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.

ОСНОВИ ПРОГРАМУВАННЯ НА МОВІ PYTHON(частина 2) ”

з дисципліни “Комп’ютерна лінгвістика”

Виконала:

студентка групи ПРЛм-11

Свіляк О.Ю.

Прийняв:

викладач

Дупак Б.П.

Львів-2015

**Мета роботи:**

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

**Теоретичні відомості**

Регулярні вирази є більш потужним і гнучким методом опису шаблонів символів, які необхідно виявити у послідовностях символів. Регулярний вираз (вислів) (в [програмуванні](http://uk.wikipedia.org/wiki/%D0%9F%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D1%83%D0%B2%D0%B0%D0%BD%D0%BD%D1%8F)) — це [рядок](http://uk.wikipedia.org/wiki/%D0%A0%D1%8F%D0%B4%D0%BE%D0%BA_%28%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D1%83%D0%B2%D0%B0%D0%BD%D0%BD%D1%8F%29) що описує або збігається з [множиною](http://uk.wikipedia.org/wiki/%D0%9C%D0%BD%D0%BE%D0%B6%D0%B8%D0%BD%D0%B0) рядків, відповідно до набору спеціальних [синтаксичних](http://uk.wikipedia.org/wiki/%D0%A1%D0%B8%D0%BD%D1%82%D0%B0%D0%BA%D1%81%D0%B8%D1%81) правил. Регулярні вислови використовуються в багатьох [текстових редакторах](http://uk.wikipedia.org/wiki/%D0%A2%D0%B5%D0%BA%D1%81%D1%82%D0%BE%D0%B2%D0%B8%D0%B9_%D1%80%D0%B5%D0%B4%D0%B0%D0%BA%D1%82%D0%BE%D1%80) та допоміжних інструментах для пошуку та зміни тексту на основі заданих шаблонів.

Для роботи з регулярними виразами у Python потрібно імпортувати бібліотеку re скориставшись: import re.

Синтаксис регулярних висловів залежить від інтерпретатора, що використовується для їх обробки. Однак, із незначними відхиленнями, майже всі поширені механізми інтерпретатори регулярних висловів мають спільні правила.

Найпростіший регулярний вислів, з якого формуються складні, є звичайний символ. Більшість символів, включаючи усі літери та цифри, є регулярними висловами, що співпадають із відповідними символами в рядках.

Символ ”.” універсальний символ , якому відповідає будь-який один символ. Нехай потрібно знайти слова з восьми літер, де j – третя літера та t – шоста літера. При створенні регулярного виразу у місцях де може бути будь-який символ вказується крапка. Символ ”^” вказує на початок стрічки.

Символ “ ? “ вказує на те що попередній символ не є обовязковим.

Перша частина виразу, «^[ghi]», вказує на початок слова після якого йде один з символів g, h, чи i.

Символ + може використовуватися як після окремих літер так і після послідовності літер у квадратних дужках.

Символ + означає одне або більше повторення.

Оператор ^ виконує іншу функцію, якщо його записати першим символом в квадратних дужках. Наприклад вираз «[^aeiouAEIOU]» встановлює відповідність до будь-яких символів крім голосних.

Символ зворотної похилої риски (backslash) означає що наступний метасимвол втрачає своє спеціальне.

Вираз re.search(regexp, w) дозволяє знаходити слова w , які відповідають регулярному виразу regexp . Регулярні вирази також можна використовувати для виявлення фрагментів слів, або для модифікації слів різними способами.

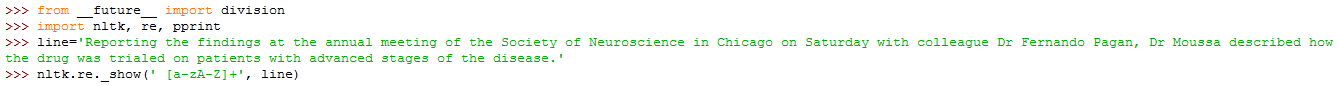
Метод re.findall() ("знайти всеl") дозволяє знайти всі відповідності даному регулярному виразу.

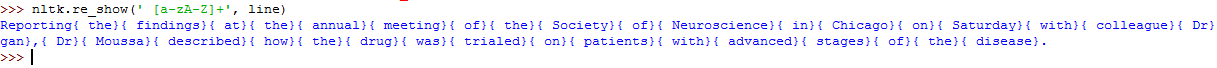
**Тексти програм на мові Python**

**Варіант 15**

1.Описати, які класи стрічок відповідають наступному регулярному виразу. [a-zA-Z]+.

Результати перевірити використовуючи nltk.re\_show()

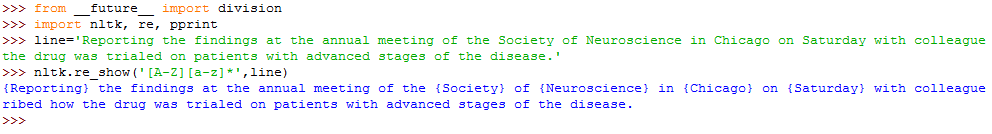




Цьому виразу відповідають стрічки довільної довжини, які складаються з великих та (або) малих літер.

2. Описати, які класи стрічок відповідають наступному регулярному виразу. [A-Z][a-z]\*.

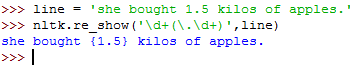
Результати перевірити використовуючи nltk.re\_show()



Цьому виразу відповідають стрічки, які складаються з однієї великої літери та 0 або більше малих літер.

3. Описати, які класи стрічок відповідають наступному регулярному виразу. \d+(\.\d+)?.

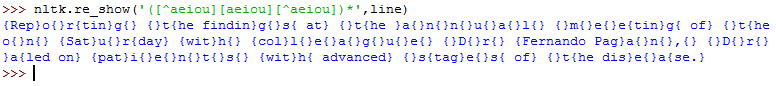
Результати перевірити використовуючи nltk.re\_show()



Цьому виразу відповідають усі десяткові числа.

4.Описати, які класи стрічок відповідають наступному регулярному виразу.

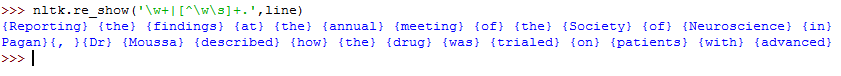
([^aeiou][aeiou][^aeiou])\*. Результати перевірити використовуючи nltk.re\_show()



Цьому виразу відповідають 0 або більше послідовностей не голосна-голосна-не голосна.

5. Описати, які класи стрічок відповідають наступному регулярному виразу. \w+|[^\w\s]+..

Результати перевірити використовуючи nltk.re\_show()



Виразу \w+|[^\w\s]+. відповідають стрічки, які складаються з довільної к-сті букв або цифр (\w+), або з довільної к-сті не букв, цифр, пробілів ([^\w\s]+) і будь-якого символу (.).

6. Описати, які класи стрічок відповідають наступному регулярному виразу. p[aeiou]{,2}t

Результати перевірити використовуючи nltk.re\_show()

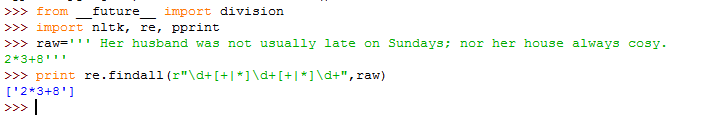


Цьому виразу відповідають стрічки, які складаються з букви ‘p’, не більше двох голосних і букви ‘t’.

8. Написати регулярний вираз, який встановлює відповідністьнаступному класу

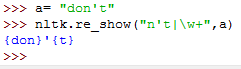
стрічок:арифметичний вираз з цілими значеннями і, який містить операції множення та

додавання *(2\*3+8*).



12. Написати регулярний вираз для токенізації такого тексту, як don't до do та n't? Пояснити

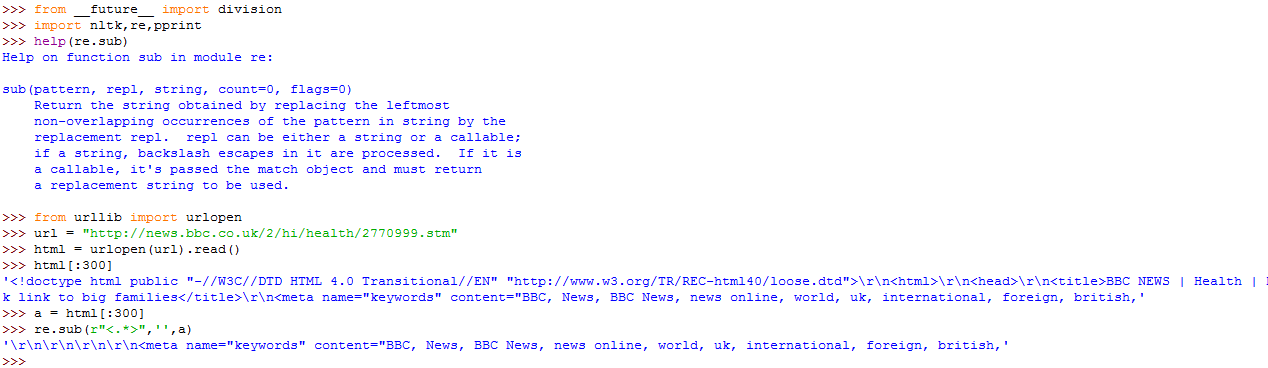
чому цей регулярний вираз не працює: «n't|\w+».



14. Прочитати файл допомоги про функцію re.sub() використовуючи help(re.sub) .

Використовуючи re.sub напишіть програму видалення HTML розмітки замінивши її на

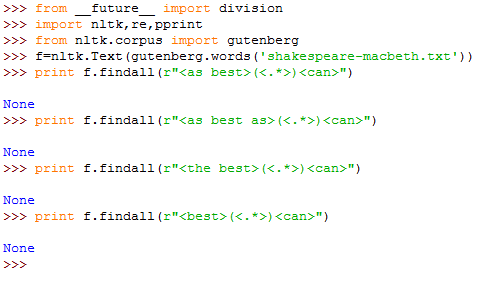
пробіли.



Функція re.sub() виконує заміну рядків на основі регулярних виразів.

15. Прочитати Додаток А. Дослідити явища описані у Додатку А використовуючи корпуси

текстів та метод findall()для пошуку в токенізованому тексті.



**Висновок:** на цій лабораторній роботі я розглянула поняття регулярних виразів та навчилася з ними працювати. Дізналася що таке метасимволи, токенізатор та розглянула основні метасимволи регулярних виразів.