**Практична робота № 10**

***Тема:*** Бібліотека NLTK мови Python**. Статистичні дослідження корпусу.**

***Мета:*** Відпрацювати практичні навички написання програм на мові Python для обробки текстових корпусів методами бібліотеки NLTK.

***Література:***

[https://studfile.net/preview/9755743/page:7/](https://studfile.net/preview/9755743/page%3A7/) структура корпусів текстів

# [https://studfile.net/preview/9755743/page:8/](https://studfile.net/preview/9755743/page%3A8/) умовний частотний розподіл. Клас ConditionalFreqDist.

[https://www.nltk.org**/**](https://www.nltk.org/)nltk Документаціянабір інструментів природної мови

**Зміст роботи:**

***Для виконання завдань можна вибрати файл з бібліотеки gutenberg***

***на свій розсуд і виконувати завдання відповідно до обраного файлу, корегуючи параметри завдань самостійно.***

1. Для здійснення статистичних досліджень тексту трагедії У. Шекспіра "Гамлет" подайте у вигляді тексту nltk файл shakespeare-hamlet.txt бібліотеки gutenberg.

2. Перегляньте контекст появи слова “Hamlet” у цій трагедії. Скільки разів зустрічається це слово?

3. Знайдіть у цьому тексті загальний контекст слів "Hamlet" та "Horatio".

4. Побудуйте графік розподілу слів "Hamlet", "Horatio", “Ghost “, та

“Polonius” у тексті драми.

5. Визначте кількість унікальних токенів у nltk-тексті та відобразіть їх список.

6. Побудуйте графік частотного розподілу слів у к nltk-тексті та кумулятивний графік частотного розподілу слів у nltk-тексті.

7. Покажіть слова, які зустрічаються у nltk-тексті лише один раз.

8. Отримайте розподіл довжин слів у nltk-тексті.

Складіть два списки слів:

 1) довжиною більше 7 символів, частота появи яких у nltk-тексті більше шести раз;

 2) довжиною менше 3 символів, що зустрічаються в тексті менше трьох раз. Порівняйте списки. Як ви вважаєте, який зі списків передає більше інформації про зміст тексту і чому?

9. Визначте, яка довжина слова найчастіше зустрічається в nltk-тексті.

10. Побудуйте діаграму розподілу частот довжин слів. Діаграму зберегти у файлі і вивести на екран.

11. Визначте, які довжини слів трапляються у тексті.

12. Виведіть на екран наявні довжини слів, із зазначенням частотності їх

появи.

13Створіть функцію, що дозволяє визначати лексичне багатство тексту (кількість слів у тексті, поділена на кількість унікальних слів).

**Методичні рекомендації.**

Для того щоб здійснювати різні статистичні дослідження тексту, необхідно подати його у вигляді **nltk-тексту** або у вигляді об'єкта класу **nltk.Text,** використовуючи як параметр список слів (токенів).

**shek=nltk.Text(gutenberg.words(fileids="ім’я файлу")).**

Метод **concordance(«слово»)** об'єкта nltk.текст дозволяє побачити варіанти появи цього слова у різноманітних контекстах.

Метод **common\_contexts (**), параметрами якого є два слова, дозволяє побачити загальний (одинаковий) контекст, у якому зустрічаються ці два слова.

Метод **dispersion\_plot ()** дозволяє побачити розподіл списку заданих як параметри слів у тексті.

У NLTK реалізовано окремий клас **FreqDist** у модулі **nltk.probability.** Об'єкт даного класу є словником, до якого можна застосувати основні методи словника.

Для класу FreqDist() вхідними даними є список, для класу **ConditionalFreqDist()** вхідними даними є список пар.

Метод **hapaxes()** об'єкта класу **FreqDist**, дає можливість отримати список слів, що зустрічаються в даному nltk-тексті лише один раз.

Контрольні запитання.

1. Яку інформацію містить корпус?
2. Чи можна створити корпус самостійно?
3. Яку команду слід виконати щоб отримати зміст корпусу?
4. Як створити текст для обробко з файлу корпусу?
5. Як визначити кількість унікальних токенів в корпусі?
6. Назвіть основні функції статистичної обробки текстів.