МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ “ЛЬВІВСЬКА ПОЛІТЕХНІКА”
ІНСТИТУТ КОМП’ЮТЕРНИХ НАУК ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра “Системи автоматизованого проектування”

Звіт

до лабораторної роботи №5

на тему: “ ВИВЧЕННЯ БІБЛІОТЕКИ ПРИКЛАДНИХ ПРОГРАМ NLTK ДЛЯ ОПРАЦЮВАННЯ ТЕКСТІВ ПРИРОДНОЮ МОВОЮ. ДОСТУП ТА РОБОТА З КОРПУСАМИ ТЕКСТІВ.”

з дисципліни “Комп’ютерна лінгвістика”

Виконала:
студентка групи ПРЛм-11
Зварич О.І.
Прийняв:
викладач
Дупак Б.П.

Львів-2015

МЕТА РОБОТА

- Вивчення основ програмування на мові *Python*.
- Вивчення методів роботи з файлами на локальних дисках та з Інтернету.
- Використання Юнікоду при обробці текстів.
- Нормалізація текстів, стемінг, лематизація та сегментація.

1. Напишіть функцію, яка приймає адресу URL, як аргумент, і повертає те що міститься за цією адресою з видаленням HTML розмітки. Використовувати `urllib.urlopen` для доступу до контенту наступним чином `raw_contents = urllib.urlopen('http://www.nltk.org/').read()`.

```
import urllib, nltk
from urllib import urlopen

def Converter(url):
    html=urlopen(url).read()
    raw=nltk.clean_html(html)
    t=nltk.word_tokenize(raw)
    return t
print Converter("http://www.bbc.com/news/world-europe-34501617")
```

```
['Turkish', 'air', 'strikes', 'on', 'Kurdish', 'PKK', 'rebels', 'as', 'mourning'
, 'continues', '-', 'BBC', 'News', 'Accessibility', 'links', 'Skip', 'to', 'cont
ent', 'Accessibility', 'Help', 'BBC', 'iD', 'BBC', 'navigation', 'News', 'News',
 'Sport', 'Weather', 'Shop', 'Earth', 'Travel', 'Capital', 'iPlayer', 'Culture',
 'Autos', 'Future', 'TV', 'Radio', 'CBBC', 'CBeebies', 'Arts', 'Make', 'It', 'Di
gital', 'Food', 'iWonder', 'Bitesize', 'Music', 'Nature', 'Earth', 'Local', 'Tra
vel', 'Menu', 'Search', 'the', 'BBC', 'BBC', 'News', 'News', 'navigation', 'Sect
ions', 'Home', 'Video', 'World', 'selected', 'UK', 'Business', 'Tech', 'Science'
, 'Magazine', 'Entertainment', '&', 'amp', ';', 'Arts', 'Health', 'In', 'Picture
s', 'Also', 'in', 'the', 'News', 'Special', 'Reports', 'Explainers', 'The', 'Rep
orters', 'Have', 'Your', 'Say', 'World', 'selected', 'Africa', 'Asia', 'Australi
a', 'Europe', 'selected', 'Latin', 'America', 'Middle', 'East', 'US', '&', 'amp'
, ';', 'Canada', 'Europe', 'Europe', 'Turkish', 'air', 'strikes', 'on', 'Kurdish
', 'PKK', 'rebels', 'as', 'mourning', 'continues', '12', 'October', '2015', 'Fro
m', 'the', 'section', 'Europe', 'Media', 'caption', 'Crowds', 'gathered', 'in',
```

2.Збережіть деякий текст у файлі `corpus.txt`. Визначити функцію `load(f)` для читання файлу, назва якого є її аргументом і повертає стрічку, яка містить текст з файлу.

```
def load(f):
    p=open(f)
    for line in p:
            print line.strip()
print load('corpus.txt')
```

```
Three Israelis have been killed and more than 20 injured in shooting and stabbin
g attacks in Jerusalem and central Israel, Israeli police say.
Two were killed when two assailants, who were identified as Arabs, shot and stab
bed passengers on a bus in Jerusalem before being shot by police.
Another Israeli died after being run down and stabbed elsewhere in the city.
Near-daily stabbings by Palestinians have left dozens of Israelis dead and wound
ed over the past fortnight.
Several attackers and at least 17 other Palestinians have been killed in the ups
urge of violence.
Israeli Prime Minister Benjamin Netanyahu convened an emergency session of the s
ecurity cabinet to discuss how to prevent further attacks.
```

3.Перепишіть наступний цикл як list comprehension:

```
>>> sent = ['The', 'dog', 'gave', 'John', 'the', 'newspaper']
>>> result = []
>>> for word in sent:
```

```
...     word_len = (word, len(word))
...     result.append(word_len)
>>> result
[('The', 3), ('dog', 3), ('gave', 4), ('John', 4), ('the', 3), ('newspaper',
9)]
```

```
sent=['The', 'dog', 'gave', 'John', 'the', 'newspaper']
result = []
for word in sent:
    word_len=(word, len(word))
    result.append(word_len)
print result
sent=['The', 'dog', 'gave', 'John', 'the', 'shoes']
print [(word, len(word)) for word in sent]
```

```
>>>
[('The', 3), ('dog', 3), ('gave', 4), ('John', 4), ('the', 3), ('newspaper', 9)]
[('The', 3), ('dog', 3), ('gave', 4), ('John', 4), ('the', 3), ('shoes', 5)]
>>>
```

4.Перевірити різницю між стрічками і цілим виконавши наступні дії: `"3" * 7` та `3 * 7`. Спробуйте здійснити конвертування між стрічками і цілими використавши `int("3")` та `str(3)`.

```
print '3' * 7
print 3 * 7
print int("3") * 7
print str(3) * 7
```

```
3333333
21
21
3333333
```

5.Що станеться, коли стрічки форматування `%6s` та `%-6s` використовується для відображення стрічки довшої ніж 6 символів?

```
print '%6s' % 'automobile'
print '%6s' % 'car'
print '%-6s' % 'automobile'
print '%-6s' % 'car'
```

```
>>>
automobile
   car
automobile
car
>>>
```

7. Створіть файл, який буде містити слова та їх частоту записані в окремих рядках через пробіл (fuzzy 53). Прочитайте цей файл використовуючи open(filename).readlines(). Розділіть кожну стрічку на дві частини використовуючи split(), і перетворіть число в ціле значення використовуючи int(). Результат повинен бути у вигляді списку: [['fuzzy', 53], ...].

```
>>> f = open('E:\Text\words.txt').readlines()
>>> f
['monday 15\n', 'tuesday 12 \n', 'wednesday 9\n'
, 'thursday 7 \n', 'friday 3 \n', 'saturday 26\n
', 'sunday 19\n']
>>> a=[]
>>> for i in f:
        b=i.split(' ')
        a+=[[b[0]]+[int(b[1])]]


>>> a
[['monday', 15], ['tuesday', 12], ['wednesday',
9], ['thursday', 7], ['friday', 3], ['saturday',
26], ['sunday', 19]]
>>>
```

```
>>> f = open('E:\Text\words.txt').readlines()
>>> f
['monday 15,\n', 'tuesday 12, \n', 'wednesday 9,\n', 'thursday 7, \n', 'friday 3
, \n', 'saturday 26, \n', 'sunday 19\n']
>>> for line in f:
         print line.split()


['monday', '15,']
['tuesday', '12,']
['wednesday', '9,']
['thursday', '7,']
['friday', '3,']
['saturday', '26,']
['sunday', '19']
```

8. Напишіть програму доступу до вебсторінки і вилучення з неї деякого тексту.

```
from __future__ import division
import nltk
import nltk, re, pprint
from urllib import urlopen
url = "http://news.bbc.co.uk/2/hi/health/2284783.stm"
html =urlopen(url).read()
raw = nltk.clean_html(html)
tokens = nltk.word_tokenize(raw)
tokens
```

```
['Turkish', 'air', 'strikes', 'on', 'Kurdish', 'PKK', 'rebels', 'as', 'mourning'
, 'continues', '-', 'BBC', 'News', 'Accessibility', 'links', 'Skip', 'to', 'cont
ent', 'Accessibility', 'Help', 'BBC', 'iD', 'BBC', 'navigation', 'News', 'News',
 'Sport', 'Weather', 'Shop', 'Earth', 'Travel', 'Capital', 'iPlayer', 'Culture',
 'Autos', 'Future', 'TV', 'Radio', 'CBBC', 'CBeebies', 'Arts', 'Make', 'It', 'Di
gital', 'Food', 'iWonder', 'Bitesize', 'Music', 'Nature', 'Earth', 'Local', 'Tra
vel', 'Menu', 'Search', 'the', 'BBC', 'BBC', 'News', 'News', 'navigation', 'Sect
ions', 'Home', 'Video', 'World', 'selected', 'UK', 'Business', 'Tech', 'Science'
, 'Magazine', 'Entertainment', '&', 'amp', ';', 'Arts', 'Health', 'In', 'Picture
s', 'Also', 'in', 'the', 'News', 'Special', 'Reports', 'Explainers', 'The', 'Rep
orters', 'Have', 'Your', 'Say', 'World', 'selected', 'Africa', 'Asia', 'Australi
a', 'Europe', 'selected', 'Latin', 'America', 'Middle', 'East', 'US', '&', 'amp'
, ';', 'Canada', 'Europe', 'Europe', 'Turkish', 'air', 'strikes', 'on', 'Kurdish
', 'PKK', 'rebels', 'as', 'mourning', 'continues', '12', 'October', '2015', 'Fro
m', 'the', 'section', 'Europe', 'Media', 'caption', 'Crowds', 'gathered', 'in',
```

10. Модуль random включає функцію choice(), яка випадковим чином вибирає елементи послідовності. Наприклад, choice("aehh ") буде вибирати один з чотирьох символів. Напишіть програму генерації стрічки з 500 випадково вибраних символів "aehh ". Для

поєднання елементів в стрічку використовуйте ".join() . Нормалізуйте отриманий результат використовуючи split() та join().

```
import random
a=[]
for i in range(500):
    a.append(random.choice("aehh "))
    ss=''.join(a)
print ss
```

```
eaeaehheaehhhahhheh hhehahhhaahahh hhe h ehaheaah a h h  eeeaahe a  hhahe ahha
h ehe  hhhhehhheheah  hhha heh hahahehhhahee ah hhhhah ahaea ahe  haae hhe ae eh
heaee  hhh aahheheeahhe  hheahhhhhhehh aae h ahe ahhhehha  e eae  eh ehhh  aehhh
a eehhaeea hhh eaeheeh h  hehaahhehahhhhh he h eahe hhaaeeahaa hhha h haehe  ea
heehh e ee  eehhh heha aehh  e hhha eaa  hh eeheahhhaaeeaahhhhaaaaahheeh  e ehhe
hha  hhhehhhha e eahheehe hee h hehaheahhe  hahhaeaah h h hhahha e hheeaae  hhhh
eaaehhahahhaaehahh h
```

14. Доступіться до текстів ABC Rural News та ABC Science News з корпуса (nltk.corpus.abc). Знайдіть значення для оцінки читабельності текстів (аналогічно до задачі №12). Використовуйте Punkt для поділу тексту на окремі речення.

```
from __future__ import division
import nltk, re, pprint
from nltk.corpus import abc
l=0
n=0
for w in nltk.corpus.abc.words():
    l+=len(w)
    n+=1

m1=l/n
print m1
m2=len(nltk.corpus.abc.words())/len(nltk.corpus.abc.sents())
print m2
ari=4.71*m1+0.5*m2-21.43
print ari
sent_tokenizer=nltk.data.load('tokenizers/punkt/english.pickle')
text=nltk.corpus.abc.raw('rural.txt')
sents=sent_tokenizer.tokenize(text)
pprint.pprint(sents[1:5])
text=nltk.corpus.abc.raw('science.txt')
sents=sent_tokenizer.tokenize(text)
pprint.pprint(sents[1:5])
```

```
4.39119446332
26.0785267311
12.2917892878
['Letters from John Howard and Deputy Prime Minister Mark Vaile to AWB have been
 released by the Cole inquiry into the oil for food program.',
 'In one of the letters Mr Howard asks AWB managing director Andrew Lindberg to
remain in close contact with the Government on Iraq wheat sales.',
 "The Opposition's Gavan O'Connor says the letter was sent in 2002, the same tim
e AWB was paying kickbacks to Iraq though a Jordanian trucking company.",
 'He says the Government can longer wipe its hands of the illicit payments, whic
h totalled $290 million.']
["That's the conclusion of two studies published in this week's issue of The New
 England Journal of Medicine.",
 'They found that inhaling a mist with a salt content of 7 or 9% improved lung f
unction and, in some cases, produced less absenteeism from school or work.',
 'Cystic fibrosis, a progressive and frequently fatal genetic disease that affec
ts about 30,000 young adults and children in the US alone, is marked by a thicke
ning of the mucus which makes it harder to clear the lungs of debris and bacteri
a.',
 'The salt water solution "really opens up a new avenue for approaching patients
 with cystic fibrosis and how to treat them," says Dr Gail Weinmann, of the US N
ational Heart, Lung, and Blood Institute, which sponsored one of the studies.']
>>>
```

Висновок: я освоїла основи програмування на мові *Python*,вивчила методів роботи з файлами на локальних дисках та з Інтернету. Навчилася використовувати Юнікод при обробці текстів. Ознайомилась з нормалізацією текстів, стемінгом, лематизацією та сегментацією.