Московский авиационный институт

(национальный исследовательский университет)

Факультет информационных технологий и прикладной математики

Кафедра вычислительной математики и программирования

Лабораторная работа №0 по курсу «Искусственный интеллект»

(6-й семестр)

Студент: Сизонов А. А.

Группа: М8О-408Б

Дата:

Оценка:

Подпись:

**Москва, 2019**

1. **Тема:**

Получение и обработка данных.

1. **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.

1. Датасеты должны быть уникальны
2. Исходный код должен быть написан в одном код стайле
3. Должен быть указан источник.
4. Данные

Корпус документов: <http://academictorrents.com/details/cd339bddeae7126bb3b15f3a72c903cb0c401bd1>

Данные для классификации и регрессии:

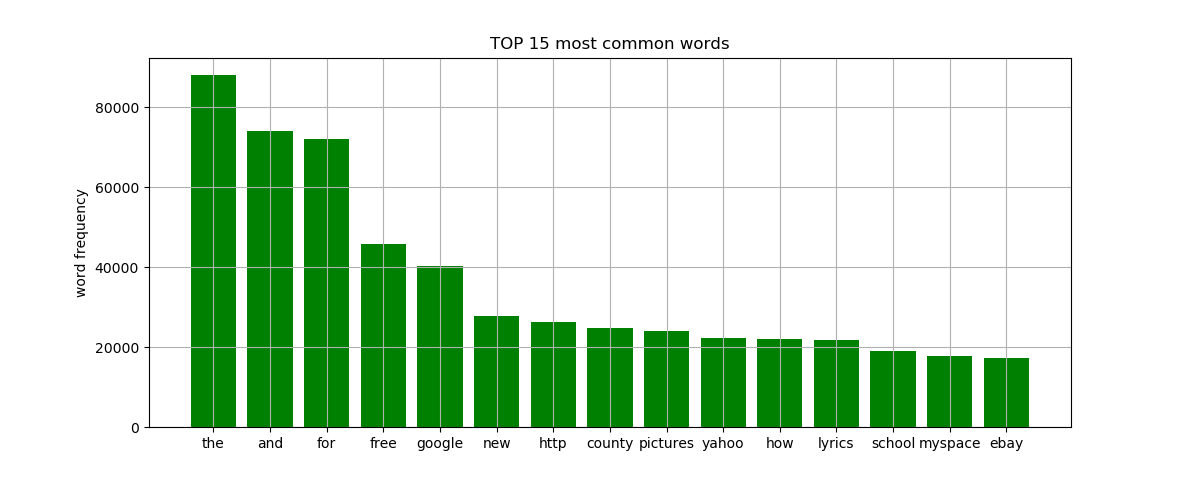
<https://vincentarelbundock.github.io/Rdatasets/csv/Ecdat/Computers.csv>

**Описание решения:**

1. **Корпус документов**

Используется база поисковых запросов AOL за 2006 год. Был произведен подсчёт вхождений каждого из слов.

15 самых встречающихся слов:



Данный график составлен по базе поисковых запросов размером 200 мб. Из графика видно, что люди часто переходили из AOL поисковика в более известные поисковики, такие как google или yahoo. Также видно, что сервисы ebay и myspace были достаточно популярны уже в 2006 году.

1. **База для регрессии и классификации**
2. Описание базы

Используется база компьютеров, которые были в продаже с 1993 года по 1996 год.

В базе около 6300 экземпляров о цене определенного компьютера в определенный месяц. Компьютеры классифицируются по техническим характеристикам и наличии/отсутствии некоторых признаков.

Атрибуты компьютера:

Price – цена

Trend – месяц c 1993 года

Speed – частота процессора

Hd – размер жесткого диска

Ram – объем оперативной памяти

Cd – имеется ли cd привод

Multi – имеется ли медиа система

Premuim – произведен ли комьютер фирмой IBM или COMPAQ

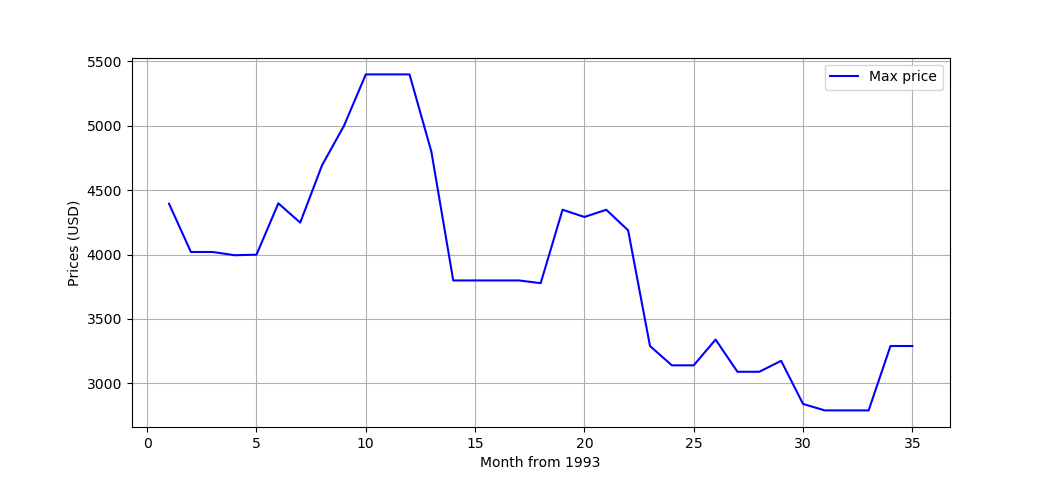
Ads – количество прайс листов за каждый месяц

1. Анализ базы

Первый график:

Для каждого месяца найдем компьютер, имеющий максимальную цену.

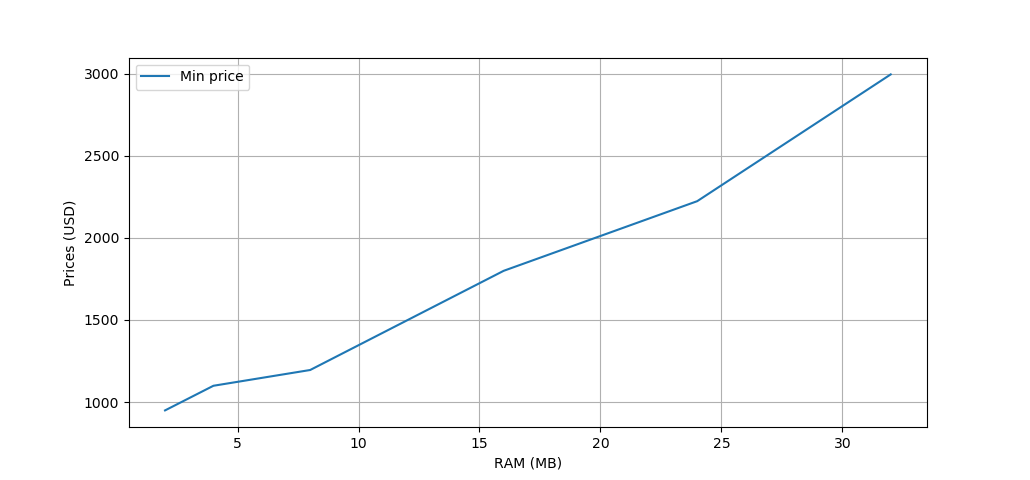
Построив график становится видно, что цены меняются в зависимости от месяца. Но в целом цены компьютеров упали за 3 года.



Второй график:

Для каждого объема оперативной памяти найдем компьютер с минимальной стоимостью.

Построив график становится видно, что чем больше объем оперативной памяти – тем выше цена. Причём на цену не влияют другие параметры компьютера.



1. **Выводы:**

В рамках данной лабораторной работы я получил опыт работы с обработкой данных на python и научился анализировать датасеты с помощью визуализации, используя построение графиков средствами библиотеки matplotlib.