НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ім. ІГОРЯ СІКОРСЬКОГО”

КАФЕДРА АВТОМАТИЗОВАНИХ СИСТЕМ ОБРОБКИ ІНФОРМАЦІЇ І УПРАВЛІННЯ

Комп’ютерний практикум № 3

з дисципліни

“Комп’ютерна лінгвістика ”

Виконала:

студентка групи ІС-71

Вознюк О.В

Перевірив:

доцент

Фіногенов О.Д.

Київ-2021

**Мета**

Метою роботи є вивчення методів визначення подібних (близьких за текстом) документів, зокрема для знаходження плагіату.

**Варіант завдання та дані про авторів**

Клас 0 – коронавірус

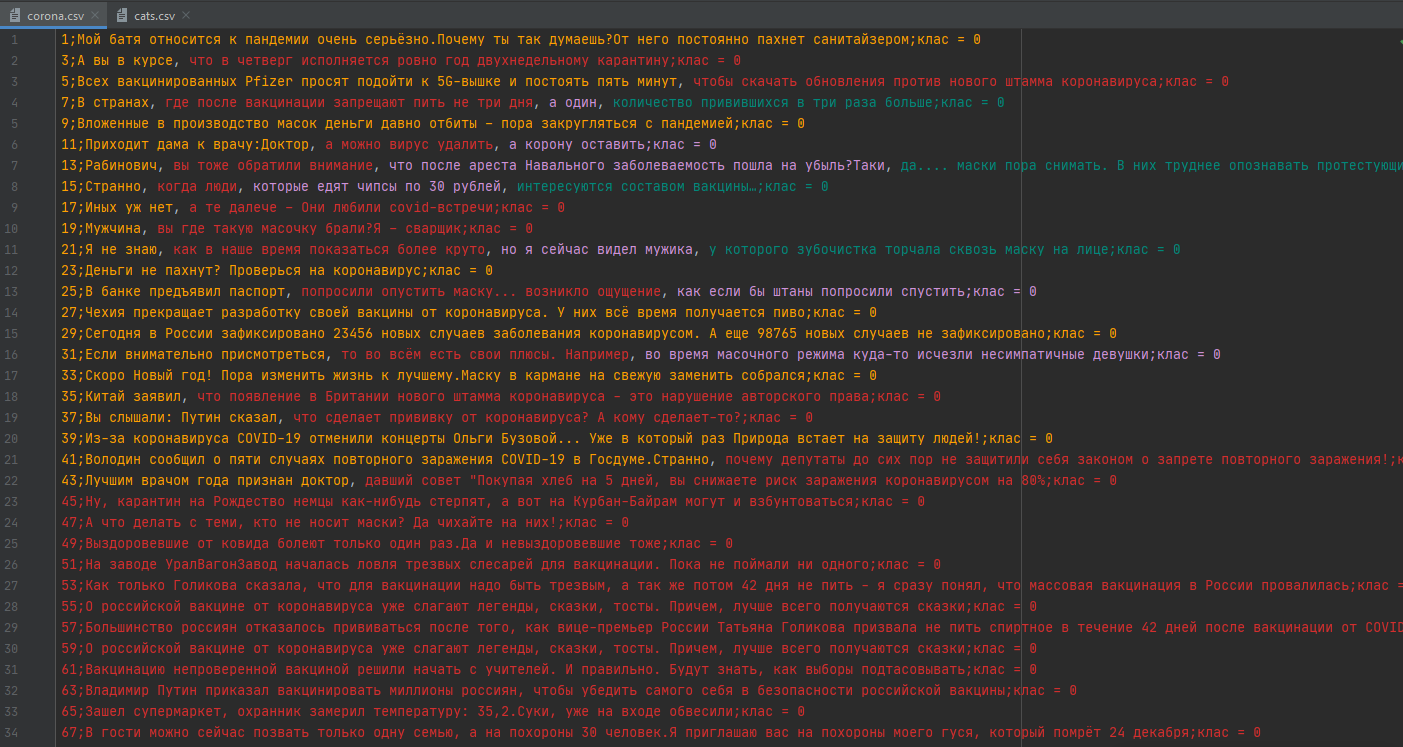


Рисунок 1

Клас 1 – коти

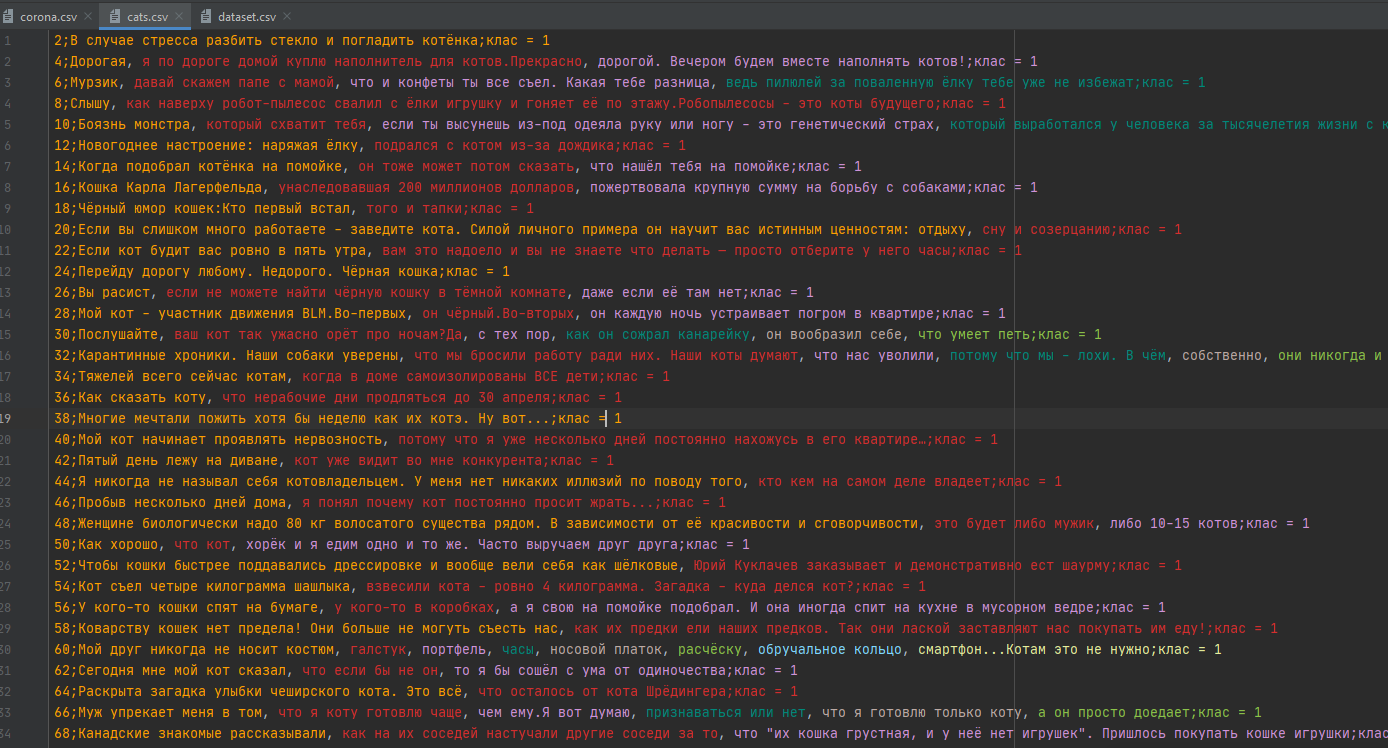


Рисунок 2

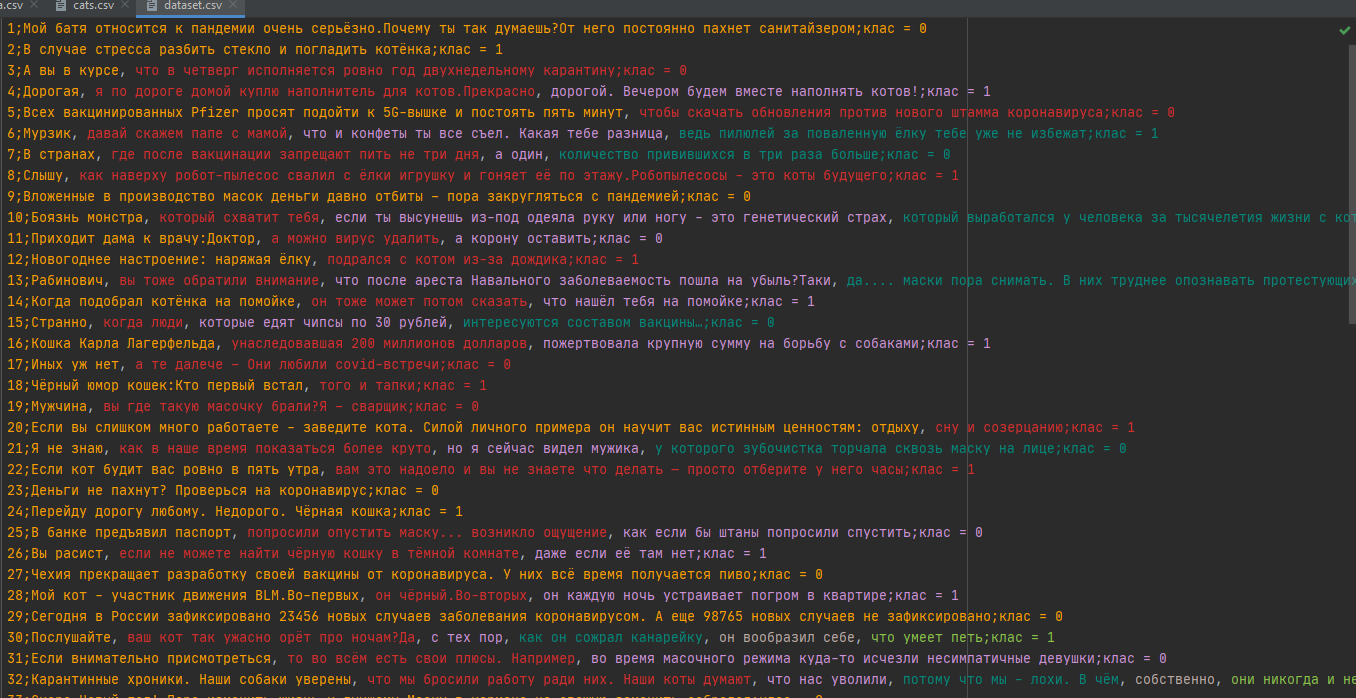


Рисунок 3 – спільний датасет

**Короткі теоретичні відомості**

Визначення тональності тексту є однією з типових задач, що має велике практичне значення. Найбільше розповсюдження дана задача отримала в галузі електронної комерції при аналізі відгуків на товари або послуги в соціальних мережах або інтернет-магазинах, визначенні вподобань цільової аудиторії та її відгук на політичні події, рекламні та передвиборчі "слогани" тощо.

Дана задача відноситься до задач класифікації в галузі лінгвістики з наперед заданим набором класів.

Одним з найпростіших методів оцінювання тональності тексту є метод машинного навчання на основі наївного баєсіва класифікатора (НБК) [3].

В основі НБК лежить теорема Баєса (2.1)

|  |  |
| --- | --- |
|  |  |
|  | (2.1) |
|  |  |

де,

– P(c|d) — імовірність, що документ d належить класу c (саме цю імовірність необхідно розрахувати);

– P(d|c) — імовірність зустріти документ d серед усіх документів класу c;

– P(c) — безумовна імовірність зустріти документ класу c в корпусі документів;

– P(d) — безумовна імовірність документа d в корпусі документів.

Метою класифікації є встановлення приналежності документу до якогось класу. Для цього необхідно не визначення безпосередньо імовірності, а визначення найбільш імовірного класу.

|  |  |
| --- | --- |
|  |  |
|  | (2.2) |
|  |  |

Тобто, необхідно обрахувати імовірності для кожного класу і обрати той, який має найбільшу імовірність. Так як знаменник (P(d)) є константою, то він не впливає на ранжування класів за імовірністю та може не враховуватися при обчисленнях (2.3).

|  |  |
| --- | --- |
|  |  |
|  | (2.3) |
|  |  |

**Припущення умовної незалежності**

В природній мові імовірність появи слів сильно залежить від контексту. НБК навпаки розглядає документ як набір слів, які умовно не залежать одне від одного (bag of words model – модель "мішку" слів). Внаслідок такого припущення, умовна імовірність документу (P(d|c)) може бути представлена як добуток умовних імовірностей всіх слів, що входять до документу (2.4).

|  |  |
| --- | --- |
|  |  |
|  | (2.4) |
|  |  |

Такий підхід має назву Unigram Language Model та грає важливу роль в задачах обробки природних мов.

Підставив вираз (2.4) в (2.3) отримаємо (2.5):

|  |  |
| --- | --- |
|  |  |
|  | (2.5) |
|  |  |

**Проблема арифметичного переповнення**

При досить великій довжині документа доведеться множити велику кількість дуже маленьких чисел. Для того щоб при цьому уникнути арифметичного переповнення знизу часто користуються властивістю логарифма добутку log(ab) = log(a) + log(b). Так як логарифм функція монотонна, її застосування до обох частин виразу змінить тільки його чисельне значення, але не параметри при яких досягається максимум. При цьому, логарифм від числа близького до нуля буде числом від’ємним, але в абсолютному значенні істотно більшим ніж вихідне число, що робить логарифмічні значення імовірностей більш зручними для аналізу. Формула (2.5) з використанням логарифму приймає вигляд (2.6):

|  |  |
| --- | --- |
|  |  |
|  | (2.6) |
|  |  |

**Оцінка параметрів Баєсової моделі**

Оцінка імовірностей  і  здійснюється на навчальній виборці. Імовірність класу можна оцінити як:

|  |  |
| --- | --- |
|  |  |
|  | (2.7) |
|  |  |

де,

­–  – кількість документів, що належать класу c,

– D – загальна кількість документів в навчальній вибірці.

Оцінка імовірності слова в класі може здійснюватись декількома способами, наприклад, з використанням мультиномінальної баєсовської моделі (2.8)

|  |  |
| --- | --- |
|  |  |
|  | (2.8) |
|  |  |

де

–  – кількість разів, скільки і-е слово зустрічається в документах класу с;

– k – словник корпусу документів (перелік всіх унікальних слів).

Тобто, чисельник показує скільки разів слово зустрічається в документах класу (включаючи повторення), а знаменник – це сумарна кількість слів у всіх документах цього класу.

**Проблема невідомих слів**

При використанні формули (2.8) виникає проблема з врахування частоти слів, які не зустрічалися у навчальній виборці, так як для них значення , і, відповідно,  дорівнюватиме 0. Тобто, наявність слова, яке відсутнє у навчальній виборці унеможливлює класифікацію документа. Для вирішення цієї проблеми, зазвичай використовують адитивне згладжування (згладжування Лапласа): додавання деякої позитивної величини α (α > 0, зазвичай α=1) до частоти кожного слова. Це дозволяє надати словам, що не зустрічались в навчальному наборі мале, але ненульове значення імовірності. Головною властивістю такого підходу є деяке зміщення в бік менш імовірних результатів оцінки імовірності.

Кінцева формула для вибору класу з врахуванням згладжування Лапласа приймає вигляд:

|  |  |
| --- | --- |
|  |  |
|  | (2.9) |
|  |  |

або при α=1 (2.10):

|  |  |
| --- | --- |
|  |  |
|  | (2.10) |
|  |  |

**Реалізація класифікатора**

Для реалізації Баєсівского класифікатора необхідно сформувати навчальну вибірку, в якій співставленні текстові документи та їх класи. На етапі класифікації буде необхідна наступна статистика з вибірки:

* відносна частота класів в корпусі документів (як часто зустрічаються документи того, чи іншого класів);
* сумарна кількість слів в документах кожного класу;
* відносна частота слів а межах кожного класу;
* розмір словника вибірки (кількість унікальних слів у вибірці).

Модель класифікатора – це сукупність даної інформації. На етапі класифікації необхідно для кожного класу обчислити значення (2.11) та обрати максимальне з них (2.10).

|  |  |
| --- | --- |
|  |  |
|  | (2.11) |
|  |  |

де:

*  – кількість документів в навчальній вибірці, що належать класу с;
*  – загальна кількість документів в навчальній вибірці;
*  – кількість унікальних слів у всіх документах навчальної вибірки;
*  – сумарна кількість слів в документах класу с в навчальній вибірці;
*  – скільки разів, і-е слово зустрічається в документах класу с в навчальній вибірці.

**Опис програми**

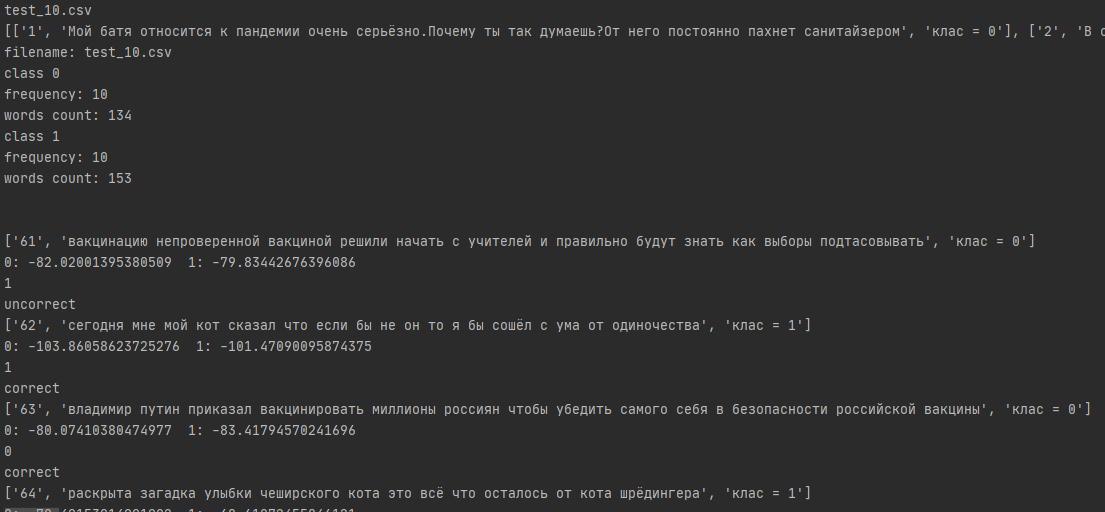


Рисунок 4 – приклад виводу результату в режимі дебаг

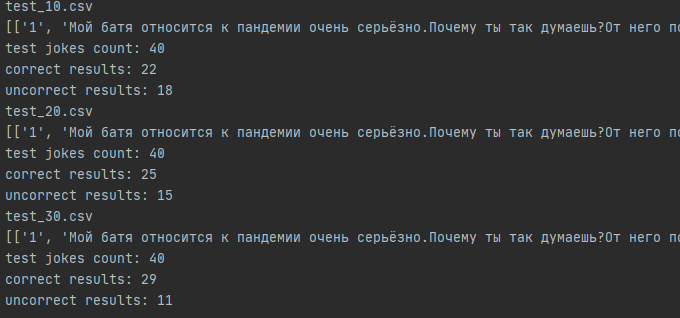


Рисунок 5- приклад виводу результату без дебаг

**Гістограми частоти 10-ти найбільш часто вживаних слів**

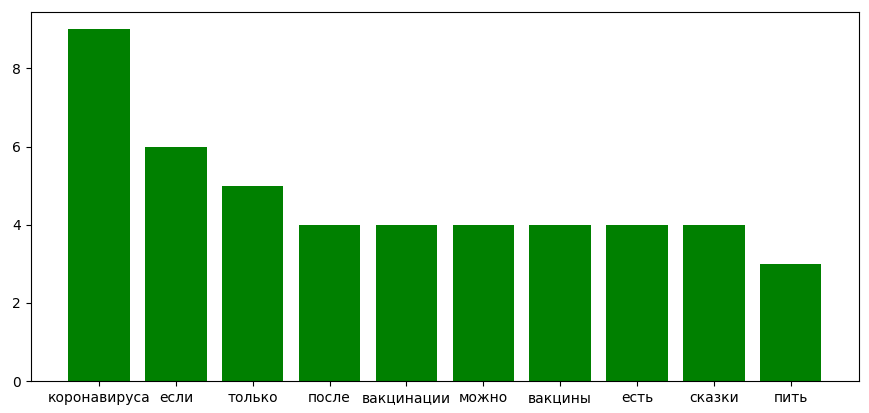


Рисунок 6 – гістограма частот для класу 0 (коронавірус)

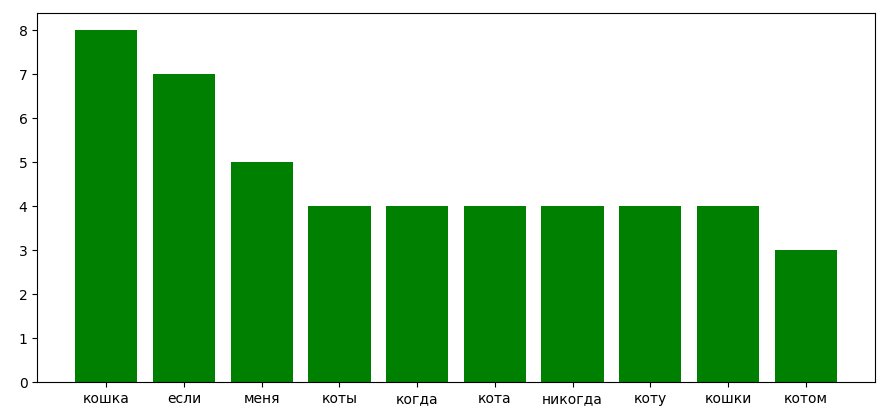


Рисунок 7 – гістограма частот для класу 1 (коти)

Стоп слів (<=3) не виявлено.

**Графік залежності точності класифікації від розміру навчальної вибірки**

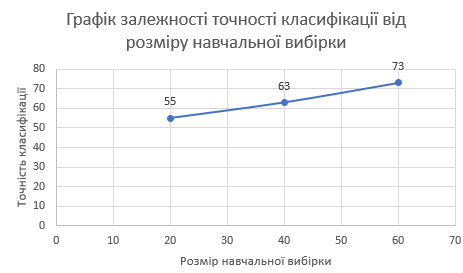


Рисунок 8 - Графік залежності точності класифікації від розміру навчальної вибірки

**Висновки щодо особливостей та ефективності використання НБК**

Для завдань з невеликою кількістю даних для навчання, НБК може досягти кращих результатів, ніж інші класифікатори, тому що він має низьку схильність до перенавчання. Займана площа пам’яті є скромною, так як ці операції не вимагають одночасного зберігання всього набору даних в оперативній пам'яті. Слід зазначити, що через припущення про незалежність ознак, ймовірності класів, виведені наївним Байєса, можуть бути досить неточними. З іншого боку, незважаючи на свою наївність, наївний Байес часто робить досить хорошу роботу з підбору правильного класу - він, можливо, не такий гарний в оцінці абсолютних ймовірностей, але він досить хороший в вимірі відносних ймовірностей.