**Практична робота №9**

***Тема:*** Бібліотека NLTK мови Python**.** Викорстання n- грам для проведення аналізу текстів. Візуалізація результатів.

***Мета:*** Відпрацювати практичні навички написання програм на мові Python для обробки текстової інформації використовуючи методи бібліотеки NLTK.

***Література:*** <http://nlpx.net/archives/57>

<https://python-scripts.com/matplotlib#5>

**Зміст роботи:**

**Завдання 1.**

Підготувати вхідний текст до 1000 слів, українською мовою. Провести попередню підготовку тексту ( вилучити розділові знаки, вилучити стоп-слова, провести токенізацію по словам). Самостійно створити список стоп-слів. Список стоп-слів можна отримати за посиланнями: <https://www.marazm.org.ua/windows/50_141.html> <https://github.com/skupriienko/Ukrainian-Stopwords>

**Завдання 2.**

Побудувати біграму тексту. Підрахувати кількість входжень кожної біграми в текті.

**Завдання 3.**

Визначити 5-10 найбільших входжень, та побудувати графік/діаграму для ілюстрації результатів.

**Завдання 4.**

Визначити чи містить ваш текст колокації. Аргументуйте вашу відповідь.

**Методичні рекомендації.**

Метод **ngram() –** для побудови n-грам.

**nltk.FreqDist()** – модуль використовується для кодування "частотних розподілів", які підраховують кількість разів, коли кожен результат експерименту виникає. Завдяки використанню класу FreqDist ми можемо отримати частоти кожного токена.

**most\_common())** - метод повертає найпоширеніші елементи із лічильника. Якщо ми не надаємо значення «n», то відсортований словник повертається з найменш поширених елементів.

**Контрольні запитання.**

1. Поясніть призначення n-грам.
2. Що є результатом роботи методу **ngram()**?
3. Як можна використати побудовані біграми?
4. Опишіть спосіб оцінки ймовірностей n-грам.
5. Що таке колокація?
6. Назвіть основні види колокацій?
7. Як використовують колокації?
8. Як побудувати діаграму в Matplotlib?