МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ “ЛЬВІВСЬКА ПОЛІТЕХНІКА”

ІНСТИТУТ КОМП’ЮТЕРНИХ НАУК ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

*Кафедра “Системи автоматизованого проектування”*



Звіт

до лабораторної роботи №11

на тему: «Вивчення бібліотеки прикладних програм NLTK для опрацювання текстів природною мовою. Автоматичний синтаксичний аналіз (частина2)»

з дисципліни “Комп’ютерна лінгвістика”

Виконала:

студентка групи ПРЛм-11

Павлів І.О.

Прийняв:

викладач

Дупак Б.П.

Львів-2015

**Мета роботи:**

* Вивчення основ програмування на мові *Python*.
* Ознайомлення з автоматичним синтаксичним аналізом в NLTK.

**КОРОТКІ ТЕОРЕТИЧНІ ВІДОМОСТІ**

## Деякі граматичні дилеми

### Лінгвістичні дані та необмежені можливості.

При виконанні попередніх лабораторних робіт значна увага приділялася обробці та аналізу текстових корпусів текстів та наголошувалось на проблемах обробки природної мови в зв’язку з лавиноподібною кількість лінгвістичних даних. Припустимо, що можна побудувати корпус, який буде містити все що було сказано чи написано англійською (чи будь-якою іншою мовою) за останні 50 років. Чи справедливо назвати такий корпус – корпусом сучасної англійської мови? Очевидно, що легко знайти багато причин щоб дати негативну відповідь. Якщо здійснити пошук словосполучення «the of» то виявляється можна знайти дуже багато прикладів його вживання, наприклад New man at the of IMG (<http://www.telegraph.co.uk/sport/2387900/New-man-at-the-of-IMG.html>). (<http://www.telegraph.co.uk/sport/2387900/New-man-at-the-of-IMG.html>). Спеціалісти, які володіють англійською мовою назвуть такі приклади помилкою, і скажуть що ці приклади не належать англійській мові. Відповідно, не можна вважати «сучасною англійською» велику кількість послідовностей слів з нашого уявного корпуса. Носії мови можуть розглядати такі послідовності і відкидати деякі з них як такі що не є граматичними (не відповідають граматиці природної мови). Звичайно, можна побудувати нове речення і знайти носіїв мови, які скажуть що це речення належить мові.

Наприклад, цікавою властивістю речень є те що вони можуть вкладатися в більші речення. Розглянемо наступні речення:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| (1) |  | |  |  |  | | --- | --- | --- | | a. |  | Usain Bolt broke the 100m record |  |  |  |  | | --- | --- | --- | | b. |  | The Jamaica Observer reported that Usain Bolt broke the 100m record |  |  |  |  | | --- | --- | --- | | c. |  | Andre said The Jamaica Observer reported that Usain Bolt broke the 100m record |  |  |  |  | | --- | --- | --- | | d. |  | I think Andre said the Jamaica Observer reported that Usain Bolt broke the 100m record | |

Якщо замінити перше речення на символ S, то наступні речення будуються за шаблонами такими як Andre said S та I think S. Ці шаблони та подібні до них шаблони (S but S, та S when S) дозволяють на основі одного речення побудувати більше речення

Мета граматики – дати явний опис природної мови. Щоб описати мову потрібно визначитись що вважати природною мовою та вивчити основні підходи до її представлення.

В лабораторній роботі розглядається формальне представлення породжуючої граматики, згідно якої мова представляється, як множина всіх граматично вірних речень, а граматика це формальна система , яка може бути використана для генерації елементів цієї

## Контекстно-вільна граматика

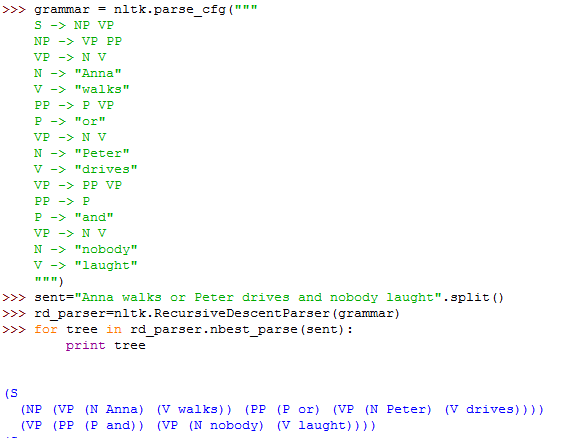
Згідно означення, першим символом зліва в першому правилі граматики є спеціальний початковий символ S, і всі дерева повинні мати цей символ , як корінь. В NLTK, контекстно-вільна граматика визначається в модулі nltk.grammar .

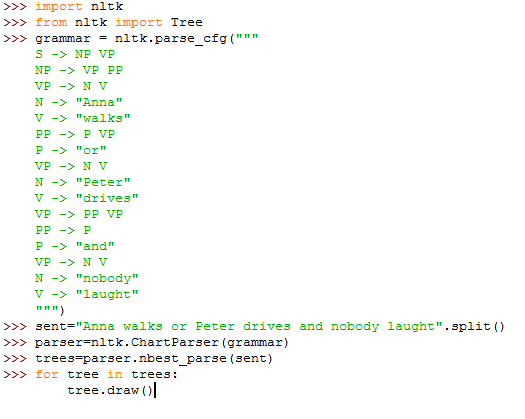
Таблиця 1. Синтаксичні категорії

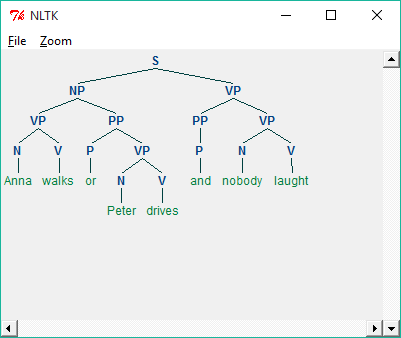
| **Символ** | **Значення** | **Приклад** |
| --- | --- | --- |
| S | sentence | the man walked |
| NP | noun phrase | a dog |
| VP | verb phrase | saw a park |
| PP | prepositional phrase | with a telescope |
| Det | determiner | the |
| N | noun | dog |
| V | verb | walked |
| P | preposition | in |

**Варіант №13**

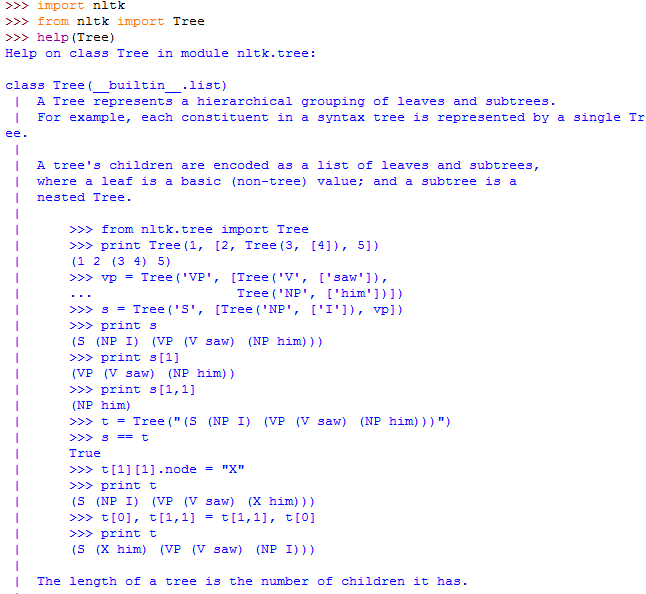
1. Здійснити аналіз речення Kim arrived or Dana left and everyone cheered. Знайти частини речення які охоплюють сполучники and та or. Побудувати дерева, які відповідають цим двом інтерпретаціям.



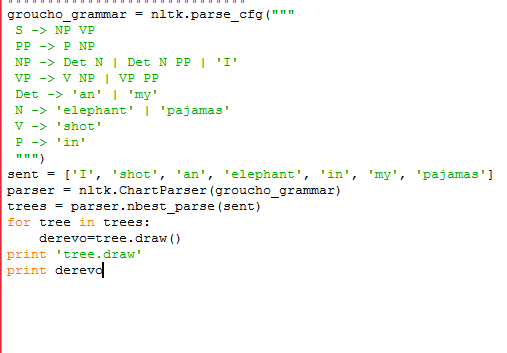


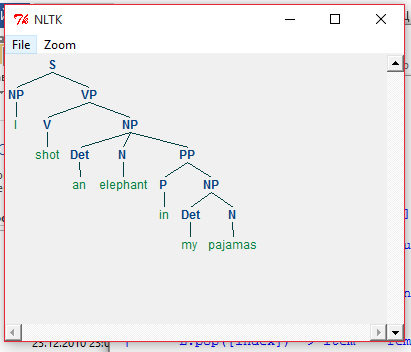


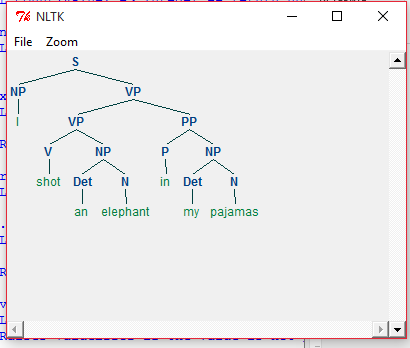
1. В класі Tree реалізовано різноманітні корисні методи. Переглянути файл допомоги  Tree з документації та описати основні з цих методів (import Tree,  help(Tree).

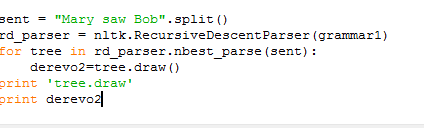


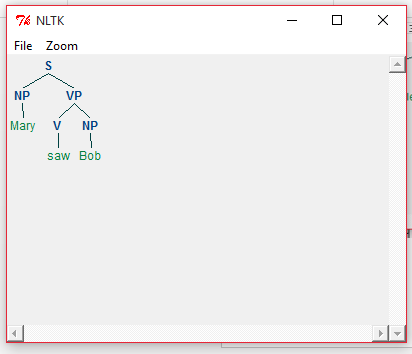
1. Перетворити всі дерева , які зустрічаються в методичних вказівка і зображені за допомогою дужок використовуючи  nltk.Tree() . Використовувати draw() для побудови графічного зображення дерева.

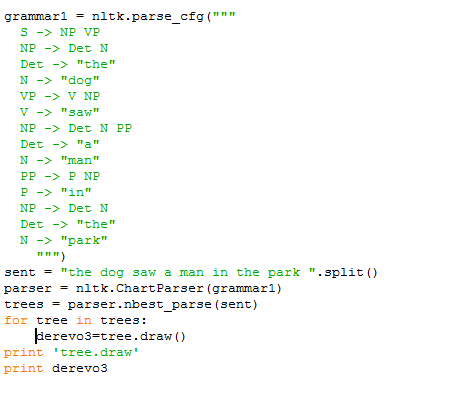


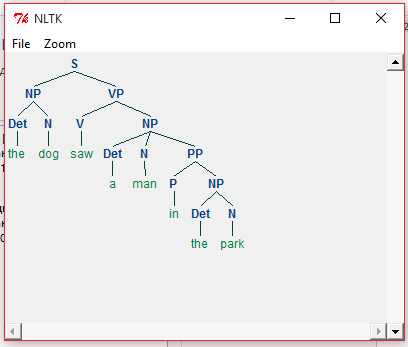


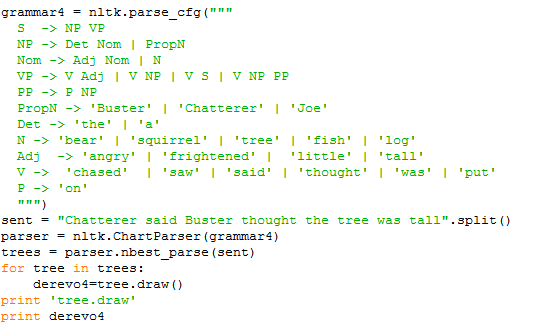


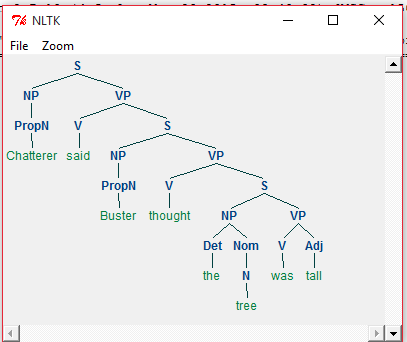


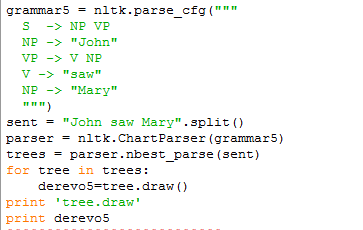


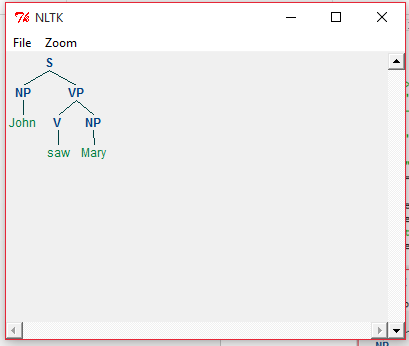


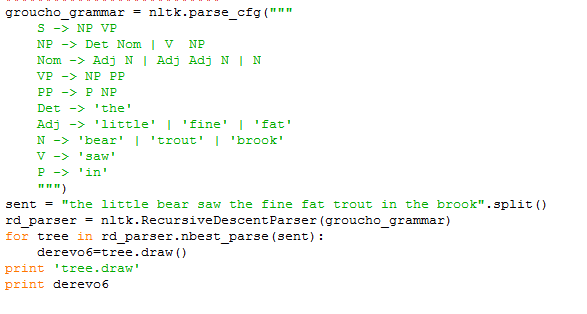


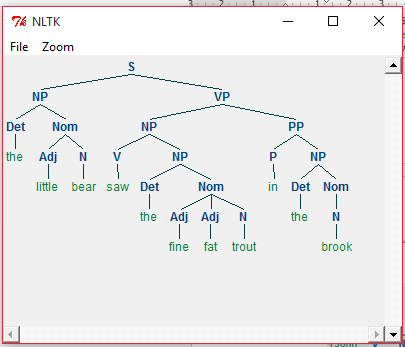




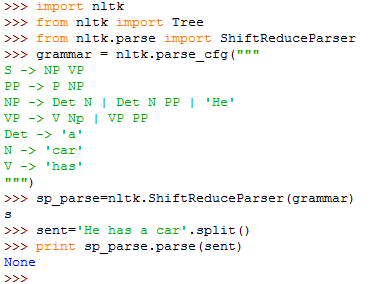




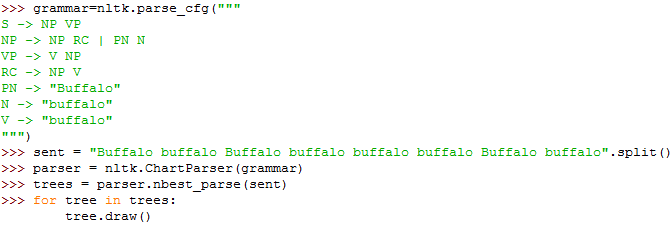


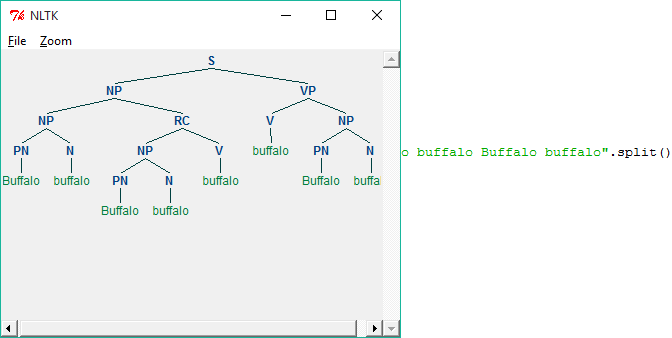


1. Використовуючи ручку і папір, здійснити всі кроки роботи аналізаторів рекурсивного спуску та переміщення згортання для довільного речення використовуючи просту контексно-вільну граматику.



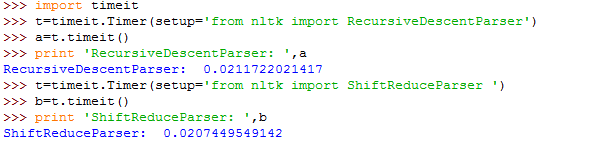
1. Здійснити аналіз послідовності слів: Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo. Оскільки, згідно з http://en.wikipedia.org/wiki/Buffalo\_buffalo\_Buffalo\_buffalo\_buffalo\_buffalo\_Buffalo\_buffalo це граматично правильне речення, напишіть контексно-вільну граматику на основі дерева наведеного на цій сторінці з Інтернету. Здійсніть нормалізацію слів (lowercase), для моделювання ситуації коли слухач сприймає це речення на слух. Скільки дерев розбору може мати це дерево в такому випадку?

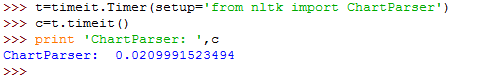






1. Написати програму порівняння швидкодії всіх аналізаторів, які згадувалися в методичних. Використовувати timeit функцію для визначення часу синтаксичного аналізу одного і того самого речення різними аналізаторами.





**Висновок:**

на цій лабораторній роботі я ознайомилася з автоматичним синтаксичним аналізатором NLTK, писала програми побудов дерев для речень, здійснювала аналіз послідовності слів, писала контексно-вільну граматику на основі наведеного дерева, здійснювала модифікацію граматики в демонстраційній програмі синтаксичного аналізу згідно алгоритму рекурсивного спуску.