*МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ*

*НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ*

*„КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ”*

*НАВЧАЛЬНО-НАУКОВИЙ КОМПЛЕКС*

*„ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ”*

***Лабораторна робота №1***

*з курсу «Інтелектуальний аналіз даних»*

*Тема: Програмна розробка методу класифікації»*

*Виконала:*

*студентка III курсу*

*групи ДА-42*

*Балан Катерина*

*Київ*

*2017*

**Мета роботи:** Здобути навички програмної розробки методу класифікації.

**Завдання:** Розробити програму для класифікації даних.

Власноруч розробити програмну реалізацію одного з методів класифікації – «Наївний класифікатор Байєса».

**Виконання роботи**

Нехай вирішується задача класифікації листів з електронної скриньки на наявність спаму. Маємо початковий файл з якого формується словник 1.txt та файл для зчитування контрольних примірників 2.txt.

*1.txt*

0 New meeting tomorrow file  
0 Corporate party tomorrow  
1 Free sales party  
1 Free file for you  
0 New greeting text  
1 Free file upload

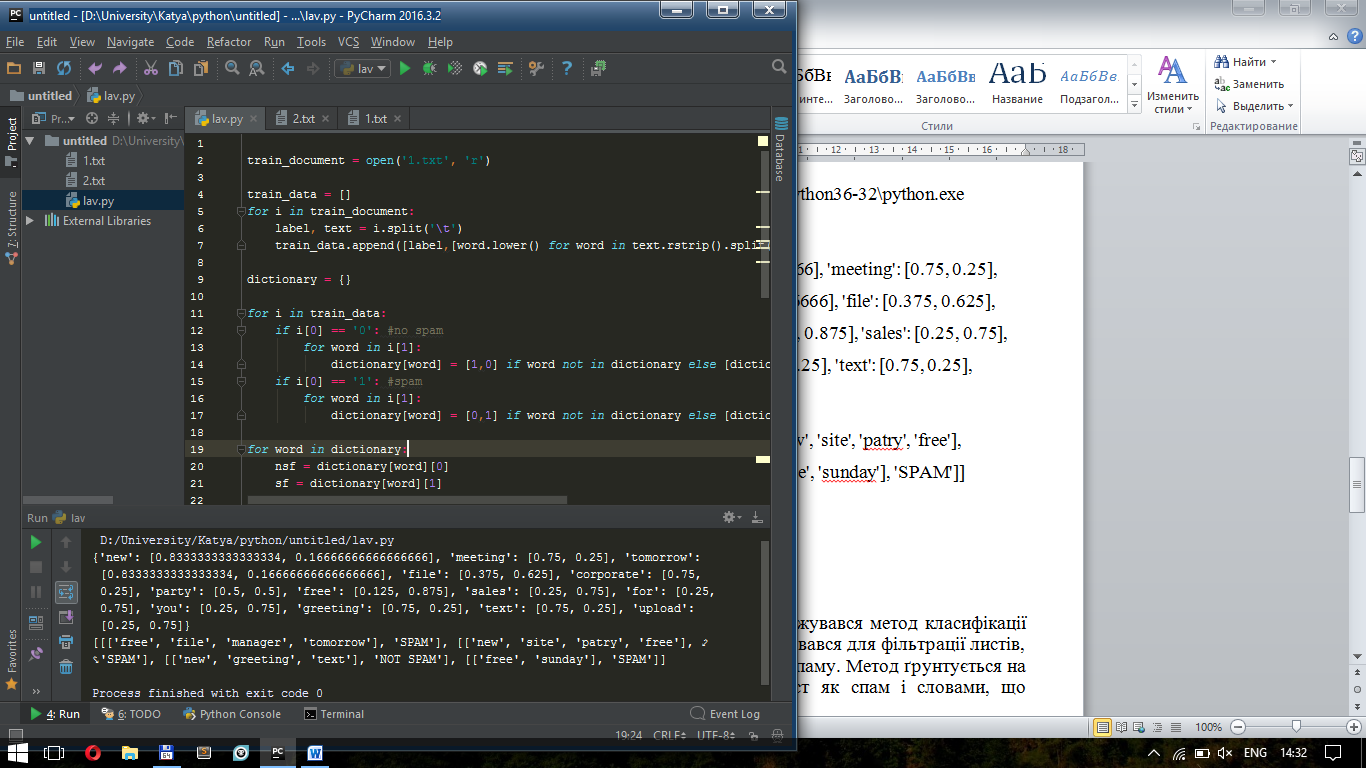
*2.txt*

Free file manager Tomorrow  
New site patry free  
New greeting text  
Free Sunday

*Код програми*

train\_document = open('1.txt', 'r')  
  
train\_data = []  
*for* i *in* train\_document:  
 label, text = i.split('\t')  
 train\_data.append([label,[word.lower() *for* word *in* text.rstrip().split()]])  
  
dictionary = {}  
  
*for* i *in* train\_data:  
 *if* i[0] == '0': #no spam  
 *for* word *in* i[1]:  
 dictionary[word] = [1,0] *if* word *not in* dictionary *else* [dictionary[word][0]+1, dictionary[word][1]]  
 *if* i[0] == '1': #spam  
 *for* word *in* i[1]:  
 dictionary[word] = [0,1] *if* word *not in* dictionary *else* [dictionary[word][0], dictionary[word][1]+1]  
  
*for* word *in* dictionary:  
 nsf = dictionary[word][0]  
 sf = dictionary[word][1]  
  
 dictionary[word][0] = (((sf + nsf) \* nsf)/(nsf + sf) + 0.5)/((sf + nsf) + 1)  
 dictionary[word][1] = (((sf + nsf) \* sf)/(nsf + sf) + 0.5)/((sf + nsf) + 1)  
  
  
print(dictionary)  
  
test\_document = open('2.txt', 'r')  
  
test\_data = []  
*for* i *in* test\_document:  
 sentence = ([word.lower() *for* word *in* i.rstrip().split()])  
 sp = nsp = 0.5  
 *for* word *in* sentence:  
 *if* word *in* dictionary:  
 nsp \*= dictionary[word][0]  
 sp \*= dictionary[word][1]  
 label = 'NOT SPAM' *if* nsp >= sp *else* 'SPAM'  
 test\_data.append([sentence, label])  
  
  
print(test\_data)

Результат виконання програми з даними словниками.



***Висновки:***

В ході виконання лабораторної роботи досліджувався метод класифікації «Наївний класифікатори Байєса». Метод застосовувався для фільтрації листів, що надходять до електронної пошти на наявність спаму. Метод ґрунтується на словниках зі словами, що характеризують лист як спам і словами, що характеризують як не спам.

В програмі використовуються таблиці ймовірностей спам / не спам для слів, що входять до листа. Для цього використовуються словники, що мають формат файла txt і зчитуються на початку виконання програми.

Всі тестові комбінації були вірно розподілені по класам( 4/4 ).