Міністерство освіти та науки України

Національний університет “Львівська політехніка”

інститут комп’ютерних наук та інформаційних технологій

Кафедра “Системи автоматизованого проектування”

Звіт

до лабораторної роботи № 2

з курсу «Комп’ютерна лінгвістика»

на тему:

«Вивчення бібліотеки прикладних програм NTLK, для опрацювання текстів природною мовою. 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*. Ознайомлення з контрольними структурами та класом FreqDist.

**Короткі теоретичні відомості**

Програми, які згадувалися в лабораторній роботі №1, доволі прості, але дозволяють здійснювати операції над послідовностями слів і зокрема над кожним із них. Ми працювали зі списками, які містять декілька слів , але є потреба працювати з такими списками, які містять тисячі слів . Більшість мов програмування дозволяють виконання окремих блоків програми, коли використовуються умовні вирази або if твердження. If твердження, так як і For твердження - це керуючі (управляючі) структури, які керують виконанням програми. Вони вибирають, яка частина коду в тілі програми буде виконуватися. Стрічка з операторами If та For повинна завершуватися двокрапкою. Якщо потрібно, щоб щось відбувалося, коли умовний вираз не справджується, використовується оператор else в If – твердженні.

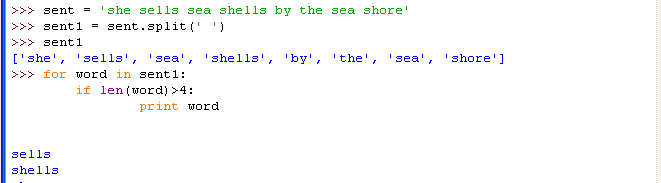
Python підтримує такий тип даних, як словник, і він використовується для встановлення відповідності (відображення) між довільними типами даних. Визначимо pos як пустий словник і додамо до нього значення, які встановлюють частиномовну приналежність певних слів . Значення додаються із використанням квадратних дужок.

Для автоматичного визначення слів, які є найбільш інформативними для текстів певного жанру або певної тематики спочатку інтуітивно виникає думка побудувати частотний список або частотний розподіл. Частотний розподіл вказує на частоту з якою в тексті зустрічається кожне зі слів. Такий частотний список називають розподілом тому, що він вказує яким чином загальна кількість слів розподіляється між словниковими статями (оригінальні слова) в тексті. Враховуючи що побудова частотних розподілів часто необхідна при обробці природної мови в NLTK реалізовано окремий клас FreqDist в модулі nltk.probability .

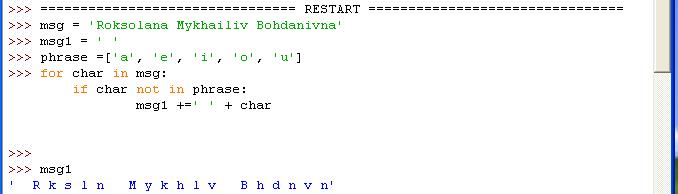
Колокація - це словосполучення яке зустрічається дуже часто. Для того щоб побудувати колокації спочатку потрібно побудувати на основі тексту пари слів, або біграми.

20 варіант

* 1. Створіть змінну *sentence* і присвойте їй значення *‘she sells sea shells by the sea shore’* та напишіть фрагмент програми для виведення на екран всіх сліва, довжина яких більша ніж 4 символи.



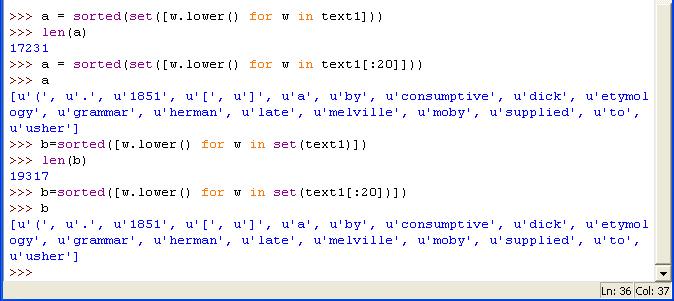
* 1. Напишіть програму, яка видаляє всі голосні зі стрічки, яка відповідає імені, по батькові та прізвищу студента. Програма повинна здійснювати наступну послідовність дій: створення початкової стрічки; створення стрічки, у якій буде зберігатися результат; *for* цикл для обробки стрічки символ за символом і запису неголосних символів в результуючу стрічку.



* 1. Виконати наступні приклади і пояснити чому отримані різні результати (різні значення змінних)

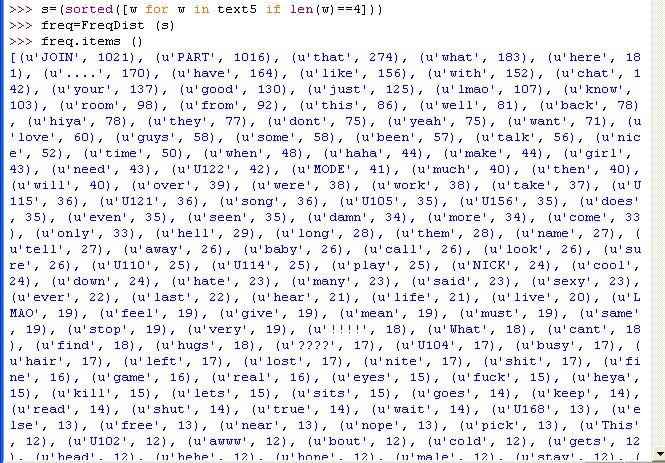
sorted(set([w.lower() for w in text1]))

sorted([w.lower() for w in set(text1)])

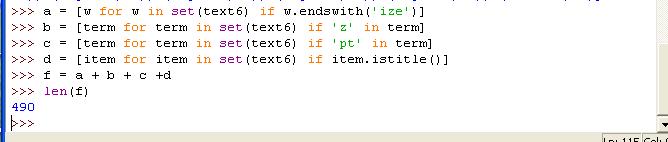


Результати різні тому, що у змінній а спочатку відбувається пошук малих букв, а тоді виконується команда set, яка видаляє ті ж самі елементи. У змінній b спочатку виконується команда set (text1), яка видаляє ті ж самі елементи (S і s – це два різні елементи), а тоді повертається до маленьких літер, тому і величина більша.

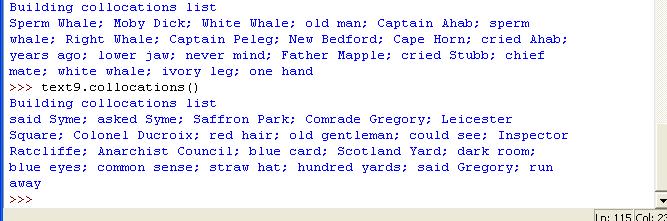
3.9.Знайдіть в тексті № 5 всі слова довжина яких дорівнює 4 і побудуйте для них частотний розподіл.



3.11. Напишіть вираз для знаходження в тексті №6 всіх слів які відповідають наступним вимогам: закінчуються на ize; містять літеру z; містять послідовність літер pt; написані з великої літери . Результат представити, як список слів.



3.19. Побудуйте колокації для текстів №1 та №9. Результати порівняйте.



**Висновок:** на лабораторній роботі я засвоїла основи програмування на мові Python та ознайомилася з контрольними структурами та класом FreqDist.