МІНІСТЕРСТВО ОСВІТИ ТА НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ “ЛЬВІВСЬКА ПОЛІТЕХНІКА”

ІНСТИТУТ КОМП’ЮТЕРНИХ НАУК ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ



**Кафедра САП**

**Звіт**

**Лабораторної роботи №3**

***ВИВЧЕННЯ БІБЛІОТЕКИ ПРИКЛАДНИХ ПРОГРАМ NLTK ДЛЯ ОПРАЦЮВАННЯ ТЕКСТІВ ПРИРОДНОЮ МОВОЮ.***

***ДОСТУП ТА РОБОТА З КОРПУСАМИ ТЕКСТІВ.***

Виконала:

студентка групи ПРЛм-12

Артим Л.Й.

Перевірив:

Дупак Б.П.

Львів 2015

1. **Мета.**

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

1. **Теоретичні відомості.**

При роботі з корпусами важливо мати засоби доступу як до окремих текстів 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()

Функція .raw()дозволяє доступитися до вмісту файла без будь-якої його попередньої лінгвістичної обробки. Тому використання len(gutenberg.raw('blake-poems.txt') дозволяє встановити скільки символів (разом з пробілами) є в тексті. Функція sents() ділить текст на оремі речення і кожне речення представляється, як список стрічок, де стічки – окремі слова.

Для роботи з менш формальною мовою, NLTK містить набір текстів з Інтернету: тексти з форуму, тексти з фільму Пірати карибського моря, тексти особистих оголошень,телефонні розмови, огляд новин:

>>> from nltk.corpus import webtext

>>> for fileid in webtext.fileids():

Також в NLTK входить корпус повідомлень з чатів, створений в Naval Postgraduate School для досліджень з метою автоматичного виявлення Інтернет злочинців. Цей корпус містить 10000 анонімних повідомлень в яких імена користувачів замінені за шаблоном "UserNNN" а також видалена інша персональна інформація. Корпус організований ,як 15 окремих файлів, кожен з яких містить декілька сотень повідомлень з певною датою створення та вікових даних авторів (підлітки, 20ти, 30ти та 40ка річні, дорослі). Назва файла містить інформацію про дату, вікову групу та кількість повідомлень, наприклад файл 10-19-20s\_706posts.xml містить 706 повідомлень двадцятирічних дописувачів від 19 жовтня 2006 року.

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

>>> from nltk.corpus import brown

>>> brown.categories()

['adventure', 'belles\_lettres', 'editorial', 'fiction', 'government', 'hobbies',

'humor', 'learned', 'lore', 'mystery', 'news', 'religion', 'reviews', 'romance',

'science\_fiction']

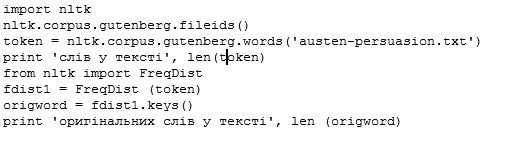
Корпус Reuters. Засобами NLTK можна звернутися до тем, яких торкаються в одному або декількох текстах або навпаки дізнатися весь перелік текстів, які належать до певної категорії.

Найпростіший корпус текстів не має структури, це набір текстів. Інші корпуси це набори текстів поділених за категоріями мови, жанру, автора. В багатьох випадках категорії текстів можуть перетинатися між собою, оскільки тексти можуть належати різним категоріям. Окремий випадок це коли набори текстів розподілені за часовими параметрами.

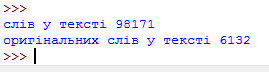
Якщо тексти в корпусі поділені на різні категорії, (за жанром, тематикою, авторами) то можна побудувати частотні розподіли для кожної з категорій. Такі дані дозволяють досліджувати відмінності між жанрами. Умовний частотний розподіл це набір частотних розподілів, кожен з яких відповідає певній «умові».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

1. **Виконання завдань.**

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

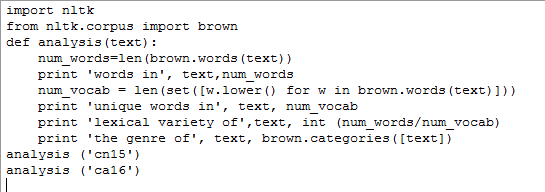


Текст програми 1

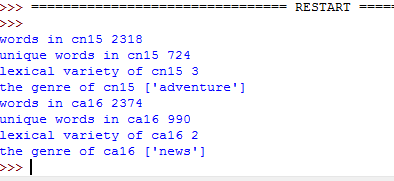


Результат програми

Виберіть пару текстів і дослідіть відмінності між ними (кількість оригінальних слів, багатство мови, жанр).

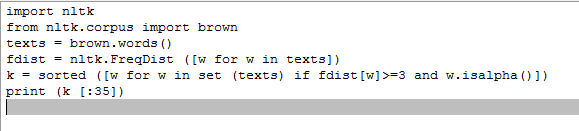


Текст програми 5

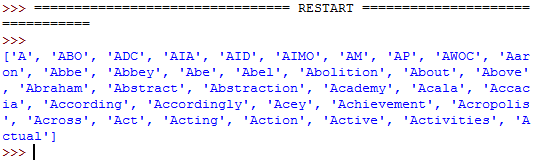


Результат програми 5

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

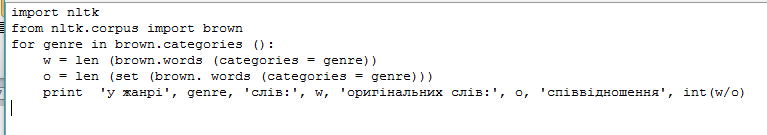


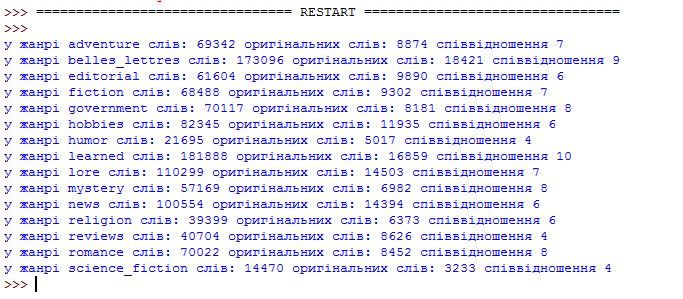
Текст програми 7



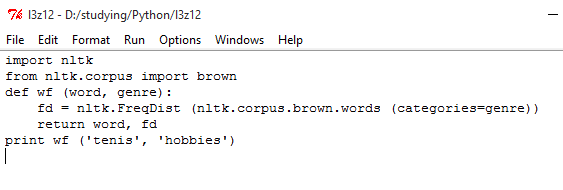
Результат програми 7

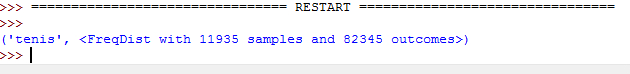
Напишіть програму генерації таблиці відношень кількість слів/кількість оригінальних слів для всіх жанрів корпуса Brown. Проаналізуйте отримані результати та поясніть їх.



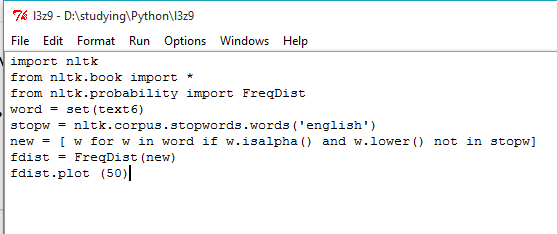


Напишіть функцію word\_freq(), яка приймає слово і назву частини корпуса Brown як аргументи і визначає частоту слова в заданій частині корпуса.





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



**4. Висновок.**

У ході виконання лабораторної роботи вивчила основи програмування на мові Python; вивчила методів доступу до корпусів текстів; вивчила клас ConditionalFreqDist.