Тегирование компонентов адресов

Дорофей «Котяра» Пролесковский

Москва, 2014

Почему?

https://github.com/nextgis/skoroproverka

- Список адресов
- Структурированные
- Много
- Реальные

Что?

```
103287, г.Москва, Петровско-Разумовский Старый пр-д, д.1/23
103287 г. Москва, пр. Петровско-Разумовский Стар. Д.1/23
103287, Москва г, Петровско-Разумовский Старый пр-д, д.1/23, д.1/23, стр.1 Б
103287, Москва, Петровско-Разумовский Проезд, 28
10328, г.Москва, ул. Башиловская, 25, 8
103309, Москва, улица Маршала Рокоссовского бульвар, д. 8, корпус4
103375, г.Москв, ул.тверская, 7, Краснодарский край, г. Сочи, Курортный проспек
103460, Москва, Зеленоград, 1213, 135
103473, г. Москва, 1-й Самотечный пер., д. 18/1
103473 г.Москва, 2-ой Волконский пер., д.1/10, стр.7, комн.правления
103473, г. Москва, Бабушкинский, пл.Суворовская, д. 1
103473, г. Москва, пл. Суворовская, д. 1
103473, г. Москва, Суворовская пл., д. 1
103473, г.Москва, Суворовская пл., д.1
103473, г.Москва, Суворовская пл., д. 1
103473, г. Москва, Суворовская площадь, д. 1
103473 г.Москва, ул.Самотечная, д.13, стр.1
103473, Москва, 1-й Волконский переулок, д. 9
103473, Москва, 4-й Самотечный переулок, д. 9
103473, Москва г, г. Москва, пл. Суворовская, д. 1
```

Популярные значения

```
 129 ул. Перова, 2А
 131 0
 132 ХМАО-Югра, г.Нижневартовск, Нижневартовский район
 150 Смоленская область
 153 г. Астрахань
 182 Забайкальский край, г. Чита
 188 г. Горно-Алтайск
 200 г. Саратов
 259 Российская Федерация
 289 РБ
 412 59
 442 в соответствии с фактическими адресами
1526 В соответствии с фактическим адресом
1529 Пермский край
3577 в соответствии с фактическим адресом
4360 нет
9688 В СООТВЕТСТВИИ С ФАКТИЧЕСКИМ АДРЕСОМ
```

Подход

- протегировать части без базы
- растаскать теги на соседние слова
- привести каждый элемент в норму
- посмотреть в базе

```
elif token in ('строение', 'сооружение'):
    tags.add('housenumber')
    tags.add('marker')
    tags.add('typo')
elif token in ('тц', 'центр', 'рынок', 'комплекс', 'аэропорт', 'завод',
'"', '«', '»', 'дворец'):
    tags.add('housename')
    tags.add('marker')
elif token in ('новосибирск','новокузнецк','новгород', 'зеленоград',
'калининград', 'псков', 'ярославль', 'луки'):
    tags.add('place')
    tags.add('city')
elif token in ('петербург','москва','спб', 'санкт-петербург', 'мо'):
    tags.add('region')
elif token in ('область', 'обл', 'край', 'республика', 'респ', 'округ',
'области', 'облать', 'облась', 'ао', 'о'):
    tags.add('region')
    tags.add('marker')
elif token in ('хмао','югра','янао', 'хмао-югра', 'округ-югра', 'рб',
'коми', 'рк', 'ненецкий'):
    tags.add('region')
elif token in ('мкр','микрорайон','квартал', 'мкр-н', 'мкрн', 'м-н'):
    tags.add('quarter')
    tags.add('marker')
elif token in ('этаж','эт'):
    tags.add('level')
    tags.add('marker')
```

```
if token in ('г', 'пгт', 'дер', 'п', 'с', 'пос', 'с/п', 'ст-ца', 'рп',
'пст', 'ст', 'снт', 'гп', 'сп', 'гор'):
    tags.add('shortened')
    tags.add('place')
    tags.add('garbage')
    tags.add('marker')
elif token in ('город', 'посёлок', 'поселок', 'деревня', 'село',
'городок', 'поселение', 'хутор', 'станица', 'товарищества'):
    tags.add('place')
    tags.add('garbage')
    tags.add('marker')
elif token in ('ул', 'пр', 'ш', 'пл', 'пер', 'просп', 'наб', 'б-р', 'пр-
т', 'пр-кт', 'пр-д', 'бульв', 'бул', 'туп', 'пр-зд'):
    tags.add('street')
    tags.add('shortened')
    tags.add('marker')
elif token in ('мкад', 'неглинная'):
    tags.add('street')
    tags.add('shortened')
elif token in ('улица', 'бульвар', 'площадь', 'проезд', 'переулок',
'шоссе', 'проспект', 'тракт', 'тупик', 'аллея', 'дорога', 'набережная',
'линия', 'вал'):
    tags.add('street')
    tags.add('marker')
elif token in ('россия', 'беларусь', 'федерация', 'рф'):
    tags.add('country')
```

105064, г.Москва, ул. Старая Басманная, д.14/2, стр.4

postcode 105064

region москва

street Старая Басманная улица

housenumber 14/2 стр 4

105064 г. Москва, ул. Новая Басманная, д. 4/6, подъезд 14

postcode 105064

region москва

street Новая Басманная улица

housenumber 4/6

door подъезд 14

105064(nopoison,digit,first-digit,postcode) , (separator,garbage) **г**(marker,shortened,place,garbage) . (shortened-left,garbage) **москва**(region)

,(separator,garbage) **ул**(marker,shortened,street) .(shortened-left,garbage) **старая**(street,*guessed*) **басманная**(street,*guessed*) ,(separator,garbage)

д(marker,shortened,housenumber,garbage) .(shortened-left,garbage) **14/2**(unsure,*guessed*,housenumber,first-digit) , (separator,garbage) **стр**(marker,shortened,housenumber) . (shortened-left,garbage) **4**(unsure,digit,*guessed*,housenumber,first-digit)

TODO

- найти ещё данных
- нормализация полей
  - streetmangler
  - смотреть в базе
  - ...идеи?

- проверить наличие адресов в OSM
- добавить в OSM индексы

Где брать?

- https://github.com/Komzpa/addresstagger
- https://github.com/nextgis/skoroproverka

me@komzpa.net