**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**

**НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ “ЛЬВІВСЬКА ПОЛІТЕХНІКА”**

**іНСТИТУТ КОМП’ютерних НАУК та ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ**

### Кафедра “Системи автоматизованого проектування ”

### gerbnational

Звіт

до лабораторної роботи № 5

**на тему: «ВИВЧЕННЯ БІБЛІОТЕКИ ПРИКЛАДНИХ ПРОГРАМ NLTK, ДЛЯ ОПРАЦЮВАННЯ ТЕКСТІВ ПРИРОДНОЮ МОВОЮ.ПОЧАТКОВА ОБРОБКА ТЕКСТІВ ПРИРОДНОЮ МОВОЮ»**

з дисципліни:

«Комп’ютерна лінгвістика»

Виконала:

студентка групи ПРЛм-11

Турчиняк О.І.

Прийняв:

викладач Дупак Б.П.

Львів 2015

1. Напишіть функцію, яка приймає адресу URL, як аргумент, і повертає те що міститься за цією адресою з видаленням HTML розмітки. Використовувати urllib.urlopen для доступу до контенту наступним чином raw\_contents = urllib.urlopen('http://www.nltk.org/').read().





2.Збережіть деякий текст у файлі corpus.txt. Визначити функцію load(f) для читання файлу, назва якого є її аргументом і повертає стрічку, яка містить текст з файлу.





3.Перепишіть наступний цикл як list comprehension:

**>>> sent = ['The', 'dog', 'gave', 'John', 'the', 'newspaper']**

**>>> result = []**

**>>> for word in sent:**

**... word\_len = (word, len(word))**

**... result.append(word\_len)**

**>>> result**

**[('The', 3), ('dog', 3), ('gave', 4), ('John', 4), ('the', 3), ('newspaper', 9)]**





4.Перевірити різницю між стрічками і цілим виконавши наступні дії: "3" \* 7 та 3 \* 7. Спробуйте здійснити конвертування між стрічками і цілими використавши int("3") та str(3).



5.Що станеться, коли стрічки форматування %6s та %-6s використовується для відображення стрічки довшої ніж 6 символів?



7. Створіть файл, який буде містити слова та їх частоту записані в окремих рядках через пробіл ( fuzzy 53). Прочитайте цей файл використовуючи open(filename).readlines(). Розділіть кожну стрічку на дві частини використовуючи split(), і перетворіть число в ціле значення використовуючи int(). Результат повинен бути у вигляді списку: [['fuzzy', 53], ...].





8. Напишіть програму доступу до вебсторінки і вилучення з неї деякого тексту.





10. Модуль random включає функцію choice(), яка випадковим чином вибирає елементи послідовності. Наприклад, choice("aehh ") буде вибирати один з чотирьох символів. Напишіть програму генерації стрічки з 500 випадково вибраних символів "aehh ". Для поєднання елементів в стрічку використовуйте ''.join() . Нормалізуйте отриманий результат використовуючи split() та join().





14. Доступіться до текстів ABC Rural News та ABC Science News з корпуса (nltk.corpus.abc). Знайдіть значення для оцінки читабельності текстів (аналогічно до задачі №12). Використовуйте Punkt для поділу тексту на окремі речення.





Висновок: я освоїла основи програмування на мові Python, вивчила методи роботи з файлами на локальних дисках та з Інтернету. Навчилася використовувати Юнікод при обробці текстів. Ознайомилась з нормалізацією текстів, стемінгом, лематизацією та сегментацією.