Кейс «Анализ геоданных»

Команда «Nova solis»

## 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

1. 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. Используемый инструмент: ленивый алгоритм или бустинг.

## 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)

## Экономический эффект

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

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.

## Информация о реализации решения

(Сроки, стоимость, порядок внедрения)

## Масштабируемость решения

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

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

## 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