МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ “ЛЬВІВСЬКА ПОЛІТЕХНІКА”

ІНСТИТУТ КОМП’ЮТЕРНИХ НАУК ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

### *Кафедра “Системи автоматизованого проектування”*



Звіт

до лабораторної роботи №10

на тему: ВИВЧЕННЯ БІБЛІОТЕКИ ПРИКЛАДНИХ ПРОГРАМ NLTK, ДЛЯ ОПРАЦЮ­ВАННЯ ТЕКСТІВ ПРИРОДНОЮ МОВОЮ.АВТОМАТИЧНИЙ МОРФОЛОГІЧНИЙ АНАЛІЗ.”

з дисципліни “Комп’ютерна лінгвістика”

Виконала:

студентка групи ПРЛм-11

Свіляк О.Ю.

Прийняв:

викладач

Дупак Б.П.

Львів-2015

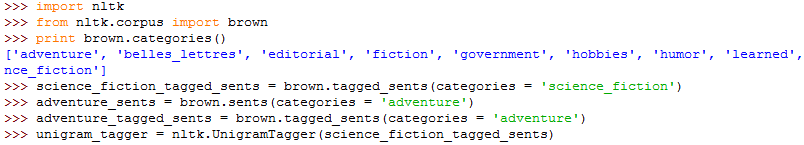
**Мета роботи:**

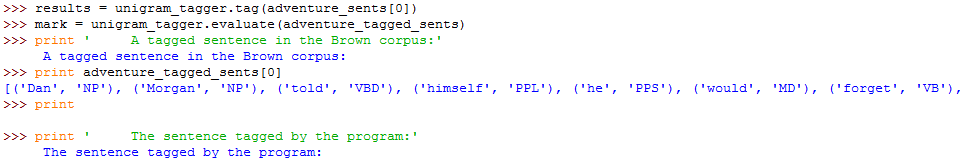
* Вивчення основ програмування на мові *Python*.
* Ознайомлення з автоматичним морфологічним аналізом в NLTK.

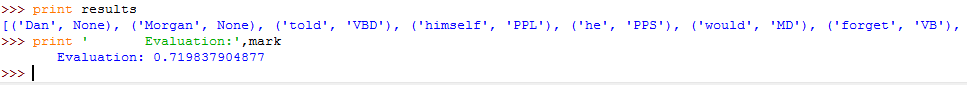
**Тексти програм на мові Python**

**Варіант 15**

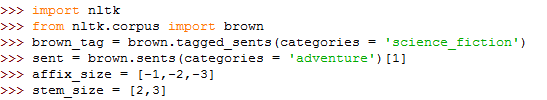
1. Здійсніть тренування юніграм аналізатора на основі частини корпуса, який відповідає першій або другій літері прізвища студента та виконайте аналіз тексту з частини корпуса, яка відповідає першій або другій літері імені студента. Результати поясніть. Чому для деяких слів не встановлені теги.

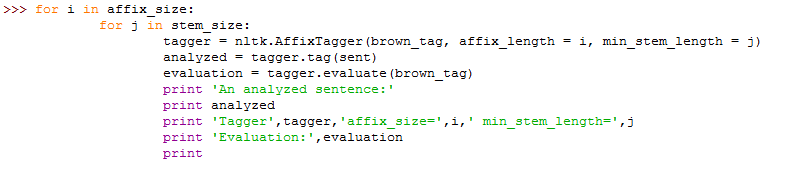


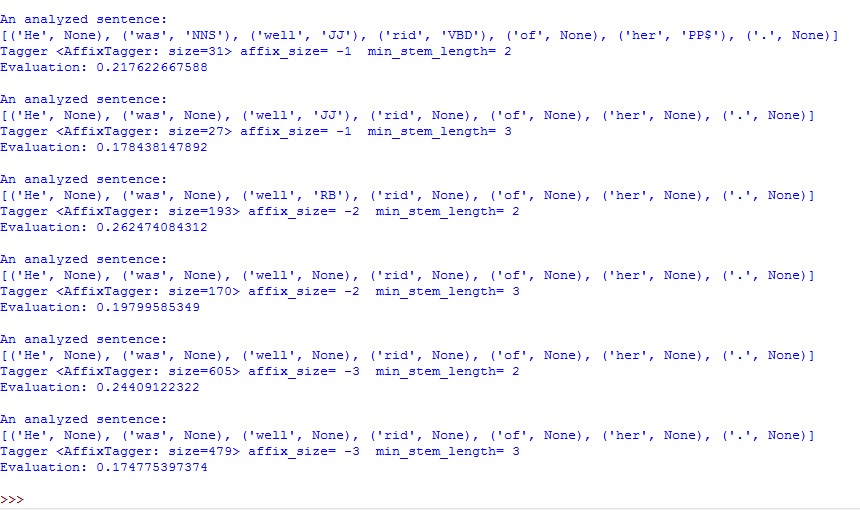




2. Прочитати файл допомого про морфологічний аналізатор на основі афіксів (help(nltk.AffixTagger)). Напишіть програму, яка викликає аналізатор на основі афіксів в циклі, з різними значеннями довжини афіксів і мінімальними довжинами слів. При яких значеннях можна отримати кращі результати.

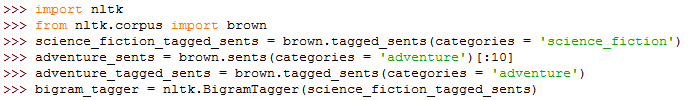


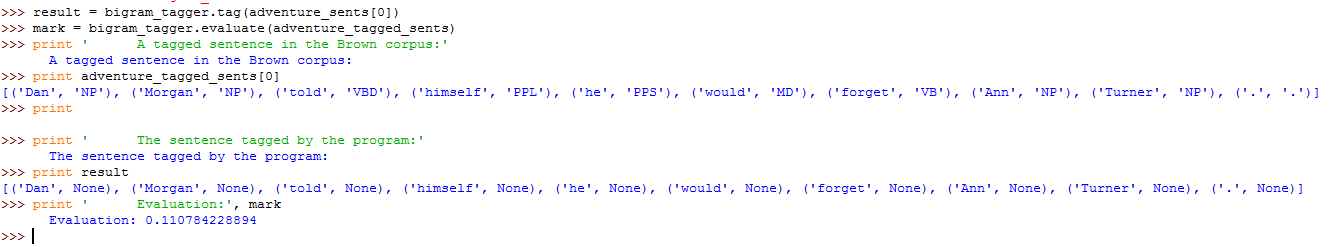




В результаті виконання програми бачимо, що найбільшу точність отримано у випадку, коли довжина афікса дорівнює 2 і мінімальна довжина кореня теж дорівнює 2. Ці показники зможемо використати в інших програмах.

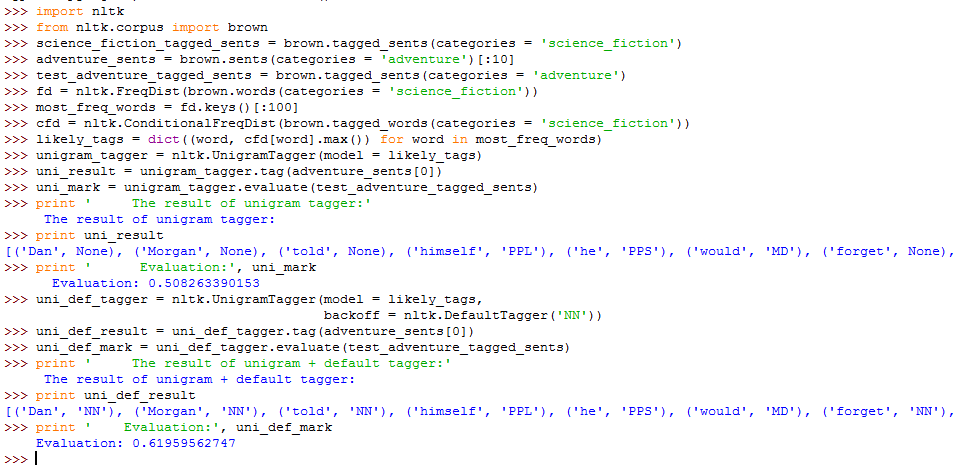
3. Здійсніть тренування біграм аналізатора на частинах корпуса з вправи 3.1 без backoff аналізатора. Перевірте його роботу. Що відбулося з продуктивністю аналізатора? Чому?



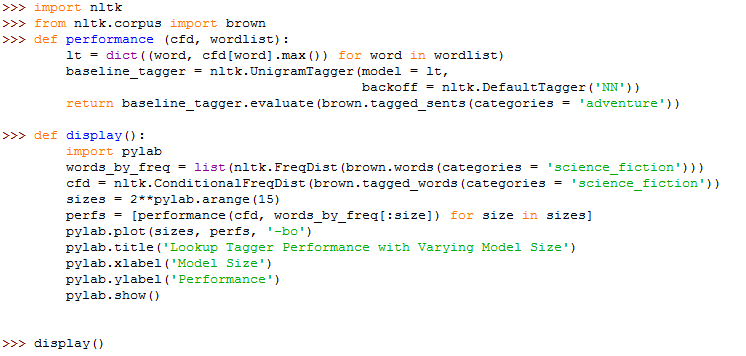


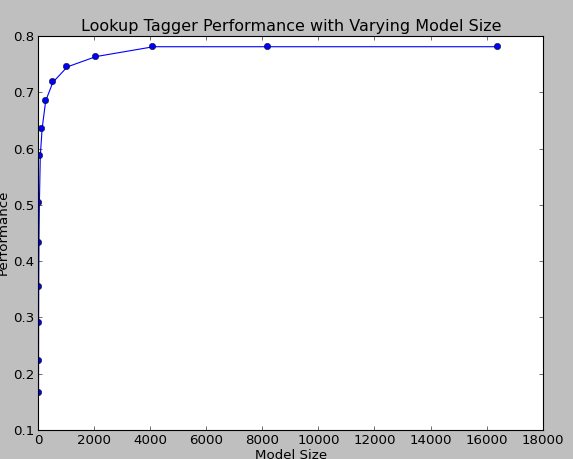
Коли у тексті зустрілося нове слово, аналізатор не може для нього встановити тег. Так само аналізатор не маркує наступне слово, навіть якщо воно зустрічалося при тренуванні, оскільки це слово в даних для тренування ніколи не зустрічалося після слова, тег якого None. Це приводить до того що і всі наступні слова в реченні не маркуються.

4. Дослідити наступні проблеми. що виникають при роботі з аналізатором на основі підстановок: що відбудеться з продуктивністю аналізатора, якщо опустити backoff аналізатор (дослідити на частині броунівського корпусу, яка відповідає першій або другій літері прізвища студента); на основі рис.1. та відповідного фрагмента програми встановити точку максимальної продуктивності незважаючи на розмір списку (об’єм оперативної пам’яті) і точку достатньої продуктивності при мінімальному розмірі списку.



При застосуванні backoff аналізатора продукривність вища, ніж при застосуванні тільки юніграм аналізатора.





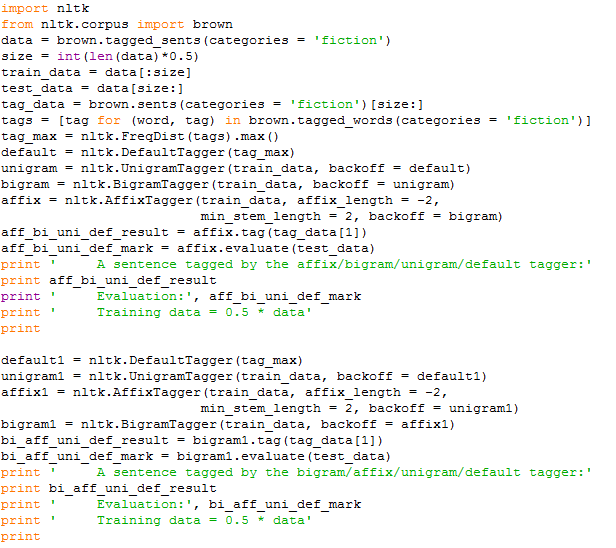
5. Знайдіть розмічені корпуси текстів для інших мов які вивчаєте або володієте (українська, польська, німецька, російська, італійська, японська). Здійсніть тренування та оцініть продуктивність роботи різних аналізаторів та комбінацій різних аналізаторів. Точність роботи аналізаторів порівняйте з точністю роботи аналізаторів для англійських корпусів. Результати поясніть.



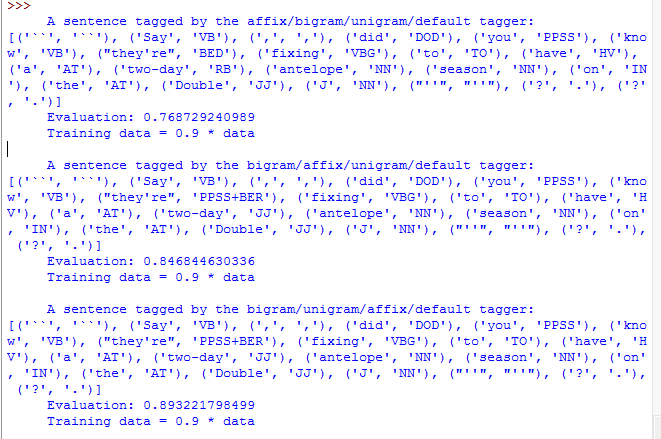


Найбільш продуктивною є комбінація “біграм аналізотор —> юніграм аналізатор —> аналізатор за замовчуванням”.

6. Створити аналізатор по замовчуванню та набір юніграм і n-грам аналізаторів. Використовуючи backoff здійсніть тренування аналізаторів на частині корпуса з вправи 3.2. Дослідіть три різні комбінації поєднання цих аналізаторів. Перевірте точність роботи аналізаторів. Визначіть комбінацію аналізаторів з максимальною точністю аналізу. Змініть розмір даних на яких проводилось тренування. Повторіть експерименти для змінених даних для тренування. Результати порівняйти і пояснити.

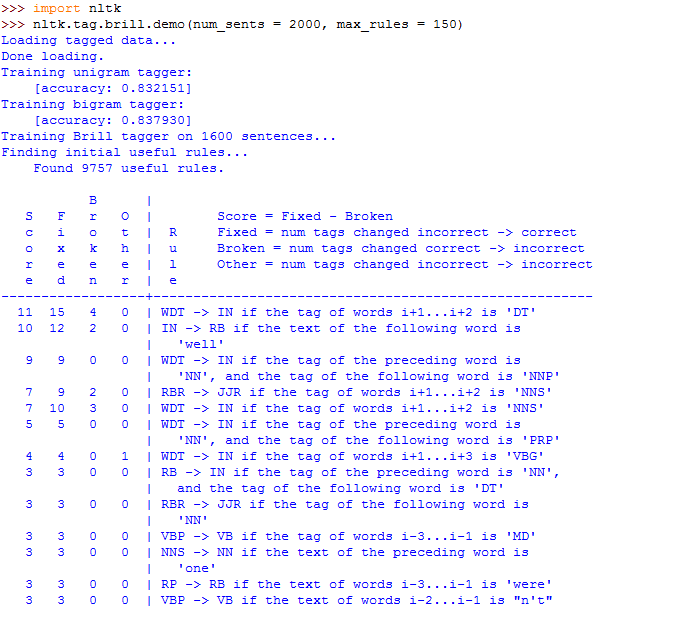


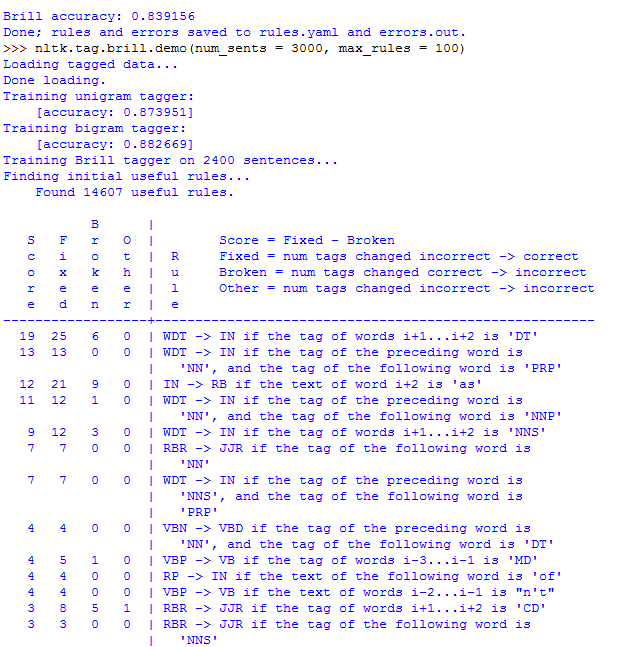


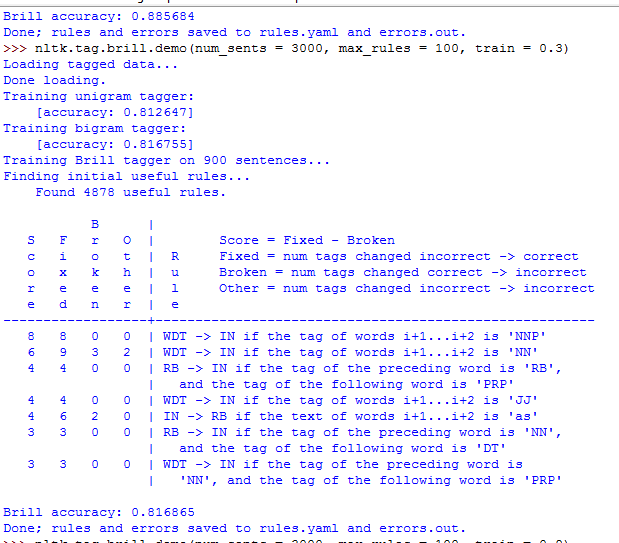


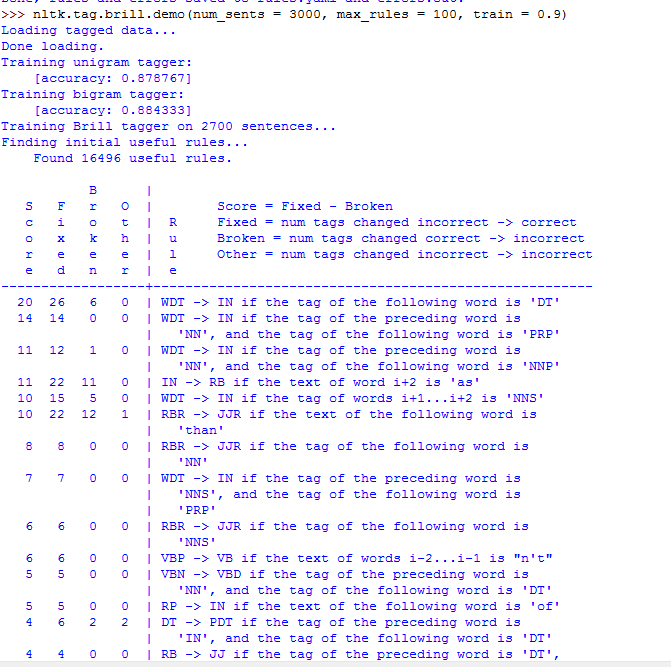
Якщо більший розмір даних для тренування, то вища продуктивність аналізатора. Комбінація “біграм аналізатор —> юніграм аналізатор —> аналізатор на основі афіксів —> аналізатор за замовчуванням” виявилася найбільш продуктивною. Найточніший результат (0.893) отримано в результаті виконання даної комбінації при розмірі даних для тренування 0.9 від загального розміру даних.

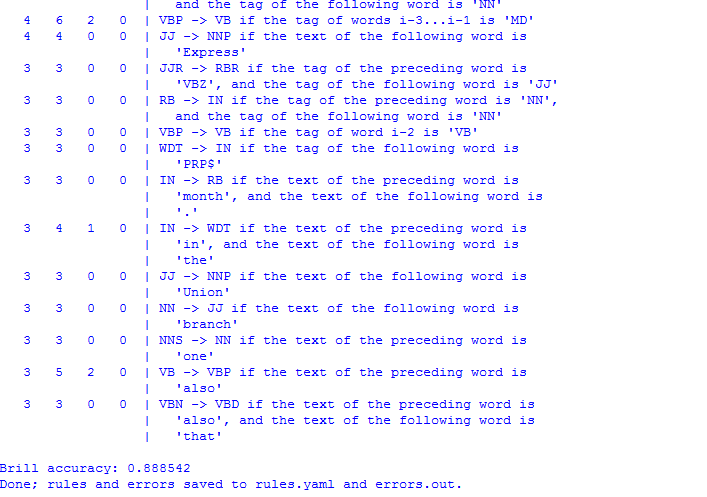
7. Прочитати стрічку документування функції demo Brill аналізатора. Здійснити експерименти з різними значення параметрів цієї функції. Встановити який взаємозв’язок є між часом тренування (навчання аналізатора) і точністю його роботи.











Якщо встановлений вищий показник точності, то час тренування програми більший, а продуктивність аналізаторів вища.

**Висновок:** на цій лабораторній роботі я вивчила основи структурного програмування мовою Python та ознайомилася з автоматичним морфологічним аналізом.