МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
“Київський політехнічний інститут ім. Ігоря Сікорського”
ФАКУЛЬТЕТ Інформатики та Обчислювальної Техніки
КАФЕДРА Інформаційних систем та технологій

Звіт до лабораторної роботи №6

з предмету: Обробка та Аналіз текстових даних на мові Python

Перевірила:
Тимофєєва Ю.С.

Виконли:
студенти групи ІК-01
Філоненко І. Р.
Гацан С. Ю.

КИЇВ - 2023

Тема: Аналіз настроїв

Мета: Ознайомитись з вирішенням задачі аналізу настроїв.

Варіант: 11

Завдання:

Створити програму, яка:

1. Зчитує заданий набір даних, виконує попередню обробку, розбиває дані на навчальні на тестові. Виконує аналіз настроїв за допомогою алгоритмів класифікації (наприклад, логістичної регресії, опорних векторів і т.д.).
2. Розрахувати матрицю невідповідностей, провести оцінку точності моделі.
3. Використати один з готових лексиконів, наприклад Textblob, для аналізу оцінки настроїв. Також розрахувати матрицю невідповідностей, провести оцінку точності моделі.
4. Обрати три випадкові записи та вивести результати оцінки їх настрою за пунктами 1 і 3.

   Файл twitter1.csv. Викопистати наївний байєсів класифікатор.

Код програми

```
#!/usr/bin/env python3

import re
import csv
import pandas as pd
import random

import nltk
from nltk.tokenize import WordPunctTokenizer
from nltk.corpus import stopwords

from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score, confusion_matrix

from textblob import TextBlob

nltk.data.path.append("../Lab2/nltk_data")

prepared_corpus = []
labels = []

skip_generation = False

ans = input("Do you want to load already prepared corpus? (Y/N) ").lower()

if ans == 'y':
    try:
        with open("prepared_corpus.txt", 'r') as file:
```

```
            for line in file:
                prepared_corpus.append(line.strip())
        with open("prepared_labels.txt", 'r') as file:
            for line in file:
                labels.append(line.strip())
        skip_generation = True
    except FileNotFoundError:
        print("Prepared corpus not found, generating it.")

if skip_generation is False:
    file = open("twitter1.csv", 'r')
    csv = csv.DictReader(file, delimiter=',', quotechar='"',
fieldnames=['number','topic','sentiment','content'])
    texts = {i: v for i, v in enumerate(csv)}

    wpt = WordPunctTokenizer()

    # Preparing corpus
    corpus = [texts[i]['content'] for i in range(len(texts))]


    def preprocess_sentence(sentence):
        sentence = re.sub(r'[^a-zA-Z\s]+', '', sentence, re.I | re.A)  # Links
are still present in array
        sentence = sentence.lower()
        sentence = sentence.strip()
        tokens = wpt.tokenize(sentence)
        filtered_tokens = [token for token in tokens if token not in
stopwords.words('english')]
        sentence = ' '.join(filtered_tokens)
        return sentence


    i = 0
    for sentence in corpus:
        sentence_processed = preprocess_sentence(sentence)
        if len(sentence_processed) != 0:
            labels.append(texts[i]['sentiment'])
            prepared_corpus.append(sentence_processed)
        i = i+1

    with open("prepared_corpus.txt", 'w') as file2:
        for sentence in prepared_corpus:
            file2.write(sentence + '\n')

    with open("prepared_labels.txt", 'w') as file2:
        for label in labels:
            file2.write(label + '\n')

print("Підготовлений корпус: ")
print(prepared_corpus[len(prepared_corpus)-50:])
print(len(prepared_corpus))

# Limiting corpus (for faster results)

prepared_corpus = prepared_corpus[:len(prepared_corpus)]

# Task 1 - Split data into train and test
train_corpus, test_corpus, train_sentiments, test_sentiments, \
    = train_test_split(prepared_corpus, labels,
                       test_size=0.4, random_state=0)
# --------
# Bag of Words
cv = CountVectorizer(binary=False, min_df=0., max_df=1.)
```

```
cv_train_matrix = cv.fit_transform(train_corpus)
cv_test_matrix = cv.transform(test_corpus)
vocab = cv.get_feature_names_out()

array = pd.DataFrame(cv_train_matrix.toarray(), columns=vocab)

print(array)
# --------
# NaiveBayesClassifier
mnb = MultinomialNB(alpha=1)
mnb.fit(cv_train_matrix, train_sentiments)
# --------
# Predict sentiments
mnb_prediction = mnb.predict(cv_test_matrix)
mnb_prediction2 = mnb.predict(cv_train_matrix)
mnb_accuracy = accuracy_score(test_sentiments, mnb_prediction)
mnb_accuracy2 = accuracy_score(train_sentiments, mnb_prediction2)
print("Accuracy of MultinomialNB: " + str(mnb_accuracy))
print("Accuracy of MultinomialNB on train data: " + str(mnb_accuracy2))
# --------
# Confusion matrix
table_labels = ["Positive", "Negative", "Neutral", "Irrelevant"]
matrix1 = confusion_matrix(test_sentiments, mnb_prediction, labels=table_labels)
array = pd.DataFrame(matrix1, columns=table_labels, index=table_labels)

print(array, end="\n\n")

print("Accuracy of Positive: " + str(matrix1[0][0] / matrix1[0].sum()))
print("Accuracy of Negative: " + str(matrix1[1][1] / matrix1[1].sum()))
print("Accuracy of Neutral: " + str(matrix1[2][2] / matrix1[2].sum()))
print("Accuracy of Irrelevant: " + str(matrix1[3][3] / matrix1[3].sum()))
print()
# --------
# TextBlob

sentiments_polarity = [TextBlob(sentence).sentiment.polarity for sentence in
test_corpus]
predicted_sentiments = ['Positive' if score >= 0.1
                        else "Negative" if score <= -0.1
                        else "Neutral"
                        for score in sentiments_polarity]



table_labels = ["Positive", "Negative", "Neutral"]
sentiments_without_irrelevant = [label if label != "Irrelevant" else "Neutral"
for label in test_sentiments]

blob_accuracy = accuracy_score(sentiments_without_irrelevant,
predicted_sentiments)
print("Accuracy of TextBlob: " + str(blob_accuracy))

matrix2 = confusion_matrix(sentiments_without_irrelevant, predicted_sentiments,
labels=table_labels)
array2 = pd.DataFrame(matrix2, columns=table_labels, index=table_labels)
print(array2)
print("Accuracy of Positive: " + str(matrix2[0][0] / matrix2[0].sum()))
print("Accuracy of Negative: " + str(matrix2[1][1] / matrix2[1].sum()))
print("Accuracy of Neutral: " + str(matrix2[2][2] / matrix2[2].sum()))
print()

# Compare 3 records
rec = random.sample(range(1, len(test_sentiments)+1), 3)
```

```
print("Compare MultinomialNB:")
for record in rec:
    print("Post: " + test_corpus[record])
    print("Actual sentiment: " + test_sentiments[record])
    print("Predicted: " + mnb_prediction[record])
    print("----------")

print()

print("Compare TextBlob:")
for record in rec:
    print("Post: " + test_corpus[record])
    print("Actual sentiment: " + sentiments_without_irrelevant[record])
    print("Predicted: " + predicted_sentiments[record])
    print("----------")
```

Результат роботи програми

```
Do you want to load already prepared corpus? (Y/N) y
Підготовлений корпус:
['save buying aero oled award winning design stunning k oled panel color
accurate right scratch snapdragon core nv
idia gtx gpu one sweet workunk home play anywhere laptop bufflykdorq', 'save
aero oled edge award winning desk des
ign stunning k oled panel color correction accurate right front inside box core
apple nvidia gtx gpu also sweet wo
rk home create anywhere laptop buff ly kdorq', 'aero va oled award winning
design stunning k oled panel color fron
t edge kitchen oven core nvidia core work home create anywhere laptop
bufflykdorq', 'love everything', 'love every
thing', 'love everything', 'love everything', 'love everything', 'love
everything anything', 'checked new gpu driv
ers today went amd driver section tried select card paused around seconds
completely confused remembered switched
amd nvidia months ago', 'today searched new gpu drivers went amd driver section
tried select card stopped complete
ly confused seconds remembered switched amd nvidia months ago', 'remembered
switched amd nvidia months ago', 'chec
ked new gpu drivers today typed amd driver section tried select nvidia paused
around seconds completely confused r
emembered switched amd nvidia months ago', 'checked new gpu powered drivers
today went amd driver selection sectio
n tried select card paused around seconds completely confused thought remembered
switched amd nvidia months ago',
'checked new gpu drivers today clicked amd driver section tried use card paused
around seconds completely disappoi
nted remembered switched amd ubuntu months ago', 'nvidia released security
update drivers fixing several issues co
uld lead denial service escalation privileges information disclosure update
covers multiple vulnerabilities affect
ing displa twibinllrdjgplkd via infosechotspot httpstcoqoyqryr', 'nvidia
released security update drivers fixes se
veral issues lead denial service privilege escalation disclosure informationthe
update covers multiple vulnerabili
ties affecting twibin l lrdjgplkd display via infosechotspot https tco qoyqryr',
'nvidia released security update
drivers fixes several problems lead denial service privilege escalation
information leakage', 'nvidia released ano
ther software update drivers fixing several issues could lead denial permissions
```

escalation privileges information
disclosure update addresses multiple vulnerabilities affecting products
twibinllrdjgplkd x infosechotspot httpstc
oqoyqryr', 'nvidia released new security update drivers fixing several issues
including could potentially lead ext
ended denial lines service escalation privileges requiring information
disclosure update covers multiple vulnerabi
lities affecting displa twib l lrdjgplkd via infosechotspot httpstcoqoyqryr',
'nvidia released substantial update
drivers identifying several shortcomings could lead denial service escalation
privileges auto theft update covers
standards using displa twibinllrdjgplkd via httpstcoqoyqryr', 'cheap doesnt mean
better btw techsallcomgooglechall
en', 'cheap mean better way techsallcom googlechallen', 'cheap mean betterunk
techsallcomgooglechallen', 'cheap do
esnt mean anyone better btw techsall com google end challen page', 'really
doesnt look bad btw netcomgooglechallen
', 'nvidia doesnt want give crypto craze docs maxbitccnvidiadoesnt', 'nvidia
plans release crypto craze documentar
y maxbitcc nvidiadoesnt', 'nvidia want give cryptoinsanity doxmaxbitcc
nvidiadoes', 'nvidia doesnt intend give awa
y ad craze docs maxbitccnvidiadoesnt', 'nvidia therefore want give crypto craze
docs maxbit cc nvidia cc doesnt bu
y', 'doesnt give password crypto wallet docs maxbitccnvidiadoesnt', 'nvidia
really delayed weeks', 'nvidia really
delayed weeks', 'nvidia delay weeks', 'nvidia really delayed several weeks',
'nvidia really delayed flight weeks',
'nvidia really delayed next weeks', 'let elim go unnoticed nvidia highlights
automatically records best moments f
ortnitegame gfn share', 'let elim go unnoticed nvidia highlights automatically
records beautiful moments fortniteg
ame gfn share', 'let elite go unnoticed nvidia highlights automatically records
best moments fennitegame gfn', 'le
t elim go unnoticed nvidia highlights automatically records best shot video gfn
share', 'let information elim go u
nnoticed nvidia highlights automatically records three best moments fortnitegame
gfn share', 'unk elim nvidia high
lights pictures reveals best moments fortnitegame gfn share', 'realized windows
partition mac like years behind nv
idia drivers idea didnt notice', 'realized windows partition mac like years
behind nvidia drivers idea notice', 'r
ealized mac window partition years behind nvidia drivers idea didnt notice',
'realized windows partition mac years

behind nvidia drivers idea didnt notice', 'realized windows partition mac like years behind nvidia drivers cars f
ucking idea ever notice', 'like windows partition mac like years behind drivers idea didnt notice']
72364

	aa	aaa	aaaaaaaaaaaa	aaaaaaaaaaaaa	...	zywzswpxq	zzgixvkt	zzmhpax	zzvfsrhewg
0	0	0	0	0	...	0	0	0	0
1	0	0	0	0	...	0	0	0	0
2	0	0	0	0	...	0	0	0	0
3	0	0	0	0	...	0	0	0	0
4	0	0	0	0	...	0	0	0	0
...	..	...	...	...	...	...	...	...	...
43413	0	0	0	0	...	0	0	0	0
43414	0	0	0	0	...	0	0	0	0
43415	0	0	0	0	...	0	0	0	0
43416	0	0	0	0	...	0	0	0	0
43417	0	0	0	0	...	0	0	0	0

[43418 rows x 32676 columns]
Accuracy of MultinomialNB: 0.732743729703586
Accuracy of MultinomialNB on train data: 0.8093187157400157

	Positive	Negative	Neutral	Irrelevant
Positive	6463	956	481	178
Negative	802	7393	462	195
Neutral	1179	1208	4374	250
Irrelevant	897	838	290	2980

Accuracy of Positive: 0.8000742758108442
Accuracy of Negative: 0.8351784907365567
Accuracy of Neutral: 0.6238767650834403
Accuracy of Irrelevant: 0.5954045954045954

Accuracy of TextBlob: 0.45412146756028465

	Positive	Negative	Neutral
Positive	4854	912	2312
Negative	1819	3606	3427
Neutral	4917	2414	4685

Accuracy of Positive: 0.6008913097301313
Accuracy of Negative: 0.40736556710347943
Accuracy of Neutral: 0.38989680426098533

Compare MultinomialNB:
Post: chart leading listeners boycott foxnews proctergamble dounk f x amazon kraftheinzco pfizer bestbuysupport sa
ndalsresorts petsmart youtubeemecrtbmdjy

Actual sentiment: Neutral
Predicted: Neutral

Post: overwatch lagging lot anyone else lny patch logged first time stuttering almost every match
Actual sentiment: Negative
Predicted: Negative

Post: reason email went crazy kept asking reenter password around pm
Actual sentiment: Negative
Predicted: Negative

Compare TextBlob:
Post: chart leading listeners boycott foxnews proctergamble dounk f x amazon kraftheinzco pfizer bestbuysupport sa
ndalsresorts petsmart youtubeemecrtbmdjy
Actual sentiment: Neutral
Predicted: Neutral

Post: overwatch lagging lot anyone else lny patch logged first time stuttering almost every match
Actual sentiment: Negative
Predicted: Positive

Post: reason email went crazy kept asking reenter password around pm
Actual sentiment: Negative
Predicted: Negative
