МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ "ЛЬВІВСЬКА ПОЛІТЕХНІКА"
ІНСТИТУТ КОМП'ЮТЕРНИХ НАУК ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра "Системи автоматизованого проектування"

Звіт

до лабораторної роботи №3

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
Гарбуз Л.В.
Прийняв:
викладач
Дупак Б.П.

Львів 2015

Мета роботи: вивчення основ програмування на мові Python. Вивчення методів доступу до корпусів текстів. Вивчення класу ConditionalFreqDist.

Теоретичні відомості:

Вирішення задач обробки текстів природною мовою передбачає використання великих об'ємів лінгвістичних даних, або інишими словами передбачає роботу з корпусами текстів. Виконання даної лабораторної роботи допоможе знайти відповідь на наступні питання: які є відомі корпуси текстів та лексичні ресурси і як отримати до них доступ використовуючи Python; які корисні конструкції має Python для виконання цієї роботи.

Корпус текстів це великий набір текстів. Багато корпусів розроблені їз збереженням балансу між текстами різних жанрів, або авторів. В попередній лабораторній роботі ми працювали з промовами президентів США, які є частиною корпуса US Presidential Inaugural Addresses. З промовами ми працювали, як з одним текстом не зважаючи на те, що кожна промова має окремого автора.Обробку ми здійснювали . При роботі з копусами важливо мати засоби доступу як до окремих тексті так і до окремих частин цих тексів а також і до окремих слів.

В NLTK входить невелика частина текстів з електронного архіву текстів Project Gutenberg , який містить 25000 безкоштовних електронних книжок різних авторів (http://www.gutenberg.org/). Тексти творів в окремих файлах. Для одержання назв файлів (ідентифікаторів файлів) в яких зберігаються текстів потрібно використати наступну функцію:

```
>>> import nltk
>>> nltk.corpus.gutenberg.fileids()
```

Корпус Brown – це перший корпус англійської мови об'ємом один мільйон слів було створено в 1961-1964 роках в університеті Brown. Цей корпус містить тексти з 500 різних джерел, які відповідають різним жанрам. В Табл.1. наведено приклади для кожного з жанрів.

Таблиця 1

Приклади текстів для кожного з жанрів корпуса Brown.

ID	Файл	Жанр	Опис тексту
A16	ca16	news	Chicago Tribune: *Society Reportage*
B02	cb02	editorial	Christian Science Monitor: *Editorials*
C17	cc17	reviews	Time Magazine: *Reviews*
D12	cd12	religion	Underwood: *Probing the Ethics of Realtors*
E36	ce36	hobbies	Norling: *Renting a Car in Europe*

Використовуючи засоби NLTK можна отримати доступ до цього корпусу, як до списку слів або списку речень (кожне речення – список слів). Також доступна можливість вибору текстів окремої категорії або з окремого файлу.

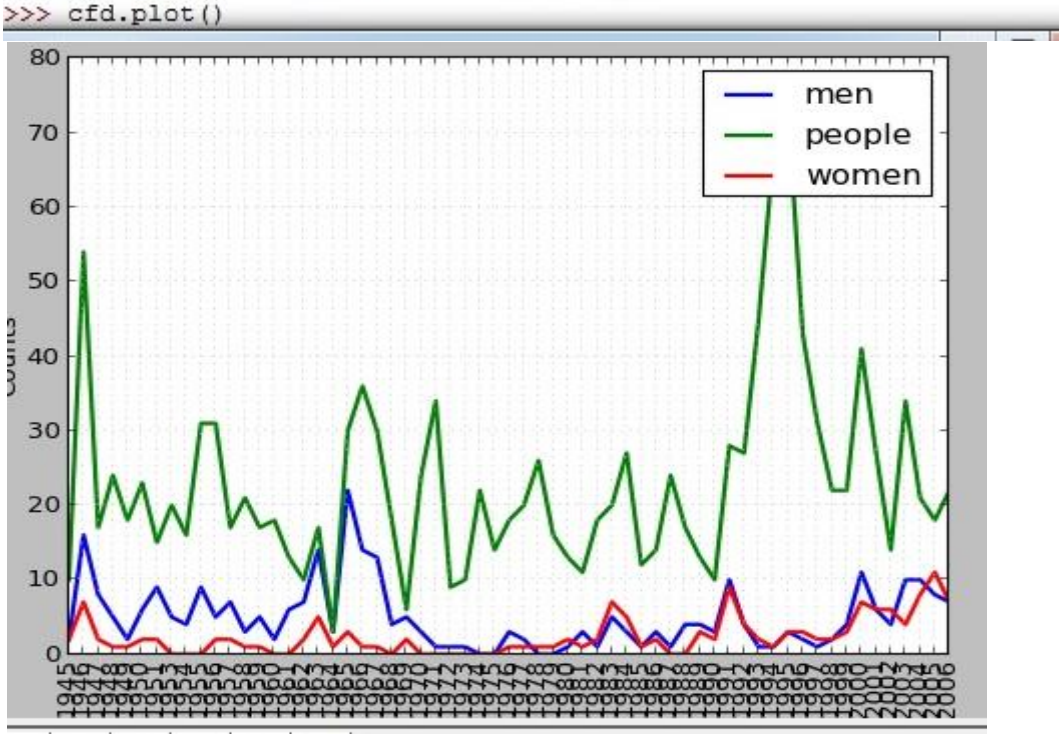
Корпус Brown – зручний ресурс для систематичного вивчення відмінностей між жанрами, або іншими словами для дослідження стилістики текстів. Спробуємо порівняти жанри і встановити , яким чином в текстах різних жанрів використовуються модальні дієслова. Для цього потрібно зробити підрахунки вживання різних модальних дієслів для різних жанрів.

Тексти програм на мові *Python*.

Варіант – 3

3. Прочитайте тексти з корпусу State of the Union addresses використовуючи state_union модуль читання. Визначити частоту вживання слів men, women, people в кожному з документів. Як змінилася частота вживання цих слів з часом?

```
>>> import nltk
>>> from nltk.corpus import state_union
>>> cfd=nltk.ConditionalFreqDist(
        (target, fileid[:4])
        for fileid in state_union.fileids()
        for w in state_union.words(fileid)
        for target in ['men', 'women', 'people']
        if w.lower().startswith(target))
>>> cfd.plot()
```



5. Виберіть пару текстів і дослідіть відмінності між ними (кількість оригінальних слів, багатство мови, жанр). Знайдіть слова, які мають різний зміст в цих текстах, подібно до слова monstrous в Moby Dick та у Sense and Sensibility.

```
import nltk
from nltk.corpus import brown
files = ['ck04','cm01']
for s in files:
    genre = brown.categories(fileids=s)
    words = brown.words(fileids=[s])
    set_s = len(set(words))
    len_s = len(words)/set_s
    print 'name_of text: '+s+'.txt\n',' genre:'+genre[0]+'\n','unique words: ' +str(set_s)+'\n','wealth_of_language: ' +str(len_s)+'\n *****'
```

```
name_of text: ck04.txt
 genre:fiction
unique words: 753
wealth_of_language: 3
 *****
name_of text: cm01.txt
 genre:science_fiction
unique words: 874
wealth_of_language: 2
 *****
```

7. Напишіть програму для знаходження всіх слів в корпусі Brown, які зустрічаються не менш ніж три рази.

```
import nltk
from nltk.corpus import brown
texts = brown.words()
fdist = nltk.FreqDist([w for w in texts])
a=sorted([w for w in set(texts) if fdist [w]>= 3 and w.isalpha()])
print (a [:50])
```

```
>>>
['A', 'ABO', 'ADC', 'AIA', 'AID', 'AIMO', 'AM', 'AP', 'AWOC', 'Aaron',
'Abbe', 'Abbey', 'Abe', 'Abel', 'Abolition', 'About', 'Above', 'Abraham
', 'Abstract', 'Abstraction', 'Academy', 'Acala', 'Accacia', 'According
', 'Accordingly', 'Acey', 'Achievement', 'Acropolis', 'Across', 'Act',
'Acting', 'Action', 'Active', 'Activities', 'Actual', 'Actually', 'Ada'
, 'Adam', 'Adams', 'Add', 'Additional', 'Additionally', 'Adelia', 'Aden
auer', 'Adios', 'Adjusted', 'Adjustment', 'Adlai', 'Adler', 'Administra
tion']
>>>
```

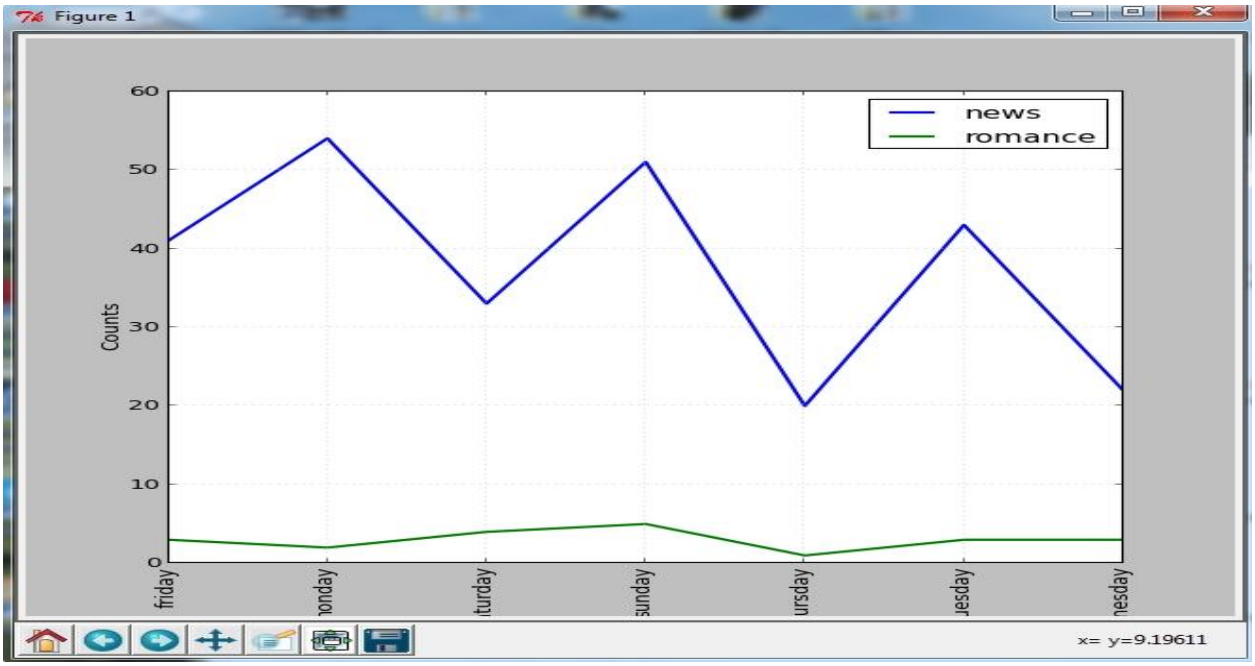
8. Напишіть програму генерації таблиці відношень кількість слів/кількість оригінальних слів для всіх жанрів корпуса Brown. Проаналізуйте отримані результати та поясніть їх.

```
import nltk
from nltk.corpus import brown
for style in brown.categories():
    num_words = len (brown.words (categories = style))
    num_original = len(set(brown.words(categories = style)))
    print num_words, num_original, int(num_words/num_original), style
```

```
>>>
69342 8874 7 adventure
173096 18421 9 belles_lettres
61604 9890 6 editorial
68488 9302 7 fiction
70117 8181 8 government
82345 11935 6 hobbies
21695 5017 4 humor
181888 16859 10 learned
110299 14503 7 lore
57169 6982 8 mystery
100554 14394 6 news
39399 6373 6 religion
40704 8626 4 reviews
70022 8452 8 romance
14470 3233 4 science_fiction
>>>
```

11. Напишіть програму для створення таблиці частот слів для різних жанрів. Знайдіть слова чия присутність або відсутність є характерною для певних жанрів (подібно до модальних дієслів).

```
import nltk
from nltk.corpus import brown
days = ["monday", "tuesday", "wednesday", "thursday", "friday", "saturday", "sunday"]
genres = ["news", "romance"]
cfd = nltk.ConditionalFreqDist(
    (genre, day)
    for genre in genres
    for day in days
    for word in brown.words(categories=genre) if word.lower() == day)
cfd.tabulate(conditions=genres, samples=days)
cfd.plot()
>>> ================================ RESTART ====================
>>>
         monday tuesday wednesday thursday friday saturday sunday
   news      54      43        22       20     41       33     51
romance       2       3         3        1      3        4      5
```



12. Напишіть функцію word_freq(), яка приймає слово і назву частини корпуса Brown як аргументи і визначає частоту слова в заданій частині корпуса.

```
>>> import nltk
>>> from nltk.corpus import brown
>>> from nltk.corpus import brown
>>> def word_freq (word, genre):
        fd = nltk.FreqDist (nltk.corpus.brown.words (categories = genre))
        return word, fd [word]
>>> word_freq('world', 'news')
('world', 37)
>>> word_freq('war', 'news')
('war', 20)
>>> word_freq('love','romance')
('love', 32)
>>> word_freq('time','fiction')
('time', 99)
```

Висновок: на цій лабораторній роботі я вивчила основи програмування на мові Python, також вивчила методи доступу до корпусів текстів і клас ConditionalFreqDist.