# КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА ШЕВЧЕНКА

# ФАКУЛЬТЕТ РАДІОФІЗИКИ, ЕЛЕКТРОНІКИ ТА КОМП'ЮТЕРНИХ СИСТЕМ

## КАФЕДРА КОМП'ЮТЕРНОЇ ІНЖЕНЕРІЇ

Лабораторна робота №4

3 дисципліни «Захист інформації у комп'ютерних системах» «Дослідження частотних характеристик української мови»

Виконав Юрченко П. А. Студент 4 курсу СА **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

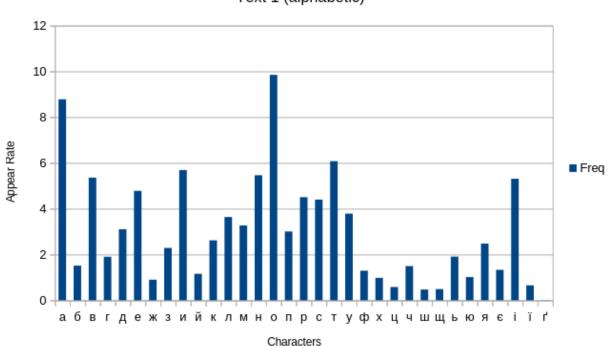
Створіть програму (будь-якою зручною для вас мовою), яка в якості аргументів приймає перелік текстових файлів, та аналізуючи їх вміст обраховує частоти появи літер українського алфавіту. Для спрощення аналізу з тексту виключаються всі знаки пунктуації окрім "пробілу".

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.

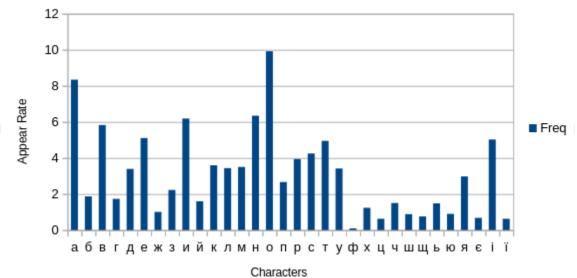
Посилання на репозиторій з кодом.

діаграму відсортовану в алфавітному порядку



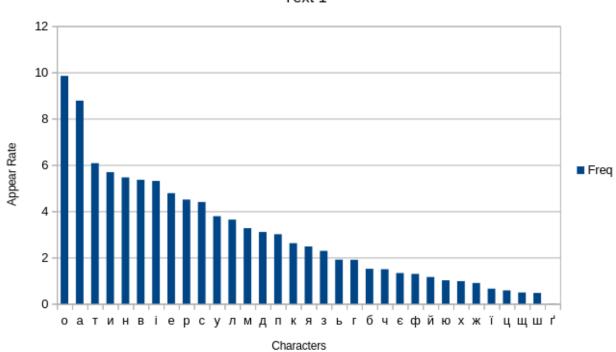
Text 1 (alphabetic)

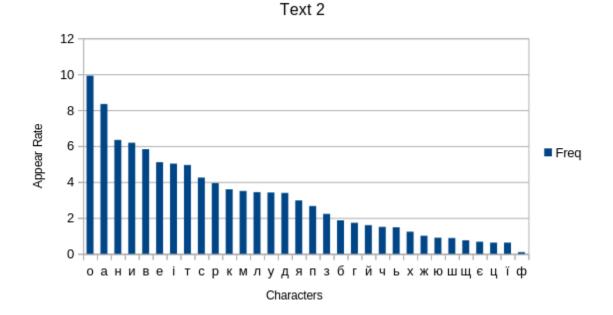




діаграму відсортовану по частотам появи літер

Text 1





послідовність літер по мірі спадання частоти появи

### Text 1:

оатинвіерсулмдпкязьгбчєфйюхжїцщшг

#### Text 2:

## оанивеітсркмлюдяпзбгйчьхжюшщєцїф

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.

## Hаписав 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.

Char	Freq (Text 1)
ÇŢ	2.13403219767877
на	1.55746911269188
ЛО	1.46012729314863
Вį	1.45263946087608
po	1.30288281542493
QB	1.28790715087982
Ţа	1.25795582178959
TO	1.25795582178959
及	1.22800449269936
ΟĹ	1.08573567952078
ΓQ	1.05578435043055
ар	1.01834518906776
TN	1.01834518906776
QM.	1.00336952452265
ИT	0.988393859977537
BO	0.950954698614751
еp	0.906027704979408
ри	0.906027704979408
дp	0.883564208161737
уc	0.883564208161737
ни	0.87607637588918
ąє	0.868588543616623
не	0.838637214526395
ÇЯ	0.838637214526395
3 <u>a</u>	0.808685885436166
<u>9</u> Л	0.808685885436166
до	0.793710220891052
įд	0.786222388618495
jç	0.786222388618495
до	0.771246724073381

Char	Freq (Text 2)
на	1.73747794217456
не	1.3302565494774
ŢQ	1.13343287634044
ві	1.1198588299172
po	1.11307180670558
QB	1.09949776028234
го	1.07234966743586
ав	1.051988597801
ТИ	1.04520157458939
СТ	1.01805348174291
ΟΓ	1.00447943531967
ВИ	0.984118365684811
Нį	0.977331342473191
ко	0.970544319261572
ен	0.929822179991856
ПО	0.916248133568617
во	0.902674087145378
НИ	0.868738971087281
НО	0.868738971087281
та	0.868738971087281
до	0.855164924664042
ал	0.848377901452423
ди	0.848377901452423
ДB	0.848377901452423
рa	0.848377901452423
еp	0.841590878240804
QM	0.828016831817565
ка	0.821229808605945
до	0.814442785394326
як	0.787294692547849

послідовність з 30-ти найбільш імовірних біграм

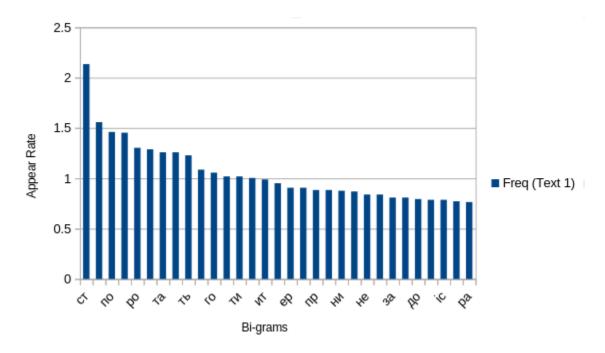
### **Текст 1:**

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

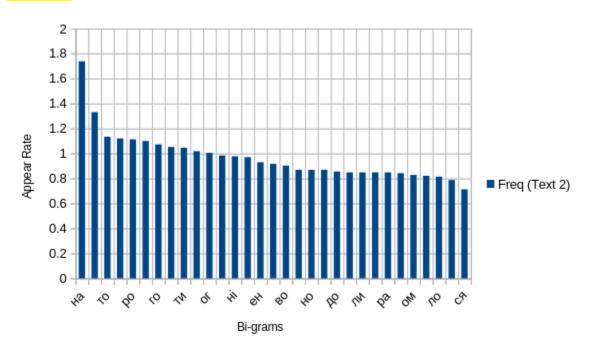
### **Текст 2:**

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:

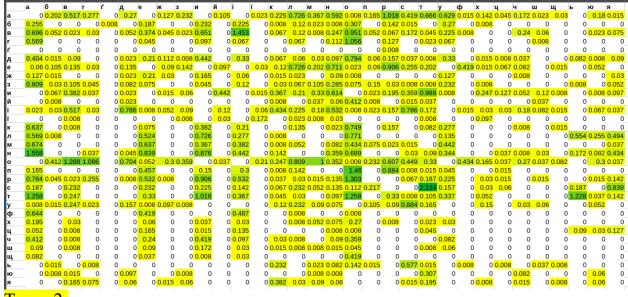


## Текст 2:

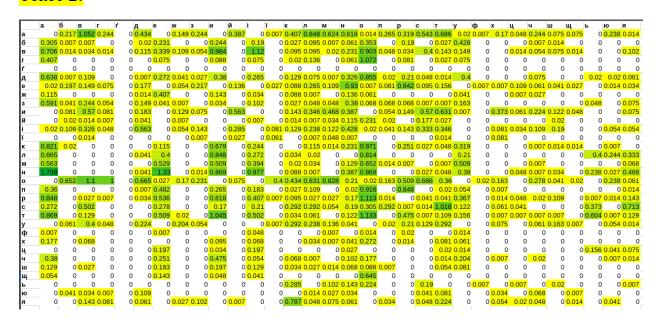


матрицю частот появи біграм (імовірність позначена кольором)

## **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:

```
iшy,0.02811094452773613
       iшi,0.009370314842578711
       iщa,0.009370314842578711
       iще,0.009370314842578711
       iюв,0.009370314842578711
       iют,0.009370314842578711
       iюч,0.009370314842578711
       iюю,0.009370314842578711
       іял,0.009370314842578711
       іят,0.009370314842578711
       i∈B,0.009370314842578711
       iiB,0.009370314842578711
       їва,0.009370314842578711
       їзд,0.009370314842578711
       їла,0.02811094452773613
       їна,0.018740629685157422
       їно,0.009370314842578711
       їнс,0.009370314842578711
       їть,0.009370314842578711
2290
       їхн,0.04685157421289355
```

#### Текст 2:

```
2630
       iui,0.008705493166187865
       iшa,0.026116479498563595
       iше,0.043527465830939326
       іши,0.043527465830939326
       ішк,0.008705493166187865
       ішн,0.043527465830939326
       iшo,0.026116479498563595
      iшv.0.008705493166187865
       iшi,0.043527465830939326
      іяк,0.026116479498563595
      іям,0.01741098633237573
      іят,0.008705493166187865
      i∈B,0.008705493166187865
      iεω, 0.043527465830939326
      i∈ï,0.07834943849569079
      їзд,0.008705493166187865
      їло,0.008705493166187865
      їна,0.008705493166187865
       їть,0.01741098633237573
       їхн,0.008705493166187865
```

послідовність з 30-ти найбільш імовірних триграм

#### Текст 1:

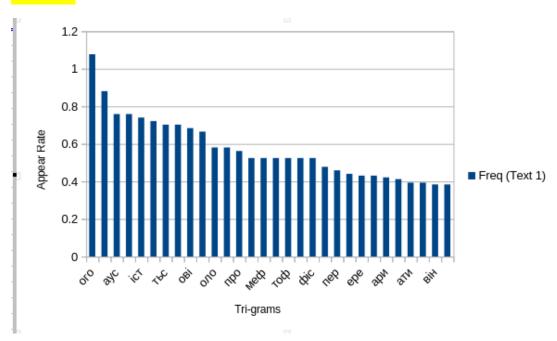
ого уст аус фау іст сто тьс ься ові від оло ста про ефі меф офе тоф фел фіс  $\epsilon$ ть пер ому ере лов ари ель ати йог він при роз

## **Текст 2**:

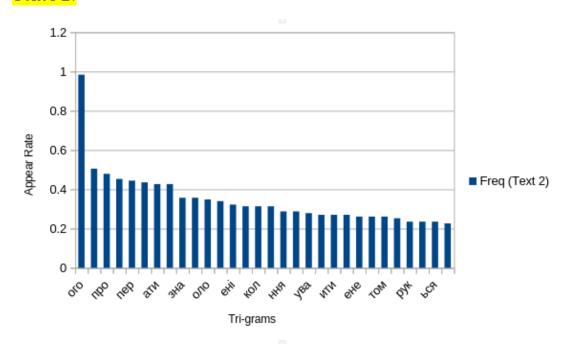
ого мен про від пер ому ати при зна так оло ере ені віт кол ові ння сво ува вся ити ово ене сам том сто рук тьс ься ага енн

діаграму відсортовану по частотам появи 30-ти найбільш імовірних триграм

### **Текст 1**:



## **Текст 2:**



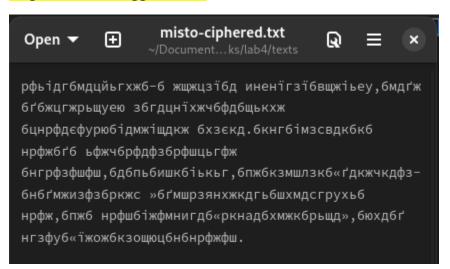
## 4. Криптоаналіз

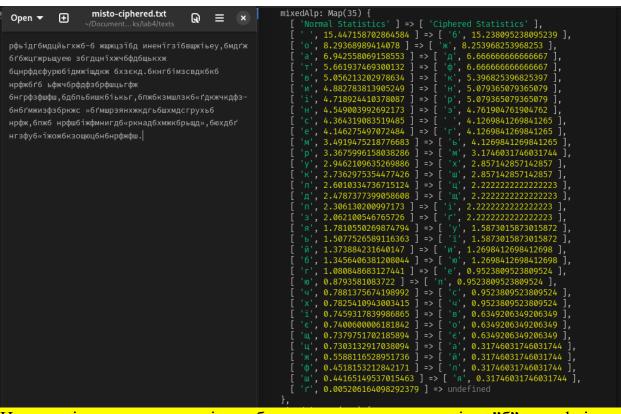
Зашифруйте за допомогою афінного шифру (а та b оберіть випадковим чином) фрагмент тексту та надайте отриманий шифротекст колегам.

#### Зашифрував та надав колегам.

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

### Отриманий шифротекст.





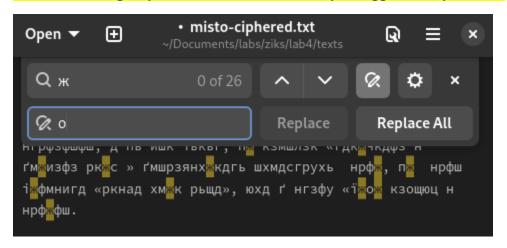
На основі частотного аналізу роблю припущення, що замість "б" у шифрі має бути пробіл.

```
kedAlp: Map(35)

    misto-ciphered.txt

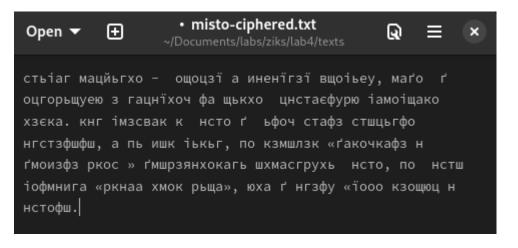
             \oplus
                                                              Q)
                                                                  ≡
Open ▼
                                                                                          рфьідг мдцйьгхж - жщжцзї д иненїгзї вщжіьеу, мдґж ґ
жцгжрьщуею з гдцнїхжч фд щькхж цнрфдєфурю ідмжіщдкж
хзєкд. кнг імзсвдк к нрфж ґ ьфжч рфдфз рфшцьгфж
нгрфзфшфш, д пь ишк іькьг, пж кзмшлзк «ґдкжчкдфз н
іжфмнигд «ркнад хмжк рьщд», юхд ґ нгзфу «їжож кзощюц н
                                                                                           3.4919475218776683
3.3675996158038286
                                                                                                                                     3.1746031746031744
                                                                                           2.6010334736715124
                                                                                           2.4787377399058608 ] =>
2.306130200997173 ] => [
2.062100546765726 ] => [
                                                                                           1.5077526589116363 ] =>
1.5077526589116363 ] =>
1.373884231640147 ] => [
1.3456406381208044 ] =>
1.080848683127441 ] => [
                                                                                          0.7459317839986865 ] =>
0.74006000006181842 ] =>
0.7379751702185894 ] =>
                                                                                                                                    0.6349206349206349
0.6349206349206349
0.31746031746031744
                                                                                          0.7303132917038094
                                                                                                                                    0.31746031746031744
```

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

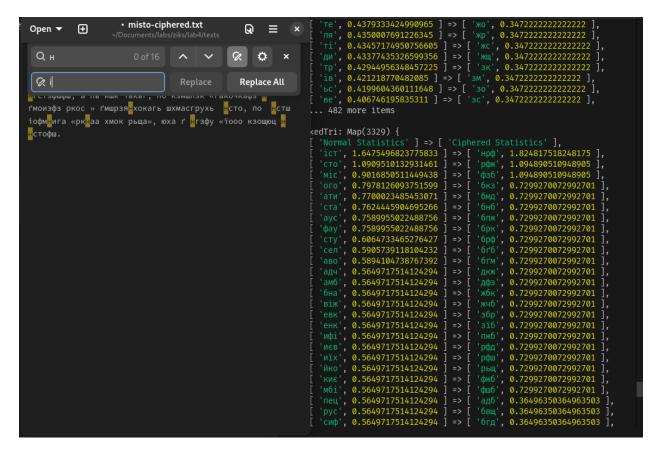


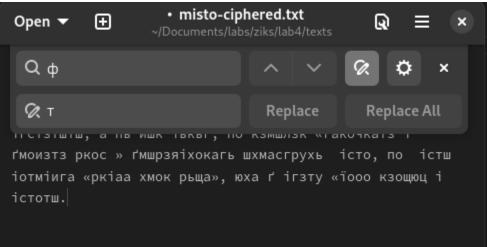
Цього разу зроблю припущення по частотах появи біграм, що "рф" це "ст".

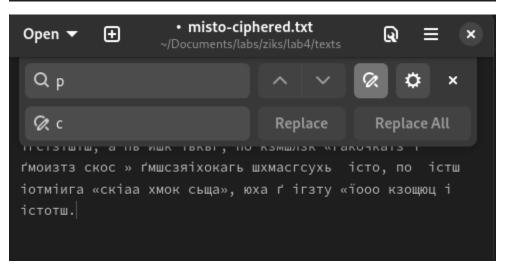
Також зі статистики біграм можна зрозуміти що "дб" це закінчення слів, бо "б" це пробіл. Також перевіривши частоту появи літер можна припустити, що "д" - "а".



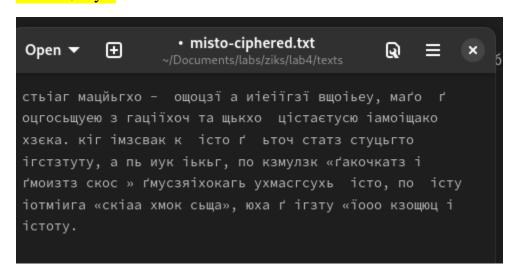
На основі триграм припускаю що "нрф" це "іст", відповідно "н" це "і".



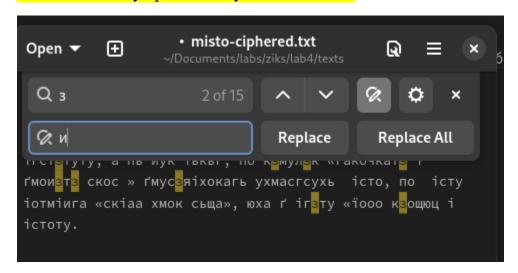


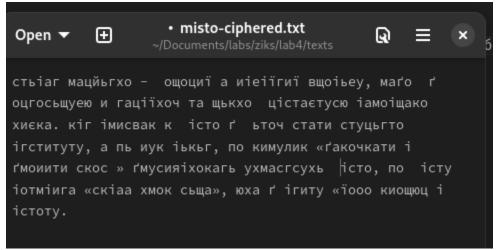


Бачу слово "істотш" це або "істота", але "а" нам уже відоме, тому це "істоту" - "ш" це "у".

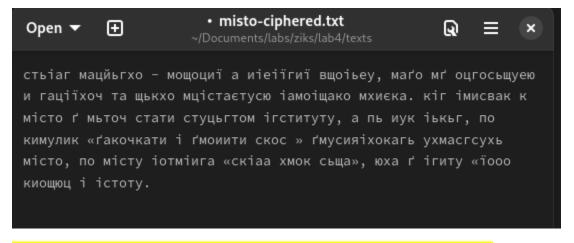


#### Також з аналізу тригам бачу що "з" це "и"

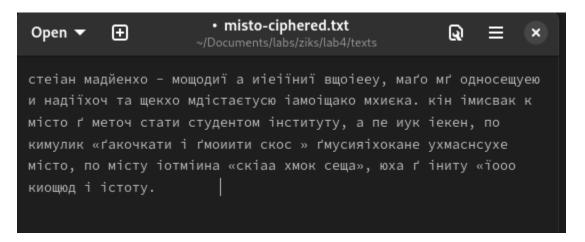




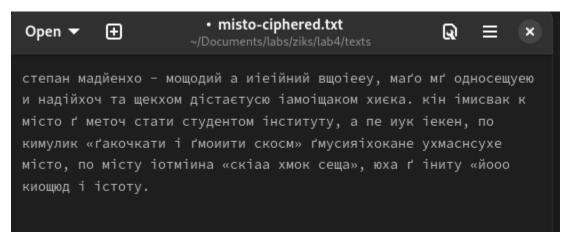
Бачу слово "місто" де пробіл виступає у ролі "м" (трохи криво замінив тут, тому деякі слова попливли).



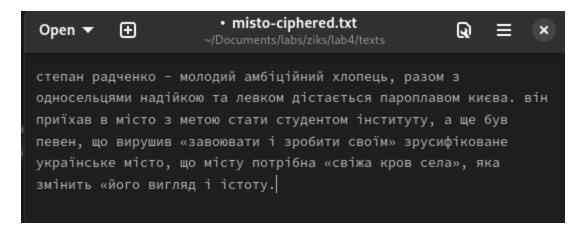
Бачу слова "стуцьгтом ігституту" це "студентом інституту".



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



Фінальний результат.



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.