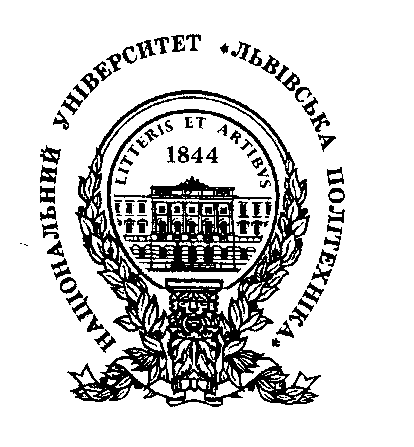
Міністерство освіти та науки України

Національний університет «Львівська Політехніка»

Інститут комп’ютерних наук та інформаційних технологій

Кафедра САПР



Лабораторна робота №3

на тему:

ВИВЧЕННЯ БІБЛІОТЕКИ ПРИКЛАДНИХ ПРОГРАМ NLTK, ДЛЯ ОПРАЦЮВАННЯ ТЕКСТІВ ПРИРОДНОЮ МОВОЮ. ДОСТУП ТА РОБОТА З КОРПУСАМИ ТЕКСТІВ.

Виконав:

студент ПРЛм-12

Білик Д.Г.

Перевірив:

Дупак Б.П.

Львів 2015

**МЕТА РОБОТИ**

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

**КОРОТКІ 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.

1. Доступ до корпусів текстів.

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.

1. 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.

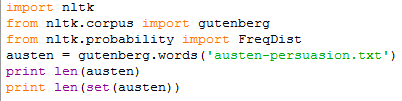
1. Використання текстового редактора при створенні програм.

Інтерпретатор виконує оператори програми відразу після їх введення. Часто виникає потреба спочатку написати програму , яка складається з багатьох рядків використовуючи текстовий редактор а потім запустити її на виконання. Використовуючи IDLE можна це зробити відкривши нове вікно за допомогою пункту меню File, зберігши текст програми у файлі \*py і запустивши програму на виконання за допомогою команди Run Module пункту меню Run.

**ТЕКСТИ ПРОГРАМ НА МОВІ PYTHON**

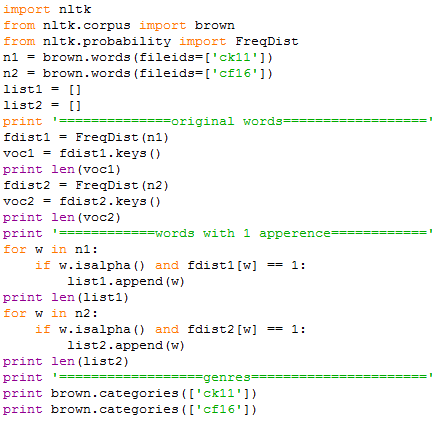
**Варіант 13**

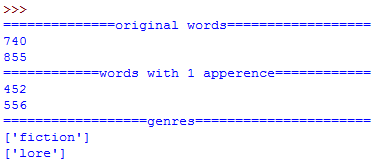
3.1. Використовуючи модуль corpus прочитайте текст austin-persuasion.txt. Визначить скільки tokens (слів) і type (унікальних слів)містить ця книжка.



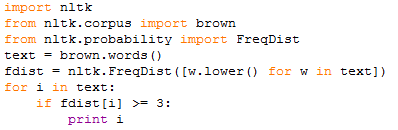


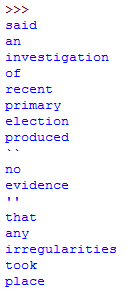
3.5. Виберіть пару текстів і дослідіть відмінності між ними (кількість оригінальних слів, багатство мови, жанр). Знайдіть слова, які мають різний зміст в цих текстах, подібно до слова monstrous в Moby Dick та у Sense and Sensibility.



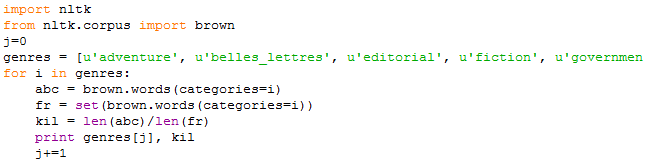


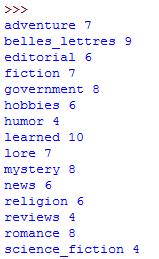
3.7. Напишіть програму для знаходження всіх слів в корпусі Brown, які зустрічаються не менш ніж три рази.



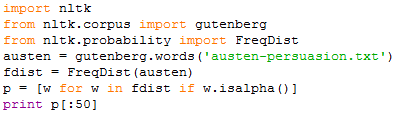


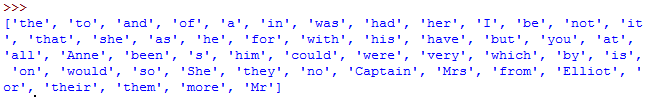
3.8. Напишіть програму генерації таблиці відношень кількість слів/кількість оригінальних слів для всіх жанрів корпуса Brown. Проаналізуйте отримані результати та поясніть їх.



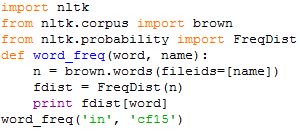


3.9. Напишіть програму для знаходження 50 найчастотніших слів в тексті, за виключенням незначущих слів.





3.12. Напишіть функцію word\_freq(), яка приймає слово і назву частини корпуса Brown як аргументи і визначає частоту слова в заданій частині корпуса.



****

**ВИСНОВОК**

На цій лабораторній роботі, ми вивчили основи програмування на мові Python, методи доступу до корпусів текстів та ознайомились з класом ConditionalFreqDist.