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.

ОСНОВИ ПРОГРАМУВАННЯ НА МОВІ PYTHON(частина 2) ”

з дисципліни “Комп’ютерна лінгвістика”

Виконала:

студентка групи ПРЛм-11

Свіляк О.Ю.

Прийняв:

викладач

Дупак Б.П.

Львів-2015

**Мета роботи:**

* Вивчення методів доступу до корпусів текстів.
* Вивчення класу ConditionalFreqDist.
* Вивчення основ програмування на мові *Python*.

**Теоретичні відомості**

Вирішення задач обробки текстів природною мовою передбачає використання великих об’ємів лінгвістичних даних, або іншими словами передбачає роботу з корпусами текстів.

Основні функції NLTK для роботи з корпусами.

|  |  |
| --- | --- |
| **Приклад використання функції** | **Опис** |
| fileids() | Файли корпуса |
| fileids([categories]) | Файли корпуса, що відповідають цій категорії |
| categories() | Категорії корпуса |
| categories([fileids]) | Категорії корпуса, що відповідають цим файлам |
| raw() | Корпус, як послідовність символів |
| raw(fileids=[f1,f2,f3]) | Послідовність символів з наступних файлів |
| raw(categories=[c1,c2]) | Послідовність символів з наступних ка  тегорій |
| words() | Слова корпуса |
| words(fileids=[f1,f2,f3]) | Слова з наступних файлів |
| words(categories=[c1,c2]) | Слова з наступних категорій |
| sents() | Речення корпуса |
| sents(fileids=[f1,f2,f3]) | Речення корпуса з наступних файлів |
| sents(categories=[c1,c2]) | Речення корпуса з наступних категорій |
| abspath(fileid) | Місцезнаходження даного файлу на диску |
| encoding(fileid) | Кодування файлу (якщо відоме) |
| open(fileid) | Відкриття файла з корпуса для читання |
| root() | Шлях до місця де встановлено корпус |
| readme() | Вміст файла README корпуса текстів |

Умовний частотний розподіл. Клас ConditionalFreqDist.

Якщо тексти в корпусі поділені на різні категорії, (за жанром, тематикою, авторами) то можна побудувати частотні розподіли для кожної з категорій. Такі дані дозволяють досліджувати відмінності між жанрами. Умовний частотний розподіл це набір частотних розподілів, кожен з яких відповідає певній «умові». Такою умовою може бути категорія тексту.

Умови і події

Частотний розподіл визначає числові значення для кожної події (подіями можемо вважати вживання слів в тексті). Умовний частотний розподіл поєднує в пари кожну подію та умову. Замість обробки послідовності слів ( #1) обробляються послідовності пар (#2).

>>> text = ['The', 'Fulton', 'County', 'Grand', 'Jury', 'said', ...] #1

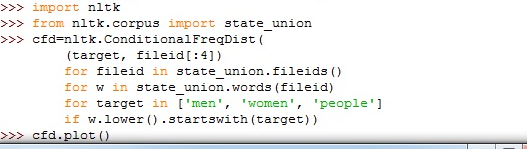
>>> pairs = [('news', 'The'), ('news', 'Fulton'), ('news', 'County'), ...] #2

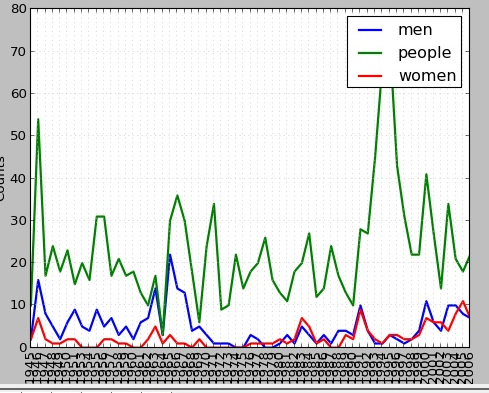
Кожна пара відповідає шаблону (condition, event).  Якщо розглядати корпус Brown за жанрами то отримаємо 15 умов (одна для жанру) і 1161192 подій (одна для слова).

**Тексти програм на мові Python**

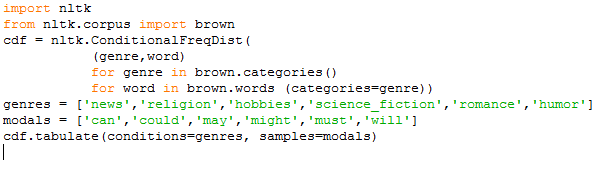
**Варіант 15**

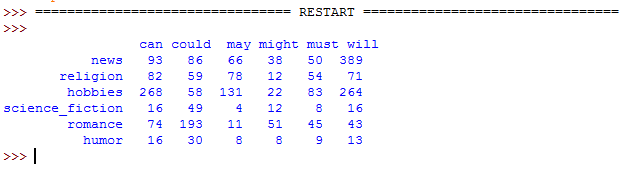
3.3. Прочитайте тексти з корпусу *State of the Union addresses* використовуючи *state\_union* модуль читання. Визначити частоту вживання слів *men, women, people*  в кожному з документів. Як змінилася частота вживання цих слів з часом?

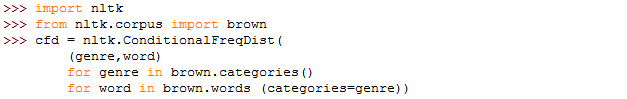




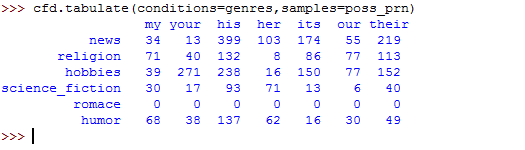
3.6.Проаналізуйте таблицю частот модальних дієслів для різних жанрів. Спробуйте її пояснити. Знайдіть інші класи слів вживання яких також відрізняються в різних жанрах.







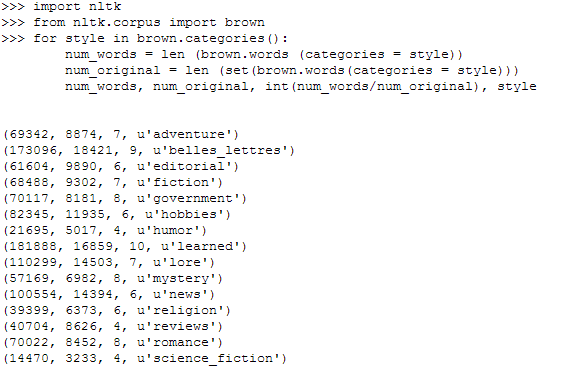




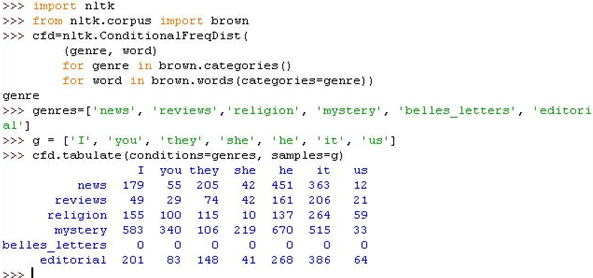
3.7. Напишіть програму для знаходження всіх слів в корпусі Brown, які зустрічаються не менш ніж три рази.



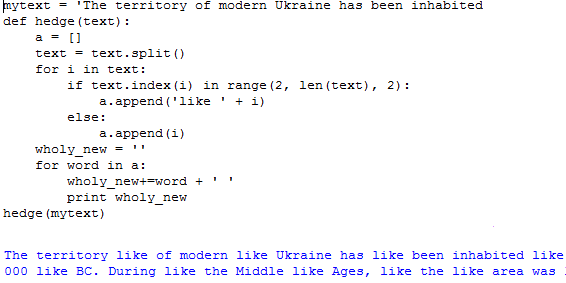
3.8.Напишіть програму генерації таблиці відношень  кількість слів/кількість оригінальних слів для всіх жанрів корпуса Brown. Проаналізуйте отримані результати та поясніть їх.



3.11. 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).



3.13. Визначити функцію hedge(text), яка обробляє текст і створює нову версію цього тексту додаючи слово ‘like’ перед кожним третім словом.



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