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 №4

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: Класифікація тестових даних

Мета: Ознайомитись з класифікацією документів за допомогою моделей машинного навчання.

Варіант: 11

Завдання:

Файл news.csv. В якості текстової моделі використати модель “Сумка слів”. Виконати класифікацію за допомогою алгоритмів наївний байєсів класифікатор та випадкові ліси, порівняти їх точність. Спробувати покращити модель випадкові ліси за допомогою GridSearchCV.

Код програми

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

import re
import csv
import numpy as np
import pandas as pd

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, cross_val_score
from sklearn.naive_bayes import MultinomialNB
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import accuracy_score

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

file = open("news.csv", 'r')
next(file)  # To skip headers row
csv = csv.DictReader(file, delimiter=',', quotechar='"', fieldnames=['text',
'label'])
texts = {i: v for i, v in enumerate(csv)}
labels = [texts[i]['label'] for i in range(len(texts))]

wpt = WordPunctTokenizer()

# Preparing corpus
corpus = [texts[i]['text'] for i in range(len(texts)) if len(texts[i]['text']) !
= 0]

def preprocess_sentence(sentence):
    sentence = re.sub(r'[^a-zA-Z\s]|http\S+', '', sentence, re.I | re.A)
    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
```

```
prepared_corpus = []

for sentence in corpus:
    prepared_corpus.append(preprocess_sentence(sentence))

print("Підготовлений корпус: ")
print(prepared_corpus)

# Task 1 - Split data into train and test
train_corpus, test_corpus, train_label_nums, test_label_nums, \
    = train_test_split(prepared_corpus, labels,
                       test_size=0.3, random_state=0)
# --------
# Task 2 - Bag of Words
cv = CountVectorizer(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)
# --------
# Task 3 - NaiveBayesClassifier
mnb = MultinomialNB(alpha=1)
mnb.fit(cv_train_matrix, train_label_nums)
# --------
# Task 4 - RandomForestClassifier
rfc = RandomForestClassifier(n_estimators = 10, random_state = 0)
rfc.fit(cv_train_matrix, train_label_nums)
# --------
# Task 5 - Compare accuracy
mnb_prediction = mnb.predict(cv_test_matrix)
mnb_accuracy = accuracy_score(test_label_nums, mnb_prediction)
print("Accuracy of MultinomialNB: " + str(mnb_accuracy))

rfc_prediction = rfc.predict(cv_test_matrix)
rfc_accuracy = accuracy_score(test_label_nums, rfc_prediction)
print("Accuracy of RandomForestClassifier: " + str(rfc_accuracy))
# --------
# Task 6 - GridSearchCV
param_grid = {
    'n_estimators': [100, 200, 500],
    'max_depth': [10, 20, 30, None],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    # 'bootstrap': [True, False],
    'warm_start': [True]
}

grid_search = GridSearchCV(rfc, param_grid=param_grid, cv=5, n_jobs=-1)
grid_search.fit(cv_train_matrix, train_label_nums)

best_params = grid_search.best_params_
print("Best hyperparameters for Random Forest Classifier:", best_params)

best_rfc_classifier = grid_search.best_estimator_
best_rfc_predictions = best_rfc_classifier.predict(cv_test_matrix)
best_rfc_accuracy = accuracy_score(test_label_nums, best_rfc_predictions)
print("Accuracy of RandomForestClassifier with best hyperparameters:",
best_rfc_accuracy)
# --------
```

Результат роботи програми

Підготовлений корпус:

... 'rare move private equity firm cdampr bought million riskiest part debt buyout cornerstone building brands limiting potential losses banks agreed fund deal', 'italian bond futures plunge lows day euro dips back parity dollar italys prime minister mario draghi said would resign', 'jpmorgan credit costs hit billion dimon predicts turmoil', 'libor jumps since traders bet bigger fed hikes', 'prospect global recession sending money managers scrambling rework portfolios especially emerging asian bonds already hammered year accelerating inflation slowing growth', 'australias biggest investors piling bonds bet backing central bank governor lowes assessment series robust rate hikes help get inflation control next year', 'south africas indebted power utility may need borrow extra billion purchase diesel pay inflationbeating salaries workers according sampp global ratings', 'us treasury yields jumped led shortdated tenors another hotterthanexpected inflation report kindled bets fed could raise rates full percentage point month', 'municipal bonds left behind fixedincome sectors move electronic trading', 'spread countries european periphery like italy continue pay higher price borrow money bond market germany thanks part ecb things remain tame compared last decade', 'treasury curve inversion deepens level last seen', 'divergence stress levels chinas offshore local credit markets widening defaults overseas notes hit record even calm reigns onshore', 'debt defaults going roll across emerging markets countries cant cope sudden increase borrowing costs according man group runs one bestperforming funds industry', 'bond investors demanding premium hold australias sovereign debt burned central banks failure provide reliable guidance inflation interest rates', 'goldman sachs agreed loan million latin american ecommerce giant mercadolibres fintech arm firm plans expanding credit offering two key markets', 'crypto lender celsius network repaid loans monday company continues battle insolvency', 'petrobras debuted green financing market billion sustainabilitylinked loan brazils oil giant looks court environmentally conscious investors', 'even standards argentinas debt crisis default nations sovereign bonds trading unprecedented lows', 'us treasury yields risen rapidly year billionaire investor leon cooperman thinks theyre still low said hed choose stocks bonds', 'despite richest corporate debt yields years worries economic downturn might turn investors risks highyield bonds look surprisingly balanced', 'italian debt sells draghi government teeters brink', 'uk strips mn greensill loans sanjeev gupta state guarantee', 'equitiesfirst unnamed debtor troubled crypto firm celsius', 'lenders take control vue cinem

a chain bn restructuring', 'china developers face bn wall dollar bond payments second half', 'kfw credit line uniper could raised bln eur h
andelsblatt', 'kfw credit line uniper could raised bln eur handelsblatt', 'russian sells bln roubles oneyear repo auction', 'global esg bon
d issuance posts h dip supranationals cut back', 'brazils petrobras says signed billion sustainability loan']

	aa	aad	aaic	aaii	aal	aalberts	aalbf	aam	aampp	...	zurichbased	zurn	zvia	zyme	zymeworks	zynx	zyus	zyversa
0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
...	..	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
11888	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
11889	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
11890	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	0	0	0
11891	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
11892	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0

[11893 rows x 19602 columns]

Accuracy of MultinomialNB: 0.7531881498920934

Accuracy of RandomForestClassifier: 0.7272905630763195

Best hyperparameters for Random Forest Classifier: {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 500, 'warm_start': True}

Accuracy of RandomForestClassifier with best hyperparameters: 0.765940749460467