*Київський національний університет імені Тараса Шевченка*

*Изображение выглядит как рисунок

Автоматически созданное описание*

*Факультет інформаційних технологій*

*Кафедра інтелектуальних технологій*

**Лабораторна робота №4**

З дисципліни «Технології обробки текстової інформації»

**Виконала студентка**

групи АнД-41:

Кліменкова Ніна Анатоліївна

Варіант 7

**Викладач:**

Золотухіна Оксана Анатоліївна

**Київ-2021**

**Завдання:**

1. Провести аналіз програмних засобів, які можна використовувати для обробки текстів українською мовою (наприклад, NLTK, SpaCy, pymorphy2 інші засоби).
2. Використовуючи вибрані засоби виконати наступні операції з українським текстом (текст із завдання лабораторної роботи №3):
   * провести токенізацію;
   * провести стемінг слів тексту;
   * виконати лематизацію слів;
   * видалити стоп-слова (словник стоп-слів можна визначити самостійно, виходячи із контексту).
3. Провести аналіз отриманих наборів токенів, стемів та лем, визначити, для яких задач результати операцій є задовільними, а в яких задачах буде потрібна додаткова обробка даних.
4. Сформувати звіт заступного змісту:
   * аналіз використаних програмних засобів для обробки текстів українською мовою;
   * код програми;
   * результати обробки тексту та висновки по ньому.

Аналіз програм

1. nltk.word\_tokenize – проводить токенізацію як на українському тексті, так і на англійському
2. бібліотека re – дає змогу якісно видалити пунктуацію
3. стоп слова – були розширені відповідно до особливостей тексту
4. UkStemmer – спеціально розроблений модуль для стеммінгу українських слів і на очищеному тексті показує гарні результати.
5. pymorphy2 – дає змогу провести якісну лематизацію для українських слів, але погано працює на англійському тексті.

**Хід роботи**

Розглянемо текст українською мовою

Text

Description automatically generated

Рис.1 Текст

Далі проводимо токенізацію за допомогою nltk, так як дана бібліотека дозволяє це для українських текстів. А також видаляємо розділові знаки за допомогою регулярних виразів.

Graphical user interface, text, application, email

Description automatically generated

Рис. 2 Результат токенізації та видалення розділових знаків

Далі ми бачимо, що в тексті наявні зайві символи, такі як наприклад цифри. Тому при формуванні списків стоп-слів, ми використали російські стоп-слова та розширили їх українськими та внесли особливості тексту.

Graphical user interface, text, application, email

Description automatically generated

Рис. 3 Результат очищення від стоп-слів

Для стемінгу використовуємо UkStemmer раніше розроблену. Так як ми зробили більш обширне очищення ніж у лабораторній 3, то стемінг показав кращі результати.

Graphical user interface, text, application, email

Description automatically generated

Рис.4 Результат стеммінгу

Для лематизації я використала pymorphy2, яка дозволяє працювати з українським текстом.

Graphical user interface, text, application, email

Description automatically generated

Рис.5 Результат лематизації

Також у рамках додаткового завдання, я вирішили перевірити роботу pymorphy2 на англійському тексті. Початкові етапи не відрізняються від лабораторної 3, але був додатково розширений список стоп-слів.

Graphical user interface, text, application

Description automatically generated

Рис.6 Результат токенізації, видалення пунктуації та стоп-слів

Проводимо стеммінг з використанням PorterStemmer та LancasterStemmer

Graphical user interface, text

Description automatically generated

Рис.7 Результат стеммінгу

Далі перевіримо як пройде лематизація за допомогою pymorphy2

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Рис. 8 Результат лемматизації

**Висновок**

Для обробки українського тексту були використані nltk, регулярні вирази, UkStemmer та pymorphy2. Також було додатково розширено список-стоп слів. Варто зазначити, що розроблені стеммер та лематайзер гарно впорались із своєю задачею. Особливо маю відзначити як спрацював pymorphy2 і дав досить обширний список лем слова. Але його робота не показала гарних результатів на англійському тексті.