Московский авиационный институт
(национальный исследовательский университет)

Факультет информационных технологий и прикладной математики

Кафедра вычислительной математики и программирования

Лабораторная работа №0 по курсу “Искусственный интеллект”

Студент: Баранов А.А.
Группа: М8О-307Б
Дата:
Оценка:
Подпись:

Москва, 2019

Тема:

Получение и обработка данных.

Задание:

Требуется сформировать/получить два набора данных соответствующие следующим критериям:

1) Один из датасетов должен представлять собой корпус документов. Язык, источник и тематика произвольна

2) Второй датасет должен содержать категориальные, количественные признаки. Для данного датасета определить предсказываемые признаки (для задачи регрессии и классфикации). Если такого признака нет, спроектировать

Данные датасеты будут в дальнейшем использованы в оставшихся лабораторных работах.

По каждому датасету построить распределения признаков (в случае корпуса документов – построить распределение слов) и объяснить имеющуюся картину. Вычислить статистические характеристики признаков. Обнаружить и решить возможные проблемы с данными. Если решить данную проблему невозможно, объяснить почему.

Требования:

- Датасеты должны быть уникальны
- Исходный код должен быть написан в одном код стайле
- Должен быть указан источник.

Файлы проекта:

https://github.com/NoruNoruBim/Python/tree/master/AI_MAI/Lab0

Источники данных:

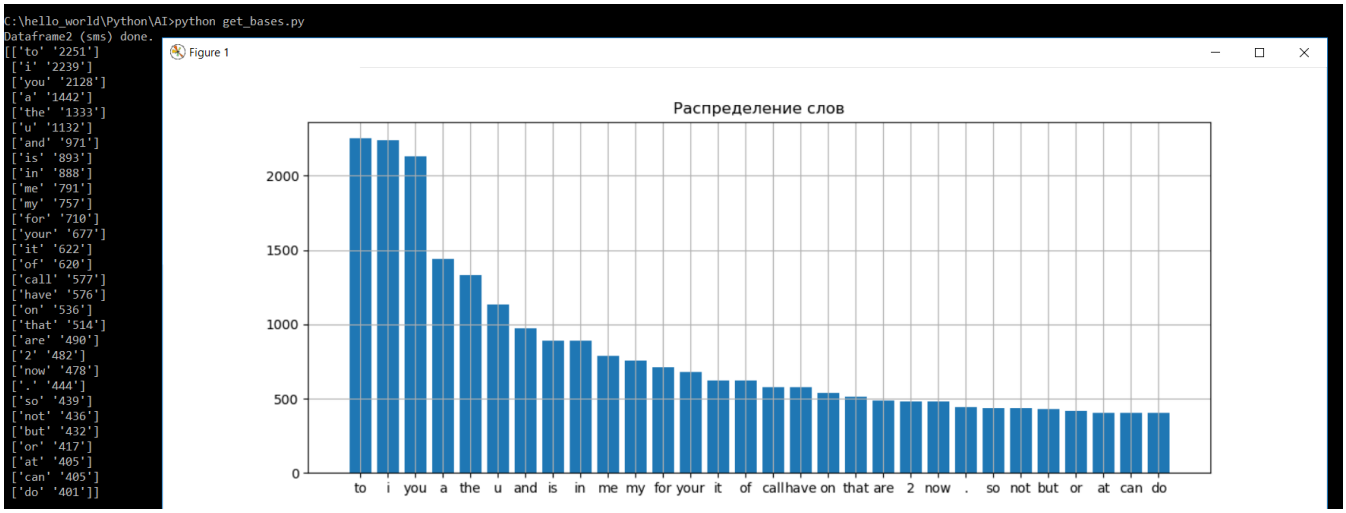
1. Корпус документов - https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection
2. Для регрессии и классификации - https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset#

Описание решения:

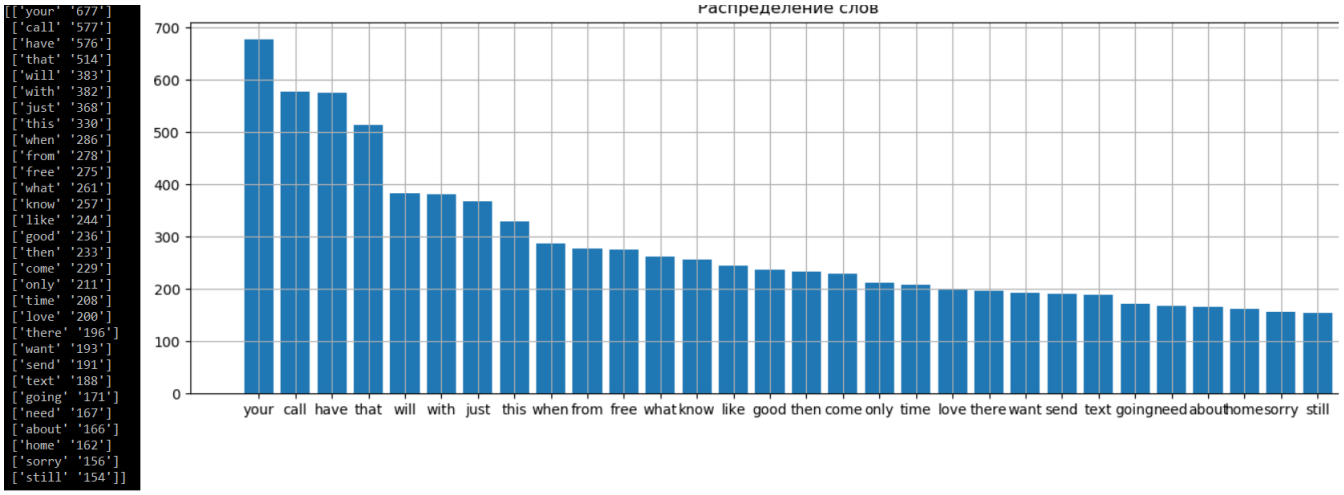
1.Корпус документов

Я взял базу с смс-текстами. Отделил тексты от меток. Далее создал словарь, где ключ – это слово, а значение – это кол-во вхождений этого слова в тексты. Далее вывел статистику.

Стоит учитывать специфику текста, так как это не научная статья, то присутствует большое количество сленга и сокращений. Также, как и в любом тексте, самые встречаемые слова – это предлоги, но в случае с смс-текстами, например слово “the”, в диалоге часто опускается, поэтому оно стоит не на первом месте (как обычно должно).



После удаления некоторого вида слов картина немного меняется:



2. Регрессия и классификация

Для классификации и регрессии я взял базу, в которой описывается статистика онлайн покупок.

Всего 12331 экземпляров и 18 признаков, некоторые из которых можно использовать как метки. На картинке ниже показаны первые 5 строк таблицы.

```
>>> df1.head()
                0                        1             2   \
0  Administrative  Administrative_Duration  Informational
1               0                        0             0
2               0                        0             0
3               0                        0             0
4               0                        0             0

                     3              4                        5   \
0  Informational_Duration  ProductRelated  ProductRelated_Duration
1                       0               1                        0
2                       0               2                       64
3                       0               1                        0
4                       0               2              2.666666667

           6           7           8           9      10                11  \
0  BounceRates  ExitRates  PageValues  SpecialDay  Month  OperatingSystems
1          0.2        0.2           0           0    Feb                 1
2            0        0.1           0           0    Feb                 2
3          0.2        0.2           0           0    Feb                 4
4         0.05       0.14           0           0    Feb                 3

       12       13          14                15       16       17
0  Browser  Region  TrafficType       VisitorType  Weekend  Revenue
1        1       1           1  Returning_Visitor    FALSE    FALSE
2        2       1           2  Returning_Visitor    FALSE    FALSE
3        1       9           3  Returning_Visitor    FALSE    FALSE
4        2       2           4  Returning_Visitor    FALSE    FALSE
>>>
```

Датасет содержит 10 числовых и 8 категориальных атрибутов. Они были сформированы по годичным наблюдениям за онлайн покупками пользователей.

"Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" и "Product Related Duration" показывают количество страниц разных типов, посещенных пользователем за одну сессию и то, сколько на них было проведено времени.

"Bounce Rate" (показатель отказов), "Exit Rate" (показатель выходов) и "Page Value" (~вклад страницы) показывают метрики, измеренные с помощью Google Analytics для каждой страницы на сайте.

"Special Day" говорит о близости «особенного», возможно скидочного, дня

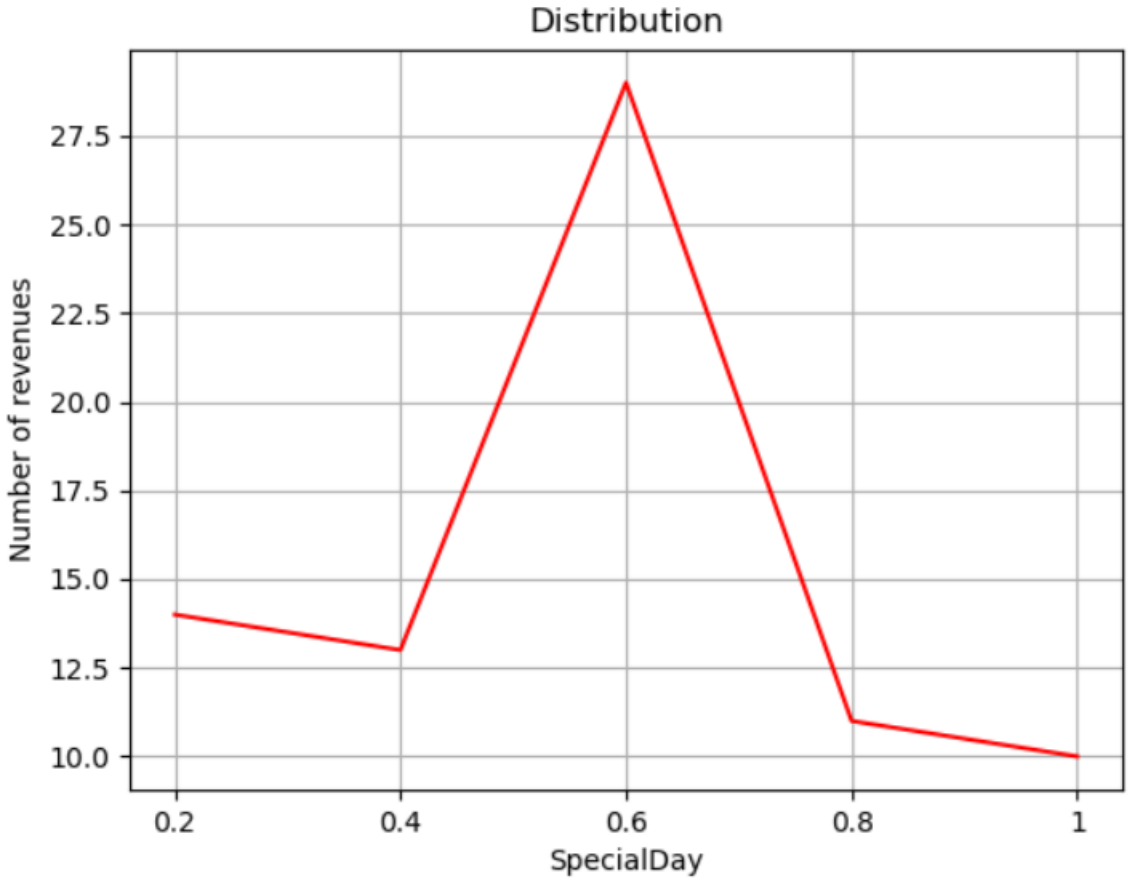
или праздника.

Также в таблице есть информация о браузере, регионе, месяце в которую была сессия и т.д.

3. Распределение признаков.

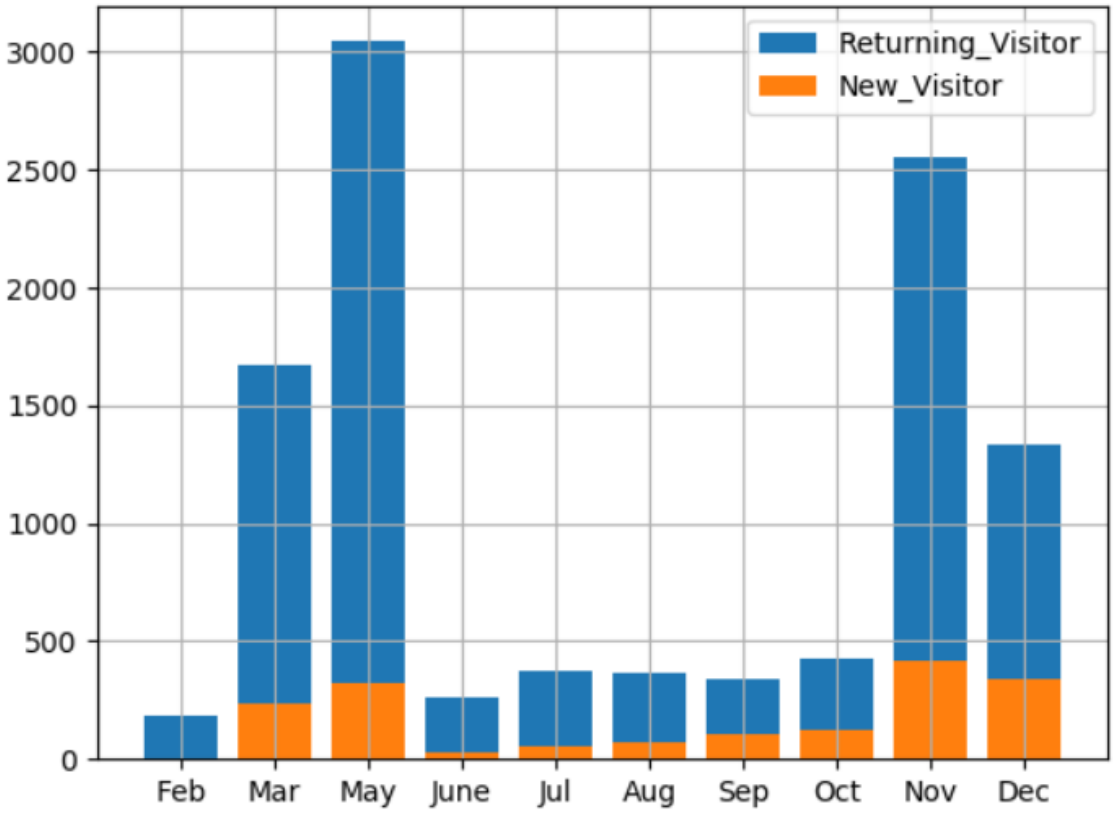
Ниже представлены распределения некоторых признаков, чтобы можно было провести анализ.

Из нижеследующего графика ясно, что люди стараются делать покупки накануне праздников (в график не включен случай когда до праздника очень много времени или он только закончился), а не в самый последний момент.



0 – вне праздничных дней, 1 – праздничные дни.

На этом рисунке показано соотношение новых и старых покупателей по месяцам. Отсюда можно сделать вывод, что время года влияет на количество как новых покупателей, так и покупателей вообще, так как весной и в конце осени / начале зимы покупателей значительно больше чем в остальные времена года, что может быть вызвано спецификой предоставляемых товаров, колебанием цен, а также погодными условиями и т.д.



Также в этом датасете есть несколько параметров, которые не пригодятся (такие как тип браузера) и я думаю их удалить.

Выводы:

Благодаря этой лабораторной я освежил знания в работе с pandas и matplotlib, что очень полезно.

Также я поработал с датасетом и сделал визуализацию ряда параметров, провел анализ. Ранее я так делал нечасто, но теперь, думаю, буду чаще визуализировать данные, так как это помогает упростить понимание данных, хранящихся в таблице и раскрыть картину во всей красе.