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: В. И. Лобов
Преподаватель: С. Х. Ахмед
Группа: М8О-306Б

Дата:
Оценка:
Подпись:

Москва, 2019

Лабораторная работа №0

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

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:

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

Выбранные датасеты:

1. Корпус документов - статьи BBC по 5 жанрам: business, entertainment, politics, sport, tech

   http://mlg.ucd.ie/datasets/bbc.html

2. Данные котировок акций американских компаний(NYSE, NASDAQ)

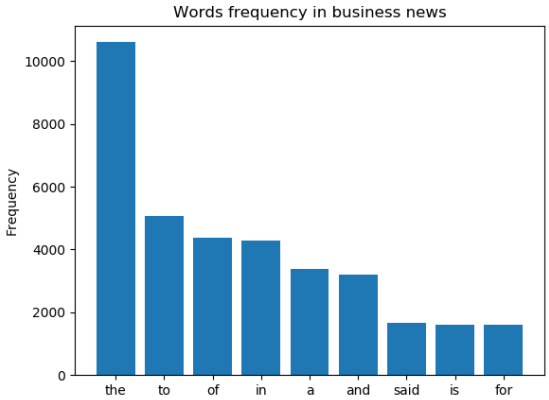
   https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs

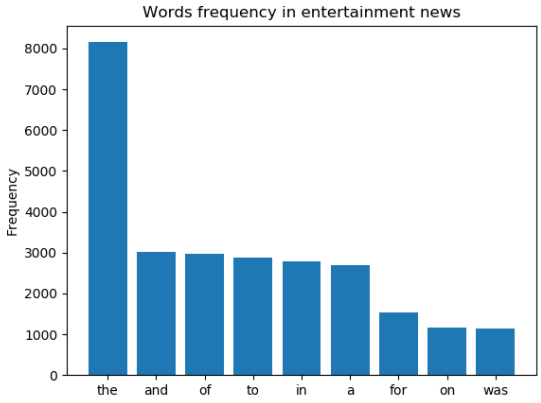
1 Описание

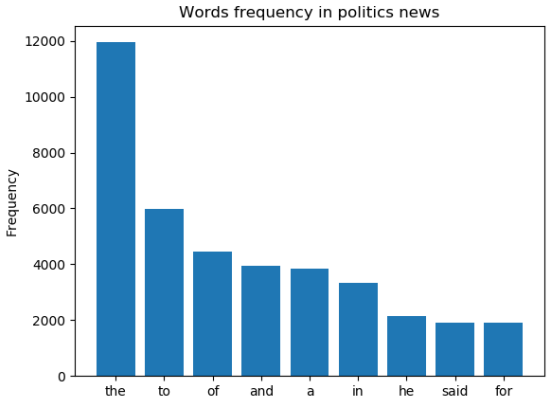
Корпус документов.

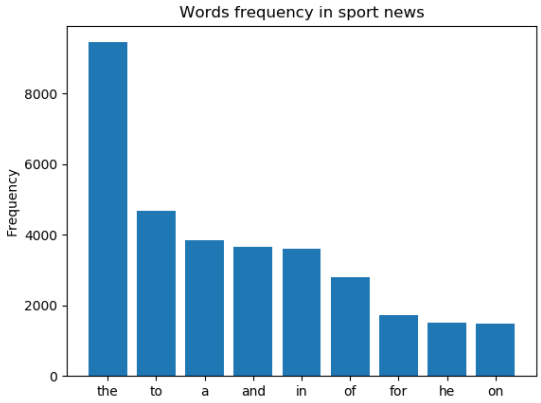
Так как стоит задача построения распределения слов в исходных текстах, необходимо устранить такие проблемы данных, как наличие знаков препинания и заглавных символов - выбросим все вхождения знаков препинания в текст и приведём его к

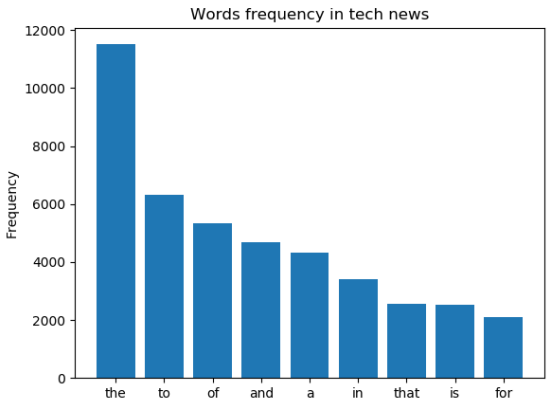
нижнему регистру. Построим распределение слов в полученном обработанном тексте для каждого жанра, а затем выберем 10 наиболее встречающихся слов - получим следующую картину:











Проанализировав полученные результаты, можно сделать вывод, что вне зависимости от жанра текста наиболее популярными словами будут короткие (3-4 буквы)

слова-связки, а также слова, обозначающие косвенную/прямую прямую речь (said), что свойственно для текстов типа новостей.

Листинг 1: src/bbc.py

```
import matplotlib.pyplot as plt
from os import listdir
import numpy as np

curDir = "C:\\Users\\pkmixer\\Downloads\\AI\\bbc\\"

for directory in listdir(curDir):
    dictionary = {}
    frequency = []
    for file in listdir(curDir + directory):
        with open(curDir + directory + "\\" + file, "r") as f:
            for line in f.readlines():
                line = line.lower()

                chars = [".", ",", "\'"]
                for char in chars:
                    line = line.replace(char, "")

                line = line.split()

                for word in line:
                    if word not in dictionary:
                        dictionary[word] = 0
                    else:
                        dictionary[word] += 1

    for word in dictionary:
        frequency.append((word, dictionary[word]))

    frequency.sort(key=lambda x: x[1])

    X = []
    Y = []
    for word in frequency[:-10:-1]:
        X.append(word[0])
        Y.append(word[1])

    y_pos = np.arange(len(X))
    plt.bar(y_pos, Y)
    plt.xticks(y_pos, X)
    plt.ylabel("Frequency")
    plt.title("Words frequency in {} news".format(directory))
    plt.savefig(curDir + "..\\bbc_pics\\" + directory + ".png")
    plt.close()
```

Данные котировок акций.

Исходные данные о котировках акций за некоторый промежуток времени представлены в виде отдельных файлов - каждый файл характеризует конкретную акцию. Можно соединить все акции в один .csv файл, при этом общий размер увеличится. Если сделать предположение, что котировки одной акции не зависят от котировок других(а так в большинстве случаев и есть), то можно рассматривать каждый файл, как отдельный датасет.

При рассмотрении датасета выяснилось, что некоторые файлы пусты, эти акции необходимо удалить, так как они не несут никакой смысловой нагрузки.

Листинг 2: src/a.py

```
import os
curDir = "C:\\Users\\pkmixer\\Downloads\\AI\\Stocks\\"

for fileName in os.listdir(curDir):
    if os.stat(curDir + fileName).st_size is 0:
        print(fileName)
        os.remove(curDir + fileName)
```

Каждая акция содержит 7 признаков:

1. Date - день котировки
2. Open - цена на открытии биржи
3. High - наибольшая цена за день
4. Low - наименьшая цена за день
5. Close - цена на закрытии биржи
6. Volume - суммарное количество операций за день
7. OpenInt(Open Interest) - сумма открытых позиций

В случае акций, величина Open Interest всегда равна 0(имеет ненулевое значение для фьючерсов и других срочных контрактов), поэтому её стоит исключить из обработки данных. Volume - не слишком информативный признак, так как на предсказание цены акции влияет её история и экономическая ситуация, но никак не количество торгуемых акций, поэтому её так же можно исключить.

Построим распределение признаков акций, найдём среднее, дисперсию, медиану и некоторые квантили и изобразим наиболее "интересные"из них.

Листинг 3: src/statistics.py

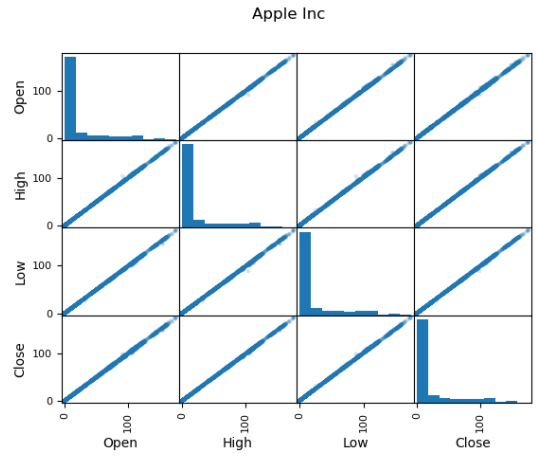
```
import matplotlib.pyplot as plt
import pandas as pd
from pandas.plotting import scatter_matrix
import os


curDir = "C:\\Users\\pkmixer\\Downloads\\AI\\Stocks\\"

names = {}
indicesDir = "C:\\Users\\pkmixer\\Downloads\\indices\\"

for fileName in os.listdir(indicesDir):
    with open(indicesDir + fileName, "r") as indiceName:
        indiceName.readline()
        for line in indiceName:
            pair = line.strip().split("\t")
            if len(pair) == 2:
                ticker, name = pair
            names[ticker.lower()] = name

skipColumns = ["Volume", "OpenInt"]

for fileName in os.listdir(curDir):
    data = pd.read_csv(curDir + fileName, usecols=lambda x: x not in skipColumns)

    index = fileName.split(".")[0]
    if not names.get(index):
        names[index] = index

    scatter_matrix(data, alpha=0.2)
    plt.suptitle(names[index])
    plt.savefig(curDir + "pics\\" + fileName + ".png")
    plt.close()
```

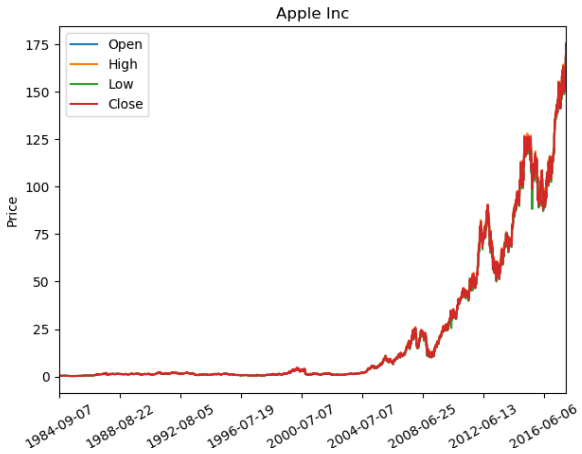
Листинг 4: src/stats.py

```
import matplotlib.pyplot as plt
import pandas as pd
import os


curDir = "C:\\Users\\pkmixer\\Downloads\\AI\\Stocks\\"

names = {}
indicesDir = "C:\\Users\\pkmixer\\Downloads\\indices\\"

for fileName in os.listdir(indicesDir):
```

```
    with open(indicesDir + fileName, "r") as indiceName:
        indiceName.readline()
        for line in indiceName:
            pair = line.strip().split("\t")
            if len(pair) == 2:
                ticker, name = pair
            names[ticker.lower()] = name

skipColumns = ["Date", "OpenInt", "Volume"]

for fileName in os.listdir(curDir):
    data = pd.read_csv(curDir + fileName, usecols=lambda x: x not in skipColumns)

    index = fileName.split(".")[0]
    if not names.get(index):
        names[index] = index

    with open(curDir + "..\\" + fileName + ".stat", "w") as file:
        file.write(data.describe().to_csv())
```

Классический пример - акции Apple.





Метод scatter_matrix строит распределение признаков, а также зависимость одного признака от другого. Как видно из графика котировок, акция в среднем росла, что и отражено в виде прямой с углом наклона $\frac{\pi}{4}$.

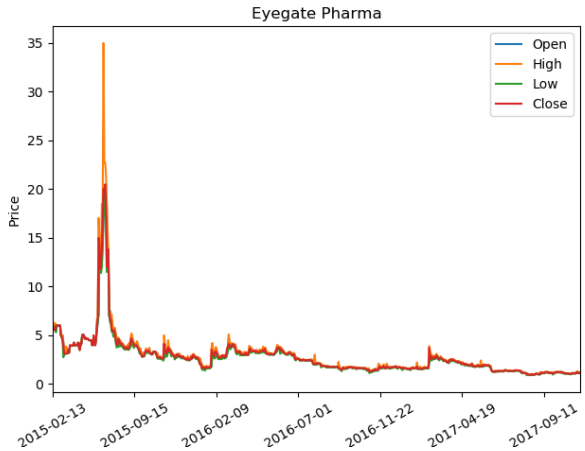
Числовые характеристики:

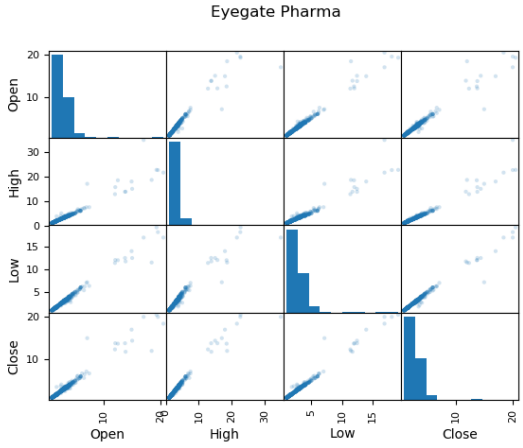
Листинг 5: src/aapl.csv

```
Name,Open,High,Low,Close
count,8364.0,8364.0,8364.0,8364.0
mean,22.284350151841224,22.49586662362506,22.054243920373025,22.281018027259687
std,37.76340245539372,38.05773267058666,37.44743243649739,37.76446946661696
min,0.23305,0.23564000000000002,0.23051,0.23051
```

```
25%,1.1371,1.1642,1.1128,1.1371
50%,1.6328,1.6634,1.6006,1.62825
75%,23.739,23.930500000000002,23.335749999999997,23.6945
max,175.11,175.61,174.27,175.61
```

Более интересная картина наблюдается с акциями Eyegate Pharma:





Характерный пик в стоимости акций отражается на зависимости одного признака от другого, а также в статистических характеристиках.

Листинг 6: src/eyeg.csv

```
Name,Open,High,Low,Close
count,645.0,645.0,645.0,645.0
mean,2.78324992248062,2.9571705426356587,2.6390708527131785,2.773861395348837
std,2.1770594051018417,2.7123703578832044,1.9780161518489183,2.2149790485909358
min,0.919,0.95,0.9,0.93
25%,1.63,1.69,1.57,1.62
50%,2.43,2.49,2.36,2.42
75%,3.28,3.41,3.15,3.2648
max,20.45,35.0,19.25,20.5
```

В данном датасете можно попробовать, например, предсказать цену акции на открытии торгов по данным, полученным за другой период, или максимальное падение акции за день в течение торгов.