**Лабораторна робота № 8**

***Тема:*** Бібліотека NLTK мови Python**. Визначення тональності текстів. n-грами. Колокації.**

***Мета:*** Відпрацювати практичні навички написання програм на мові Python для визначення тональності текстів та знаходження колокацій.

***Література:***

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

<https://lang.org.ua/uk/dictionaries/> словник тональності української мови

**Зміст роботи:**

**Завдання 1.** ( за матеріалами лекції 9)

Взяти текст в розмірі 30 речень на українській мові. Провести процедуру токенізації, вилучення розділових знаків та стоп слів.

Визначити тональність обраного тексту використовуючи словник тональності української мови. (Словник тональності можна розробити самостійно)

**Завдання 2.** ( за матеріалами лекції 9)

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

Визначити 10 найбільших входжень

Визначити чи містить ваш текст колокації. Аргументуйте вашу відповідь.

**Методичні рекомендації.**

Підходи, засновані на словниках, використовують звані тональні словники (affective lexicons) для аналізу тексту. У простому вигляді тональний словник є список слів зі значенням тональності для кожного слова. <https://lang.org.ua/uk/dictionaries/>

Щоб проаналізувати текст, можна скористатися наступним алгоритмом: спочатку кожному слову в тексті привласнити його значенням тональності зі словника (якщо воно є условнику), а потім обчислити загальну тональність всього тексту. Обчислювати загальну тональність можна у різний спосіб. Найпростіший із них — середнє арифметичне всіх значень.

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

**nltk.FreqDist()** – модуль використовується для кодування "частотних розподілів", які підраховують кількість раз, 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.

Колокації є послідовністю слів, які проявляються разом частіше, ніж незалежно один від одного.

У деякому наближенні можна сказати, що двослівні колокації - це біграми, що часто зустрічаються. Тобто практично необхідно визначити біграми, які з'являються в тексті частіше, ніж очікується виходячи з частоти появи складових слів.

**Контрольні запитання.**

1. Дайте визначення поняттю тональність текстів.
2. Як визначити тональність текстів?
3. Які існують види тональності текстів?
4. Які операції бажано виконати перед визначенням тональності тексту?
5. Де використовують визначення тональності текстів.
6. Поясніть призначення n-грам.
7. Що є результатом роботи методу **ngram()**?
8. Як можна використати побудовані біграми?
9. Опишіть спосіб оцінки ймовірностей n-грам.
10. Що таке колокація?
11. Назвіть основні види колокацій?
12. Як використовують колокації?