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 (processed emails directory)
2. Обробка даних:
   З текстом повідомлень необхідно провести наступні операції
   - видалити цифри і спеціальні символи
   - привести до єдиного регістру
   - прибрати стоп.слова (словник може бути складений вручну - приклад стоп. слів “***a, the, to, in***”)
   - стемінг https://xapian.org/docs/stemming.html
3. Створити словник слів для кожної з двох категорій.
   Формат: Слово - скільки разів зустрічається в категорії. Зберегти в окремі файли.
4. Графічне відображення
   a. Вивести на графіках розподіл по довжині слів для кожної категорії і середню довжину слів.
   b. Вивести на графіках розподіл по довжині повідомлень для кожної категорії і середню довжину повідомлення.
   c. Провести частотний аналіз появи слів для двох категорій. Вивести на графіках 20 слів, які зустрічаються найчастіше для кожної категорії окремо.

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

Примітки:

- лабораторна робота повинна бути викладена на public git repository
- графіки та файли за результатом роботи програми, повинні бути складені в директорії output
- код буде використовуватися для роботи №2