НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ім. ІГОРЯ СІКОРСЬКОГО”

КАФЕДРА АВТОМАТИЗОВАНИХ СИСТЕМ ОБРОБКИ ІНФОРМАЦІЇ І УПРАВЛІННЯ

Комп’ютерний практикум № 1

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 автора тексту за допомогою статистичного аналізу буквосполучень.

**Варіант завдання та дані про авторів**

Автор 1: Іван Багряний - справжнє ім'я: Лозов'яга Іва́н Па́влович (також: Лозов'ягін); 19 вересня (2 жовтня) 1906, Охтирка, Харківська губернія, Російська імперія — 25 серпня 1963, Новий Ульм, ФРН — український поет, прозаїк, публіцист, політичний діяч.

* Маруся Богуславка – роман (виданий під заголовком "Буйний вітер", Мюнхен, 1957)
* Сад Гетсиманський – роман (Новий Ульм, 1950)
* Тигролови – роман (Львів–Краків, 1944)
* Людина біжить над прірвою – роман (посмертно, Новий Ульм, Нью-Йорк, 1965) – для тестових уривків

Автор 2: Михайло Старицький - (14) листопада 1839, Кліщинці — 14 (27) квітня 1904, Київ — український письменник (поет, драматург, прозаїк), театральний і культурний діяч. Один з корифеїв українського побутового театру.

* Талан – драма
* Кармелюк – роман (1927-1928)
* Облога Буші – історичний роман (1894)
* Молодість Мазепи – історичний роман (1898) – для тестових уривків

Автор 3: Марко Вовчок - ім'я при народженні — Марія Олександрівна Вілінська, за першим чоловіком — Маркович, за другим чоловіком — Лобач-Жученко; 10 (22) грудня 1833, маєток Єкатерининське Єлецького повіту Орловської губернії, тепер Липецька область РФ — 28 липня (10 серпня) 1907, Нальчик, Росія — українська і російська письменниця, перекладачка, видавниця. Її твори мали антикріпацьке спрямування та описували історичне минуле України.

* Інститутка – повість (початкова назва «Панночка»; Спб: журнал «Основа» No.3, 1862)
* Маруся - повість
* Три долі – оповідання (1862)
* Павло Чорнокрил – оповідання – для тестових уривків

**Короткі теоретичні відомості**

При проведені досліджень, а також при автоматичній обробці текстів часто виникає сукупність задач з ідентифікації автора тексту, жанру тексту, формату тексту (роман, повість, розповідь і т.д.) та інших. Одним з методів дослідження тексту є статистичний аналіз повторюваності елементів тексту (буквосполучень): грам, біграм, триграм. Кількісні характеристики тексту для кожного автора теоретично будуть відрізнятися в залежності від словникового запасу, жанру тексту, часу, коли текст був написаний та інших чинників. Ця умова не є достатньою. так як будь-який автор може скласти будь-який текст, але будемо вважати, що без накладання додаткових обмежень, авторський стиль буде давати змогу відрізнити автора від інших.

В таблиці 1.1 наведено частота символів російської та української мов, обрахована на множині текстів [1].

Таблиця 1.1 – Частота грамів російської та української мов

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № | Буква рос. мови | Частота | Буква укр. мови | Частота |
| 1 | А | 0.062 | А | 0.0807 |
| 2 | Б | 0.014 | Б | 0.0177 |
| 3 | В | 0.038 | В | 0.0535 |
| 4 | Г | 0.013 | Г, Ґ | 0.0155 |
| 5 | Д | 0.025 | Д | 0.0338 |
| 6 | Е, Ё | 0.072 | Е | 0.0495 |
| 7 | Ж | 0.007 | Ж | 0.0093 |
| 8 | З | 0.016 | З | 0.0232 |
| 9 | И | 0.062 | И | 0.0626 |
| 10 | Й | 0.010 | Й | 0.0138 |
| 11 | К | 0.028 | К | 0.0354 |
| 12 | Л | 0.035 | Л | 0.0369 |
| 13 | М | 0.026 | М | 0.0303 |
| 14 | Н | 0.053 | Н | 0.0681 |
| 15 | О | 0.090 | О | 0.0942 |
| 16 | П | 0.023 | П | 0.0290 |
| 17 | Р | 0.040 | Р | 0.0448 |
| 18 | С | 0.045 | С | 0.0424 |
| 19 | Т | 0.053 | Т | 0.0535 |
| 20 | У | 0.021 | У | 0.0336 |
| 21 | Ф | 0.002 | Ф | 0.0028 |
| 22 | Х | 0.009 | Х | 0.0119 |
| 23 | Ц | 0.004 | Ц | 0.0083 |
| 24 | Ч | 0.012 | Ч | 0.0141 |
| 25 | Ш | 0.006 | Ш | 0.0076 |
| 26 | Щ | 0.003 | Щ | 0.0056 |
| 27 | Ь, Ъ | 0.014 | І | 0.0575 |
| 28 | Ы | 0.016 | Ї | 0.0065 |
| 29 | Э | 0.003 | Ь | 0.0177 |
| 30 | Ю | 0.006 | Є | 0.0061 |
| 31 | Я | 0.018 | Ю | 0.0093 |
| 32 | Пробел | 0.175 | Я | 0.0248 |

Аналіз біграмів (буквосполучень 2-х символів) можна представити у вигляді температурної діаграми (рис. 1.1), де кольорами відображається частота використання, або у вигляді квадратної таблиці.

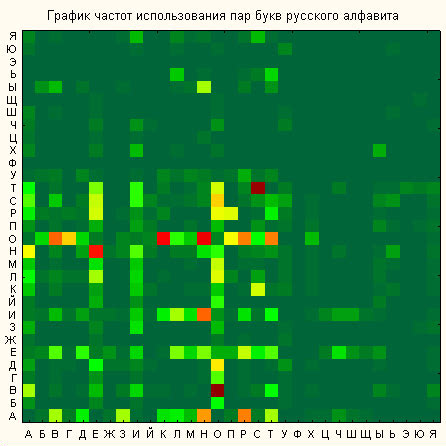


Рисунок 1.1 – Частоти використання пар букв російського алфавіту

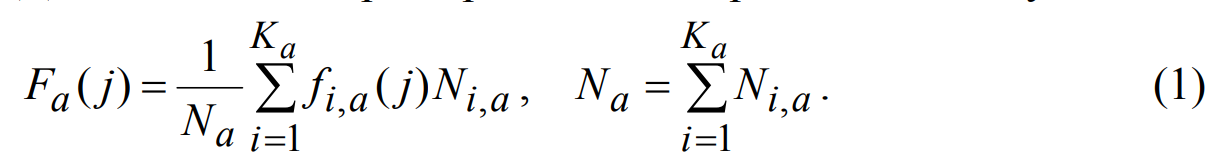
Зазвичай, пробіли, розділові знаки, цифри, букви іншого алфавіту та інші символи ігноруються, якщо аналіз не направлений саме на ці особливості.

Для уникнення неузгодженості початку відліку буквосполучень при визначенні кількості n-грамів (наприклад, для типового закінчення дієслів "-ться" початок відліку для біграмів може бути з літери "т" або з "ь" в залежності від парності/непарності кількості літер попереднього тексту), зсув при аналізі n-грамів відбувається на 1 позицію, а не на довжину n-грама. Таким чином кожна літера (за виключенням n-1 перших та останніх) зустрінеться n-грамах n разів, що дасть змогу уникнути некоректного відображення результатів.

Формалізована задача розпізнавання автора тексту може бути представлена у наступний спосіб [2]:

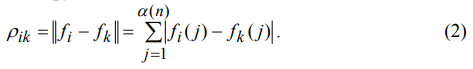
Нехай є бібліотека текстів, яка представлена у вигляді щільності функції розподілу або статистиці n-грамів (грамів, біграмів, триграмів і т.д.) для А відомих авторів. Нехай  – кількість текстів а–го автора,  – кількість символів в і-му тексті цього автора, де . Будемо вважати, що довжина кожного тексту достатня для проведення статистичного аналізу (тобто довжина тексту дає достатньо інформації про стиль автора в розрізі частоти зустрічі n-грамів).

Нехай  – функція щільності n-грамів відповідного тексту. де  (Ω – розмір алфавіту або кількість символів, що досліджуються). Для кожного автора визначимо середньозважене значення щільності функції розподілу, нехтуючи відмінністю (одиницею) між грамами, біграмами і далі, внаслідок :



Значення  – будемо вважати авторським еталоном.

Введемо «бібліотечну норму»  як відстань між щільністю функцій розподілу текстів  та  в нормі підсумованих функцій



Нехай в наявності є текст «0» невідомого автора, який необхідно ідентифікувати всередині даної бібліотеки. Автором тексту «0» вважається той з авторів «a», для якого норма  різниці між щільністю функції розподілу  тексту «0» та середньою авторською щільністю функції розподілу  мінімальна:



Дослідження російської мови показують, що найкращий результат по розпізнаванню автора відбувається при використанні 5 або 6 літерних буквосполучень, але в цьому випадку відсутня проста графічна інтерпретація результатів і тому в даній лабораторній роботі використані лише грами та біграми.

**Опис програми**

https://github.com/oleksandravozniuk/ComputationalLinguistics/tree/main/1

**Дані про кількість символів творів**

Таблиця 1.1 – Дані творів

|  |  |
| --- | --- |
| Автор  (Твір) | Обсяг(символів, укр.) |
|  |  |
| Іван Багряний (Маруся Богуславка) | 823091 |
| Іван Багряний (Сад Гетсиманський) | 1055588 |
| Іван Багряний (Тигролови) | 447273 |
| Михайло Старицький (Талан) | 28,738 |
| Михайло Старицький (Кармелюк) | 1438294 |
| Михайло Старицький (Облога Буші) | 148566 |
| Марко Вовчок (Інститутка) | 65416 |
| Марко Вовчок (Маруся) | 155357 |
| Марко Вовчок (Три долі) | 136221 |

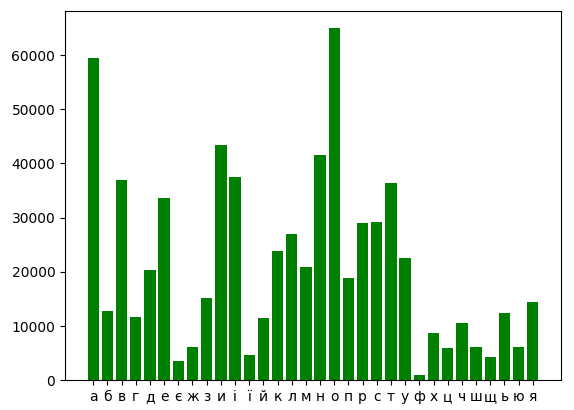


Рисунок 1 - Іван Багряний (Маруся Богуславка)

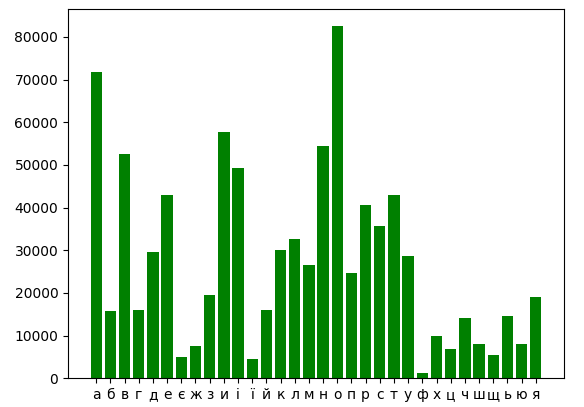


Рисунок 2 - Іван Багряний (Сад Гетсиманський)

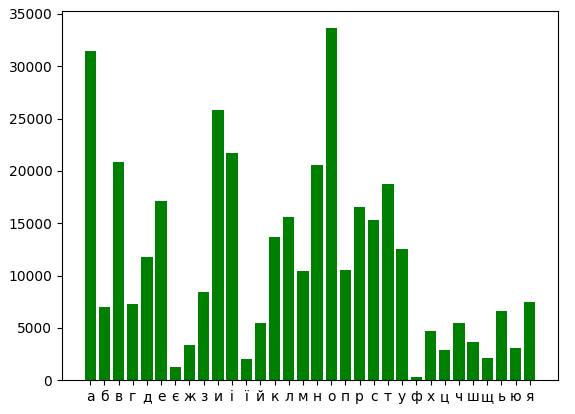


Рисунок 3 - Іван Багряний (Тигролови)

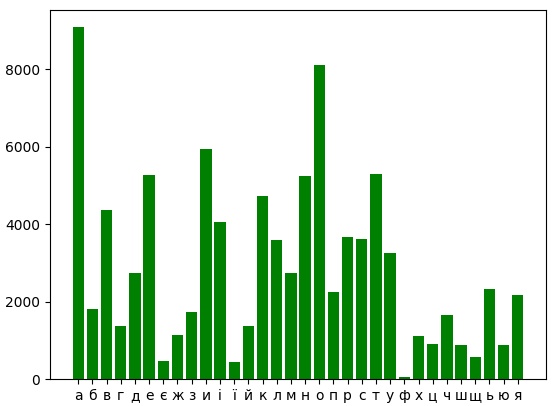


Рисунок 4 - Михайло Старицький (Талан)

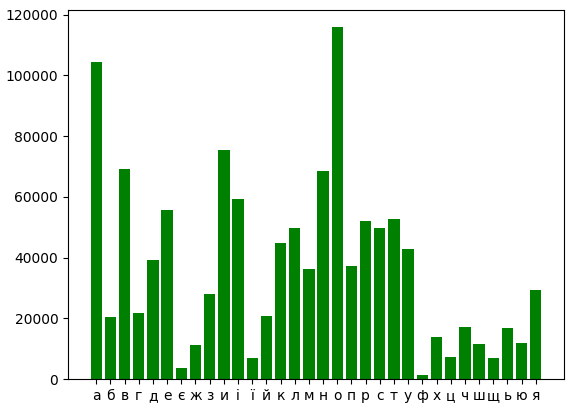


Рисунок 5 - Михайло Старицький (Кармелюк)

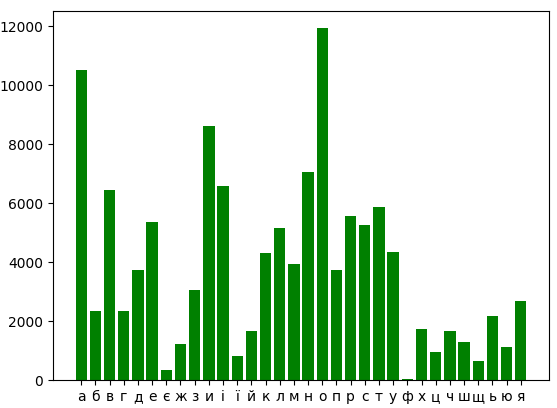


Рисунок 6 - Михайло Старицький (Облога Буші)

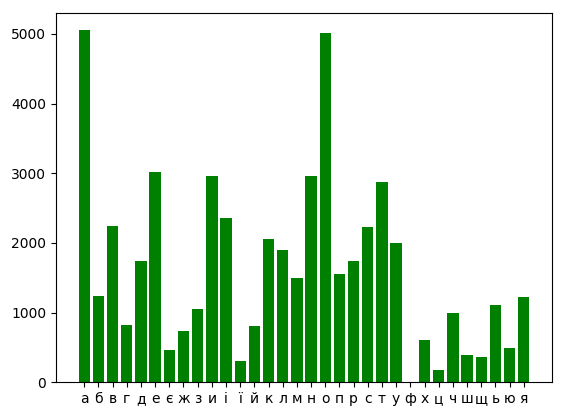


Рисунок 7 - Марко Вовчок (Інститутка)

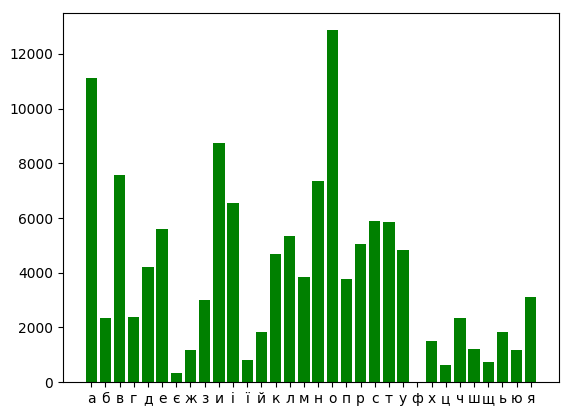


Рисунок 8 - Марко Вовчок (Маруся)

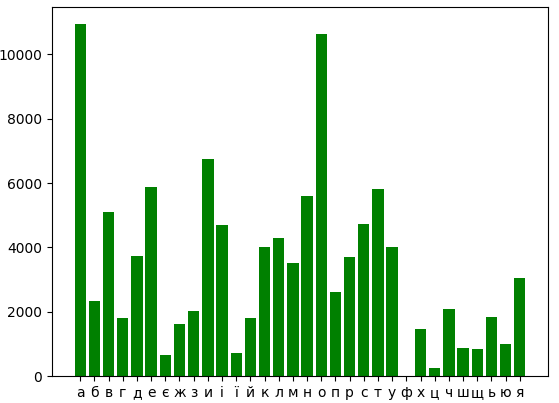


Рисунок 9 - Марко Вовчок (Три долі)

**Гістограми частот творів**

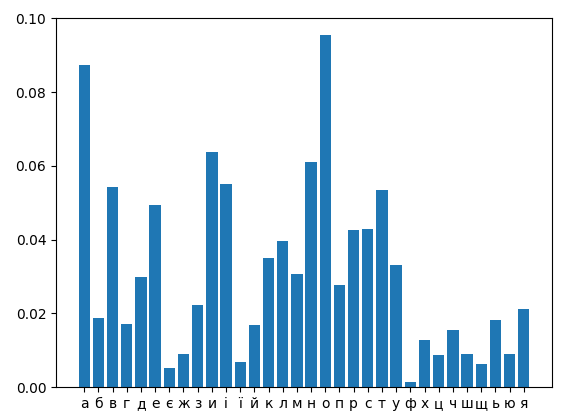


Рисунок 10 - Іван Багряний (Маруся Богуславка)

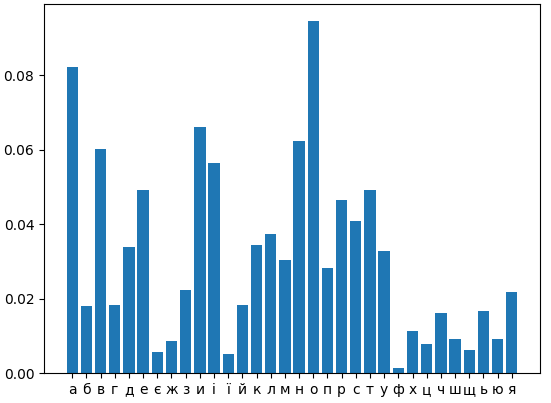


Рисунок 11 - Іван Багряний (Сад Гетсиманський)

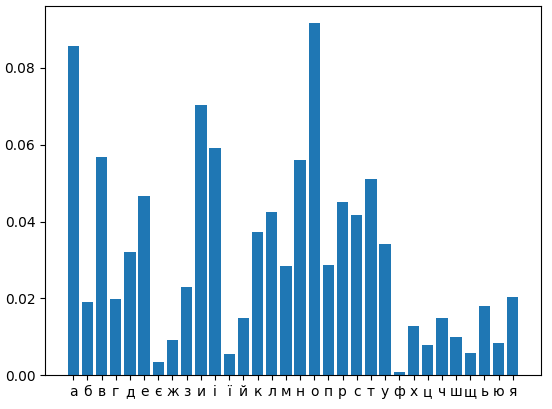


Рисунок 12 - Іван Багряний (Тигролови)

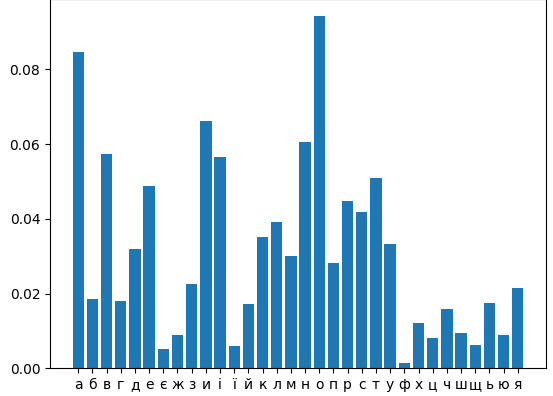


Рисунок 13 – сукупна для Івана Багряного

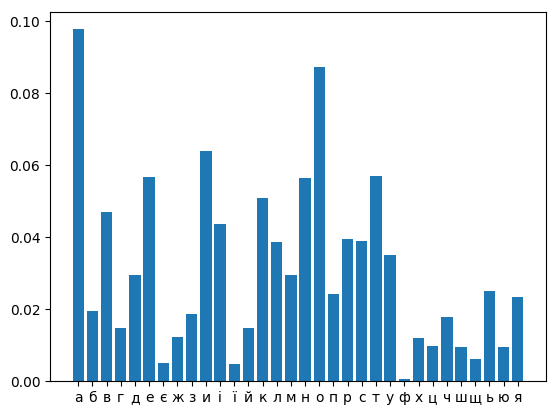


Рисунок 14 - Михайло Старицький (Талан)

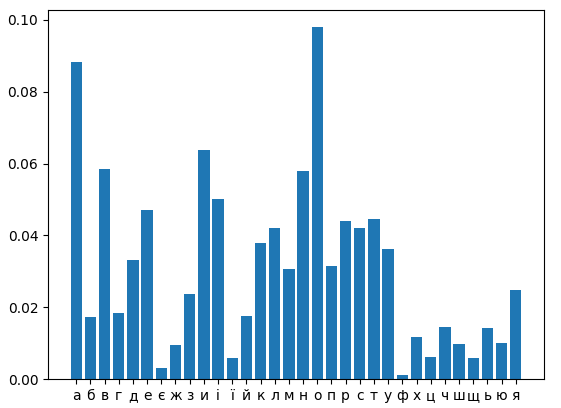


Рисунок 15 - Михайло Старицький (Кармелюк)

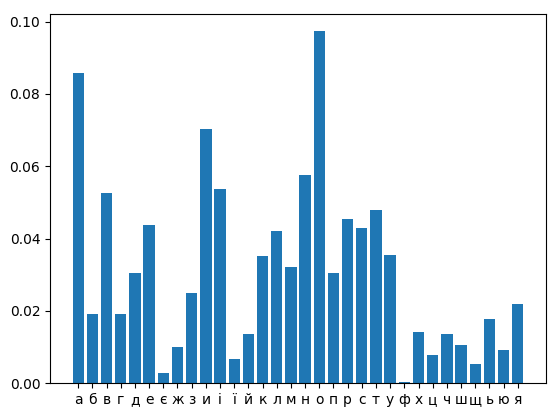


Рисунок 16 - Михайло Старицький (Облога Буші)

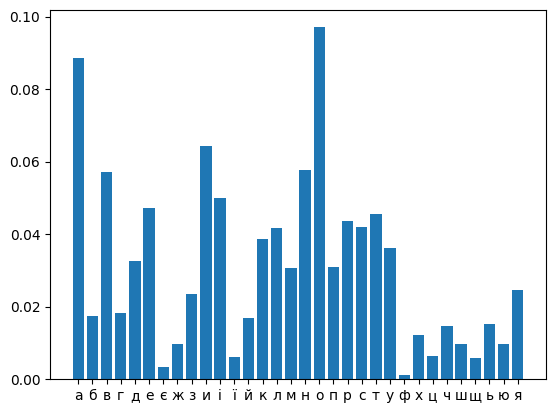


Рисунок 17 – сукупна для Михайла Старицького

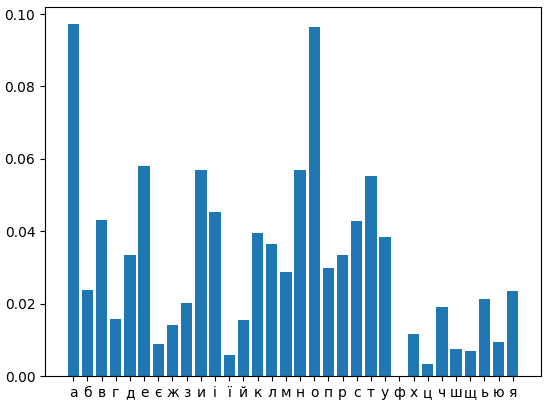


Рисунок 18 - Марко Вовчок (Інститутка)

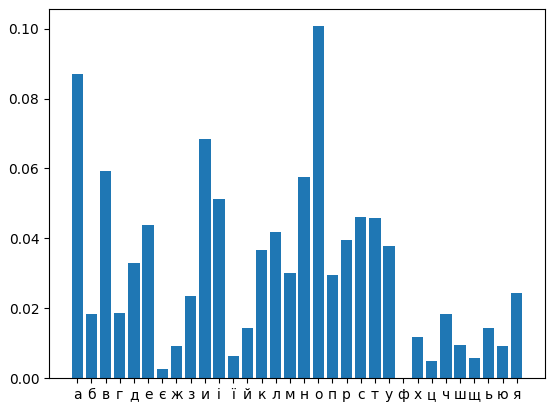


Рисунок 19 - Марко Вовчок (Маруся)

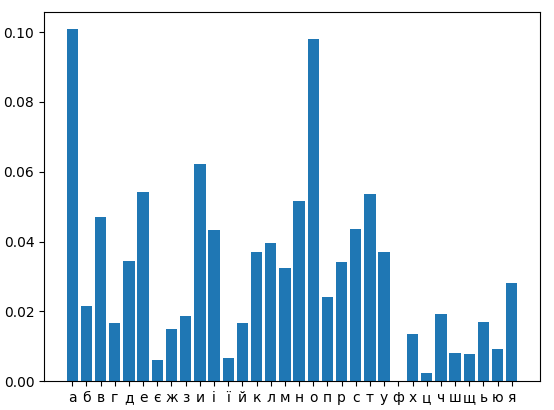


Рисунок 20 - Марко Вовчок (Три долі)

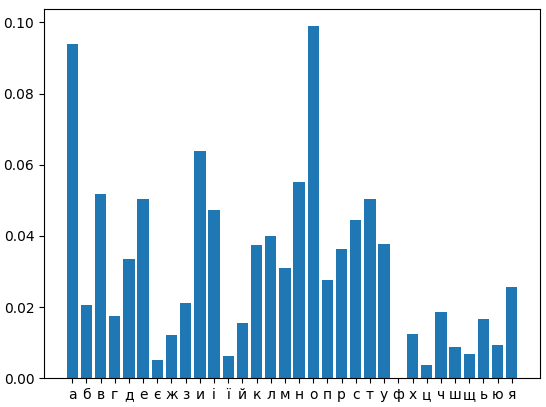


Рисунок 21 – сукупна для Марка Вовчка

**Теплові діаграми**

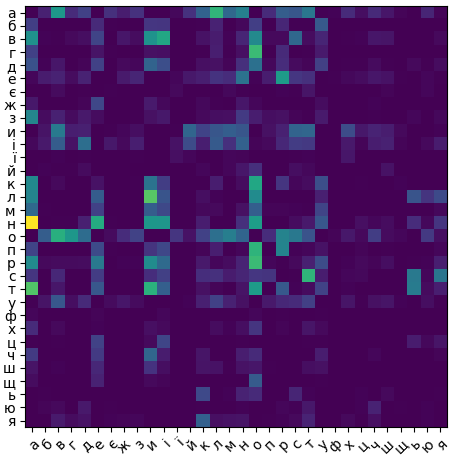


Рисунок 22 – сукупна для Івана Багряного

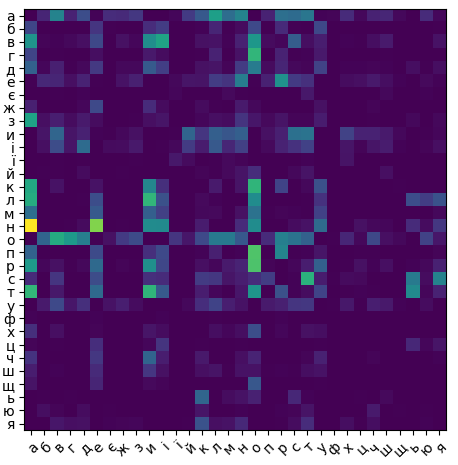


Рисунок 23 – сукупна для Михайла Старицького

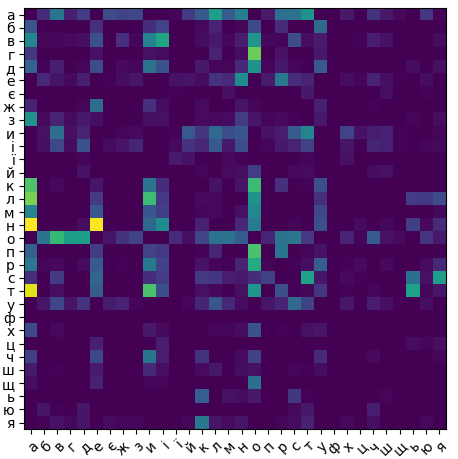


Рисунок 24 – сукупна для Марка Вовчка

Можемо бачити з теплових діаграм частот біграм, що вони є дуже подібними між собою.

**Гістограми різниці частот використання пар букв українського алфавіту для сукупного набору творів кожного автора та загальної частоти використання букв українського алфавіту**

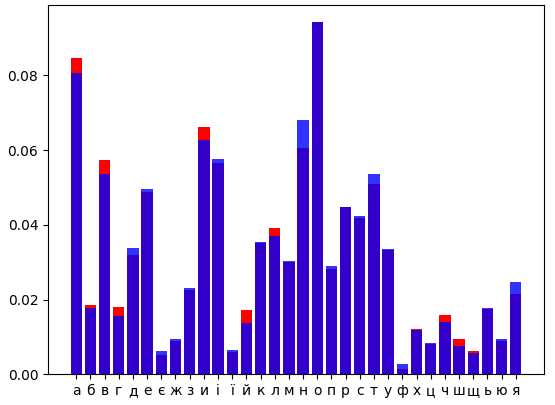


Рисунок 25 – сукупна для Івана Багряного

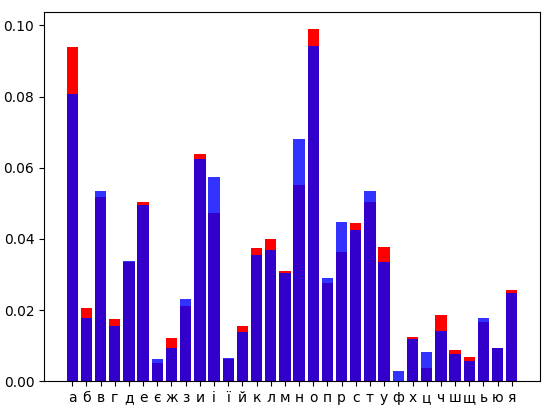


Рисунок 26 – сукупна для Михайла Старицького

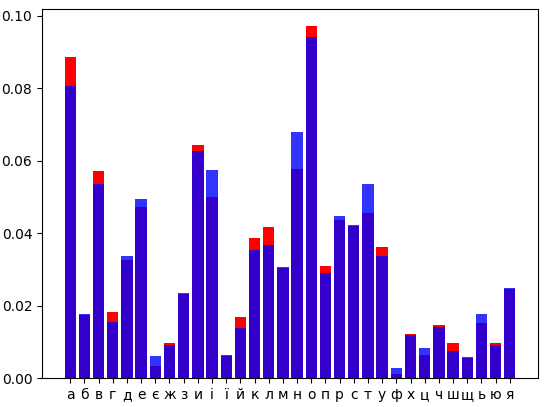


Рисунок 27 – сукупна для Марка Вовчка

Червоний колір означає, що частота використання більше ніж в українській мові, а світло синій навпаки, що менше.

**Різниці частот**

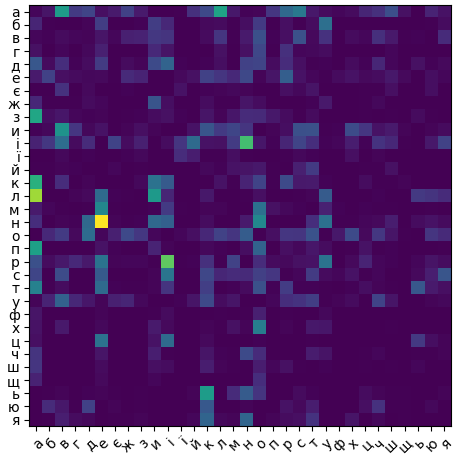


Рисунок 28 – (Багряний - Старицький)

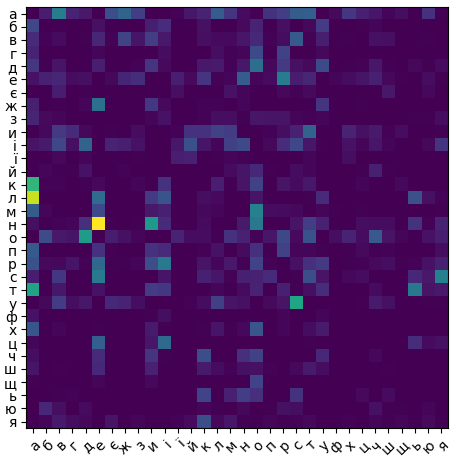


Рисунок 29 – (Багряний – Марко Вовчок)

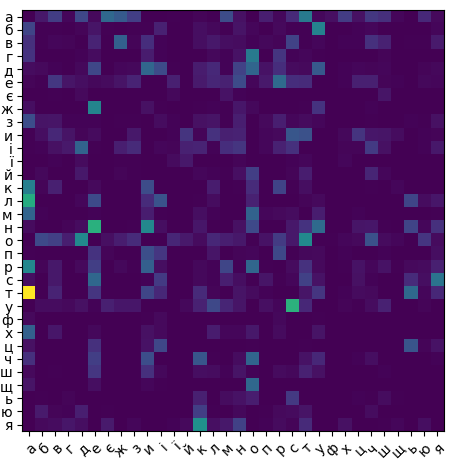


Рисунок 30 – (Старицький – Марко Вовчок)

Бачимо, що у Багряного значно більше біграм “не” ніж у Старицького та Марка Вовчка. А у Старицького більше біграм “та” ніж у Марка Вовчка. В усіх остальних випадках різниці незначні.

**Визначення автора за уривком тексту за допомогою грамів**

Таблиця 1.3 – Визначення автора невідомого тексту (Іван Багряний - Людина біжить над прірвою)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Розмір  уривку  тексту | Іван Багряний | Михайло Старицький | Марко Вовчок | Визначено автора |
| 5000 | 0.08588456238815285 | 0.09511863793207281 | 0.12988914403049304 | + |
| 10000 | 0.08312798294080517 | 0.08875748077450774 | 0.12355459650421524 | + |
| 25000 | 0.06060995889589778 | 0.086476834813784 | 0.11619349134739705 | + |
| 50000 | 0.0557129232172798 | 0.07713935525700072 | 0.11072567744785572 | + |

Таблиця 1.4 – Визначення автора невідомого тексту (Михайло Старицький - Молодість Мазепи)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Розмір  уривку  тексту | Іван Багряний | Михайло Старицький | Марко Вовчок | Визначено автора |
| 5000 | 0.2551026676952059 | 0.2615840567669273 | 0.2623401108242164 | - |
| 10000 | 0.22491311540608863 | 0.23192969662659035 | 0.23710752884811223 | - |
| 25000 | 0.20614180803781507 | 0.20705993169802708 | 0.20272432918066802 | - |
| 50000 | 0.20699372432290739 | 0.20724775222565622 | 0.20711044017505142 | - |

Таблиця 1.5 – Визначення автора невідомого тексту (Марко Вовчок - Павло Чорнокрил)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Розмір  уривку  тексту | Іван Багряний | Михайло Старицький | Марко Вовчок | Визначено автора |
| 5000 | 0.13010736022716649 | 0.11615851251778876 | 0.0860647348750019 | + |
| 10000 | 0.12150577389618854 | 0.1008008876824641 | 0.07560956826837481 | + |
| 25000 | 0.07560956826837481 | 0.0978108536067914 | 0.06778618955934665 | + |
| 50000 | 0.11462684603414341 | 0.09305271905685797 | 0.06057769044883725 | + |

Було неправильно розпізнано автора в таблиці 1.4 із-за достатньо малих різниць в значеннях і велику схожість теплових діаграм авторів в цілому. Також, можемо зробити висновок, що чим більший уривок тексту тим більша вірогідність правильно визначити автора.

**Таблиця результатів визначення автора за уривком тексту за допомогою біграмів**

Таблиця 1.6 – Визначення автора невідомого тексту (Іван Багряний - Людина біжить над прірвою)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Розмір  уривку  тексту | Іван Багряний | Михайло Старицький | Марко Вовчок | Визначено автора |
| 5000 | 0.3169472440096983 | 0.35160803599180884 | 0.4187956494512635 | + |
| 10000 | 0.2593214543973275 | 0.3030391112507171 | 0.37521928313116026 | + |
| 25000 | 0.19148697914706042 | 0.2490865167742695 | 0.33005478751989387 | + |
| 50000 | 0.17899870991834874 | 0.23846946730252963 | 0.32790717705500366 | + |

Таблиця 1.7 - Визначення автора невідомого тексту (Михайло Старицький - Молодість Мазепи)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Розмір  уривку  тексту | Іван Багряний | Михайло Старицький | Марко Вовчок | Визначено автора |
| 5000 | 0.5202601223752776 | 0.5234431435684263 | 0.5011615067359579 | - |
| 10000 | 0.469541002085491 | 0.47070886326680544 | 0.4519441717166633 | - |
| 25000 | 0.42317836279389603 | 0.41740015195347646 | 0.40192157869535916 | - |
| 50000 | 0.40412061458 | 0.39234337635 | 0.39417961850 | + |

Таблиця 1.8 - Визначення автора невідомого тексту (Марко Вовчок - Павло Чорнокрил)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Розмір  уривку  тексту | Іван Багряний | Михайло Старицький | Марко Вовчок | Визначено автора |
| 5000 | 0.3441281935778748 | 0.3394549111760917 | 0.30432789603692034 | + |
| 10000 | 0.3355282377262333 | 0.32303770260689224 | 0.25771653507048875 | + |
| 25000 | 0.32836944468018786 | 0.3003434614500422 | 0.21393200076013205 | + |
| 50000 | 0.31636462839730745 | 0.29034645560463396 | 0.19773154905817872 | + |

Для твору “Молодість Мазепи” Старицького порівняння біграм дало трохи кращі результати, але все ще на малому по розміру уривку дана техніка не змогла правильно ідентифікувати автора. В усіх інших випадках автори були визначені правильно. Результати на великих уривках тексту є кращими.

**Висновки**

За допомогою статистичного аналізу, який був проведений в даній лабораторній работі можемо зробити висновок, що не для всіх творів ми можемо правильно ідентифікувати автора, так як в даному випадку гістограми частот, теплові діаграми та гістограми порівняння творів авторів з використанням символів української мови не були дуже відмінними. Аналіз за допомогою біграмів дав твохи кращий результат у визначенні ніж аналіз тексту за допомогою грамів. Розмір уривку тексту також впливає на точність визначення автора (Чим більший уривок тим краще).