*Московский государственный технический университет им. Н. Э. Баумана*

***Факультет*** *«Информатика и системы управления»*

***Кафедра*** *«Системы обработки информации и управления»*



*Отчёт по лабораторной работе №2*

*по курсу «****Введение в машинное обучение****»*

*Исполнила: Корнеева А.П.., ИУ5-41*

*Проверил:*Гапанюк Ю*.Е.*

*Москва 2018 г.*

**Задание:**

Необходимо реализовать скрипт, выполняющий следующие действия:

**1. Скачивание 1000 последних объявлений с hh.ru**

Для выполнения этого пункта вам понадобится библиотека requests. С помощью этой библиотеки можно делать HTTP-запросы в API hh.ru. Что такое api. С API, предоставляемым hh.ru, можно ознакомиться здесь. Общая информация здесь

Вам понадобится этот метод. Чтобы тестировать запросы в api и смотреть, что они возвращают, можно использовать Postman. В качестве поискового запроса можно вводить ключевые слова, связанные с тематикой анализа данных: machine learning, data science, машинное обучение, big data, data analytics и тд. В ответе API будут интересующие нас поля: salary, area, name, employer

**2. Получить медианное значение зарплат**

Необходимо сделать обработку полученных на первом шаге данных и получить следующую структуру: Словарь, где ключом является название вакансии (как оно задано на hh.ru), а значением - медианное значение зарплаты по этой вакансии. То есть необходимо сгруппировать данные по имени вакансии. Также можно использовать другие варианты, например, сгруппировать по городу или любому другому интересному параметру из выдачи. В поле salary hh.ru отдает значения диапазона. Значением зарплаты считать среднее значение из диапазона, например, если зп от 100 до 150, то фиксировать значение 125.

**3. Получить распределение зарплат по диапазонам**

Необходимо выделить диапазоны зарплат, например:

до 80к, 80-120к, 120-150к, 150-200к, 200-300к, 300к+ Для каждого диапазона подсчитать количество предлагаемых вакансий.

**\*. Построить графики по пунктам 2 и 3.**

**Скрипт:**

import requests

import matplotlib.pyplot as plt

from pylab import \*

from matplotlib import rcParams

%matplotlib inline

#1.Скачивание 1000 последних объявлений с hh.ru

mas = []

for i in range(10):

res=requests.get('http://api.hh.ru/vacancies/?per\_page=100&page='+str(i)+'&text=data+science+OR+data+analytics')

mas.append(res.json())

print(res.json())

vac = []

for i in mas:

for vacancy in i['items']:

vac.append(vacancy)

print('\n')

#2.Получить медианное значение зарплат

#получение значения зарплат

sal\_dict = {}

for sal in vac:

if(sal['salary'] != None):

if ((sal['salary']['to'] != None) and (sal['salary']['from'] != None)):

salary = (sal['salary']['to'] + sal['salary']['from'])/2

sal\_dict[sal['name']] = salary

elif ((sal['salary']['to'] == None) and (sal['salary']['from'] != None)):

salary = sal['salary']['from']

sal\_dict[sal['name']] = salary

elif ((sal['salary']['to'] != None) and (sal['salary']['from'] == None)):

salary = sal['salary']['to']/2

sal\_dict[sal['name']] = salary

print(sal\_dict)

#получение медианного значения зарплат

arr1=[]

arr2=[]

arr3=[]

arr4=[]

arr5=[]

for key in sal\_dict:

if ('Руководитель' in key):

arr1.append(sal\_dict[key])

if ('Специалист' in key):

arr2.append(sal\_dict[key])

if ('Аналитик' in key):

arr3.append(sal\_dict[key])

if ('Программист' in key):

arr4.append(sal\_dict[key])

if ('Менеджер' in key):

arr5.append(sal\_dict[key])

print ('Руководитель', arr1)

print ('Специалист', arr2)

print ('Аналитик', arr3)

print ('Программист', arr4)

print ('Менеджер', arr5)

print('Медианные значения зарплат:')

arr1.sort()

arr2.sort()

arr3.sort()

arr4.sort()

arr5.sort()

med1=arr1[len(arr1)//2]

med2=arr2[len(arr2)//2]

med3=arr3[len(arr3)//2]

med4=arr4[len(arr4)//2]

med5=arr5[len(arr5)//2]

print (med1)

print (med2)

print (med3)

print (med4)

print (med5)

#3. Получить распределение зарплат по диапазонам

sum\_dict = {}

t80=0

t120=0

t150=0

t200=0

t300=0

t301=0

for key in sal\_dict:

if sal\_dict[key]<=80000:

t80=t80+1

if sal\_dict[key]>80000 and salary<=120000:

t120=t120+1

if sal\_dict[key]>120000 and salary<=150000:

t150=t150+1

if sal\_dict[key]>150000 and salary<=200000:

t200=t200+1

if sal\_dict[key]>200000 and salary<=300000:

t300=t300+1

if sal\_dict[key]>300000:

t301=t301+1

sum\_dict['80000']=t80

sum\_dict['120000']=t120

sum\_dict['150000']=t150

sum\_dict['200000']=t200

sum\_dict['300000']=t300

sum\_dict['>300000']=t301

print('\n')

print(sum\_dict)

Результаты:

