**Лабораторна робота №1.**

Первинна обробка даних, статистичний аналіз даних.

1. В якості вхідних даних можуть бути використані набори даних:

* sms\_spam\_corpus (csv файл, де кожен рядок позначений приналежністю до категорії spam \ ham) +
* оброблені файли з емейл повідомленнями та розбитими на категорії +

1. Обробка даних:

- видалити цифри і спеціальні символи +

- привести до єдиного регістру +

- прибрати стоп.слова (словник може бути складений вручну) +

- стемінг +

<https://xapian.org/docs/stemming.html>

* 1. Вивести на графіках розподіл по довжині слів для кожної категорії і середню довжину. +
  2. Вивести на графіках розподіл по довжині повідомлень для кожної категорії і   
     середню довжину повідомлення. +
  3. Провести частотний аналіз появи слів для двох категорій. Вивести на графіках 20 найбільш зустрічаються слів з двох категорій.

Технології та мови програмування: рекомендована python, але може буде використана будь-яка.