**Лабораторна робота №2.**

Наївний байєсівський аналізатор в задачі класифікації тексту.

1. В якості навчальної вибірки необхідно використати dataset з лабораторної роботи #1 (текстові повідомлення та електронні листи).
2. Реалізовати алгоритм наївного байєсовського класифікатора для задачі класифікації тексту.
3. Для програми розробленої для лабораторної роботи №1 додати:
   1. поле для задання повідомлення та вибору навчальної вибірки, яке необхідно класифікувати;
   2. можливість запустити аналізатор та вивести для користувача до якого класу відноситься повідомлення.

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

**Теоретичний матеріал**

“Наївність” алгоритму полягає в тому, що ми припускаємо незалежність появи слів в повідомленнях. За теоремою Байеса:

P(ham | bodyText) = (P(ham) \* P(bodyText | ham)) / P(bodyText)

P(spam | bodyText) = (P(spam) \* P(bodyText | spam)) / P(bodyText)

Нам необхідно тільки порівняти P(ham | bodyText) та P(spam | bodyText), тому P(bodyText) - рахувати не потрібно.

P(ham) = кількість повідомлень з категорії ham / загальна кількість повідомлень

P(spam) = кількість повідомлень з категорії spam / загальна кількість повідомлень

Для повідомлення довжиною N:

bodyText = [word1, word2, …, wordN]

P(bodyText | ham) = P(word1 | ham) \* P(word2 | ham) \* …

P(bodyText | spam) = P(word1 |spam) \* P(word2 | spam) \* …

,де:

P(word1 | spam) = кількість word1 в категорії spam / загальна кількість слів в spam.

Згладжування Лапласа - якщо слова немає в навчальній вибірці, ми вважаємо, що слова зустрічається один раз, але нам треба також змінити вірогідності інших слів:

P (word1 | spam) = (кількість word1 які належать категорії ham + 1) / (загальна кількість слів, які належать категорії ham +кількість слів, яких немає в навчальній вибірці)

Всі обчислення можна привести до логарифмічної форми, щоб не працювати з маленькими числами з плаваючою точкою: <http://getpopfile.org/docs/faq:bayesandlogs>