Лабораторна робота №6, Обробка та аналіз текстових даних на Python, Варіант 14

Виконав: студент групи ІП-11, Лошак Віктор Іванович

Перевірив: Юлія Тимофєєва Сергіївна

Тема роботи: Аналіз настроїв

Мета роботи: Ознайомитись з вирішенням задачі аналізу настроїв та базовими можливостями бібліотеки spaCy.

04.04.2024

Завдання:

1. У файлі twitter2.csv містяться дані в форматі: clean_text,category, де можливими значеннями category є:
   -1 – негативний коментар,
   0 – нейтральний коментар,
   1 – позитивний коментар.

Використати наївний байєсів класифікатор для sentiment analysis.
2. У файлі lab6-1.txt.
a) Знайти та вивести всі слова з тексту, які не є стоп-словами.
b) Знайти та вивести всі прикметники, які присутні у тексті.
c) Знайти та вивести організації та дати, які присутні у тексті.

Task:

1. The file twitter2.csv contains data in the format: clean_text,category, where the possible values of category are:
   -1 - negative comment,
   0 - neutral comment,
   1 - a positive comment.

Use a naive Bayesian classifier for sentiment analysis.
2. In the file lab6-1.txt.
a) Find and extract all words from the text that are not stop words.
b) Find and display all the adjectives that are present in the text.
c) Find and display organizations and dates that are present in the text.

Task 1

In [ ]:

```
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.pipeline import make_pipeline
from sklearn.metrics import classification_report
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
import pandas as pd
```

In [ ]:

```
df = pd.read_csv('twitter2.csv')
df
```

Out[ ]:

	clean_text	category
0	when modi promised “minimum government maximum...	-1.0
1	talk all the nonsense and continue all the dra...	0.0
2	what did just say vote for modi welcome bjp t...	1.0
3	asking his supporters prefix chowkidar their n...	1.0
4	answer who among these the most powerful world...	1.0
...	...	...
162975	why these 456 crores paid neerav modi not reco...	-1.0
162976	dear rss terrorist payal gawar what about modi...	-1.0
162977	did you cover her interaction forum where she ...	0.0
162978	there big project came into india modi dream p...	0.0
162979	have you ever listen about like gurukul where ...	1.0

162980 rows × 2 columns

In [ ]:

```
df.isna().sum()
```

Out[ ]:

```
clean_text    4
category      7
dtype: int64
```

In [ ]:

```
df_clean = df.dropna()
df_clean
```

Out[ ]:

	clean_text	category
0	when modi promised “minimum government maximum...	-1.0
1	talk all the nonsense and continue all the dra...	0.0
2	what did just say vote for modi welcome bjp t...	1.0
3	asking his supporters prefix chowkidar their n...	1.0
4	answer who among these the most powerful world...	1.0
...	...	...
162975	why these 456 crores paid neerav modi not reco...	-1.0
162976	dear rss terrorist payal gawar what about modi...	-1.0
162977	did you cover her interaction forum where she ...	0.0
162978	there big project came into india modi dream p...	0.0
162979	have you ever listen about like gurukul where ...	1.0

162969 rows × 2 columns

In [ ]:

```
X_clean = df_clean['clean_text']
y_clean = df_clean['category']

X_train_clean, X_test_clean, y_train_clean, y_test_clean = train_test_split(X_cl
model_clean = make_pipeline(CountVectorizer(), MultinomialNB())
model_clean.fit(X_train_clean, y_train_clean)
y_pred_clean = model_clean.predict(X_test_clean)
report_clean = classification_report(y_test_clean, y_pred_clean, target_names=['

print(report_clean)
```

```
              precision    recall  f1-score   support

    Negative       0.75      0.62      0.68      7152
     Neutral       0.92      0.60      0.73     11067
    Positive       0.68      0.92      0.78     14375

    accuracy                           0.75     32594
   macro avg       0.78      0.71      0.73     32594
weighted avg       0.78      0.75      0.74     32594
```

In [ ]:

```
conf_matrix_clean = confusion_matrix(y_test_clean, y_pred_clean)
accuracy_clean = accuracy_score(y_test_clean, y_pred_clean)

(conf_matrix_clean, accuracy_clean)
```

Out[ ]:

```
(array([[ 4409,   296,  2447],
        [  652,  6672,  3743],
        [  790,   320, 13265]], dtype=int64),
 0.7469472909124378)
```

Task 2

In [ ]:

```
from textblob import TextBlob
import numpy as np
```

In [ ]:

```
# Function to categorize sentiment based on TextBlob polarity score
def categorize_sentiment(text):
    sentiment = TextBlob(text).sentiment.polarity
    if sentiment < 0:
        return -1
    elif sentiment == 0:
        return 0
    else:
        return 1
```

In [ ]:

```
textblob_df = df_clean.copy()
# Applying TextBlob sentiment analysis to the dataset
textblob_df['textblob_category'] = textblob_df['clean_text'].apply(categorize_se

# Calculating the confusion matrix and accuracy for TextBlob
y_true_tb = textblob_df['category']
y_pred_tb = textblob_df['textblob_category']

conf_matrix_tb = confusion_matrix(y_true_tb, y_pred_tb)
accuracy_tb = accuracy_score(y_true_tb, y_pred_tb)

(conf_matrix_tb, accuracy_tb)
```

Out[ ]:

```
(array([[35509,     0,     0],
        [    0, 55211,     0],
        [    0,     2, 72247]], dtype=int64),
 0.9999877277273592)
```

Task 3

Printing all words that are not stop words. Printing all adjectives.

In [ ]:

```
from spacy.lang.en import English
from spacy.lang.en.stop_words import STOP_WORDS
from nltk.tokenize import word_tokenize
import spacy
```

In [ ]:

```
with open('lab6-1.txt', 'r') as file:
    text = file.read()
    text = word_tokenize(text)
    text = ' '.join([w for w in text if w.isalnum() ])

text
```

```
Out[ ]: 'US retail sales fell in January the biggest monthly decline since last August
        driven down by a heavy fall in car sales The fall in car sales had been expecte
        d coming after December 4 rise in car sales fuelled by generous special offers
        Excluding the car sector US retail sales were up in January twice what some ana
        lysts had been expecting US retail spending is expected to rise in 2005 but not
        as quickly as in 2004 Steve Gallagher US chief economist at SG Corporate Invest
        ment Banking said January figures were decent numbers We are not seeing the num
        bers that we saw in the second half of 2004 but they are still pretty healthy h
        e added Sales at appliance and electronic stores were down in January while sal
        es at hardware stores dropped by and furniture store sales dipped Sales at clot
        hing and clothing accessory stores jumped while sales at general merchandise st
        ores a category that includes department stores rose by These strong gains were
        in part put down to consumers spending gift vouchers they had been given for Ch
        ristmas Sales at restaurants bars and coffee houses rose by while grocery store
        sales were up In December overall retail sales rose by Excluding the car sector
        sales rose by just Parul Jain deputy chief economist at Nomura Securities Inter
        national said consumer spending would continue to rise in 2005 only at a slower
        rate of growth than in 2004 Consumers continue to retain their strength in the
        first quarter he said Van Rourke a bond strategist at Popular Securities agreed
        that the latest retail sales figures were slightly stronger than expected'
```

```
In [ ]: nlp = spacy.load("en_core_web_sm")
        doc = nlp(text)
```

```
In [ ]: # a) Extract words that are not stop words
        non_stop_words = [token.text for token in doc if token.text not in STOP_WORDS]

        # b) Find and display all adjectives
        adjectives = [token.text for token in doc if token.pos_ == "ADJ"]

        # c) Find and display organizations and dates present in the text
        organizations = [ent.text for ent in doc.ents if ent.label_ == "ORG"]
        dates = [ent.text for ent in doc.ents if ent.label_ == "DATE"]

        print("Non stop words: ", non_stop_words),
        print("Adjectives: ", adjectives),
        print("Organisations: ", organizations),
        print("Dates: ", dates)
```

```
Non stop words:  ['US', 'retail', 'sales', 'fell', 'January', 'biggest', 'monthl
y', 'decline', 'August', 'driven', 'heavy', 'fall', 'car', 'sales', 'The', 'fal
l', 'car', 'sales', 'expected', 'coming', 'December', '4', 'rise', 'car', 'sale
s', 'fuelled', 'generous', 'special', 'offers', 'Excluding', 'car', 'sector', 'U
S', 'retail', 'sales', 'January', 'twice', 'analysts', 'expecting', 'US', 'retai
l', 'spending', 'expected', 'rise', '2005', 'quickly', '2004', 'Steve', 'Gallaghe
r', 'US', 'chief', 'economist', 'SG', 'Corporate', 'Investment', 'Banking', 'sai
d', 'January', 'figures', 'decent', 'numbers', 'We', 'seeing', 'numbers', 'saw',
'second', 'half', '2004', 'pretty', 'healthy', 'added', 'Sales', 'appliance', 'el
ectronic', 'stores', 'January', 'sales', 'hardware', 'stores', 'dropped', 'furnit
ure', 'store', 'sales', 'dipped', 'Sales', 'clothing', 'clothing', 'accessory',
'stores', 'jumped', 'sales', 'general', 'merchandise', 'stores', 'category', 'inc
ludes', 'department', 'stores', 'rose', 'These', 'strong', 'gains', 'consumers',
'spending', 'gift', 'vouchers', 'given', 'Christmas', 'Sales', 'restaurants', 'ba
rs', 'coffee', 'houses', 'rose', 'grocery', 'store', 'sales', 'In', 'December',
'overall', 'retail', 'sales', 'rose', 'Excluding', 'car', 'sector', 'sales', 'ros
e', 'Parul', 'Jain', 'deputy', 'chief', 'economist', 'Nomura', 'Securities', 'Int
ernational', 'said', 'consumer', 'spending', 'continue', 'rise', '2005', 'slowe
r', 'rate', 'growth', '2004', 'Consumers', 'continue', 'retain', 'strength', 'qua
rter', 'said', 'Van', 'Rourke', 'bond', 'strategist', 'Popular', 'Securities', 'a
greed', 'latest', 'retail', 'sales', 'figures', 'slightly', 'stronger', 'expecte
d']
Adjectives:  ['retail', 'biggest', 'monthly', 'last', 'heavy', 'generous', 'speci
al', 'retail', 'retail', 'chief', 'decent', 'second', 'healthy', 'electronic', 'g
eneral', 'strong', 'overall', 'retail', 'deputy', 'chief', 'slower', 'first', 'la
test', 'retail', 'stronger']
Organisations:  ['SG Corporate Investment Banking', 'Nomura Securities Internatio
nal', 'Consumers', 'Popular Securities']
Dates:  ['January', 'monthly', 'last August', 'December 4', 'January', '2005', '2
004', 'January', 'the second half of 2004', 'January', 'Christmas', 'December',
'2005', '2004', 'the first quarter']
```

Висновок:

В ході виконання даної лабораторної роботи я ознайомився з основами аналізу настроїв у текстових даних за допомогою мови програмування Python та бібліотеки spaCy. Завдання полягало в аналізі настроїв в даних з Twitter та обробці тексту з файлу, що включало визначення настрою коментарів, виявлення слів, що не є стоп-словами, прикметників, організацій та дат. Використання наївного байєсового класифікатора дозволило провести класифікацію коментарів на негативні, нейтральні та позитивні з достатньою точністю, що демонструє ефективність цього методу для аналізу настроїв. Результати роботи показали, що методи обробки та аналізу текстових даних можуть бути ефективно застосовані для вирішення практичних завдань, таких як аналіз настроїв. Виконання цієї лабораторної роботи дало мені цінний досвід роботи з текстовими даними та їх аналізу, що буде корисним у моїй подальшій професійній діяльності.