МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «ЛЬВІВСЬКА ПОЛІТЕХНІКА»

Кафедра САПР



ЗВІТ

до лабораторної роботи № 2

на тему:

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

***ДОСТУП ТА РОБОТА З ЛЕКСИЧНИМИ РЕСУРСАМИ.***

ОСНОВИ ПРОГРАМУВАННЯ НА МОВІ PYTHON (частина 2)

з дисципліни “Комп’ютерна лінгвістика”

Виконала:

ст.гр. ПРЛм-12

Артим Л.Й.

Перевірив:

Старший викладач кафедри САПР

Дупак Б. П.

Львів 2015

***МЕТА РОБОТИ***

* *Вивчення основ програмування на мові Python.*
* *Вивчення методів доступу та роботи з лексичним ресурсами.*
* *Семантичний словник англійської мови WordNet****.***

***Поняття функції та модуля***.

При програмуванні часто необхідно частину програми виконати (використати) декілька разів. Наприклад, потрібно написати програму, яка здійснює утворення множини з однини іменників і вона буде виконуватись в різних місцях програми. Швидше ніж повторювати той самий код декілька разів і більш ефективно і надійно організувати цю роботу через функцію. Функція - це програмна конструкція, яку можна викликати з одним або більше вхідними параметрами, і отримувати результат на виході.

***Генерація випадкового тексту за допомогою біграмів.***

Умовний частотний розподіл можна використати для побудови таблиці біграмів (пар слів). Функція NLTK bigrams() , як аргумент бере список слів і повертає список послідовних пар слів.

***Лексичні ресурси NLTK.***

Лексичний ресурс або просто словник це набір слів тa/або словосполучень, які асоціюються з такою інформацією, як частина мови та опис значення. Лексичні ресурси є вторинними по відношенню до текстів і зазвичай створюються і вдосконалюються з використанням текстів. Наприклад, якщо визначити текст my\_text тоді vocab = sorted(set(my\_text)) побудує словник тексту my\_text, word\_freq = FreqDist(my\_text) визначить частоту кожного слова в тексті. vocab та word\_freq – приклад простих лексичних ресурсів. Так само конкорданс дає інформацію про використання слів і ця інформація може бути використана при побудові словників.

***Корпуси слів***

NLTK розповсюджується з деякими корпусами, які насправді є списками слів. Корпус words це файл з Unix, який використовується для перевірки правопису.

***Словник із позначенням вимови.***

Більш багатим лінгвістичним ресурсом може бути словник де кожному слову поставлена у відповідність певна інформація. NLTK включає CMU Pronouncing Dictionary американського варіанту англійської , який розроблений для використання в синтезаторах мови.

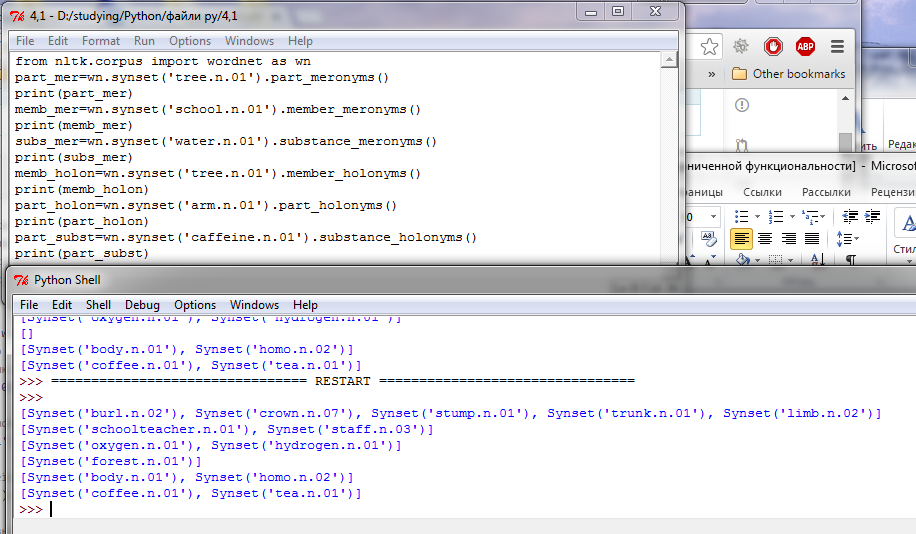
***Порівняльні (компаративні) словники.***

Інший словник в NLTK це порівняльний словник (Swadesh wordlists), який містить 200 спільних слів для 24 мов. Мови ідентифікуються за двосимвольними кодами (ISO 639).

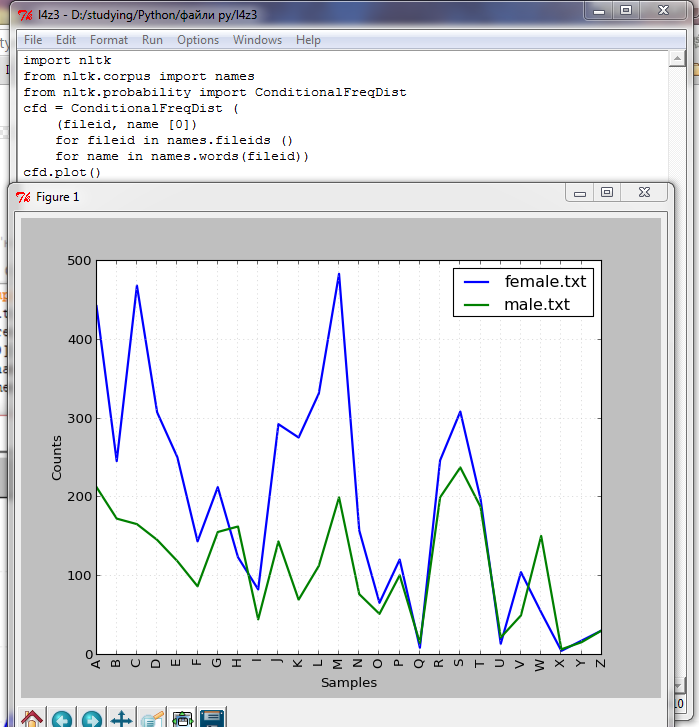
ТЕКСТИ ПРОГРАМ НА МОВІ PYTHON

ВАРІАНТ №1

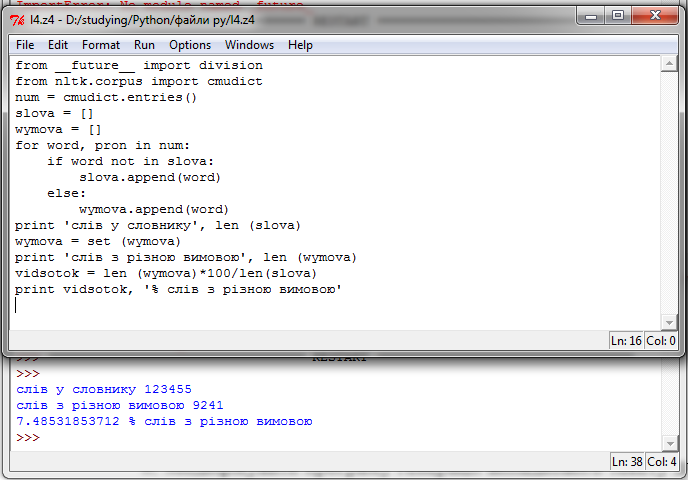
Дослідити зв’язки голонім-меронім для іменників. Знайти іменники для демонстрації наступних зв’язків: member\_meronyms(), part\_meronyms(), substance\_meronyms(), member\_holonyms(), part\_holonyms(), та substance\_holonyms().



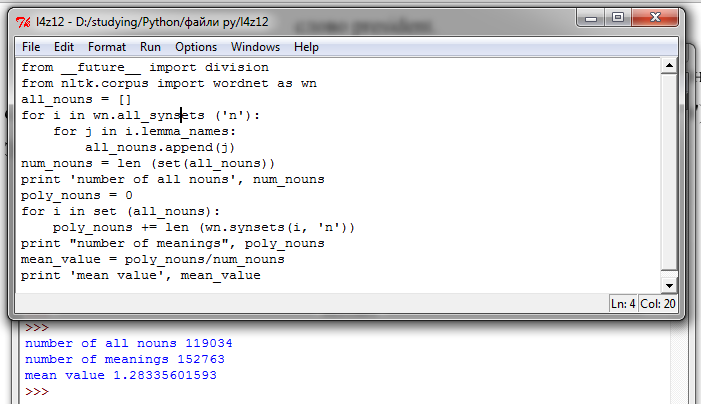
Побудувати умовний частотний розподіл для корпусу імен. Знайти які перші літери частіше використовуються в чоловічих та жіночих іменах.



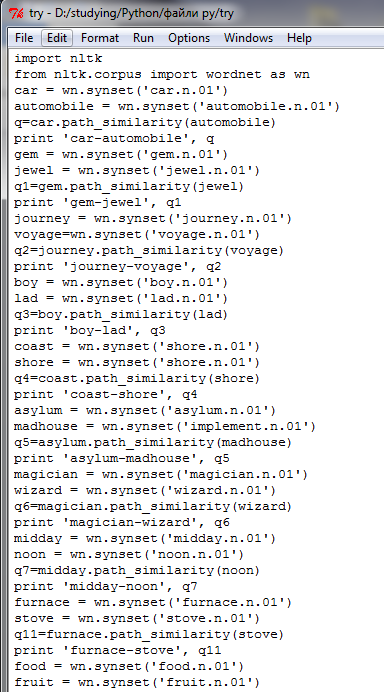
Здійснити аналіз словника вимов. Знайти скільки різних слів він містить. Який відсоток слів з цього словника можуть мати різну вимову?

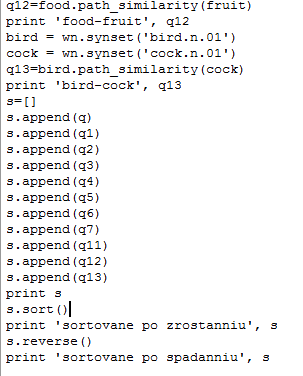


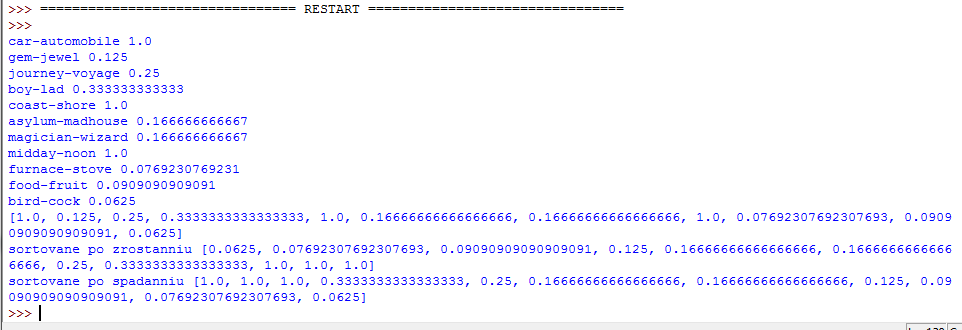
Полісемія - це явище коли одне слово має декілька значень ( іменник dog має 7 значень, кількість яких визначити можна як len(wn.synsets('dog', 'n'))). Знайдіть середнє значення полісемії для іменників.



Використовуючи один з методів визначення подібності слів побудуйте відсортований по спаданню список значень подібності для наступних пар слів: car-automobile, gem-jewel, journey-voyage, boy-lad, coast-shore, asylum-madhouse, magician-wizard, midday-noon, furnace-stove, food-fruit, bird-cock.







**Висновок:** під час лабораторної роботи я вивчила основи програмування на мові Python, а також методи доступу та роботи з лексичним ресурсами, плюс ознайомилась із семантичним словником англійської мови WordNet**.**