МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «ЛЬВІВСЬКА ПОЛІТЕХНІКА»

### Кафедра САПР

ЗВІТ

до лабораторної роботи № 9

на тему:

ВИВЧЕННЯ БІБЛІОТЕКИ ПРИКЛАДНИХ ПРОГРАМ NLTK, ДЛЯ ОПРАЦЮВАННЯ ТЕКСТІВ ПРИРОДНОЮ МОВОЮ.

АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗ (частина1)

з дисципліни “Комп’ютерна лінгвістика”

Виконала:

Студентка групи ПРЛм-12

Василишин В. В.

Перевірив:

Асистент кафедри САПР

Дупак Б. П.

Львів 2015

МЕТА РОБОТИ

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

КОРОТКІ ТЕОРЕТИЧНІ ВІДОМОСТІ

Процес класифікації слів за їх приналежністю до частини мови і їх відповідне маркування називається морфологічним аналізом (tagging, POS tagging). В загальному для назв цих груп слів вживаються терміни – класи слів, лексичні категорії, частини мови.

Перелік тегів, який використовується для цієї специфічної задачі називається набором тегів. Наголос в цій та наступній лабораторних роботах буде зроблено на вивченні використання тегів і здійснення автоматичного морфологічного аналізу. Автоматичний морфологічний аналіз це дуже важливий та цінний етап опрацювання текстів природною мовою, результати якого мають широке застосування.

Морфологічний аналізатор (POS-tagger), це програма, яка обробляє послідовність слів і ставить у відповідність до кожного з них, відповідний тег (тег відповідає певному набору морфологічних характеристик.

За домовленістю в NLTK промарковані слова (tokens) представляються з використанням типу даних – кортеж, Кортеж містить слово і тег. Для створення таких спеціальних кортежів з стандартної стрічки промаркованих слів потрібно використовувати функцію str2tuple().

Промарковані слова представляються у вигляді(word, tag), де слово асоціюється з тегом, який відповідає певним морфологічним характеристикам. Автоматичний морфологічний аналіз можна розглядати , як задачу пошуку відповідного тега для слова. Найпростіший спосіб збереження та обробки таких відповідностей (відображень) в Python це використання такого типу даних, як словник.

Python підтримує такий тип даних, як словник, і він використовується для встановлення відповідності (відображення) між довільними типами даних. Визначимо pos як пустий словник і додамо до нього значення, які встановлюють частиномовну приналежність певних слів . Значення додаються із використанням квадратних дужок.

ТЕКСТИ ПРОГРАМ НА МОВІ PYTHON

ВАРІАНТ №2

2. Проаналізувати таблицю 4.

У даній таблиці представлено лінгвістичні об’єкти та відповідності, яких вони потребують. До основних лінгвістичних даних належать індексовані документи, тезаурус, словник, порівняльний список слів та морфологічний аналізатор.

|  |  |  |
| --- | --- | --- |
| **Лінгвістичний об'єкт** | **Позначуване** | **Позначене** |
| Картка платника податків | Номер картки | Дані про особу |
| Словник антонімів | Слово | Антоніми до слова |
| Словник термінів | Слово | Визначення терміну |
| Список неправильних дієслів | Дієслово | Форми past tense та past participle. |

3. Опрацювати приклади з методичних вказівок по роботі зі словниками. Що станеться, якщо доступитися до неіснуючого запису звичайного словника та словника по замовчуванню?

*Звичайний словник*

>>> pos = {'colorless': 'ADJ', 'ideas': 'N', 'sleep': 'V', 'furiously': 'ADV'}

>>> pos

{'furiously': 'ADV', 'sleep': 'V', 'ideas': 'N', 'colorless': 'ADJ'}

>>> pos['night']

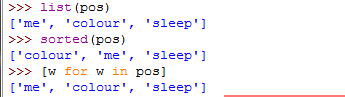
Traceback (most recent call last):

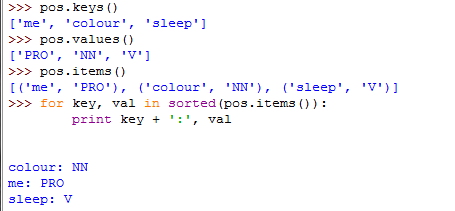
File "<pyshell#5>", line 1, in <module>

pos['night']

KeyError: 'night'

Якщо доступатися до неіснуючого запису звичайного словника, то вибиває помилку.

До словника можна застосовувати функції list() та sorted(), а також ітерацію



Функція keys() виводить значення ключів, values() - значень, items() - елементів

**Словник за замовчуванням**

>>> pos = nltk.defaultdict(list)

>>> pos['blog']

[] #пуста множина, бо не вказано ніяких параметрів для значення

>>> pos = nltk.defaultdict(lambda: 'V') #доступаюся до неіснуючого запису словника

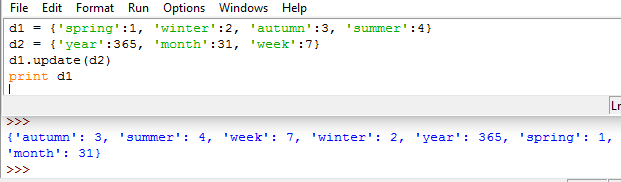
>>> pos['google'] #автоматично додається

'V'

У випадку зі словником за замовчуванням слово автоматично додається до словника.

5. Створити два словники. Що станеться зі словниками після виконання команди d1.update(d2).

Програма додає всі записи із словника d2 до словника d1.

Рис. 1. Текст програми №5.

7. Використовуючи sorted() та set()отримайте відсортований список всіх тегів корпуса Brown без їх дублювання.

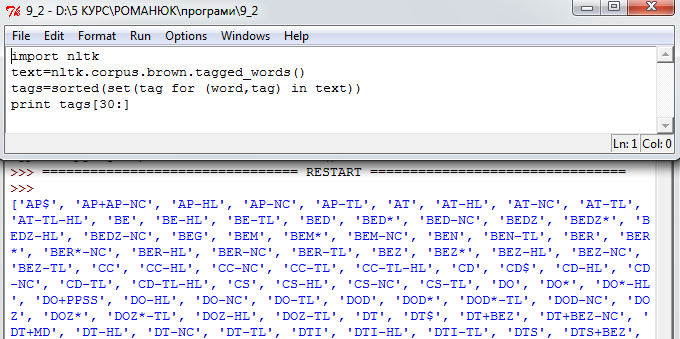


Рис. 2. Текст програми №7

11. Напишіть програму, яка обробить Brown Corpus і допоможе відповісти на

наступне запитання: які теги для маркування іменників найчастіше

використовуються і що вони означають.

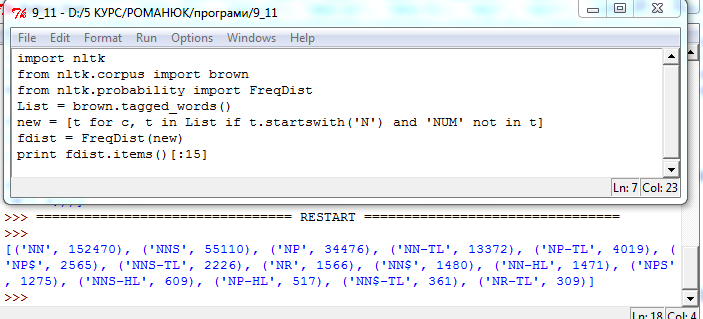


Рис. 3. Текст програми №11

13. Напишіть програму для збору статистичних даних по розмічених корпусах і відповіді на наступне запитання: скільки слів мають неоднозначності в змісті того, що маркуються двома і більше тегами.

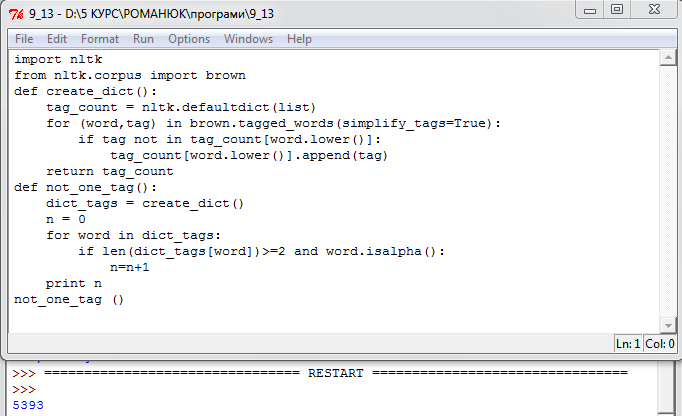


Рис. 4. Текст програми №13

17. Напишіть програми для знаходження слів та словосполучень згідно відповідних їм тегів для відповіді на наступне питання: які слова можуть бути іменниками (множина) або дієсловами (третя особа однини)(deals, flies)?

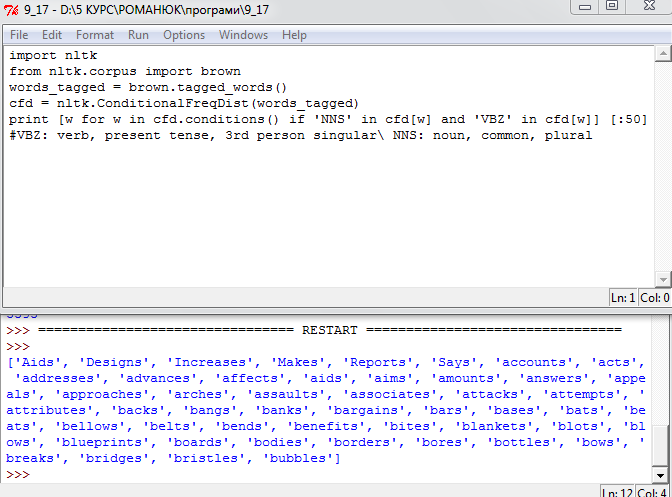


Рис. 5. Текст програми №17

21. Написати програму побудови словника, записами якого будуть набори словників. Використовуючи створений словник, збережіть у ньому набори можливих тегів, які зустрічаються після заданого слова з певним тегом, наприклад word i → tag i → tag i+1.

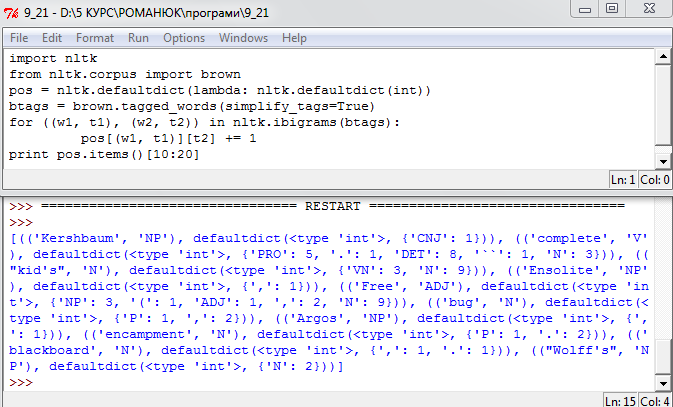


Рис. 6. Текст програми №21