

Анализ адресов объектов залога

Толюпа Евгений

Ноябрь'22



Проект ПЕРЕЗАПУСК

О себе

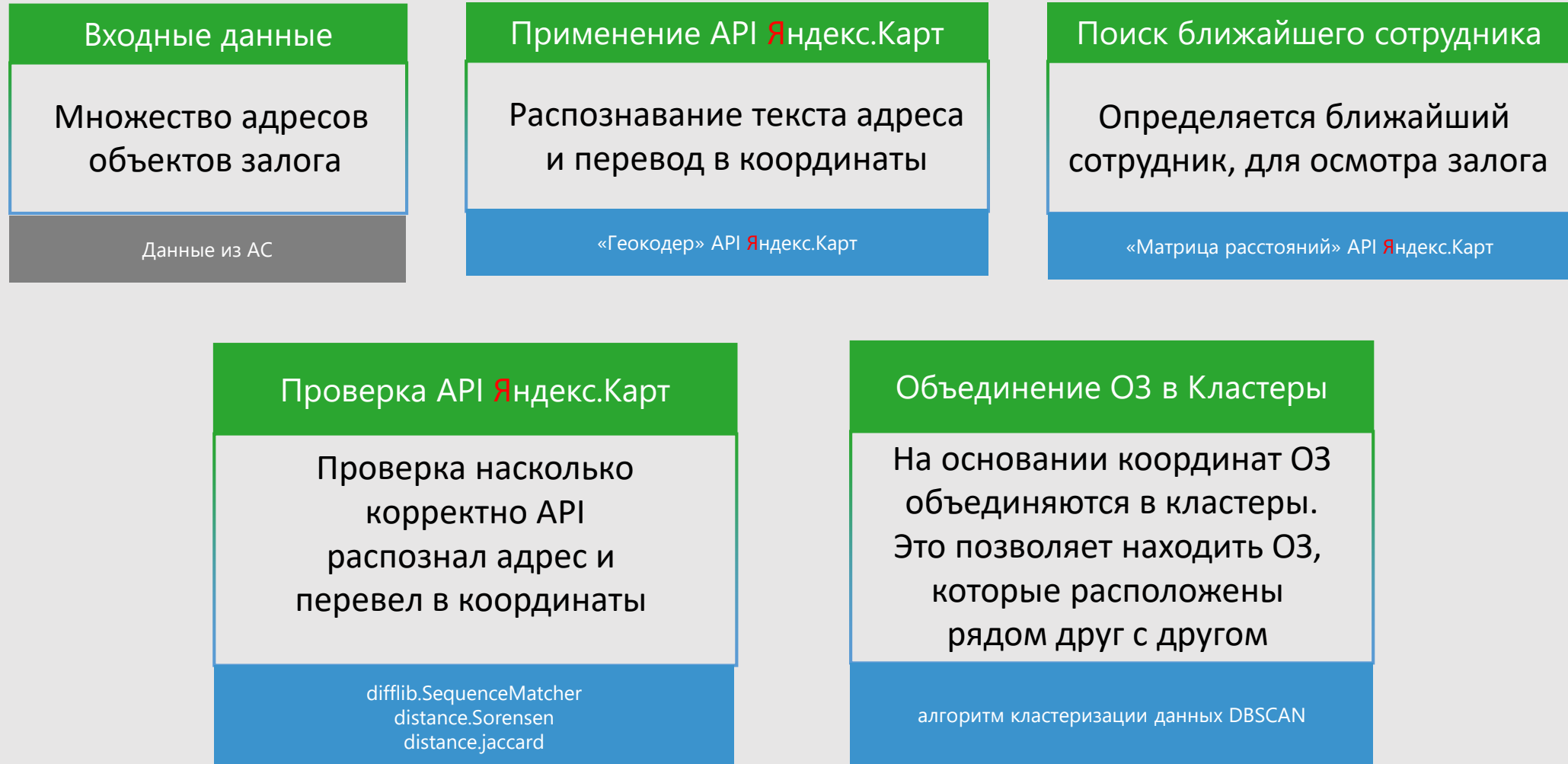
- *Толюпа Евгений Алексеевич*
- *ЯрГУ им. П.Г. Демидова, математический факультет*
- *Ведущий специалист, ЦУПФиП Кредитный мониторинг*
- *Город Ярославль, переезд затруднен*
- *Контакты: +7-915-981-6929*

Описание проекта

- На основании адресов расположения объектов залога (ОЗ) построен алгоритм объединяющий ОЗ в кластеры с расчетом расстояний;
- Использовался API Яндекс.Карт, метод DBSCAN, модули *difflib* и *distance*;
- Ссылка на [github](https://github.com/wwwZeus/python-sber-2022/tree/main/Final%20Project).

<https://github.com/wwwZeus/python-sber-2022/tree/main/Final%20Project>

Бизнес-логика



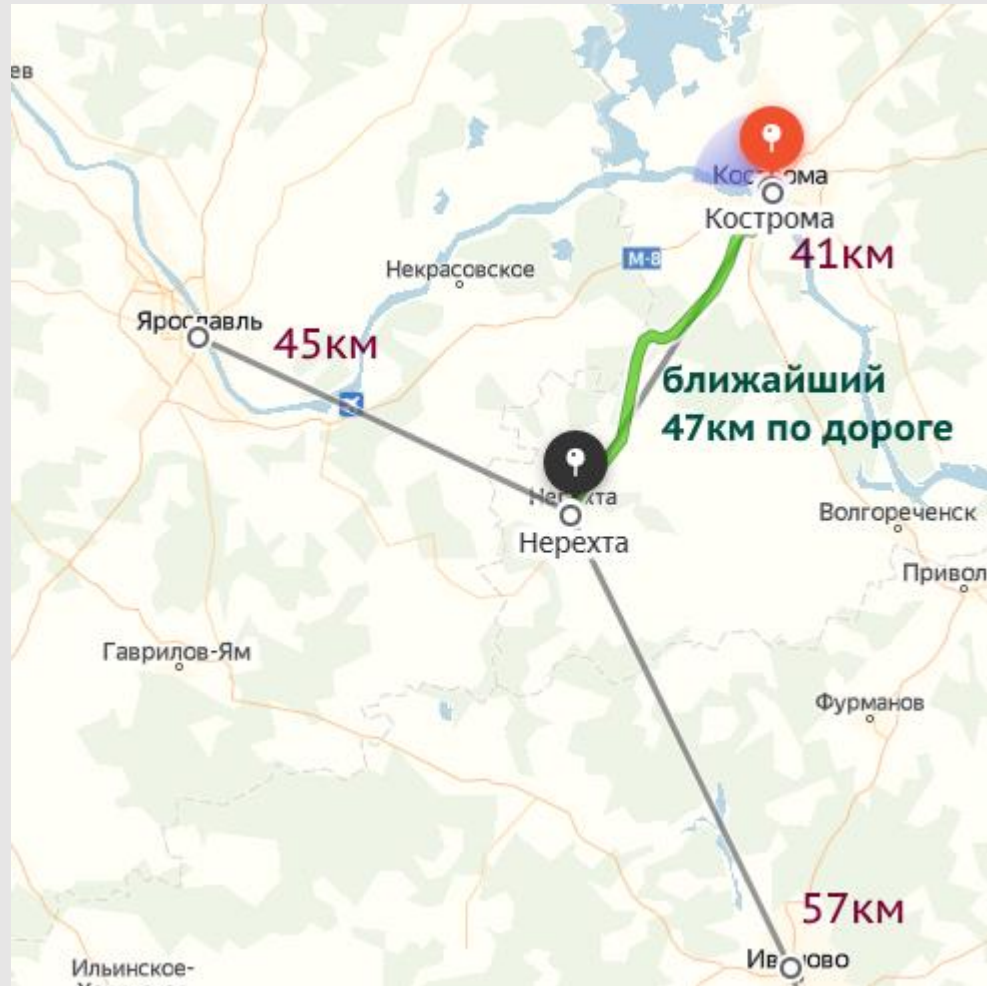
Модель данных

Адрес объекта	ID Объекта
Нерехта, Климушинская, 2а	be691f15c214afe52aa3bcdfe54ee771
г.Москва г. Москва Романов пер., д. 4, стр. 2	969f1a35bb86b73c3f6d6f3e4a6da642
Ивановская область Шуйский р-н, д.Остапово, дом 6	de49dcbdc2ec1b24510e15731934e428

Результат API Яндекс.Карт

ID адреса	be691f15c214afe52aa3bcdfe54ee771
Исходный запрос	Нерехта, Климушинская, 2а
Субъект РФ	Костромская область
Адрес объекта Яндекса	Россия, Костромская область, Нерехта, Климушинская улица, 2А
Точность поиска	exact
Что найдено?	house
Координаты	40.552827 57.465109
Расстояние до объекта, метры	46,1
Ближайший адрес сотрудника банка	г. Кострома, ул. Никитская, 33
Номер кластера	-1
Кол-во Адресов в кластере	28
difflib	0,588
sorensen	0,895
jaccard	0,810

API Яндекс.Карт Геокодер и Матрица Расстояний



Из множества сотрудников ПМЗ
найти ближайшего для выбранного
объекта залога

Адрес объекта залога	Костромская область, г. Нерехта
Адрес сотрудника ПМЗ	Ярославль, ул. Советская д. 34
Иной ближайший адрес сотрудника ПМЗ	Кострома, 47 км
Иное время в пути	45 минут
Ярославль	45 км
Иваново	57 км
Кострома	41 км

Результат: Ярославль не является ближайшим, г. Кострома ближе на 23 км

Объединение в кластеры

кластеры выделены **красным**



Для кластеризации была выбрана модель DBSCAN.

Это алгоритм кластеризации, основанный на плотности — если дан набор точек (координат) в некотором пространстве (двумерное пространство, оси широта и долгота), алгоритм группирует вместе точки, которые расположены друг от друга на заданном расстоянии.

Множество признается кластером, если в нем 2 и более объектов.

В отличие от метода **k-means** **DBSCAN** не задается количество кластеров

Цифры рядом с кластером «58, (42)» Рыбинск:

Номер кластера «58» для поиска в Excel;

Количество уникальных адресов ОЗ «(42)»



-- расположение сотрудника ПМЗ



-- ОЗ, которые не вошли в кластер

Проверка API Яндекс.Карт

Идея сравнивать входные данные с результатами API Яндекс.Карт появилась во время обучения.

- Корректный пример распознавания адреса:

Исходный адрес	Адрес объекта Яндекса
Нерехта, Климушинская, 2а	Россия, Костромская область, Нерехта, Климушинская улица, 2А

- Ошибочный пример:

Исходный адрес	Адрес объекта Яндекса
молочно – товарная ферма с. Новые Чебеньки Зианчуринского района Республики Башкортостан	Польша, Подляское воеводство, повят Бельский, сельская гмина Браньск, Ферма

- Видно, что второй пример неверно обработал исходный адрес. Убедиться в этом можно перейдя по [ссылке](#);
- Сравнить исходный адрес и результат работы API возможно с помощью классов `difflib`, `distance.sorensen` и `distance.jaccard`;
- Для ошибочного примера `difflib` возвращает результат близости текстов всего 0,2085, а для корректного примера 0,588;
- Принято решение – если `difflib` возвращает меньше 0,3, то исходный адрес необходимо корректировать.

СПАСИБО ЗА ВНИМАНИЕ!