Московский авиационный институт

(Национальный исследовательский университет)

Факультет: «Информационные технологии и прикладная математика»
Кафедра: 806 «Вычислительная математика и программирование»
Дисциплина: «Искусственный интеллект»

Лабораторная работа № 0

Тема: Получение и обработка данных.

Студент: Шевчук П.В.

Группа: М80-304Б

Преподаватель: Ахмед Самир Халид

Москва, 2019

1. Постановка задачи

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.

2. Требования

1) Датасеты должны быть уникальны

2) Исходный код должен быть написан в одном код стайле

3) Должен быть указан источник данных

3. Описание выполненной работы.

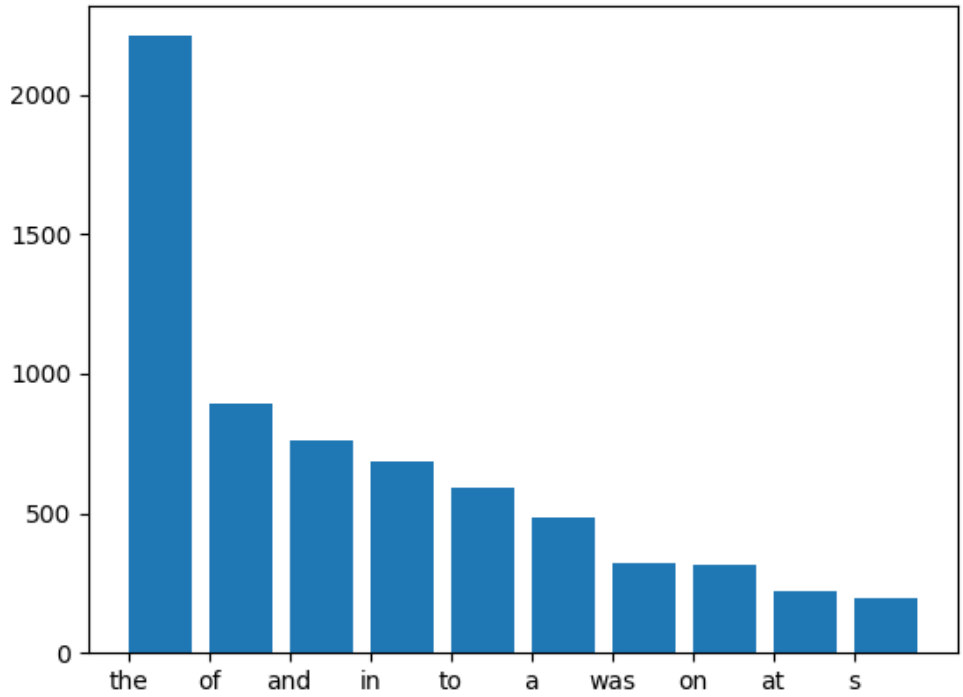
3. Решение задачи

Датасет:https://blog.einstein.ai/the-wikitext-long-term-dependency-language-modeling-dataset/, статьи из Википедии

Проблемы данных: заглавные буквы, знаки препинания и цифры, формы слова, лишние слова.

Решение: убрать большие буквы, лишние слова, знаки препинания и цифры.

В ходе решения данной задачи убраны заглавные буквы, цифры, знаки препинания и слово ‘unk’, которое часто встречается при ссылках. В абстрактный вид данных список добавлены две характеристики слово и количество, того сколько раз это слово встретилось. Выведены в форме гистограммы 10 самых распространенных слов. Среди них оказались в основном артикли, предлоги и союзы. Проблемой является то, что слова имеют разные формы и поэтому одно слово разделяется на несколько. Поэтому одним из часто встречаемых слов оказалось ‘s’, которое ставится после апострофа, что по-хорошему не является словом.



Код программы:

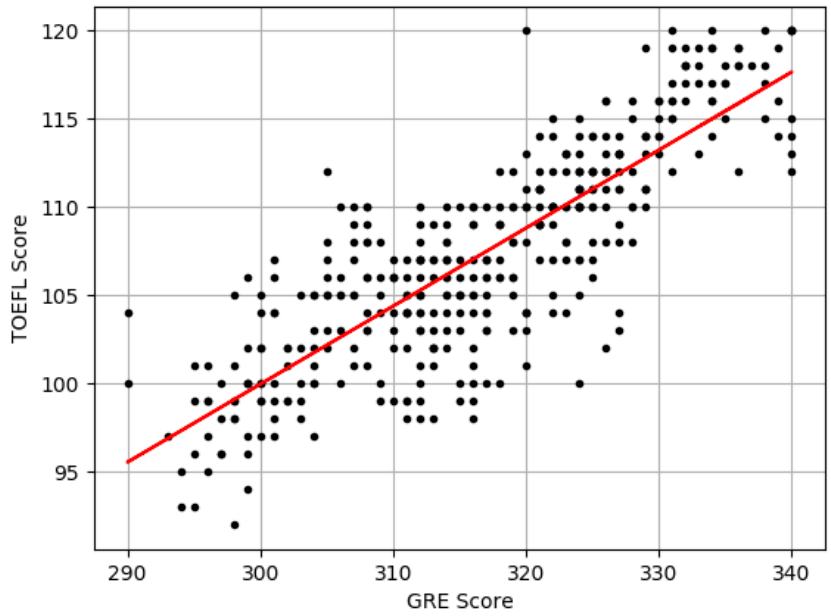
```
import pylab as plt
import string
import re

f = open('wiki.txt', 'r')
data_text = f.readlines() # считывание из файла
data_text_word = []  # словарь всех слов

for line in data_text: # считывание всех строк
    line = line.lower() # приведение к нижнему регистру
    line = re.sub('unk', '', line) # удалить слово unk
    line = "".join(i for i in line if i not in string.punctuation)# удаление всех знаков препинания
    result = "".join([i for i in line if not i.isdigit()]) # удаление всез цифр
    line = line.split() # разбить строку на части
    data_text_word.extend(line) # добавть строку

frequency = {} # считаем частоту слов
for word in data_text_word:
    count = frequency.get(word,0)
    frequency[word] = count + 1
```

```
frequency_list = frequency.keys() # представление ключей словаря
list = []
for words in frequency_list:
    list.append([words, frequency[words]])


list.sort(key=lambda x: x[1]) # сортировать по частоте


s=[]
n=[]
for i in reversed(list[-10:]):
    s.append(i[1]) # слово
    n.append(i[0]) # частота
x=range(len(s)) # количество столбцов
ax = plt.gcf()
ax.canvas.set_window_title('Самые распростанённые слова wiki.txt')
ax = plt.gca()
ax.bar(x, s, align='edge')
ax.set_xticklabels(n) # написать слово
ax.set_xticks(x) # слово под столбцом
plt.show()
```

Второй датасет - admission. Этот набор данных создан для прогнозирования приема выпускников с индийской точки зрения.

Набор данных содержит несколько параметров, которые считаются важными при подаче заявки на магистерские программы. Параметры включают в себя: 1. GRE баллов (из 340) 2. TOEFL баллов (из 120) 3. Университетский рейтинг (из 5) 4. Заявление о цели и рекомендательное письмо сила (из 5) 5. Бакалавриат Средний балл (из 10) 6. Опыт исследования (0 или 1) 7. Вероятность поступления (от 0 до 1)

Решение: Построим модель линейной регрессии между баллами GRE Score и TOEFL.



Наблюдается прямо-пропорциональная связь между баллами GRE Score и TOEFL.

Код программы:

```
import pandas as pd

from sklearn.linear_model import LinearRegression

import matplotlib.pyplot as plt

df = pd.read_csv("./admission.csv")

df.columns = ['Serial No.','GRE Score','TOEFL Score','University Rating','SOP','LOR' ,'CGPA','Research','Chance of Admit']

model = LinearRegression()

X = df[['GRE Score']].values

Y = df['TOEFL Score'].values

model.fit(X,Y)

plt.figure()

plt.xlabel('GRE Score')

plt.ylabel('TOEFL Score')

plt.plot(X, Y, 'k.')

plt.plot(X, model.predict(X), color='r')

plt.grid(True)

plt.show()
```

4. Вывод

В ходе данной лабораторной работы освоен алгоритм работы с корпусом документов, с помощью которого можно выявить самые распространенные слова. А также освоены принципы реализации линейной регрессии между количественными признаками. Вспомнили, про существование апострофа.