Міністерство освіти та науки України

Національний університет “Львівська політехніка”

інститут комп’ютерних наук та інформаційних технологій

Кафедра “Системи автоматизованого проектування”

Звіт

до лабораторної роботи № 4

з курсу «Комп’ютерна лінгвістика»

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)»

Виконала:

Студентка групи ПРЛм-11  
Михайлів Р.Б.

Прийняв:

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.

**Короткі теоретичні відомості**

***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. Визначаємо функцію, використовуючи ключове слово *def* далі потрібно дати назву функції і визначити вхідні параметри, після двокрапки записується тіло функції. Ключове слово *return* використовується для відображення значення, яке ми хочемо отримати на виході функції.

***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.***

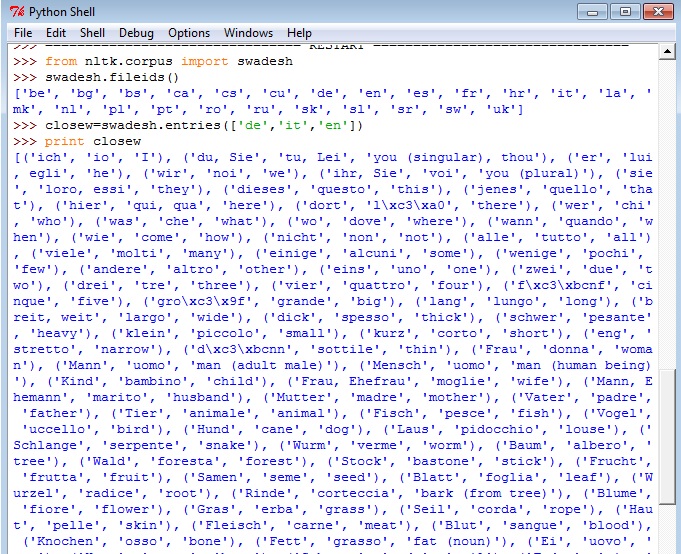
Лексичний ресурс або просто словник це набір слів тa/або словосполучень, які асоціюються з такою інформацією, як частина мови та опис значення. Лексичні ресурси є вторинними по відношенню до текстів і зазвичай створюються і вдосконалюються з використанням текстів. Наприклад, якщо визначити текст my\_text 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.

***WordNet – лексична база даних англійської мови***

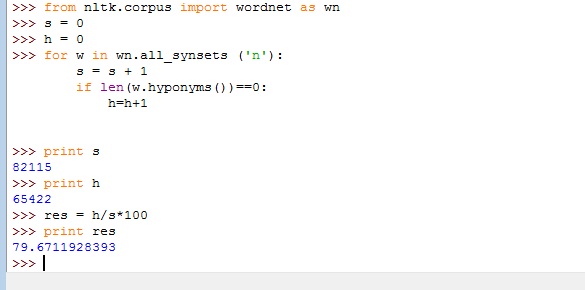
*WordNet,* це семантично орієнтований словник англійської мови, подібний до традиційних тезаурусів але з більш багатою структурою. У *WordNet* слова групуються у набори синонімів – синсети, кожен із своїм визначеннямі зв’язками з іншими синсетами. *WordNet 3.0* розповсюджується разом з NLTK і містить 155287 слів та 117659 синсетів. Хоча *WordNet* розроблявся для психолінгвістики - цей словник широко використовується в NLP та в задачах інформаційного пошуку. Розробки для інших мов проводяться на основі документації, яка наведена у <http://www.globalwordnet.org/>.

1. Варіант

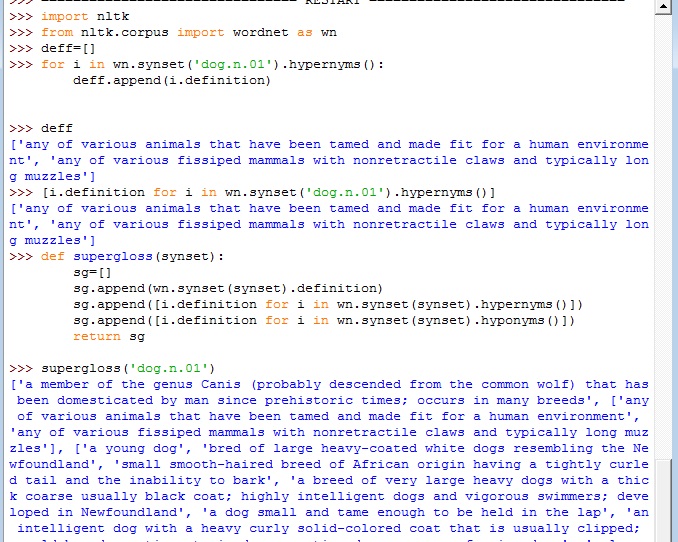
**2.Використовуючи компаративний словник знайти близькі слова для німецької, італійської та англійської мов. Чи можуть отримані результати використовуватися для здійснення перекладу?**

****

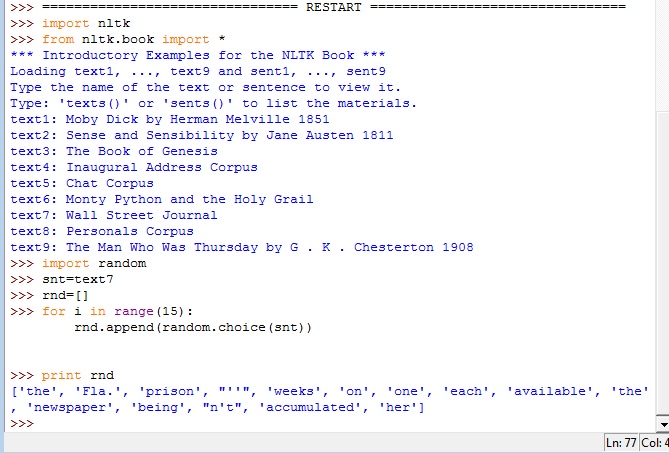
***5.Який відсоток синсетів іменників не мають гіпонімів? До всіх синсетів можна доступитися за допомогою wn.all\_synsets('n').***

******

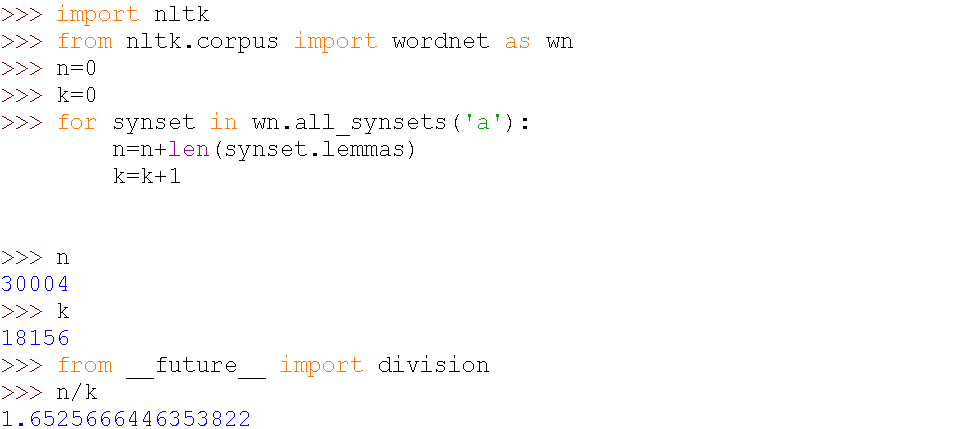
***6.Визначити функцію supergloss(s) , яка буде приймати синсет s як аргумент і повертати стрічку в якій будуть поєднані всі описи всіх значень синсету s та описи всіх гіпернімів та гіпонімів s.***

******

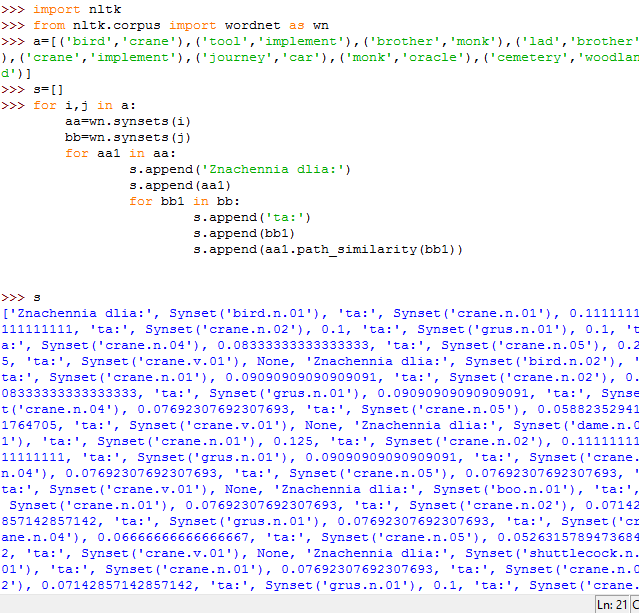
***7.Модифікувати програму генерації випадкового тексту для виконання наступного: зберігати можливі наступні слова у списку та вибирати їх за допомогою random.choice() попередньо виконавши import random.***

******

**11.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 прикметників.**

****

***15.Використовуючи один з методів визначення подібності слів побудуйте список значень подібності для наступних пар слів: bird-crane, tool-implement, brother-monk, lad-brother, crane-implement, journey-car, monk-oracle, cemetery-woodland.***

******

**Висновок:** на цій лабораторній роботі я вивчила основи програмування на мові *Python*, методи доступу та робити з лексичними ресурсами та семантичний словник WordNet.