Міністерство освіти і науки України

Національний технічний університет України

“Київський політехнічний інститут імені Ігоря Сікорського”

Факультет інформатики та обчислювальної техніки

Кафедра інформатики та програмної інженерії

Лабораторна робота №7

Аналіз текстів на мові Python

Тема: Знайомство з об’єктами бібліотеки spaCy.

Варіант: 1

Виконав
студент групи ІП-11:
Панченко С. В.

Перевірила:
Тимофєєва Ю. С

Київ 2023

ЗМІСТ

1 МЕТА ЛАБОРАТОРНОЇ РОБОТИ

Ознайомитись з вирішенням задач обробки природньої мови за допомогою бібліотеки spaCy.

2 ЗАВДАННЯ

Створити програму,яка:

1) Виконує завдання No 2 лабораторної роботи No1 за допомогою класу Matcher.
2) Виконує завдання відповідно до варіанту засобами бібліотеки spaCy.

Варіант 1.Файл lab7-1.txt.

a) Знайти та вивести стоп-слова, які присутні в тексті.
b) Знайти та вивести всі іменники, які присутні у тексті.
c) Знайти та вивести числа і дати, які присутні у тексті.

3 ВИКОНАННЯ

3.1 Завдання перше

Зчитаємо файл. with - оператор контексту, який автоматично закриває файл.

```
In [143…
with open('text1.txt', 'r') as file:
    s = ''.join(file.readlines())
print(s)
```

```
At three  o'clock 12/05/1895 precisely I was at Baker Street, but Holmes had not
yet returned (005)-456-34-23. The landlady informed me that he baker_street@here.uk had le
ft the house
shortly after eight o'clock in the morning. I sat down beside the
fire, however, with the intention of awaiting him,, however long he
might be. 145 124 245 I was already 67-56-34 deeply interested in his inquiry, for, though
it was surrounded by none of the grim and strange features which
were Watson3@gmail.com associated with the two crimes which I have already recorded,
still, the nature of  the case and the exalted station of his client
gave it a character of its own 1896/01/23.. Indeed, apart from the nature of the
investigation which my friend had on hand, there was something in his
masterly 5618 4582 8225 1471 grasp of a situation, and his (03)-8-45-34 keen, incisive rea
soning, which
made it a pleasure to me to study his system of work, and to follow the
quick, subtle 4987 1514 6555 4212 methods by which he disentangled the most inextricable
mysteries. So accustomed was I ShHolmes@mail.uk to his invariable success that the very
possibility of his failing  had ceased to enter into my head.
```

Рисунок 3.1.1 - Зчитування файлу

Імпортуємо SpaCy та словник англійської мови.

```
In [144…
import spacy
from spacy.matcher import Matcher
py_nlp = spacy.load("en_core_web_sm")
```

Рисунок 3.1.2 - Імпортування SpaCy

Покажемо розбиття тексту на токени. Бачимо, що токени з текстами складно піддати обробці за допомогою звичайних атрибутів Matcher паттерна, тому використаємо атрибут REGEX.

```
In [145…
doc = py_nlp(s)
[tok.text for tok in doc]
```

```
Out[145]: ['At',
 'three',
 ' ',
 "o'clock",
 '12/05/1895',
 'precisely',
 'I',
 'was',
 'at',
 'Baker',
 'Street',
 ',',
 'but',
 'Holmes',
 'had',
 'not',
 '\n',
 'yet',
 'returned',
 '(',
 '005)-456',
 '-',
 '34',
 '-',
 '23',
 '.',
 'The',
 'landlady',
 'informed',
 'me'
```

Рисунок 3.1.3 - Токени

Визначимо паттерн та знайдемо всі номери телефонів.

```
In [146…
patterns = [
    [{"IS_DIGIT": True, 'LENGTH': 3, 'OP': '{3,}'}],
    [{"IS_DIGIT": True, 'LENGTH': 2},
     {'TEXT': '-'},
     {"IS_DIGIT": True, 'LENGTH': 2},
     {'TEXT': '-'},
     {"IS_DIGIT": True, 'LENGTH': 2}],
    [{'TEXT': '('},
     {'TEXT': {'REGEX': r'\d\d\d[)][-]\d\d\d'}},
     {'TEXT': '-'},
     {"IS_DIGIT": True, 'LENGTH': 2},
     {'TEXT': '-'},
     {"IS_DIGIT": True, 'LENGTH': 2}],
    [{'TEXT': '('},
     {'TEXT': {'REGEX': r'\d\d[)][-]\d'}},
     {'TEXT': '-'},
     {"IS_DIGIT": True, 'LENGTH': 2},
     {'TEXT': '-'},
     {"IS_DIGIT": True, 'LENGTH': 2}]
]
matcher = Matcher(py_nlp.vocab)
matcher.add("PROPER_PHONE_NUMBER", patterns)
matches = matcher(doc)
for match in matches:
    print(match, doc[match[1]:match[2]])
```

```
(3884770101205390969, 19, 25) (005)-456-34-23
(3884770101205390969, 71, 74) 145 124 245
(3884770101205390969, 77, 82) 67-56-34
(3884770101205390969, 179, 185) (03)-8-45-34
```

Рисунок 3.1.4 - Знаходження номерів телефонів

Замінимо цифри на зірочки.

In [147…
```
tokens = [el.text for el in doc]
for match in matches:
    found_first_digit = False
    elememts = []
    tt = doc[match[1]:match[2]]
    for i, el in enumerate(tt):
        chars = list(el.text)
        for j, c in enumerate(chars):
            if c.isdigit() and not found_first_digit:
                found_first_digit = True
                continue
            elif c.isdigit() and found_first_digit:
                chars[j] = '*'
        elememts.append(''.join(chars))
    s = s.replace(tt.text, f''.join(elememts))
print(s)
```

```
At three  o'clock 12/05/1895 precisely I was at Baker Street, but Holmes had not
yet returned (0**)-***-**-**. The landlady informed me that he baker_street@here.uk had le
ft the house
shortly after eight o'clock in the morning. I sat down beside the
fire, however, with the intention of awaiting him,, however long he
might be. 1******** I was already 6*-**-** deeply interested in his inquiry, for, though
it was surrounded by none of the grim and strange features which
were Watson3@gmail.com associated with the two crimes which I have already recorded,
still, the nature of  the case and the exalted station of his client
gave it a character of its own 1896/01/23.. Indeed, apart from the nature of the
investigation which my friend had on hand, there was something in his
masterly 5618 4582 8225 1471 grasp of a situation, and his (0*)-*-**-** keen, incisive rea
soning, which
made it a pleasure to me to study his system of work, and to follow the
quick, subtle 4987 1514 6555 4212 methods by which he disentangled the most inextricable
mysteries. So accustomed was I ShHolmes@mail.uk to his invariable success that the very
possibility of his failing  had ceased to enter into my head.
```

Рисунок 3.1.5 - Заміна тексту

3.2Завдання друге

Зчитаємо файл.

In [148…
```
with open('lab7-1.txt', 'r') as file:
    s = ''.join(file.readlines())
doc = py_nlp(s)
print(s)
```

Out[148]:
```
'US retail sales fell 0.3% in January, the biggest monthly decline since last August, dri
ven down by a heavy fall in car sales.  The 3.3% fall in car sales had been expected, com
ing after December\'s 4% rise in car sales, fuelled by generous pre-Christmas special off
ers. Excluding the car sector, US retail sales were up 0.6% in January, twice what some a
nalysts had been expecting. US retail spending is expected to rise in 2005, but not as qu
ickly as in 2004.  Steve Gallagher, US chief economist at SG Corporate & Investment Banki
ng, said January\'s figures were "decent numbers".  "We are not seeing the numbers that w
e saw in the second half of 2004, but they are still pretty healthy," he added. Sales at
appliance and electronic stores were down 0.6% in January, while sales at hardware stores
dropped by 0.3% and furniture store sales dipped 0.1%. Sales at clothing and clothing acc
essory stores jumped 1.8%, while sales at general merchandise stores, a category that inc
ludes department stores, rose by 0.9%. These strong gains were in part put down to consum
ers spending gift vouchers they had been given for Christmas.  Sales at restaurants, bars
and coffee houses rose by 0.3%, while grocery store sales were up 0.5%. In December, over
all retail sales rose by 1.1%. Excluding the car sector, sales rose by just 0.3%. Parul J
ain, deputy chief economist at Nomura Securities International, said consumer spending wo
uld continue to rise in 2005, only at a slower rate of growth than in 2004. "Consumers co
ntinue to retain their strength in the first quarter," he said. Van Rourke, a bond strate
gist at Popular Securities, agreed that the latest retail sales figures were "slightly st
ronger than expected". '
```

Рисунок 3.2.1 - Зчитування файлів

Знайдемо та виведемо стоп-слова, які присутні у тексті.

In [149…
```
stop_words = [token.text for token in doc if token.is_stop]
stop_words
```

Out[149]:
```
['US',
 'in',
 'the',
 'since',
 'last',
 'down',
 'by',
 'a',
 'in',
 'The',
 'in',
 'had',
 'been',
 'after',
 "'s",
 'in',
 'by',
 'the',
 'US',
 'were',
 'up',
 'in',
 'what',
 'some',
 'had',
 'been',
 'US',
 'is',
 'to',
 'in',
 'but',
 'not',
 'as',
 'as',
 'in',
 'US',
 'at',
 "'s",
 'were',
 'We',
 'are',
 'not',
 'the',
 'that',
 'we',
 'in',
 'the',
 'of',
 'but',
 'they',
 'are',
 'still',
 'he',
 'at',
 'and',
 'were',
 'down',
 'in',
 'while',
 'at',
 'by',
 'and',
 'at',
 'and',
 'while',
 'at',
 'a',
 'that',
 'by',
 'These',
 'were',
 'in',
 'part',
 'put',
 'down',
 'to',
 'they',
 'had',
 'been',
 'for',
 'at',
```

Рисунок 3.2.2 - Стоп-слова

Знайдемо та виведемо всі іменники з тексту.

```
In [150…  nouns = [token for token in doc if token.pos_ == 'NOUN' and token.text.isalpha()]
          nouns

Out[150]: [sales,
           decline,
           fall,
           car,
           sales,
           fall,
           car,
           sales,
           rise,
           car,
           sales,
           offers,
           car,
           sector,
           sales,
           analysts,
           spending,
           economist,
           figures,
           numbers,
           numbers,
           half,
           Sales,
           appliance,
           stores,
           sales,
           hardware,
           stores,
           furniture,
           store,
           sales,
           Sales,
           clothing,
           clothing,
           accessory,
           stores,
           sales,
           merchandise,
           stores,
           category,
           department,
           stores,
           gains,
           part,
           consumers,
           gift,
           vouchers,
           Sales,
           restaurants,
           bars,
           coffee,
           houses,
           grocery,
           store,
           sales,
           sales,
           car,
           sector,
           sales,
           economist,
           consumer,
           spending,
           rate,
           growth,
           Consumers,
           strength,
           quarter,
           bond,
           strategist,
           sales,
           figures]
```

Рисунок 3.2.3 - Іменники

Виведемо числа та дати.

In [151…

```
matcher = Matcher(py_nlp.vocab)
patterns = [[{'LIKE_NUM': True}]]
matcher.add("PROPER_PHONE_NUMBER", patterns)
matches = matcher(doc)
for match in matches:
    print(match, doc[match[1]:match[2]])
```

```
(3884770101205390969, 4, 5) 0.3
(3884770101205390969, 29, 30) 3.3
(3884770101205390969, 43, 44) 4
(3884770101205390969, 69, 70) 0.6
(3884770101205390969, 90, 91) 2005
(3884770101205390969, 98, 99) 2004
(3884770101205390969, 137, 138) second
(3884770101205390969, 140, 141) 2004
(3884770101205390969, 161, 162) 0.6
(3884770101205390969, 173, 174) 0.3
(3884770101205390969, 180, 181) 0.1
(3884770101205390969, 191, 192) 1.8
(3884770101205390969, 210, 211) 0.9
(3884770101205390969, 244, 245) 0.3
(3884770101205390969, 253, 254) 0.5
(3884770101205390969, 264, 265) 1.1
(3884770101205390969, 276, 277) 0.3
(3884770101205390969, 298, 299) 2005
(3884770101205390969, 309, 310) 2004
(3884770101205390969, 320, 321) first
```

Рисунок 3.2.4 - Числа та дати

ДОДАТОК А ТЕКСТИ ПРОГРАМНОГО КОДУ

Тексти програмного коду

(Найменування програми (документа))

Жорсткий диск

(Вид носія даних)

(Обсяг програми (документа), арк.)

Студента групи ІП-113 курсу

Панченка С. В

```
with open('text1.txt', 'r') as file:
s = ''.join(file.readlines())
print(s)
import spacy
from spacy.matcher import Matcher
py_nlp = spacy.load("en_core_web_sm")
doc = py_nlp(s)
[tok.text for tok in doc]
patterns = [
[{"IS_DIGIT": True, 'LENGTH': 3, 'OP': '{3,}'}],
[{"IS_DIGIT": True, 'LENGTH': 2},
{'TEXT': '-'},
{"IS_DIGIT": True, 'LENGTH': 2},
{'TEXT': '-'},
{"IS_DIGIT": True, 'LENGTH': 2}],
[{'TEXT': '('},
{'TEXT': {'REGEX': r'\d\d\d[)][-]\d\d\d'}},
{'TEXT': '-'},
{"IS_DIGIT": True, 'LENGTH': 2},
{'TEXT': '-'},
{"IS_DIGIT": True, 'LENGTH': 2}],
[{'TEXT': '('},
{'TEXT': {'REGEX': r'\d\d[)][-]\d'}},
{'TEXT': '-'},
{"IS_DIGIT": True, 'LENGTH': 2},
{'TEXT': '-'},
{"IS_DIGIT": True, 'LENGTH': 2}]
]
matcher = Matcher(py_nlp.vocab)
matcher.add("PROPER_PHONE_NUMBER", patterns)
matches = matcher(doc)
for match in matches:
print(match, doc[match[1]:match[2]])
tokens = [el.text for el in doc]
for match in matches:
found_first_digit = False
elememts = []
tt = doc[match[1]:match[2]]
for i, el in enumerate(tt):
chars = list(el.text)
for j, c in enumerate(chars):
if c.isdigit() and not found_first_digit:
found_first_digit = True
continue
```

```
elif c.isdigit() and found_first_digit:
chars[j] = '*'
elememts.append(''.join(chars))
s = s.replace(tt.text, f''.join(elememts))
print(s)
with open('lab7-1.txt', 'r') as file:
s = ''.join(file.readlines())
doc = py_nlp(s)
print(s)
stop_words = [token.text for token in doc if token.is_stop]
stop_words
nouns = [token for token in doc if token.pos_ == 'NOUN' and token.text.isalpha()]
nouns
matcher = Matcher(py_nlp.vocab)
patterns = [[{'LIKE_NUM': True}]]
matcher.add("PROPER_PHONE_NUMBER", patterns)
matches = matcher(doc)
for match in matches:
print(match, doc[match[1]:match[2]])
```