Кейс «Анализ геоданных»

Команда «Nova solis»

## ок - Проблематика

*(какие из заявленных и выявленных проблем решены, как\за счет какого функционала решены проблемы)*

Строки почтовых адресов используются в аналитических исследованиях. Используемый для обработки адресов Нормализатор не всегда корректно распознает адреса, если они не полные или содержат ошибки.

До 25% адресов из набора данных распознаются не полностью. (вар. плашкой)

Доля нераспознанных адресов может доходить до 25%. (вар 1)

Они не могут быть использованы в гео-аналитических исследованиях. Например, в анализе географического распределении покупок определенного сегмента клиентов.

**Нераспознанные данные = упущенная выгода** (выделить)

## Решение

**Расширено:**

На вход подается json-запрос, обратно получаем ответ json с исправленным адресом.

{ "address": "'1; 115516, Москва, Бехтерева, 2, 1 house РОССИЯ МОСКВА МОСКВА 0 0 0 Улица БЕХТЕРЕВА 2 0 0'" }

1. Подготовка: Построили словарь на основе данных из «good.csv», по которому будут проверяться слова на опечатки.
2. ненормализованную строку адреса мы преобразуем в вектор с помощью метода TF-IDF, оценивая важность слова в контексте документа
3. Вар. 2. Исходную строку разбиваем на слова регулярными выражениями.
4. Цифры оставляем как есть, слова-токены проверяем на опечатки и исправляем их. Используемый инструмент: нейронная сеть BERT.
5. Формируем исправленные данные в требуемом формате с использованием ленивого алгоритма или бустинга.

Подается на вход:

20000003;117105, г. Москва, шоссе Варшавское, д. 1 стр. 1-2 ком. 3Б;

На выходе

20000003;117105, г. Москва, шоссе Варшавское, д. 1 стр. 1-2 ком. 3Б;house;РОССИЯ;МОСКВА;МОСКВА;0;0;0;Шоссе;ВАРШАВСКОЕ

**Кратко:**

На входе:

{ "address": "117105, г. Москва, шоссе Варшавское, д. 1 стр. 1-2 ком. 3Б" }

1. Построение словаря
2. Разбиение исходной строки на вектора с помощью метода TF-IDF
3. Проверка и исправление слов-токенов на опечатки. С использованием нейронной сети BERT
4. Формирование исправленных данных. С использованием ленивого алгоритма или бустинга.

На выходе:

{ "address": "117105, г. Москва, шоссе Варшавское, д. 1 стр. 1-2 ком. 3Б;house;РОССИЯ;МОСКВА;МОСКВА;0;0;0;Шоссе;ВАРШАВСКОЕ”}

## Наглядная демонстрация работы решения или частично реализованного решения

Видео

## Информация о конкурентных или существующих решениях, преимущества решения команды (доп. необяз. блок)

## Возможность интеграции в существующие решения (доп. необяз. блок)

## ок - Экономический эффект

*(влияет ли решение на уменьшение организационных\ операционных\человеческих\ресурсных затрат компании\организации)*

Повышение распознавания адресов, используемых в исследованиях, позволит более точно вести региональную маркетинговую программу *– настраивать рекламные кампании и делать предложение только для данного географического расположения*, оптимизировать доставку заказов.

В маркетинговых исследованиях: более точная региональная маркетинговая программа *– настройка рекламной кампании и предложения только для данного географического расположения*.

Оптимизация доставки заказов: уменьшение случаев доставки отправлений и приезда курьеров/экспедиторов по неверному адресу.

Это будет выгодно также для различных крупных бизнесов, использующих почтовые адреса. Например:

Оптимизация доставки заказов: уменьшение случаев доставки отправлений и приезда курьеров/экспедиторов по неверному адресу.

* позволит снизить нагрузку с операторов, которым необходимо уточнять почтовые адреса клиентов,
* Уменьшит случаи доставки отправлений и приезда курьеров/экспедиторов по неверному адресу.

## Информация о реализации решения

(Сроки, стоимость, порядок внедрения)

Решение представляет из себя веб-сервис, взаимодействие с которым происходит через REST API. Достаточно послать POST-запрос с данными в формате JSON и получить обратно исправленные адреса.

## Масштабируемость решения

*(В иные процессы\увеличение нагрузки\и др.)*

- Если писать про нагрузку, то нужны данные по нагрузке (для текущего словаря 25000 слов)

## ок - Команда

*(кто принимал участие в разработке решения: профайл и контактные данные)*

**Интересный факт о команде (?):**

**Перелыгина Вера Николаевна –** дизайнер

Петропавловск-Камчатский

vera.kamunicorn@gmail.com

+7 914 620-08-94

**Бочаров Алексей Михайлович** – Backend

Воронеж

alex.bo2018vrn@gmail.com

+7 929 007-49-64

**Белоушкин Александр Николаевич** – Team Leader, Backend

Воронеж

aleks.beloushkin@yahoo.com

+7 (900) 300-13-31

## Описание процесса (Frontend)?

В презентации не будет такого слайда

Вариант 1:

1. В поле ввода вписываем адрес, подсказка на исправление?

С описания кейса (к проблематике):

Из-за того, что Нормализатор почтовых адресов не обрабатывает корректно некоторые плохо структурированные адреса, система геокодирования Банка имеет существенную долю адресов (до 25%) без географических координат

Поэтому эти данные не могут быть использованы в гео-аналитических исследованиях. То есть теми, которые проводятся при помощи ГИС и прочих аналитических систем. Например исследование для малого и среднего бизнеса о гео-распределении покупок определенного сегмента клиентов