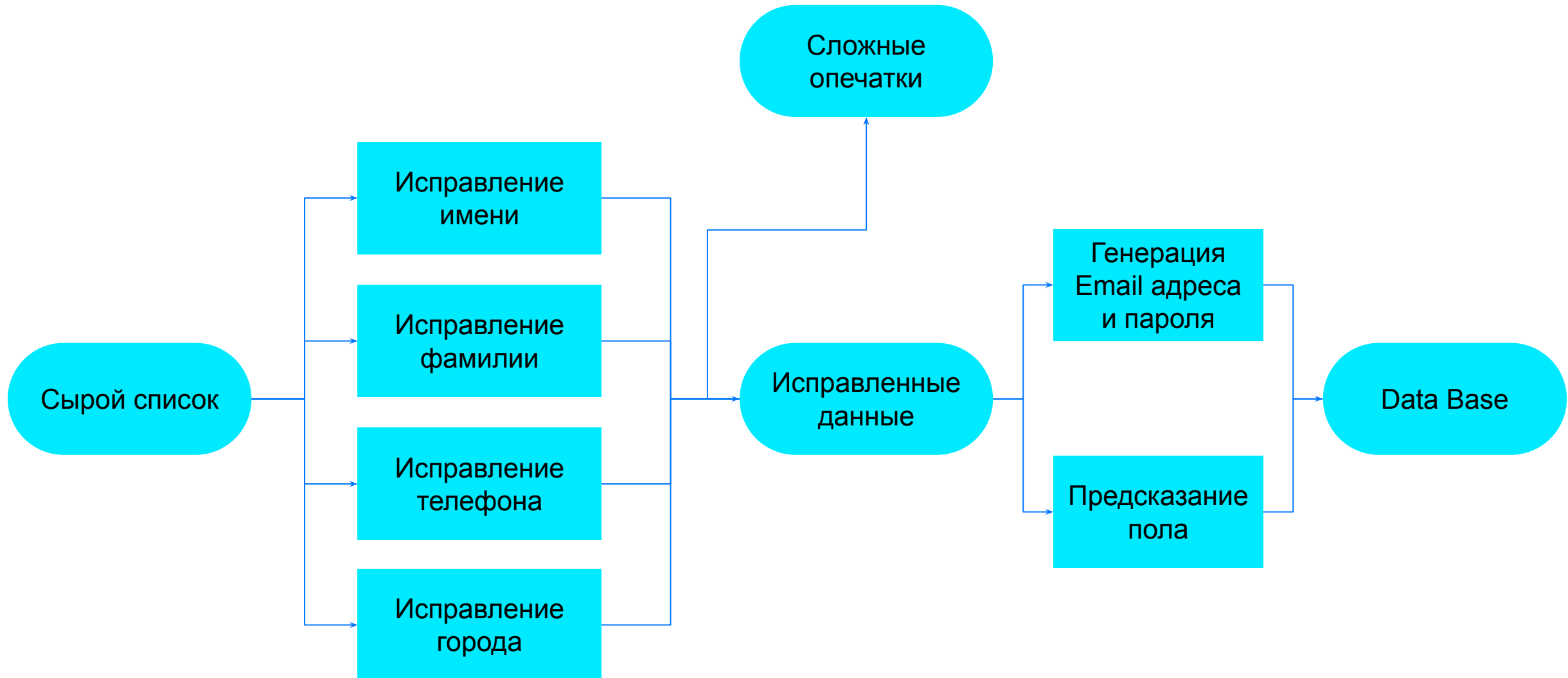
4

Определение пола по
имени и фамилии

Инструменты для реализации проекта

- Python
- PyEnchant
- Difflib
- CatBoost

Пайплайн работы программы



Данные на входе

- их мало
- слова не имеют контекста
- фамилии могут быть очень разнообразными
- нет разметки

4	Ekaterina	Ilyina	nan	St.Petersburg
5	Anastasiya	Grigoryan	1928421	Ekaterinburg
6	Andrey	Fedorov	85212384	Minsk
7	Alexey	Lisitsyn	1239532	Tver
8	Dariya	Abramova	7163908	Moscow
9	Alexandr	Evdokimov	482	Volokolamsk
10	Nataliya	Kostina	9031433	Moscow
11	Nikolay	Ermolin	8539233	St.Petersburg
12	nan	nan	nan	nan
13	Vladimir	Solovovo	4758395	St.Petersburg
14	Vladimir	Ivanov	4827594	Novosibirsk
15	Sergey	Nikolaev	1294375	Sarov
16	Ivan	Ivanov	8532354	Kazan
17	Konstantin	Semenov	8532286	Moscow
18	Grigoriy	Smirnov	3249235	Kaliningrad
19	Vasiliy	nan	7123465	Moscow
20	Alexandr	Lashko	9548324	St.Petersburg
21	sdfsdf	dwfef	9994532	Rostov





Собранные данные



транслитерация



	name	gender
0	Andrej	1.0
1	Dmitrij	1.0
2	Vladimir	1.0
3	Sergej	1.0
4	Jurij	1.0
...
2103290	Aleksandr	1.0
2103291	Aleksandr	1.0
2103292	Echavarrija	1.0
2103293	Kulikova	0.0

№ ↕	Герб	Город ↕
605		Москва
827		Санкт-Петербург
673		Новосибирск
276		Екатеринбург

транслитерация



Mosalsk
Moscow
Mozdok
Mozhaysk
Mozhga
Mtsensk
Murashi
Muravlenko
Murino
Murmansk

Исправление слова

1. удаление лишних символов
2. считается схожесть с каждым словом в словаре
 - a. если лучшая схожесть ниже заданного порога, то слово предлагается для ручного исправления
 - b. иначе заменяется на наиболее подходящее

```
suggestions = set(dictionary.suggest(woi))

for word in suggestions:
    measure = difflib.SequenceMatcher(None, woi, word).ratio()
    sim[measure] = word

if not suggestions or woi == 'Nan':
    return 1, init_word

best_sim = max(sim.keys())

if best_sim < tresh_hold:
    return 1, init_word
else:
    return 0, sim[best_sim].capitalize()
```


Пример работы

```
Enter name: Keoill
Enter last_name: Kamaev
Enter phone: 1234567
Enter city: Moskwa
Created user:
```

	EMAIL	NAME	LAST_NAME	TEL	CITY	PASSWORD
0	K.Kamaev@companyname.com	Kirill	Kamaev	1234567	Moscow	7Mc_bxfn

Определение
пола по имени
и фамилии

Определение пола по имени и фамилии



feature engineering

- имя и фамилия
- последние две буквы фамилии + флажки на ОВ, ИН, АН
- последняя буква имени
- количество букв в имени
- количество букв в фамилии

	0	1	2	3	4	5	6	7	target
0	Ababij	Andrej	ij	ej	0	6	6	0	1
1	Ababilov	Dmitrij	ov	ij	1	8	7	0	1
2	Ababilov	Vladimir	ov	ir	1	8	8	0	1
3	Ababilov	Sergej	ov	ej	1	8	6	0	1
4	Ababilov	Jurij	ov	ij	1	8	5	0	1
...
4218084	Ksenda	Grigorij	da	ij	0	6	8	0	1
4218085	Mescherjakova	Alena	va	na	0	13	5	1	0
4218086	Shestoperov	Ivan	ov	an	1	11	4	0	1
4218087	Semenova	Efrosin'ja	va	ja	0	8	10	1	0
4218088	Semenova	Efrosinija	va	ja	0	8	10	1	0

Метрики

Классификатор	Accuracy	ROC-AUC
Catboost	0.997	0.998

Команда проекта



Боков Федор



Камаев Кирилл



Спасибо
за внимание!