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: А. В. Скворцов
Группа: М8О-308Б

Москва, 2019

Условие

Требуется сформировать/получить два набора данных соответствующие следующим критериям:

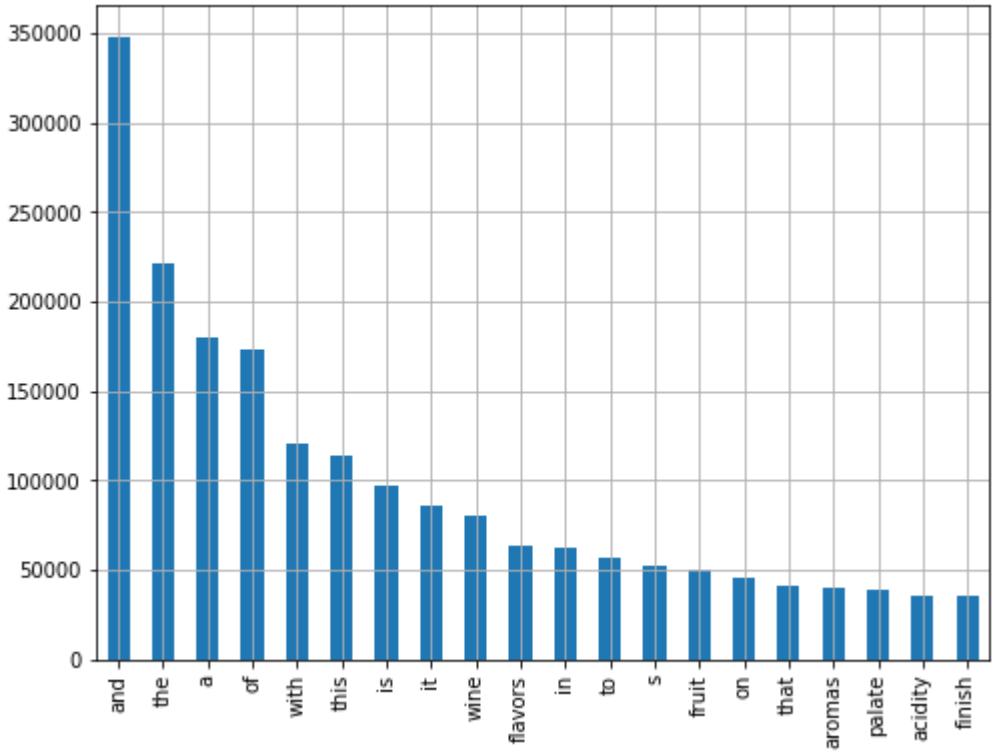
1. Один из датасетов должен представлять собой корпус документов. Язык, источник и тематика произвольна
2. Второй датасет должен содержать категориальные, количественные признаки. Для данного датасета определить предсказываемые признаки (для задачи регрессии и классфикации). Если такого признака нет, спроектировать

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.

Метод решения

Корпус документов

В качестве корпуса документов взяты обзоры вина (130 тыс.) с различными оценками и характеристиками: `https://www.kaggle.com/zynicide/wine-reviews`. Обзоры были разбиты на слова, чтобы получить их распределение. Получилась следующия картина:

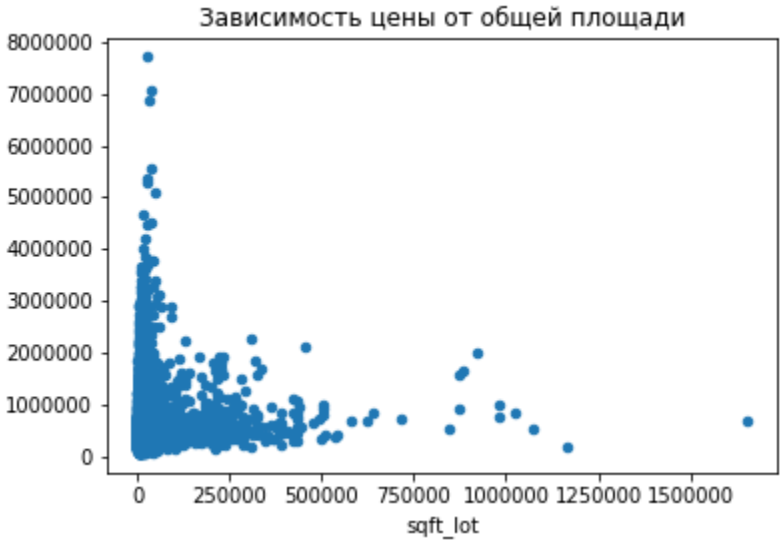


Такое распределение достаточно логично, большая часть слов в топе — союзы, предлоги и местоимения, что свойственно естественному языку. Другие слова, такие как wine, flovors, aromas, palate, finish и acidity, соответствуют теме документа.

Датасет с категориальными, количественными признаками

Для выявления категориальных признаков использован датасет с проданными домами и их характеристиками: `https://www.kaggle.com/harlfoxem/housesalesprediction`. Проведем его небольшое исследование:

То есть в целом рост жилой площади сопровождается ростом цены, что выглядит весьма логично.



Кажется, рост общей площади сказывается отрицательно на цене дома.

Похоже, что возраст дома не сильно влияет на цену.
Рассмотрим статистические характеристики цены:

1. Средняя цена дома: 540088.14
2. Среднее квадратическое отклонение: 367127.19
3. Максимальная цена за дом: 7700000.0
4. Минимальная цена за дом: 75000.0

Выводы

В данной лабораторной работе я ознакомился с основными определениями машинного обучения и научился пользоваться pandas - библиотекой для обработки и анализа данных.