СЕССИЯ 1

Исходные файлы:

1) Region.geojson (Исходные данные)
2) Сессия 1.pdf (Инструкция к 1 сессии)

Результаты работы:

1) Data.zip (Предобработанные данные)
2) Report.html + Report.ipynb (Отчет о проделанной работе)
3) Readme.txt (Дополнительные комментарии)

ВВЕДЕНИЕ

На этом чемпионате вам предстоит разработать рекомендательную систему по выявлению опасных дорожных участков и предотвращению будущих дорожно-транспортных происшествий на территории регионов Российской Федерации. Данная система диагностики будет включать исследование имеющихся открытых данных по дорожно-транспортным происшествиям портала https://dtp-stat.ru, приложение-бот для анализа опасности дорожной ситуации по введенным данным окружения и интерфейс для взаимодействия с промоботом — сервисным роботом, оказывающим помощь в сборе данных и информировании посетителей чемпионата (https://promo-bot.ru).

Данные по ДТП доступны для загрузки на странице https://dtp-stat.ru/opendata/ и представлены в формате .geojson по всем регионам. Каждый файл имеет следующую структуру:

```
{
    "id": 384094, # идентификатор
    "tags": ["Дорожно-транспортные происшествия"], # показатели с официального сайта ГИБДД
    "light": "Светлое время суток", # время суток
    "point": {"lat": 50.6039, "long": 36.5578}, # координаты
    "nearby": [ "Нерегулируемый перекрёсток неравнозначных улиц (дорог)", "Жилые дома индивидуальной
застройки"], # координаты
    "region": "Белгород", # город/район
    "address": "г Белгород, ул Сумская, 30", # адрес
    "weather": ["Ясно"], # погода
    "category": "Столкновение", # тип ДТП
    "datetime": "2017-08-05 13:06:00", # дата и время
    "severity": "Легкий", # тяжесть ДТП/вред здоровью
    "vehicles": [ # участники - транспортные средства
      {
        "year": 2010, # год производства транспортного средства
        "brand": "ВАЗ", # марка транспортного средства
        "color": "Иные цвета", # цвет транспортного средства
        "model": "Priora", # модель транспортного средства
        "category": "C-класс (малый средний, компактный) до 4,3 м", # категория транспортного
средства
        "participants": [ # участники внутри транспортных средств
          {
            "role": "Водитель", # роль участника
            "gender": "Женский", # пол участника
            "violations": [], # нарушения правил участником
```

```
            "health_status": "Раненый, находящийся...", # состояние здоровья участника
            "years_of_driving_experience": 11 # стаж вождения участника (только у водителей)
          }
        ]
      },
    ],
    "dead_count": 0, # кол-во погибших в ДТП
    "participants": [], # участники без транспортных средств (описание, как у участников внутри
транспортных средств)
    "injured_count": 2, # кол-во раненых в ДТП
    "parent_region": "Белгородская область", # регион
    "road_conditions": ["Сухое"], # состояние дорожного покрытия
    "participants_count": 3, # кол-во участников ДТП
    "participant_categories": ["Все участники", "Дети"] # категории участников
  }
```

В настоящее время накоплено достаточно много данных по дорожно-транспортным происшествиям, включающих все подробности окружения, которые можно проецировать на другие, аналогичные, условия. В результате анализа таких данных можно выявить дорожные участки и перекрёстки, требующие наибольшего внимания в плане перепроектирования или реорганизации движения на них. Предложенные решения помогут в будущем снизить или вообще предотвратить количество дорожно-транспортных происшествий. Актуальность повышения безопасности дорожного движения заложена в международной программе Vision Zero https://ru.wikipedia.org/wiki/Vision_Zero, которая активно обсуждается с 2010-х годов в России.

В рамках всего конкурсного задания вам потребуется предобработать данные, выполнить анализ данных и выявить ключевые зависимости, построить необходимые модели машинного обучения, разработать приложение-бот и интерфейс взаимодействия с роботом-промоботом для сбора данных и информирования посетителей чемпионата.

На этой сессии необходимо подготовить набор данных и произвести его предобработку для дальнейшего исследования и построения моделей обучения, а также запрограммировать промобот для сбора данных у посетителей чемпионата.

ЗАДАНИЕ

1.1 Парсинг данных
На основании файлов .geojson размещенных на странице https://dtp-stat.ru/opendata/, содержащего данные по дорожно-транспортным происшествиям в каждом регионе, необходимо построить исходный набор данных. Набор данных должен быть преобразован в единый файл формата .csv. Можно дополнить набор какими-либо другими данными, если они могут быть полезны для дальнейшего исследования.

1.2 Предобработка данных и выделение значимых атрибутов

Предобработанный набор данных должен содержать только уникальные случаи дорожно-транспортных происшествий и не содержать в своих атрибутах перечислений. Задача определения опасности дорожного участка или перекрёстка заключается в определении класса (кластера) – уровня опасности. Уровни опасности определяются произвольно таким образом, чтобы в группу наиболее опасных дорожных участков, требующих срочных решений, попали не более 10% от имеющихся в наборе данных по каждому региону. Исходя из этого, необходимо определить, какие атрибуты имеют наибольшее влияние на определение таких классов (кластеров), и оставить только их для последующего обучения. Также необходимо обосновать выбор дополнительных атрибутов и причину исключения каких-либо данных из исходного набора документов.

1.3 Описание структуры набора данных

Для каждого атрибута подготовьте описание, содержащее текстовое представление (расшифровка, перевод, назначение) и статистику распределения данных (плотность, наличие пустых значений).

1.4 Подготовка отчета

Подготовьте отчет о проделанной работе по итогам сессии в котором будут представлены результаты, выводы и обоснования выбора по каждому разделу задания. В отчете также опишите содержимое результирующих файлов архива Data.zip

1.5 Программирование промобота

Используя интерфейс промобота, определите и загрузите в лингвистическую базу ряд вопросов, позволяющих в дальнейшем сформировать набор данных для исследования. Структура собранных в результате последующего взаимодействия промобота с посетителями чемпионата данных должна быть схожа с исходным набором данных