**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ**

**ІМЕНІ ТАРАСА ШЕВЧЕНКА**

**ФАКУЛЬТЕТ РАДІОФІЗИКИ, ЕЛЕКТРОНІКИ ТА КОМП’ЮТЕРНИХ СИСТЕМ**

**КАФЕДРА КОМП’ЮТЕРНОЇ ІНЖЕНЕРІЇ**

Лабораторна робота №4

З дисципліни «Захист інформації у комп’ютерних системах»

«Дослідження частотних характеристик української мови»

Виконав   
Юрченко П. А.   
Студент 4 курсу СА

Київ 2024

**Мета**: дослідити вірогіднісні параметри появи літер української мови для різних типів текстів. Аналіз найбільш імовірних літер, біграм та триграм для використання в частотному криптоаналізі.

**Хід роботи**

1. Знаходження відносної частоти появи літер українського алфавіту

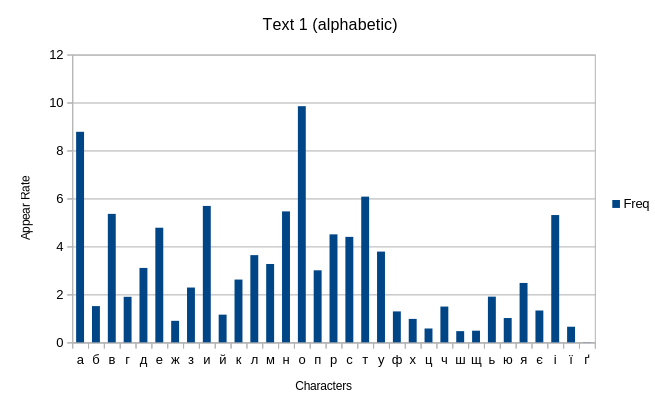
Створіть програму (будь-якою зручною для вас мовою), яка в якості аргументів приймає перелік текстових файлів, та аналізуючи їх вміст обраховує частоти появи літер українського алфавіту. Для спрощення аналізу з тексту виключаються всі знаки пунктуації окрім “пробілу”.

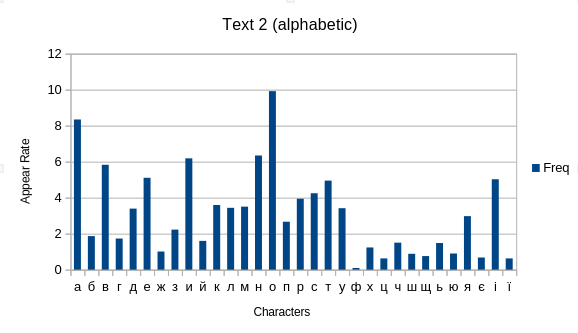
Для аналізу відносної частоти появи літер в українській мові було використано текст переказу “Фауста” та перша глава “Життєвої філософії кота Мура” (приблизно 20к символів кожен).

Код для проведення аналізу був написаний на JavaScript.

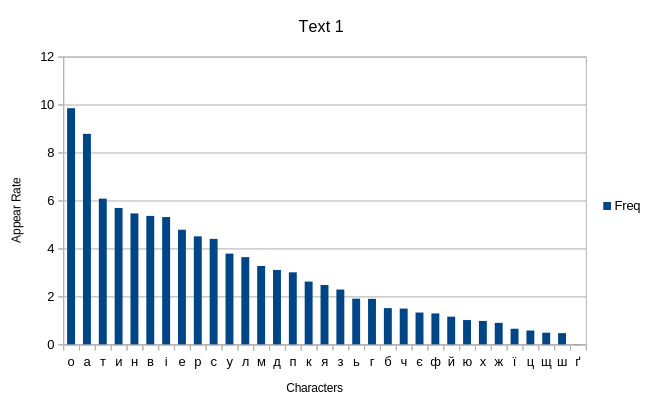
[Посилання на репозиторій з кодом.](https://github.com/psht13/ukr-lang-stats)

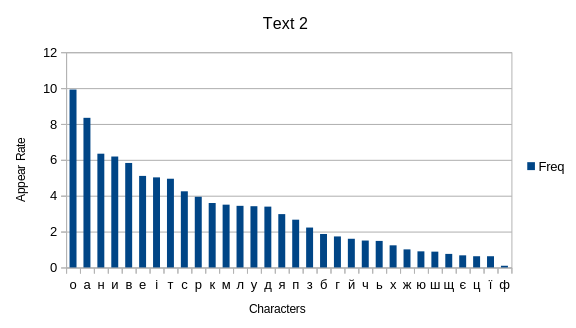
діаграму відсортовану в алфавітному порядку





діаграму відсортовану по частотам появи літер





послідовність літер по мірі спадання частоти появи

Text 1:

о а т и н в і е р с у л м д п к я з ь г б ч є ф й ю х ж ї ц щ ш ґ

Text 2:

о а н и в е і т с р к м л ю д я п з б г й ч ь х ж ю ш щ є ц ї ф

2. Знаходження відносної частоти появи біграм українського алфавіту

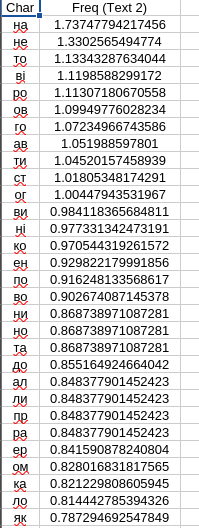
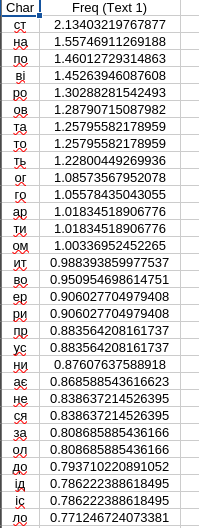
Аналогічно до першого завдання, створіть програму яка підрахує відносні частоти появи біграм української мови аналізуючи вміст текстів, що задаються.

Написав знову ж таки на JavaScript.

Для текстів, що були використані в першому завданні проаналізуйте отримані результаті та в звіті наведіть:

таблицю з відносними частотами біграм, відсортовану за спаданням частоти

На жаль, помістилися далеко не всі біграми, але наводжу певний список відсортований за зростанням частоти для обох текстів.



послідовність з 30-ти найбільш імовірних біграм

Текст 1:

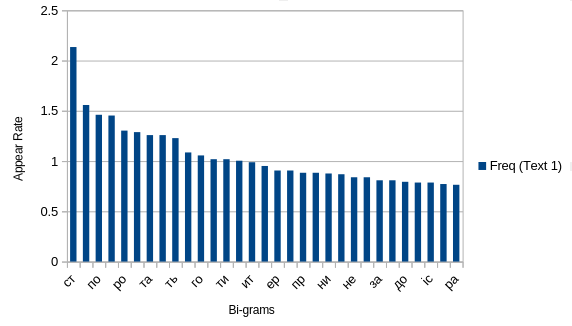
на не то ві ро ов го ав ти ст ог ви ні ко ен по во ни но та до ал ли пр ра ер ом ка ло як

Текст 2:

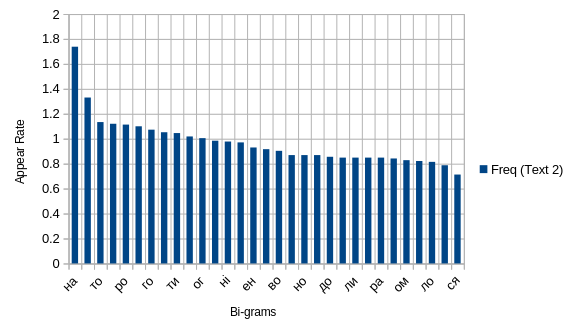
ст на по ві ро ов та то ть ог го ар ти ом ит во ер ри пр ус ни ає не ся за ол до ід іс ло

діаграму відсортовану по частотам появи 30-ти найбільш імовірних біграм

Текст 1:

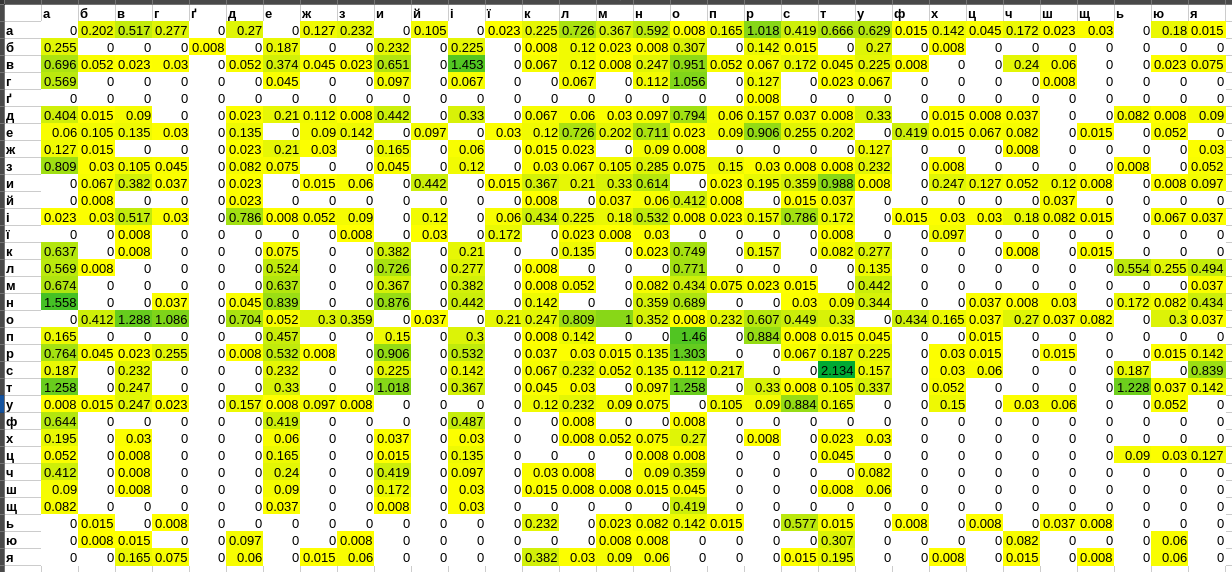


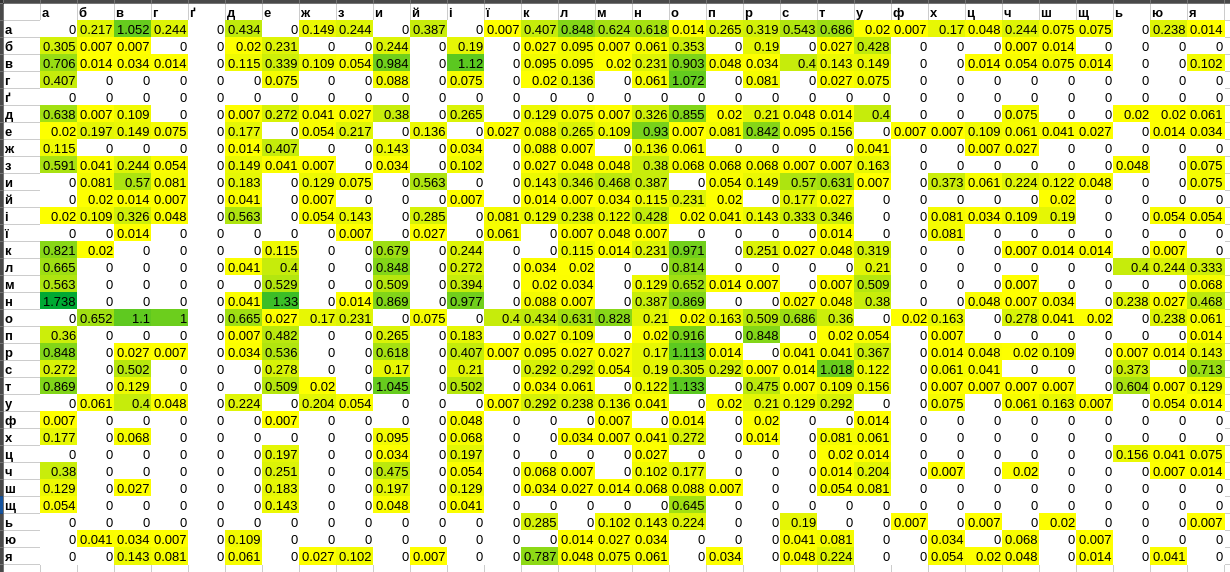
Текст 2:



матрицю частот появи біграм (імовірність позначена кольором)

Текст 1:

Текст 2:



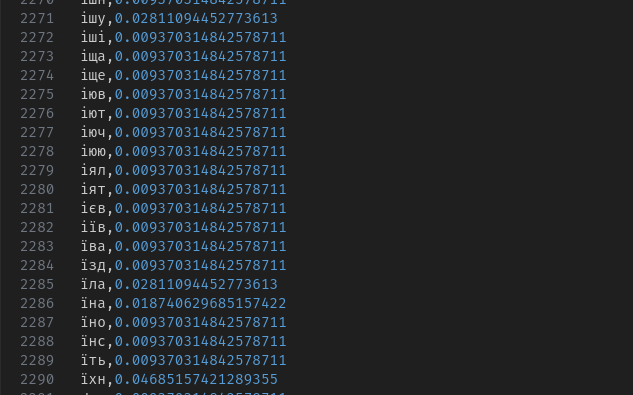
3. Знаходження відносної частоти появи триграм українського алфавіту

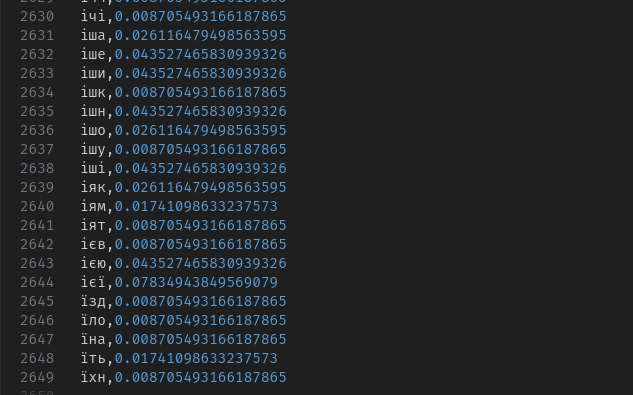
Для текстів, що були використані в першому завданні проаналізуйте отримані результаті та в звіті наведіть:

таблицю з відносними частотами триграм, відсортовану за спаданням частоти

Не думаю що це можливо вмістити 2000+ триграм в один екран таблиці тому, демонструю лише такі скріни.

Текст 1:

Текст 2:

послідовність з 30-ти найбільш імовірних триграм

Текст 1:

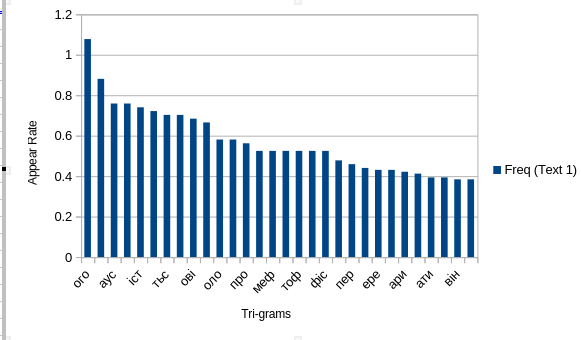
ого уст аус фау іст сто тьс ься ові від оло ста про ефі меф офе тоф фел фіс єть пер ому ере лов ари ель ати йог він при роз

Текст 2:

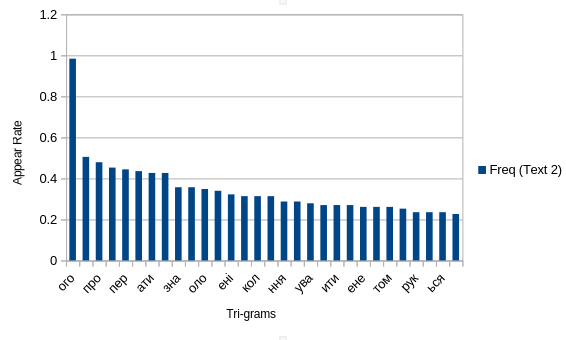
ого мен про від пер ому ати при зна так оло ере ені віт кол ові ння сво ува вся ити ово ене сам том сто рук тьс ься ага енн

діаграму відсортовану по частотам появи 30-ти найбільш імовірних триграм

Текст 1:



Текст 2:



4. Криптоаналіз

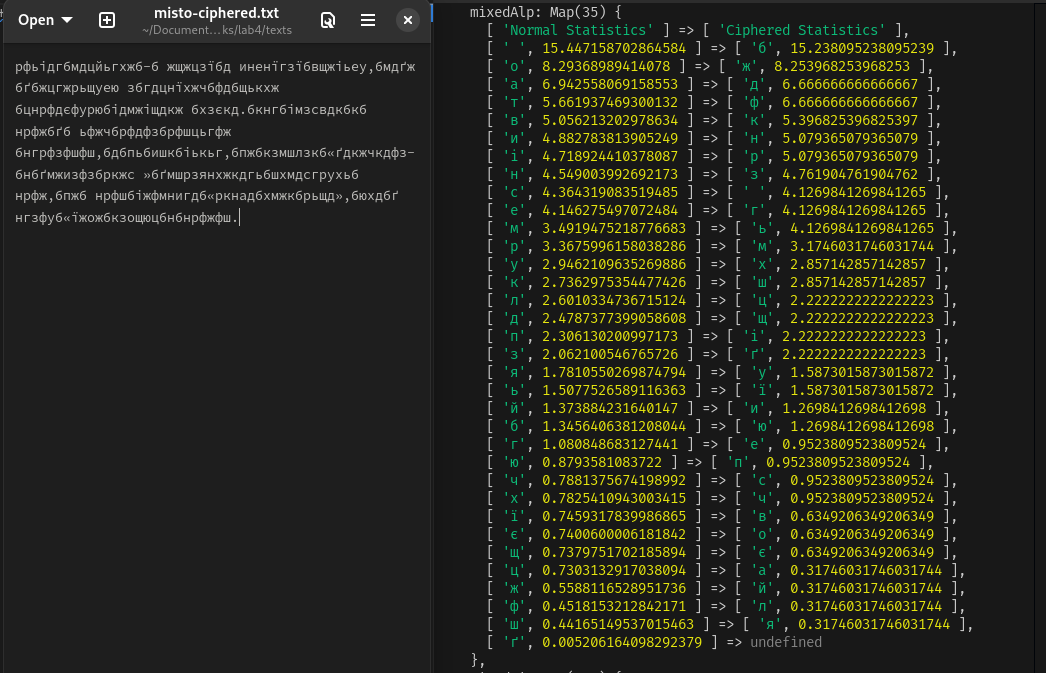
Зашифруйте за допомогою афінного шифру (a та b оберіть випадковим чином) фрагмент тексту та надайте отриманий шифротекст колегам.

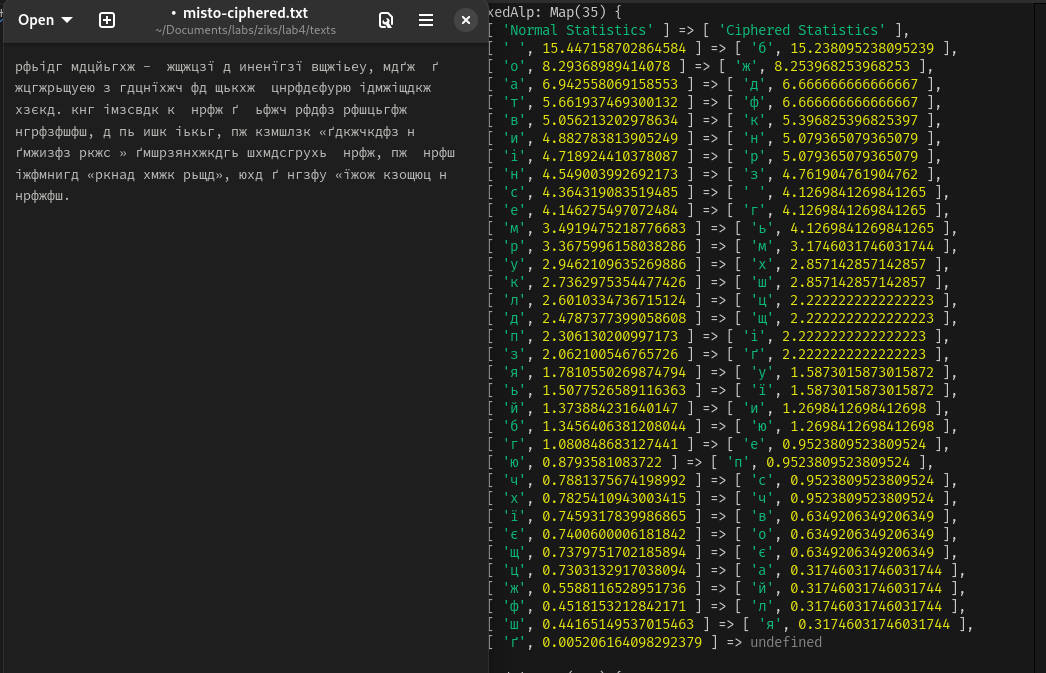
Зашифрував та надав колегам.

Отримайте шифротекст від колег, виконайте їх криптоаналіз та проілюструйте в звіті процес його проведення за допомогою частотного аналізу.

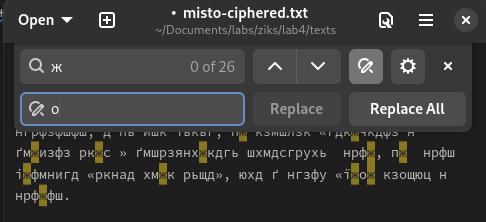
Отриманий шифротекст.



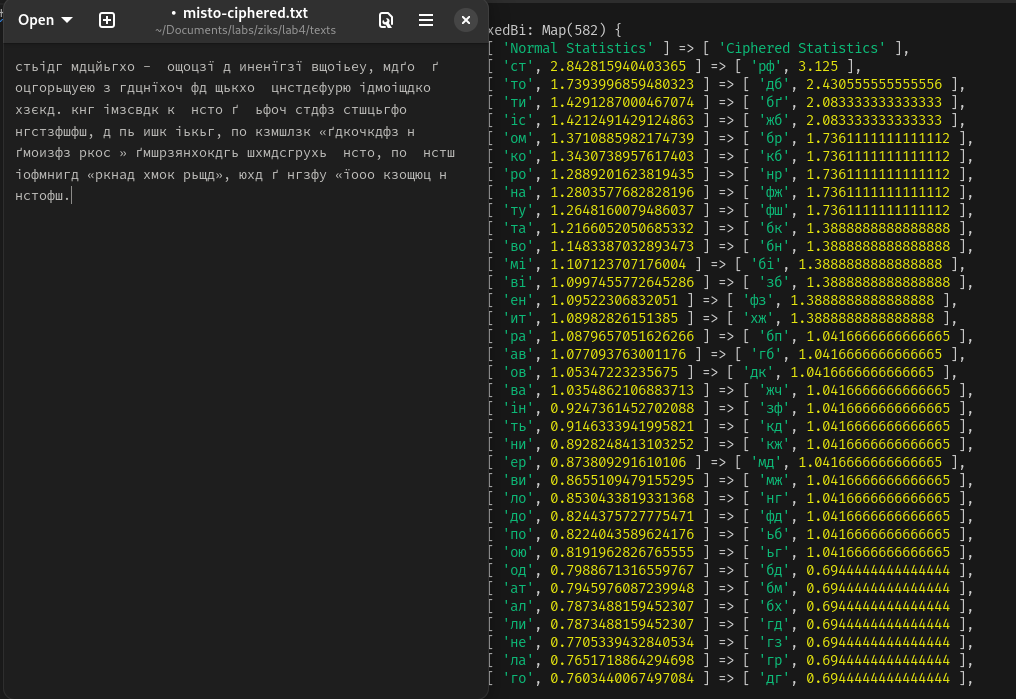
На основі частотного аналізу роблю припущення, що замість ”б” у шифрі має бути пробіл.

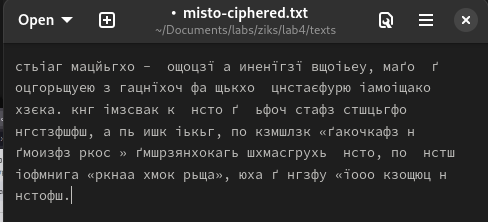


По аналогії припускаю, що замість ”ж” у шифрі має бути “о”.

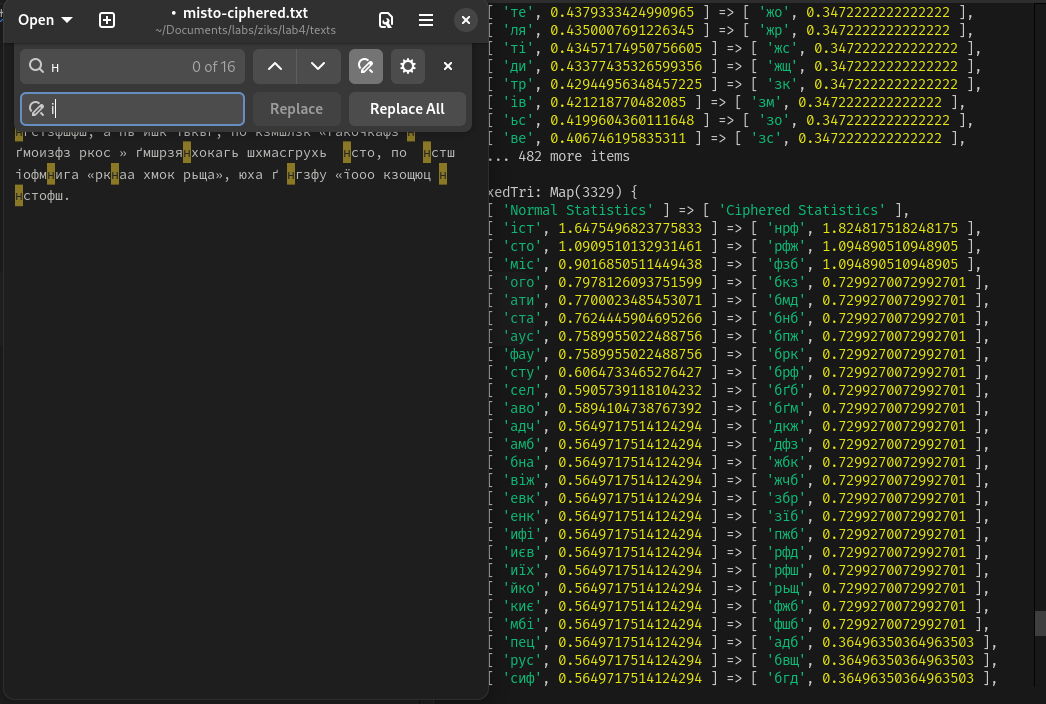


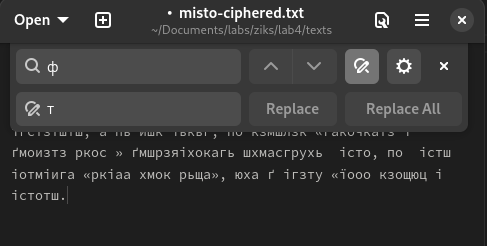
Цього разу зроблю припущення по частотах появи біграм, що “рф” це “ст”.

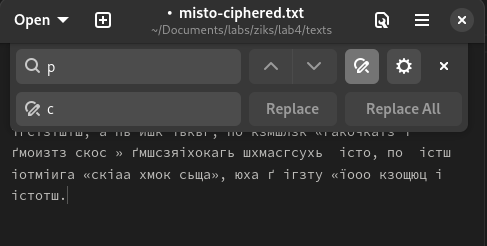
Також зі статистики біграм можна зрозуміти що ”дб” це закінчення слів, бо ”б” це пробіл. Також перевіривши частоту появи літер можна припустити, що ”д” - ”а”.



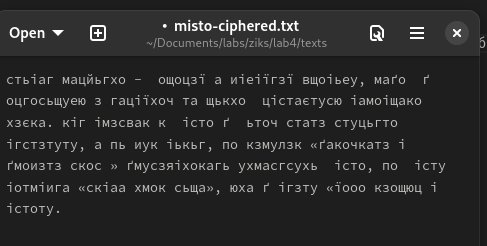
На основі триграм припускаю що “нрф” це “іст”, відповідно “н” це “і”.



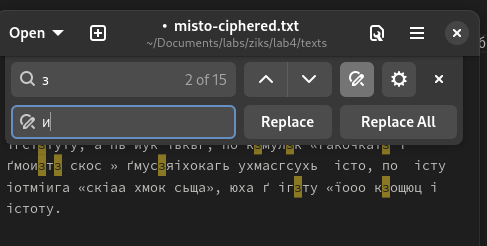


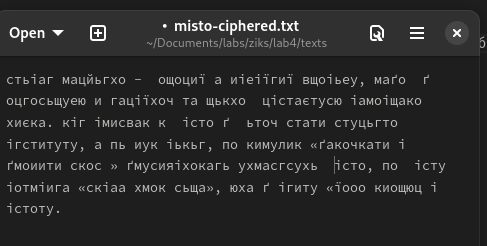


Бачу слово “істотш” це або “істота”, але “а” нам уже відоме, тому це “істоту” - “ш” це “у”.

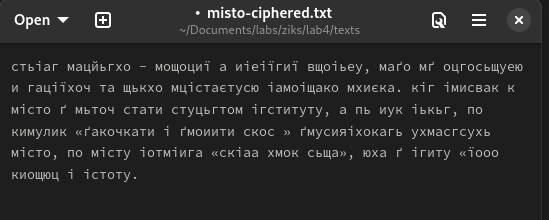


Також з аналізу тригам бачу що “з” це “и”

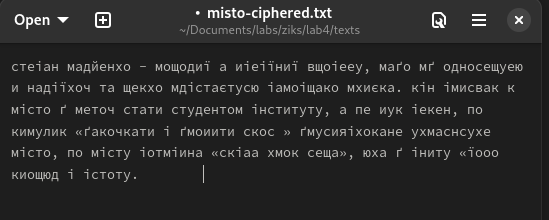




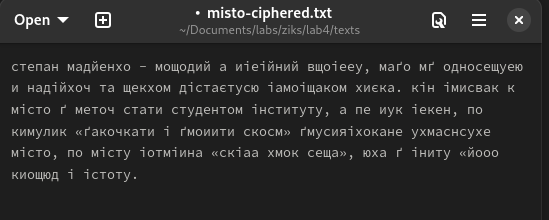
Бачу слово “місто” де пробіл виступає у ролі “м” (трохи криво замінив тут, тому деякі слова попливли).



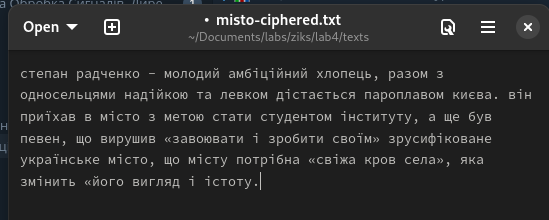
Бачу слова “стуцьгтом ігституту” це “студентом інституту”.



На даному етапі можна вибирати окремі слова, та розшифровувати до кінця.



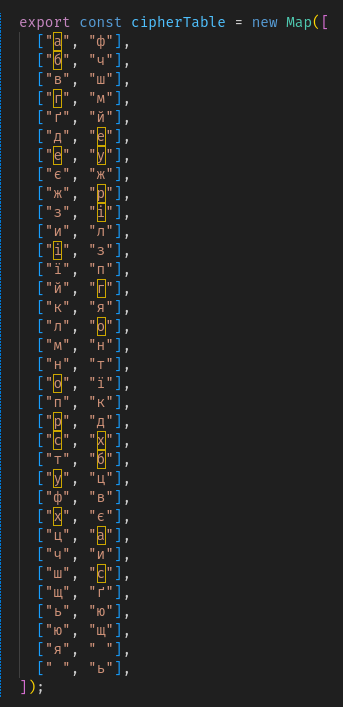
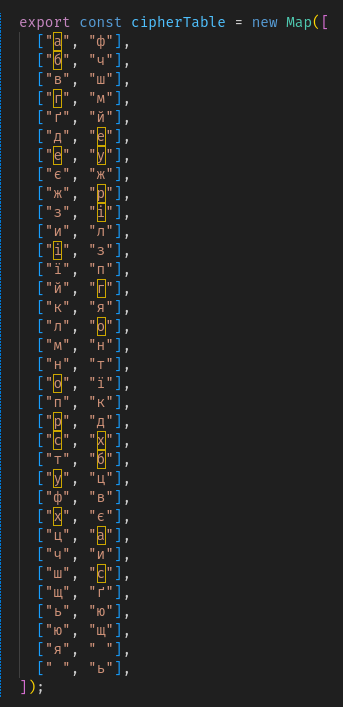
Фінальний результат.



5. Криптоаналіз довільного моноалфавітного шифру.

Згенеруйте випадкову моноалфавітну підстановку (таблицю відповідності) та зашифруте за її допомогою достатньо довгий фрагмент тексту.

Таблиця по якій відбувалося шифрування:

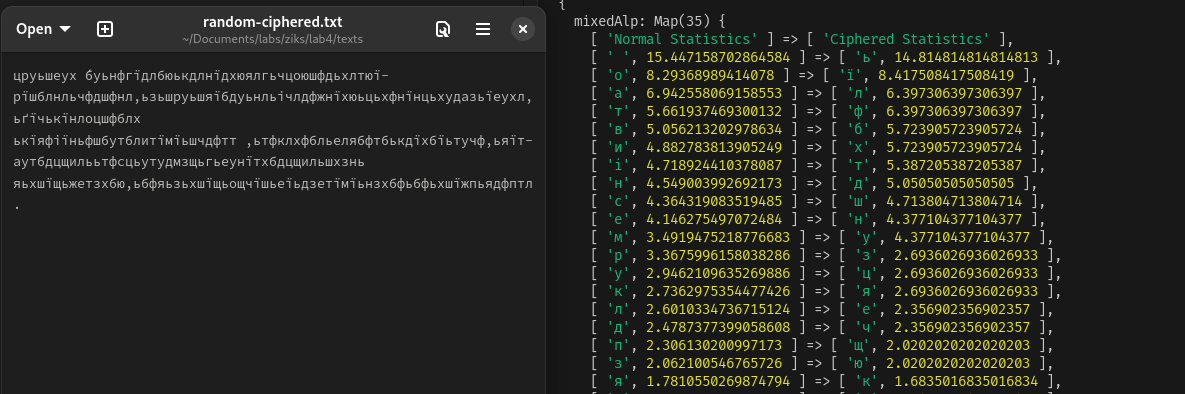


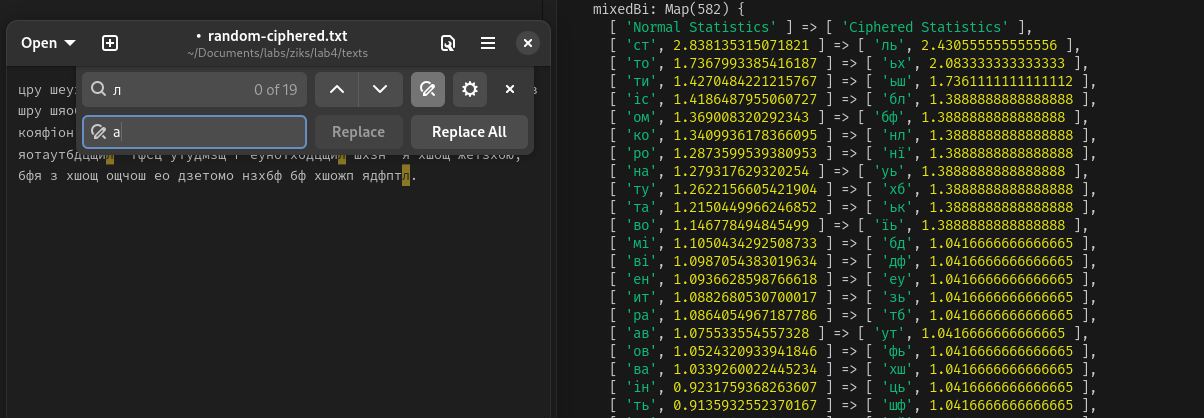
Отриманий текст:



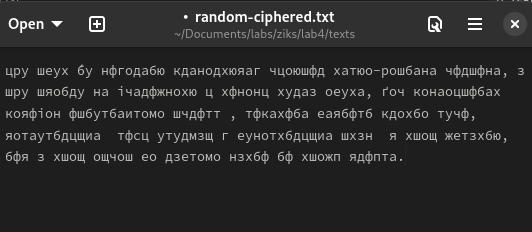
Виконайте криптоаналіз отриманого шифротексту, використовуючи частотний аналіз (в т.ч. частоти біграм, триграм та припущень).

На основі статистики зроблю припущення що «ь» та «ї це пробіл та «о» відовідно.

 По біграмах видно що «ль» це закінчення, скоріш за все голосна + пробіл, припускаю що це «а»



Ось результат, якого зміг досягти, далі туго пішло, частотні характеритиски зашифрованого тексту не зовсім відповідають моїм статистичним даним.



**Висновок.** Під час виконання лаборатоної роботи було проведено аналіз частотних характеристик літер, біграм та триграм української мови, що дозволяє ефективно застосовувати їх для частотного криптоаналізу. Частотний підхід підтвердив, що мовні особливості українських текстів можуть значно покращити точність розшифрування і підвищити ефективність криптоаналітичних методів.