**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**

**НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ “ЛЬВІВСЬКА ПОЛІТЕХНІКА”**

**ІНСТИТУТ КОМП’ЮТЕРНИХ НАУК ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ**

Кафедра “Системи автоматизованого проектування”



Звіт до лабораторної роботи №6

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

ВИКОРИСТАННЯ РЕГУЛЯРНИХ ВИРАЗІВ ДЛЯ ОБРОБКИ ТЕКСТУ» з дисципліни “Комп’ютерна лінгвістика”

Виконала:

студентка групи ПРЛм-11

Неїжмак О.А.

Прийняв:

викладач Дупак Б.П.

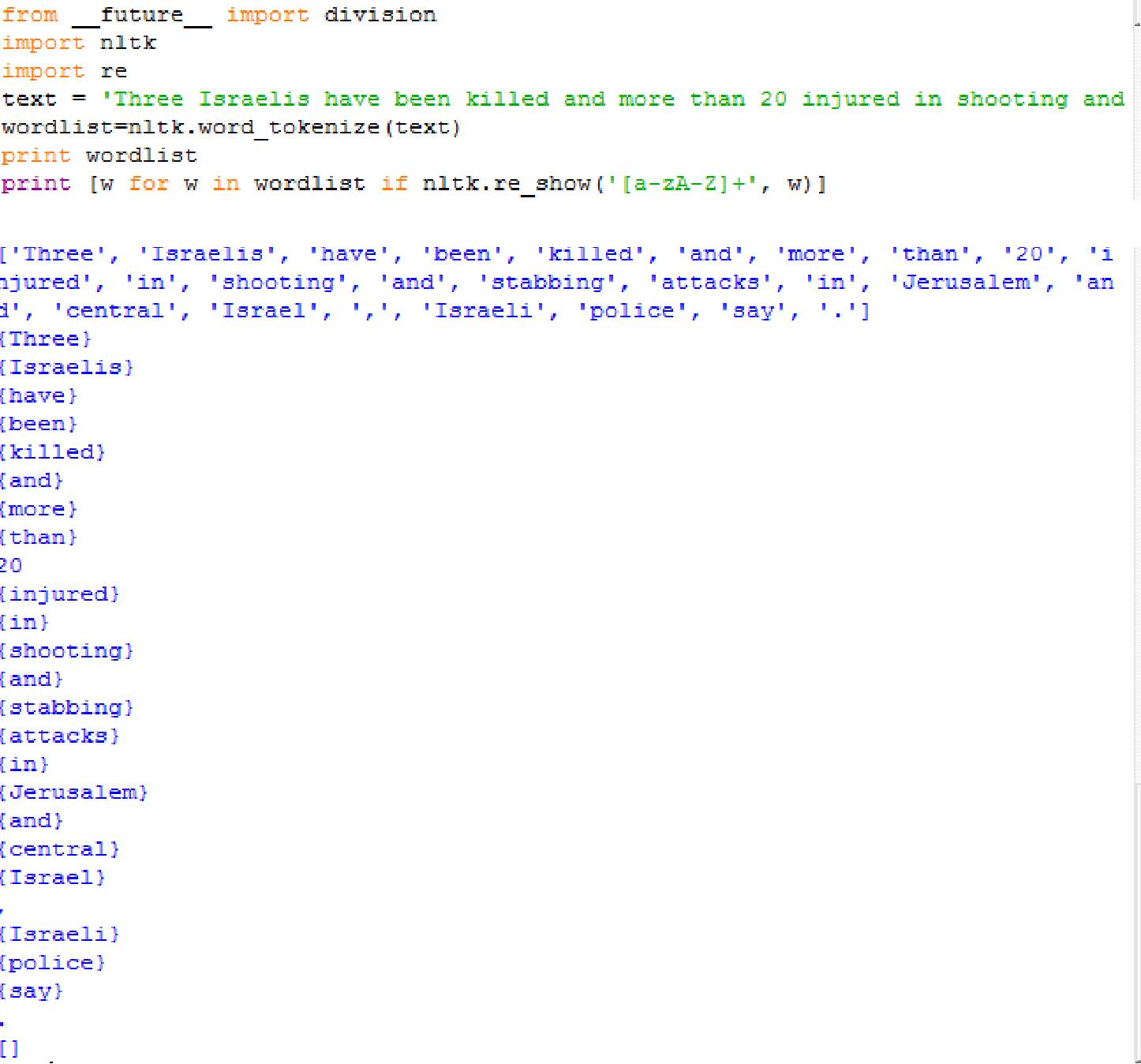
Львів-2015

**МЕТА РОБОТА**

* + Вивчення основ програмування на мові *Python*.
  + Використання регулярних виразів для обробки текстів.

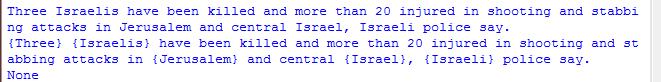
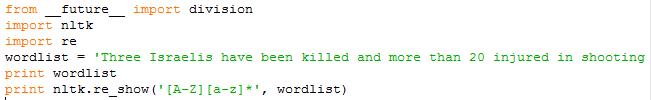
Варіант 10

* + 1. Описати, які класи стрічок відповідають наступному регулярному виразу. [a-zA-Z]+. Результати перевірити використовуючи nltk.re\_show()
* цій програмі ми шукаємо всі слова, які складаються з малих та великих літер англійського алфавіту. Слова, в яких є повторення будь-якої літери один або більше разів.



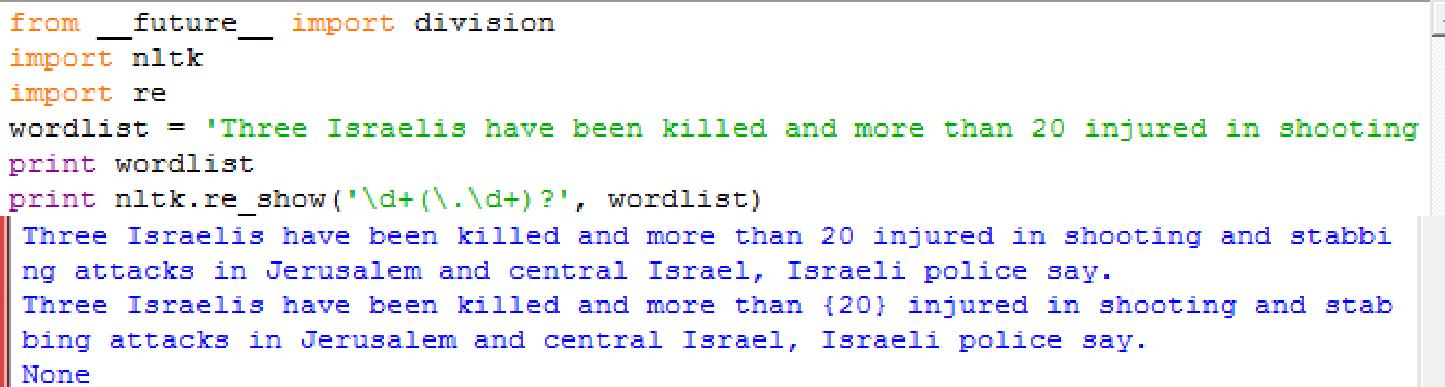
2. Описати, які класи стрічок відповідають наступному регулярному виразу. [A-Z][a-z]\*. Результати перевірити використовуючи nltk.re\_show()

У цій програмі ми шукаємо послідовності символів, що починаються з великої літери англійського алфавіту



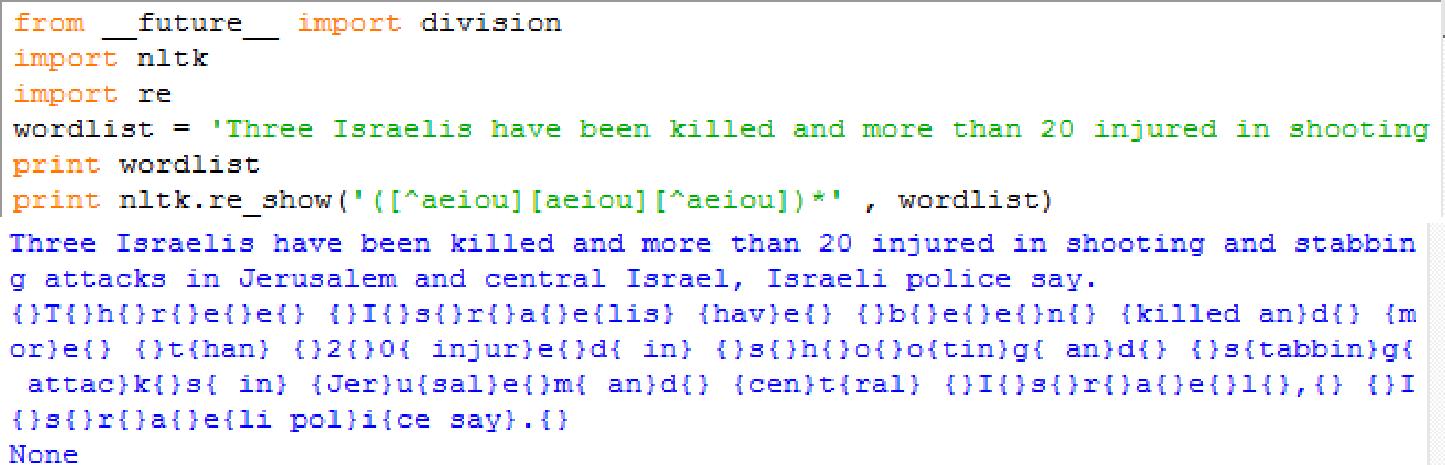
3. Описати, які класи стрічок відповідають наступному регулярному виразу. \d+(\.\d+)?. Результати перевірити використовуючи nltk.re\_show()

У цій програмі ми шукаємо послідовності символів, що складаються з цифр, які повторюються з 1 і більше разів; (\.\d+)? – послідовність цифр після крапки не є обов’язковою



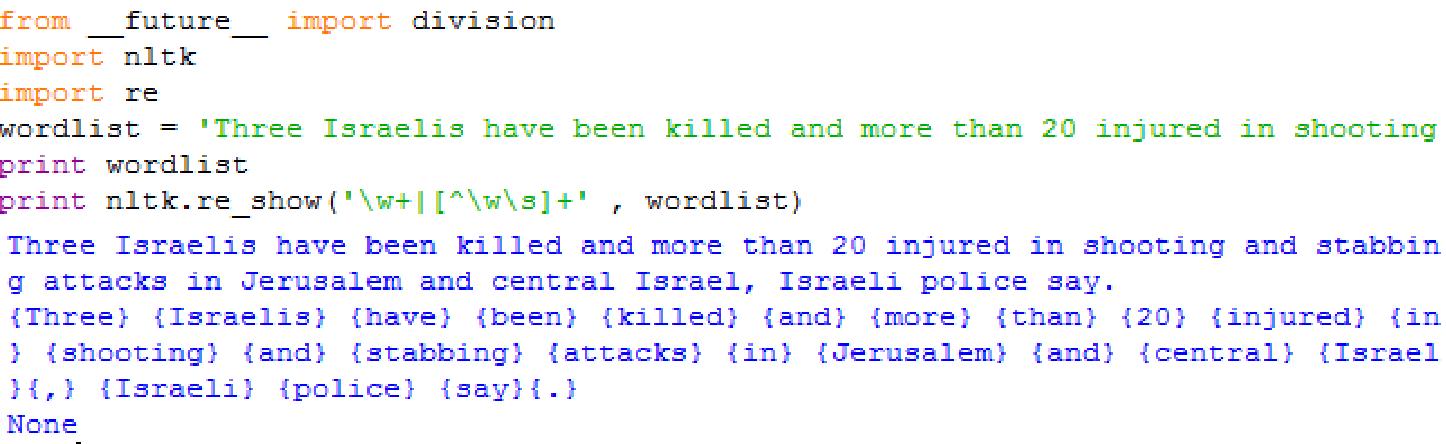
**4.** Описати,які класи стрічок відповідають наступному регулярному виразу.([^aeiou][aeiou][^aeiou])\*. Результати перевірити використовуючи nltk.re\_show()

У цій програмі ми шукаємо послідовності символів, що складаються з трьох символів, перший і третій з яких не є голосною, а другий – будь-яка голосна з набору [aeiou] і зустрічаються 0 і більше разів. Якщо послідовність не знайдено – виводиться {}



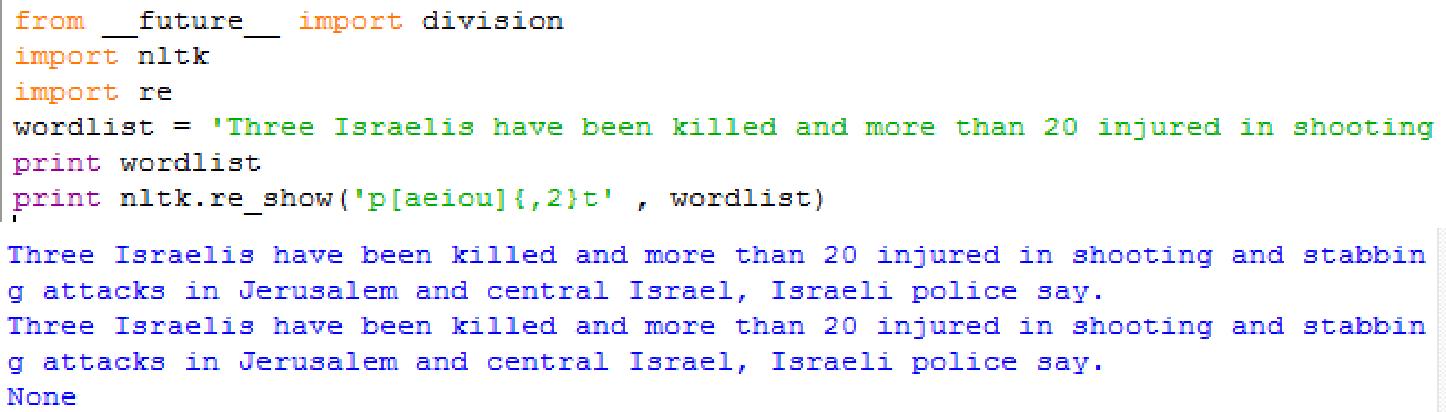
**5.** Описати,які класи стрічок відповідають наступному регулярному виразу.\w+|[^\w\s]+.. Результати перевірити використовуючи nltk.re\_show()

У цій програмі ми шукаємо послідовності, що складаються з літер чи цифр, які повторюються 1 і більше разів

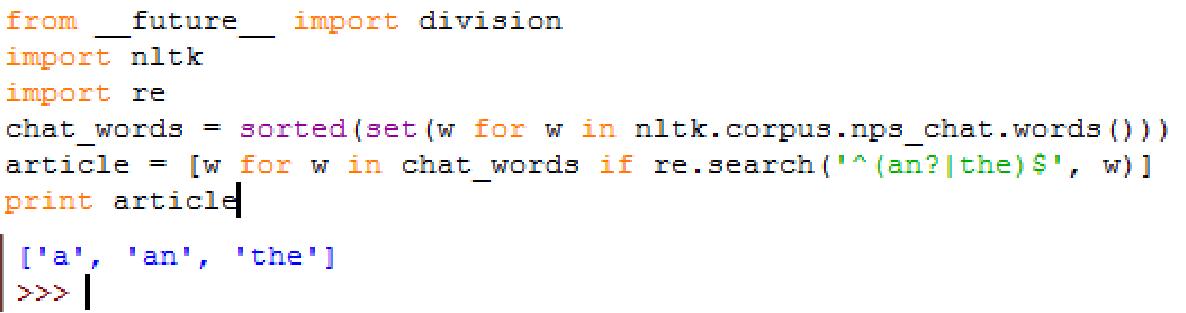


**6.** Описати,які класи стрічок відповідають наступному регулярному виразу.p[aeiou]{,2}t Результати перевірити використовуючи nltk.re\_show()

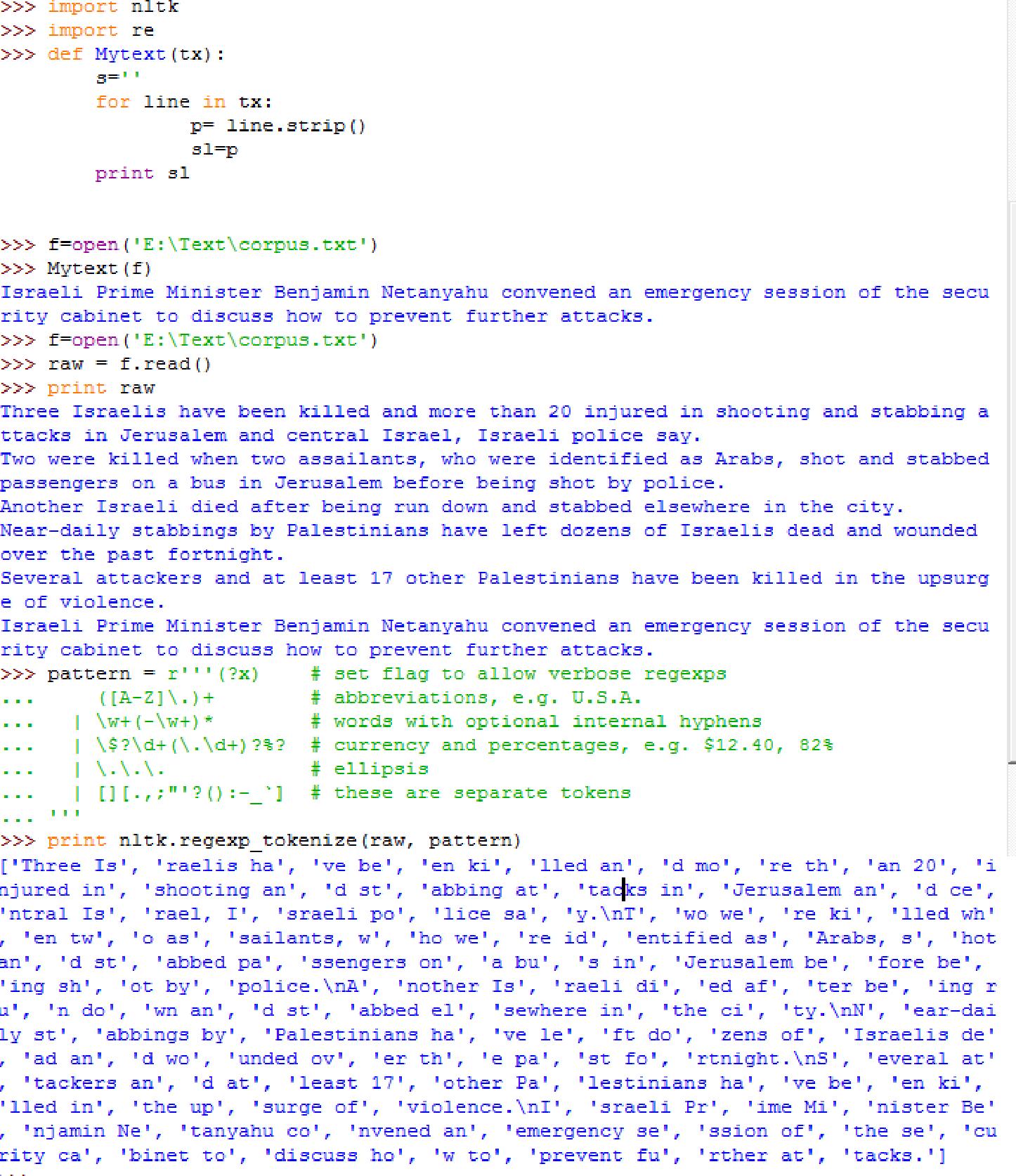
У цій програмі ми шукаємо стрічки, які складаються з букви “p”, жодної, одної або двох голосних з набору і літери “t”.



**7.** Написати регулярний вираз,який встановлює відповідність наступному класустрічок: всі артиклі *(a, an, the*).



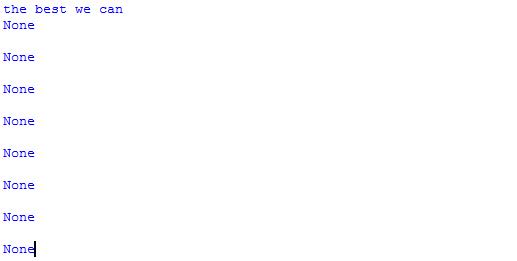
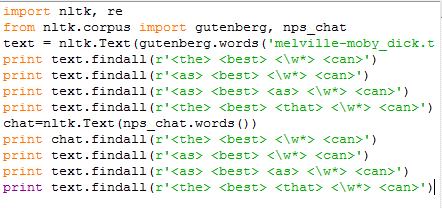
9. Зберегти довільний текст у файлі corpus.txt. Визначити функцію для читання з цього файлу (назва файлу аргумент функції) і повертає стрічку, яка містить текст з файлу. Використовуючи nltk.regexp\_tokenize() розробити токенізатор для токенізації різних типів пунктуації в цьому тексті. Використовувати багаторядковий запис регулярного виразу з коментарями та «verbose flag»



11. Написати функцію unknown(), яка приймає інтернет адресу як аргумент і повертає не відомі слова, які зустрічаються в тексті. При розробці функції використовувати re.findall() для виявлення всіх підстрічок та корпус Words Corpus (nltk.corpus.words) для виявлення не відомих слів.



15. Прочитати Додаток А. Дослідити явища описані у Додатку А використовуючи корпуси текстів та метод findall()для пошуку в токенізованому тексті.



Висновок: під час виконання цієї лабораторної роботи я навчилася використовувати регулярні вирази для обробки текстів.