Отчет по лабораторной работе №0
по курсу «Искусственный интеллект»

Выполнил студент группы М8О-308б-16 Никитин Андрей

Тема:
Получение и предобработка данных

Задача:
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 классификации). 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.

Оборудование студента:
Ноутбук Lenovo ThinkPad 13, процессор Intel® Core™ i5-7200U CPU 1.70 GHz, память 8ГБ, 64-разрядная система.

Программное обеспечение:
ОС Linux Mint 19, Python 3.6.7 (С библиотеками Pandas, Numpy и Scikit-Learn), Jupyter notebook 4.4.0

Ход работы:

Выбранные датасеты:

1. Датасет с характеристиками бриллиантов и ценами на них: https://www.kaggle.com/shivam2503/diamonds
2. Датасет документов по категориям «The 20 Newsgroups» http://qwone.com/~jason/20Newsgroups/

Весь анализ данных с построением статистических оценок, распределений, соответствующих графиков и выводами оформлен с помощью Jupyter Notebook: https://github.com/AndrewNikitin/ML_Labs