МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «ЛЬВІВСЬКА ПОЛІТЕХНІКА»

Кафедра САПР



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:

ст.гр. ПРЛм-12

Щур Н. М.

Перевірив:

Старший викладач кафедри САПР

Дупак Б. П.

Львів 2015

**МЕТА РОБОТА**

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

## Використання регулярних виразів для виявлення слів за заданими шаблонами.

Багато задач лінгвістичних досліджень передбачають встановлення відповідності заданому шаблону. Наприклад, можна знайти слова, які закінчуються на “ed” використовуючи метод endswith('ed'). Подібні методи перевірки слів перелічені в Таблиці Методичних вказівок до лабораторної роботи №2. 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. Корпус слів англійської мови Words Corpus буде використовуватися в якості лінгвістичних даних серед яких буде проводитися пошук. Попередня підготовка списку слів передбачає видалення власних імен.

**Використання основних метасимволів (операторів повтору).**

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.

## Застосування регулярних виразів

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.

**Тексти програм на мові PYTHON**

**Варіант 11**

**Завдання 1, 2, 3, 4, 5, 6, 8, 10, 13, 15**

**Завдання 1.** Описати, які класи стрічок відповідають наступному регулярному виразу. [a-zA-Z]+. Результати перевірити використовуючи nltk.re\_show()

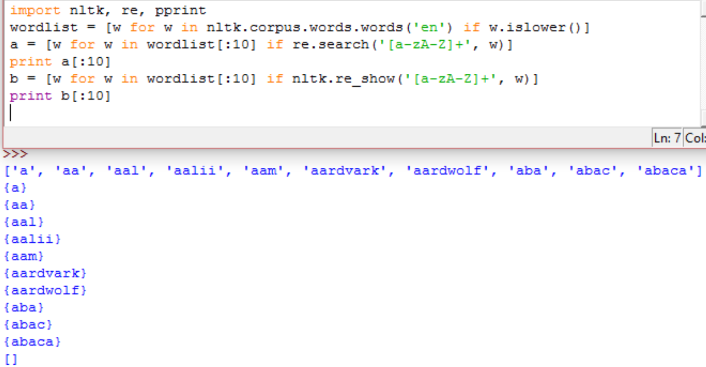


Рис.1. Завдання 1

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

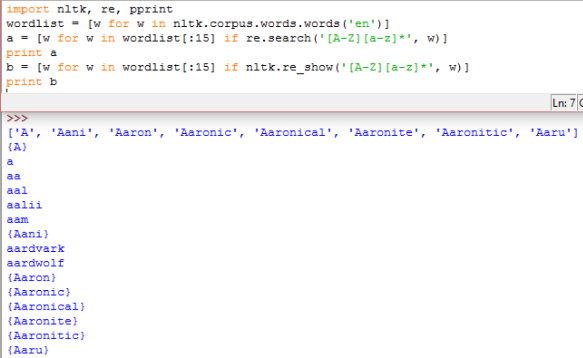


Рис. 2. Завдання 2

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

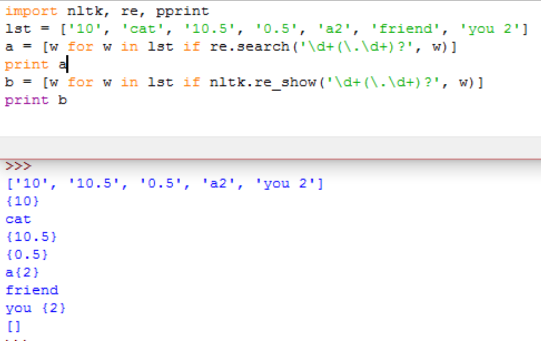


Рис. 3. Завдання 3

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

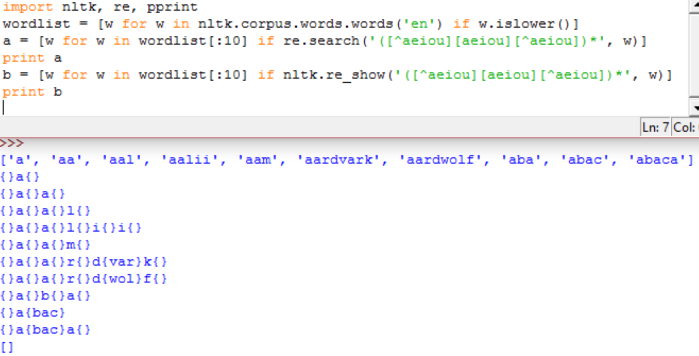


Рис. 4. Завдання 4

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

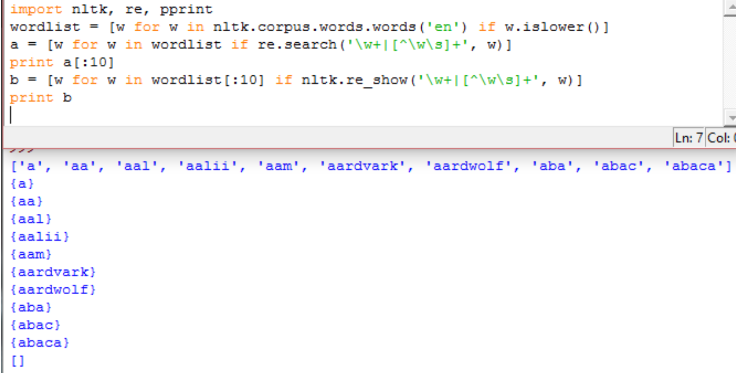


Рис. 5. Завдання 5

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

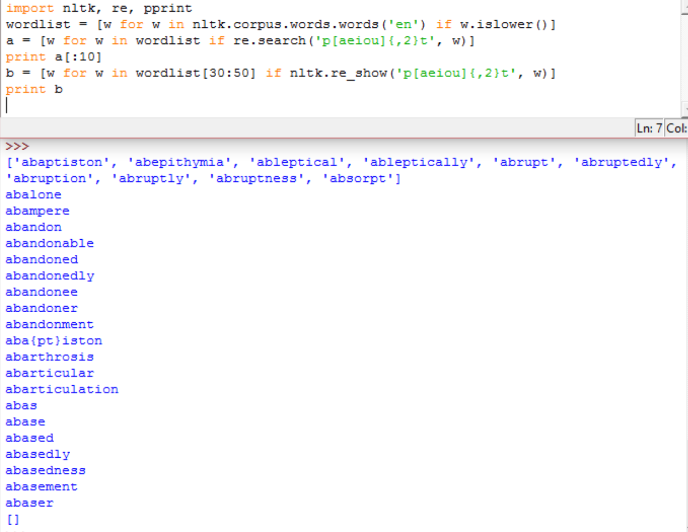
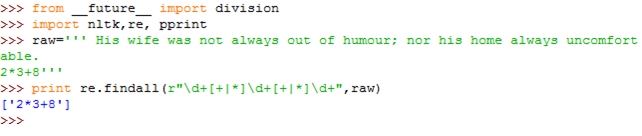


Рис. 6. Завдання 6

**Завдання 8.** Написати регулярний вираз, який встановлює відповідність наступному класу стрічок:арифметичний вираз з цілими значеннями і, який містить операції множення та додавання *(2\*3+8*).

Рис. 7. Завдання 8.

**Завдання 10.** 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 виразів: грошові одиниці, дати, імена людей та організацій. Використовувати багаторядковий запис регулярного виразу з коментарями та «verbose flag».

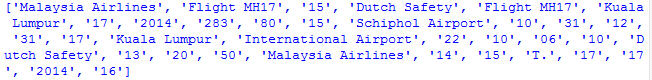
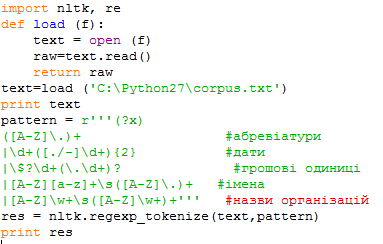


Рис. 8. Завдання 10

**Завдання 13.** Напишіть програму, яка конвертує текст в *Pig Latin. String->ingstray, idle->idleay*. (Конвертація відбувається переміщенням приголосної або групи приголосних на початок слова та додаванням до слова ay ).

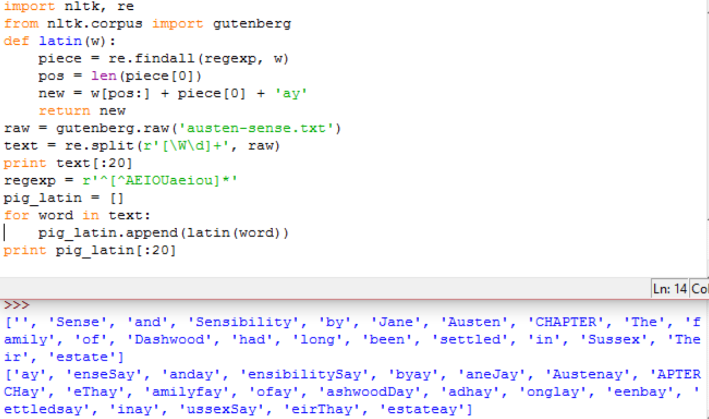


Рис. 9. Завдання 13

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

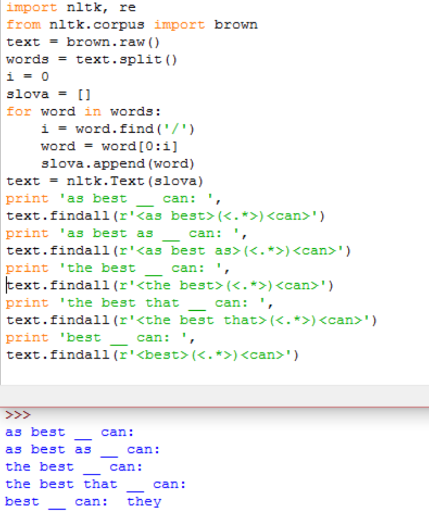


Рис. 10. Завдання 15

**Висновок:** сьогодні на занятті я вивчила основи програмування на мові *Python* та опанувала використання регулярних виразів для обробки текстів.