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 №7

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: Знайомство з об’єктами бібліотеки spaCy

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

Варіант: 11

Завдання:

Створити програму, яка:

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

   Файл *lab7-3.txt*.

   1. Знайти та вивести стоп-слова, які присутні у тексті
   2. Знайти та вивести всі іменники, які присутні у тексті
   3. Знайти та вивести всі числа і організації, які присутні у тексті.

Код програми

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

import spacy
from spacy import displacy
from spacy.matcher import Matcher

text = ""
with open("text3.txt", "r") as file:
for line in file:
# Normalize some phone numbers
line = line.replace(')', ' )').replace('(', '( ').replace('-', ' - ')
text = text + line.strip()

text2 = ""
with open("lab7-3.txt", "r") as file:
for line in file:
text2 = text2 + line.strip()

nlp = spacy.load("en_core_web_sm")
doc = nlp(text)
doc2 = nlp(text2)
print([token.text for token in doc])
print([token.text for token in doc2])

# Task 1. Match phone numbers with Matcher
matcher = Matcher(nlp.vocab)
pattern_phone = [
[
```

```
{'TEXT': '(', 'OP': '+'},
{'TEXT': {'REGEX': '^\\d{2,3}$'}},
{'TEXT': ')', 'OP': '+'},
{'TEXT': '-', 'OP': '?'},
{'TEXT': {'REGEX': '^\\d{2,3}$'}},
{'TEXT': '-', 'OP': '?'},
{'TEXT': {'REGEX': '^\\d{2,3}$'}}
],
[
# {'TEXT': '(', 'OP': '?'},
{'TEXT': {'REGEX': '^\\d{2,3}'}},
# {'TEXT': ')', 'OP': '?'},
{'TEXT': '-', 'OP': '?'},
{'TEXT': {'REGEX': '^\\d{2,3}$'}},
{'TEXT': '-', 'OP': '?'},
{'TEXT': {'REGEX': '^\\d{2,3}$'}}
]
]
matcher.add('phoneNum', pattern_phone)
matches = matcher(doc)
for match_id, start, end in matches:
m_span = doc[start:end]
print(start, end, '\t', m_span.text)
# -----------
# Task 2.
# Find and display stop-words, which are present
print("Stop words:")
for token in doc2:
if token.is_stop:
print(token.text + ", ", end='')

print()

# Find and display all nouns, which are present
print("Nouns:")
for token in doc2:
if token.tag_ == 'NN':
print(token, end=', ')

print()

# Find and display all numbers and organizations, which are present
print("Numbers and organizations:")
ner_tagged = [(word.text, word.ent_type_) for word in doc2]
for tag in ner_tagged:
if tag[1] == 'ORG' or tag[1] == 'CARDINAL':
print(tag)
```

Результат роботи програми


```
['Slowly', 'and', 'solemnly', 'he', 'was', 'borne', 'into', 'Briony', 'Lodge', '145a', 'and',
'laid', 'out', 'in', 'theprincipal', 'room', ',', 'while', 'I', '(', '094', ')', '-', '44', '-',
'33', 'still', 'observed', 'the', 'proceedings', 'from', 'my', 'postby', 'the', 'window',
'12/15/1892', '.', 'The', 'lamps', 'had', 'been', 'lit', ',', 'but', 'the', 'blinds', 'had', 'not',
'beendrawn', ',', 'so', 'that', 'I', 'could', 'see', 'Holmes', 'sh_holmes221@mail.uk', 'as', 'he',
'lay', 'upon', 'the', 'couch', '.', 'I', 'do', 'notknow', 'whether', 'he', 'was', 'seized', 'with',
'compunction', 'at', 'that', 'moment', 'for', 'the', 'parthe', 'was', 'playing', ',', 'but', 'I',
'know', 'that', 'I', 'never', 'felt', 'more', 'heartily', 'ashamed', 'ofmyself', 'in', 'my', 'life',
'34', '-', '34', '-', '32', 'than', 'when', 'I', 'saw', 'the', 'beautiful', 'creature', 'against',
'whomI', 'wston23@mymail.com', 'was', 'conspiring', ',', 'or', 'the', 'grace', 'and', 'kindliness',
'with', 'which', 'she', 'waitedupon', 'the', 'injured', 'man', '.', 'And', 'yet', 'it', 'would',
'be', 'the', 'blackest', 'treachery', 'to', '1895', '-', '11', '-', '17Holmes', 'to', 'draw',
'back', 'now', 'from', 'the', 'part', 'which', 'he', 'had', 'intrusted', 'to', 'me', 'Avenue',
'123.I', 'hardened', 'my', 'heart', 'Wtason', 'Street', '45b', ',', 'and', 'took', 'the', 'smoke',
'-', 'rocket', 'from', 'under', 'my', 'ulster', '.', 'After', 'all', ',', '344', '-', '21', '-',
'01', 'I', 'thought', ',', 'we', 'are', 'not', 'injuring', 'her', '.', 'We', 'are', 'but',
'preventingher', 'from', 'injuring', 'another', 'adlerrr@mail.com', '.']
['Gordon', 'Brown', 'has', 'issued', 'a', 'rallying', 'cry', ',', 'telling', 'supporters', 'the',
'"', 'stakes', 'are', 'too', 'high', '"', 'to', 'stay', 'at', 'home', 'or', 'protest', 'vote', 'in',
'the', 'forthcoming', 'general', 'election', '.', ' ', 'The', 'chancellor', 'said', 'the', 'vote',
'-', 'expected', 'to', 'fall', 'on', '5', 'May', '-', 'will', 'give', 'a', '"', 'clear', 'and',
'fundamental', '"', 'choice', 'between', 'Labour', 'investment', 'and', 'Conservative', 'cuts', '.',
'Speaking', 'at', 'Labour', "'s", 'spring', 'conference', 'in', 'Gateshead', ',', 'Mr', 'Brown',
'claimed', 'the', 'NHS', 'was', 'not', 'safe', 'in', 'Conservative', 'hands', '.', 'He', 'said',
'Tory', 'plans', 'to', 'cut', 'BJ35bn', 'tax', 'would', '"', 'cut', 'deep', 'into', 'public',
'service', '"', '.', ' ', 'To', 'a', 'packed', 'audience', 'at', 'Gateshead', "'s", 'Sage',
'Centre', ',', 'the', 'chancellor', 'said', 'the', 'cuts', 'proposed', 'by', 'shadow', 'chancellor',
'Oliver', 'Letwin', 'were', 'the', 'equivalent', 'of', 'sacking', 'every', 'teacher', ',', 'GP',
'and', 'nurse', 'in', 'the', 'country', ',', 'he', 'told', 'activists', '.', 'Laying', 'into',
'the', 'Conservative', "'s", 'record', 'in', 'government', 'he', 'said', ':', '"', 'I', 'give',
'you', 'this', 'promise', '-', 'with', 'Labour', ',', 'Britain', 'will', 'never', 'return', 'to',
'the', 'mistakes', 'of', 'ERM', 'and', '10', '%', 'inflation', ',', '15', '%', 'interest', 'rates',
',', 'BJ3bn', 'in', 'lost', 'reserves', ',', '250,000', 'repossessed', ',', 'one', 'million', 'in',
'negative', 'equity', 'and', 'three', 'million', 'unemployed', '.', '"', 'Never', 'again', 'Tory',
'boom', 'and', 'bust', '.', ' ', '"', 'This', 'will', 'be', 'the', 'central', 'dividing', 'line',
'at', 'the', 'election', ',', 'between', 'a', 'Conservative', 'Party', 'taking', 'Britain', 'back',
'and', 'planning', 'deep', 'cuts', 'of', 'BJ35bn', 'in', 'our', 'services', ',', 'and', 'a',
'Labour', 'government', 'taking', 'Britain', 'forward', ',', 'which', 'on', 'a', 'platform', 'of',
'stability', 'will', 'reform', 'and', 'renew', 'our', 'hospitals', ',', 'schools', 'and', 'public',
'services', 'and', ',', 'I', 'am', 'proud', 'to', 'say', ',', 'spend', 'by', '2008', 'BJ60bn',
'more', '.', '"', 'Turning', 'to', 'the', 'economy', ',', 'the', 'chancellor', 'pledged', 'to',
'continue', 'economic', 'stability', 'and', 'growth', 'in', 'a', 'third', 'term', 'in', 'power',
'.', ' ', 'He', 'said', 'after', 'seven', 'years', 'Labour', 'had', 'transformed', 'from', 'a',
'party', 'not', 'trusted', 'with', 'the', 'economy', 'to', '"', 'the', 'only', 'party', 'trusted',
'with', 'the', 'economy', '"', '.', 'It', 'was', 'now', 'a', '"', 'party', 'not', 'just', 'of',
'employees', ',', 'but', 'of', 'employers', 'and', 'managers', '"', ',', 'he', 'said', '.', 'In',
'the', 'speech', '-', 'which', 'prompted', 'a', 'standing', 'ovation', 'from', 'an', 'audience',
'clearly', '"', 'warm', '"', 'to', 'Mr', 'Brown', '-', 'he', 'also', 'promised', 'to', 'end',
'teenage', 'unemployment', 'within', 'the', 'next', 'five', 'years', '.', 'He', 'also',
'highlighted', 'plans', 'for', '100', '%', 'debt', 'relief', 'for', 'the', 'world', "'s", 'poorest',
'countries', ',', 'a', 'national', 'minimum', 'wage', 'for', '16', 'and', '17', '-', 'year', '-',
'olds', 'and', 'the', 'creation', 'of', 'a', 'network', 'of', 'children', "'s", 'centres', 'and',
```

'flexibility', 'in', 'maternity', 'leave', '.', 'The', 'prime', 'minister', 'is', 'to', 'take', 'part', 'later', 'on', 'Saturday', 'in', 'an', 'interactive', 'question', 'and', 'answer', 'session', ',', 'fielding', 'queries', 'sent', 'in', 'by', 'e', '-', 'mail', ',', 'text', 'message', 'and', 'telephone', 'as', 'part', 'of', 'Labour', "'s", 'attempt', 'to', 'engage', 'the', 'public', 'in', 'their', 'campaign', '.']

19 26 (094) - 44 - 33

96 101 34 - 34 - 32

177 182 344 - 21 - 01

Stop words:

has, a, the, are, too, to, at, or, in, the, The, the, to, on, May, will, give, a, and, between, and, at, 's, in, the, was, not, in, He, to, would, into, To, a, at, 's, the, the, by, were, the, of, every, and, in, the, he, into, the, 's, in, he, I, give, you, this, with, will, never, to, the, of, and, in, one, in, and, three, Never, again, and, This, will, be, the, at, the, between, a, back, and, of, in, our, and, a, which, on, a, of, will, and, our, and, and, I, am, to, say, by, more, to, the, the, to, and, in, a, third, in, He, after, had, from, a, not, with, the, to, the, only, with, the, It, was, now, a, not, just, of, but, of, and, he, In, the, which, a, from, an, to, he, also, to, within, the, next, five, He, also, for, for, the, 's, a, for, and, and, the, of, a, of, 's, and, in, The, is, to, take, part, on, in, an, and, in, by, and, as, part, of, 's, to, the, in, their,

Nouns:

cry, home, protest, vote, election, chancellor, vote, choice, investment, spring, conference, tax, service, audience, chancellor, equivalent, teacher, nurse, country, record, government, promise, %, inflation, %, interest, equity, boom, bust, dividing, line, election, BJ35bn, government, platform, stability, economy, chancellor, stability, growth, term, power, party, economy, party, economy, party, speech, ovation, audience, unemployment, %, debt, relief, world, minimum, wage, year, creation, network, flexibility, maternity, leave, minister, part, question, answer, session, e, -, mail, text, message, telephone, part, attempt, public, campaign,

Numbers and organizations:

('5', 'CARDINAL')

('Labour', 'ORG')

('Labour', 'ORG')

('NHS', 'ORG')

('Gateshead', 'ORG')

("'s", 'ORG')

('Sage', 'ORG')

('Centre', 'ORG')

('GP', 'ORG')

('Conservative', 'ORG')

('ERM', 'ORG')

('BJ3bn', 'ORG')

('250,000', 'CARDINAL')
('one', 'CARDINAL')
('million', 'CARDINAL')
('three', 'CARDINAL')
('million', 'CARDINAL')
('Conservative', 'ORG')
('Party', 'ORG')
('Labour', 'ORG')
('16', 'CARDINAL')
('Labour', 'ORG')