Міністерство освіти і науки України

Західноукраїнський національний університет

Факультет комп’ютерних інформаційних технологій

Кафедра інформаційно-обчислювальних систем і управління

Звіт до лабораторної роботи №10

З дисципліни “Інтелектуальний аналіз даних”

Студента групи КН-31

Ковальковського В. В.

Тернопіль 2023

**Лабораторна робота №10**

**Простий аналіз текстової інформації**

Варіант:

| 13 | 14 | ABP Majha | Marathi |

**Парсинг**

Збір даних (назв статей) я вирішив проводити по тегу <p> та класі «other-news», так як в цих параметрах прописувались назви статей



Рис. 1 – Результат виконання порсингу

**Переклад назв**

Переклад я вирішив зробити в два кроки, спочатку з маратхі на англійську, потім з англійської на українську



Рис. 2 – Перекладені назви на англійську

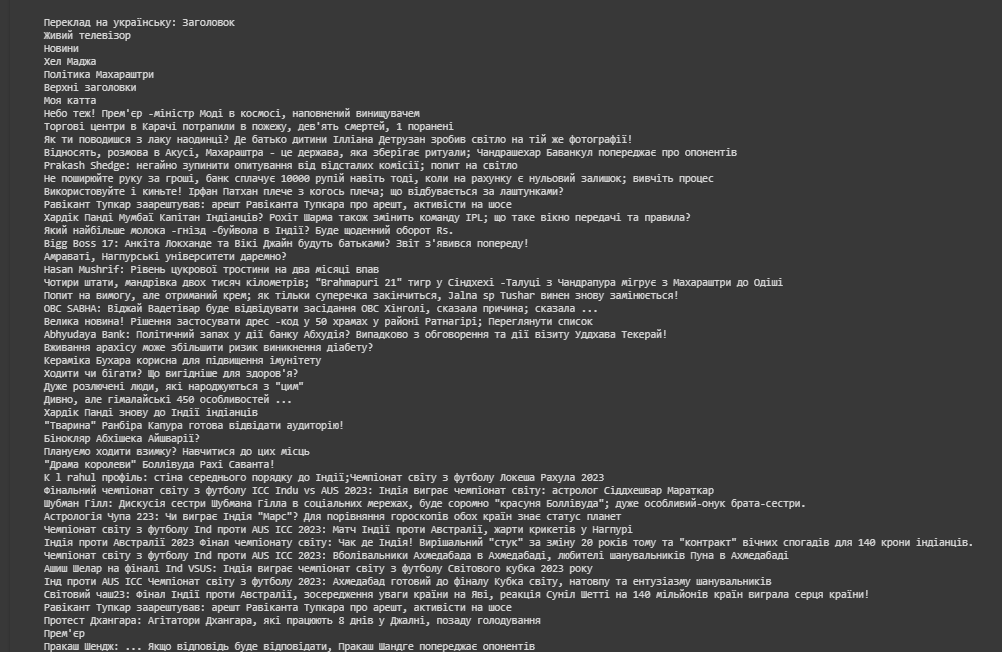


Рис. 3 – Перекладені назви на українську

**Простий аналіз слів**

У цьому пункті всі назви були розбиті на окремі слова, визначалась їх довжина та унікальність

Результати:

Загальна кількість слів: 583

Кількість унікальних слів: 453



Рис. 4 – Зображення хмар слів

Очищення тексту (вилучення спеціальних символів, цифр і інших незначущих символів):

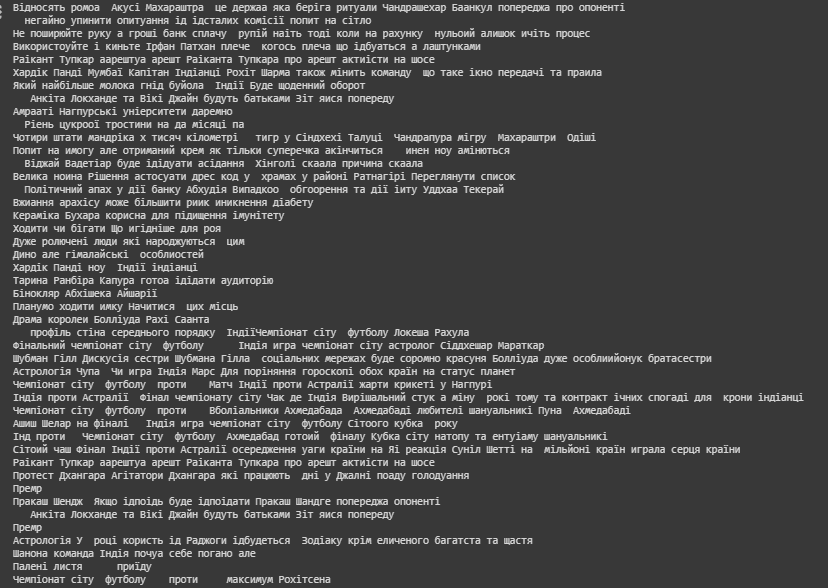


Рис. 5 – Результат виконання очищення

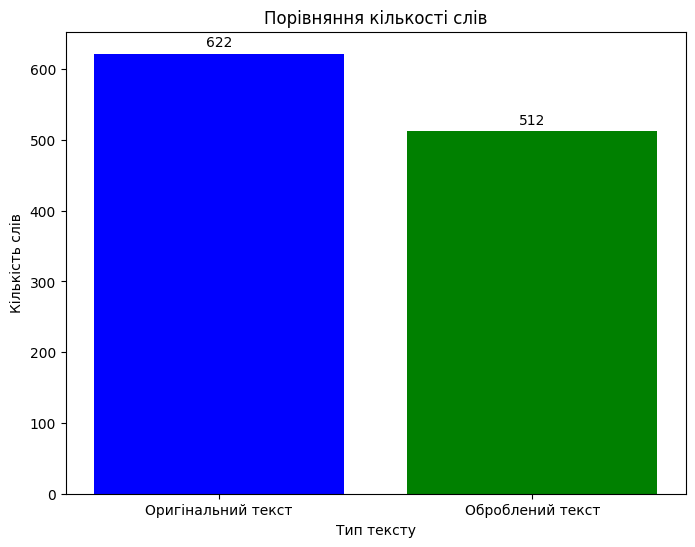


Рис. 6 – Порівняння кількоксті слів в оригіналі та після обробки

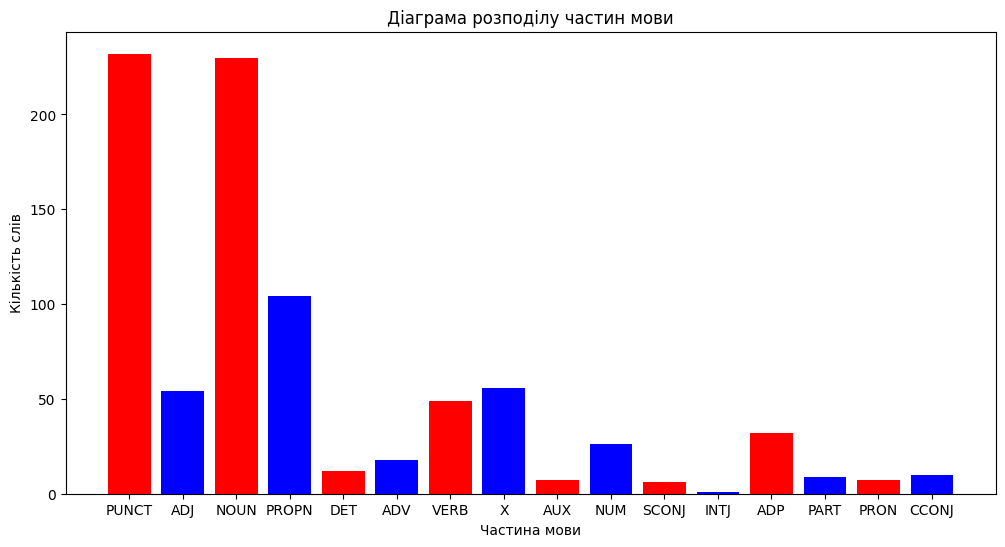


Рис. 7 – Діаграма розподілу токенізованих слів за частинами мови

На діаграмі виділені такі частини мови:

PUNCT (пунктуація) - має найвищу кількість вживань, що є типовим для текстів, оскільки знаки пунктуації використовуються для організації речень.

ADJ (прикметник) - значно менше вживається в порівнянні з пунктуацією.

NOUN (іменник) - також має досить високу кількість вживань, що свідчить про багато іменників у тексті.

PROPN (власна назва) - використовується менше, ніж загальні іменники.

DET (детермінатив) - використовується ще менше.

ADV (прислівник) - має порівняно малу кількість вживань.

VERB (дієслово) - вживається досить часто, але менше, ніж іменники.

X - може представляти інші, не класифіковані частини мови або слова, які не вдається однозначно віднести до визначеної категорії.

AUX (допоміжне дієслово), NUM (числівник), SCONJ (підрядне сполучник), INTJ (вигук), ADP (прийменник), PART (частка), PRON (займенник), CCONJ (сполучник) - вживаються ще рідше.

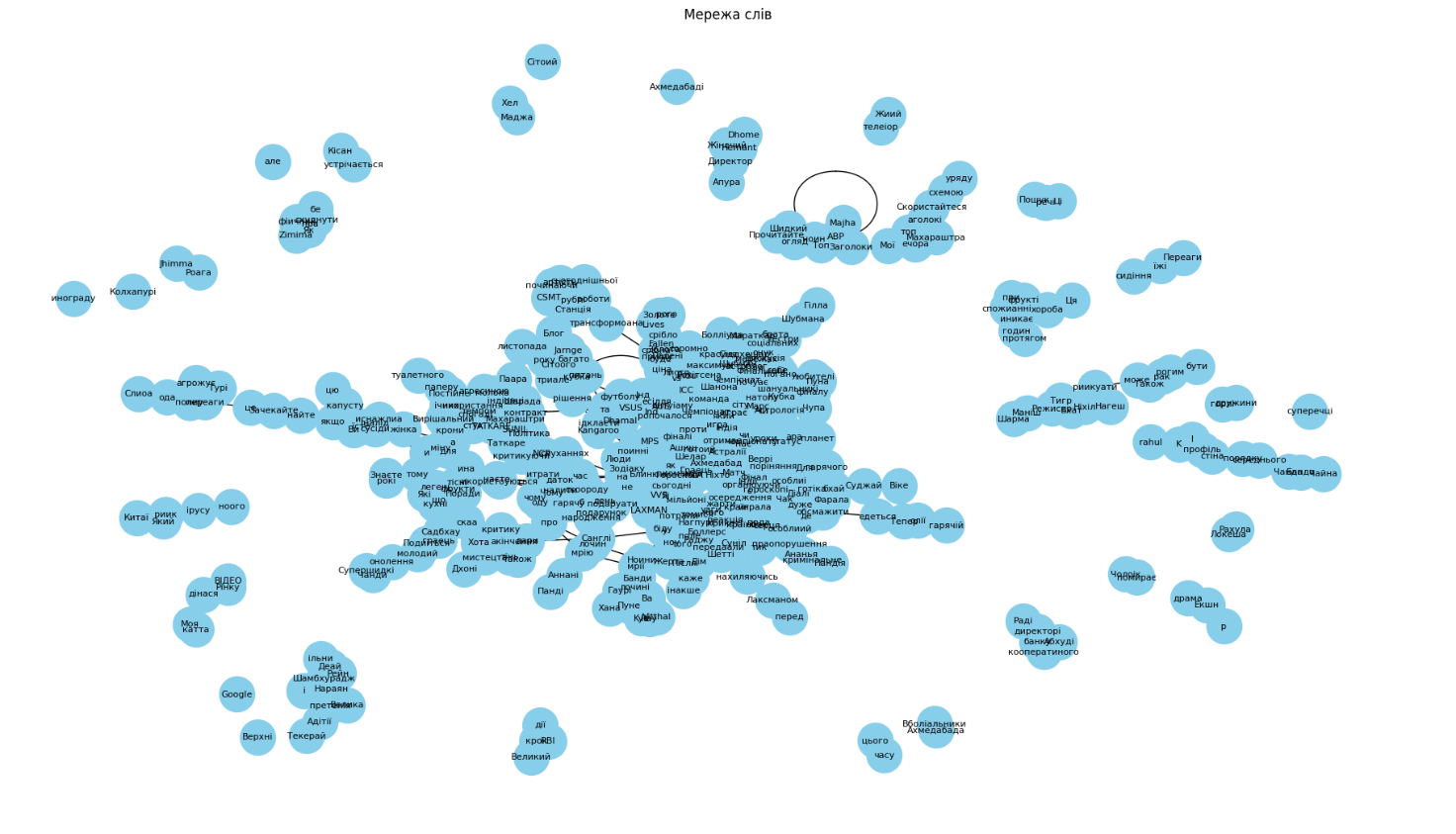


Рис. 8 – Графік зображення мережі слів.

На цьому графіку ми можемо бачити велику кількість скупчень, це означає, що на новинному ресурсі – зображаються новини не закріплені тематикою, а включають різні теми та інтереси.

Висновок: під час виконання лабораторної роботи, був проведений парсинг сайту новин, простий аналіз отриманих з парсингу даних та аналізування отриманих результатів за допомогою графічного зображення.