

**Київський національний університет імені Тараса Шевченка  
факультет радіофізики, електроніки та комп'ютерних систем**

**Лабораторна робота № 4**

**Тема: «Дослідження частотних характеристик  
української мови»**

**Роботу виконав  
студент 4 курсу КІ СА  
Голубцов Владислав Романович**

**Київ 2021**

**Тема:** Дослідження частотних характеристик української мови

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

**Короткі відомості:** Відносна частота - вірогідність появи певної літери чи  $m$ -грами в певному місці тексту.

Обчислюється як відношення числа появи  $m$ -грами в тексті до загальної кількості підрахованих  $m$ -грам. Криптоаналіз — розділ криптології, що займається математичними методами порушення конфіденційності і цілісності інформації без знання ключа. Частотний аналіз — метод криптоаналізу, який ґрунтується на частоті появи знаків шифротексту, а саме на припущенні про існування нетривіального статистичного розподілу окремих символів і їх послідовностей як у відкритому тексті, так і в шифротексті, який, з точністю до заміни символів, буде зберігатися в процесі шифрування і дешифрування.

# Хід роботи

## 1. Знаходження відносної частоти появи літер українського алфавіту

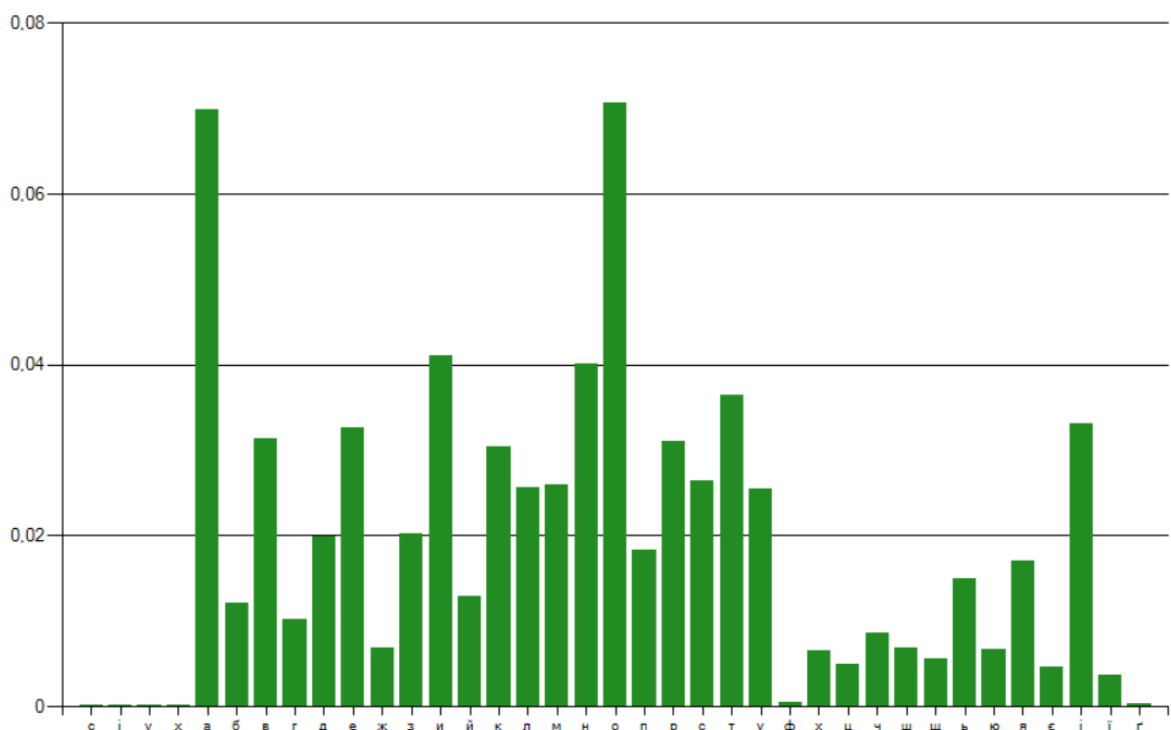
1. Створіть програму (будь-якою зручною для вас мовою), яка в якості аргументів приймає перелік текстових файлів, та аналізуючи їх вміст обраховує частоти появи літер українського алфавіту. Для спрощення аналізу з тексту виключаються всі знаки пунктуації окрім “пробілу”.

### *C#, WindowsForms*

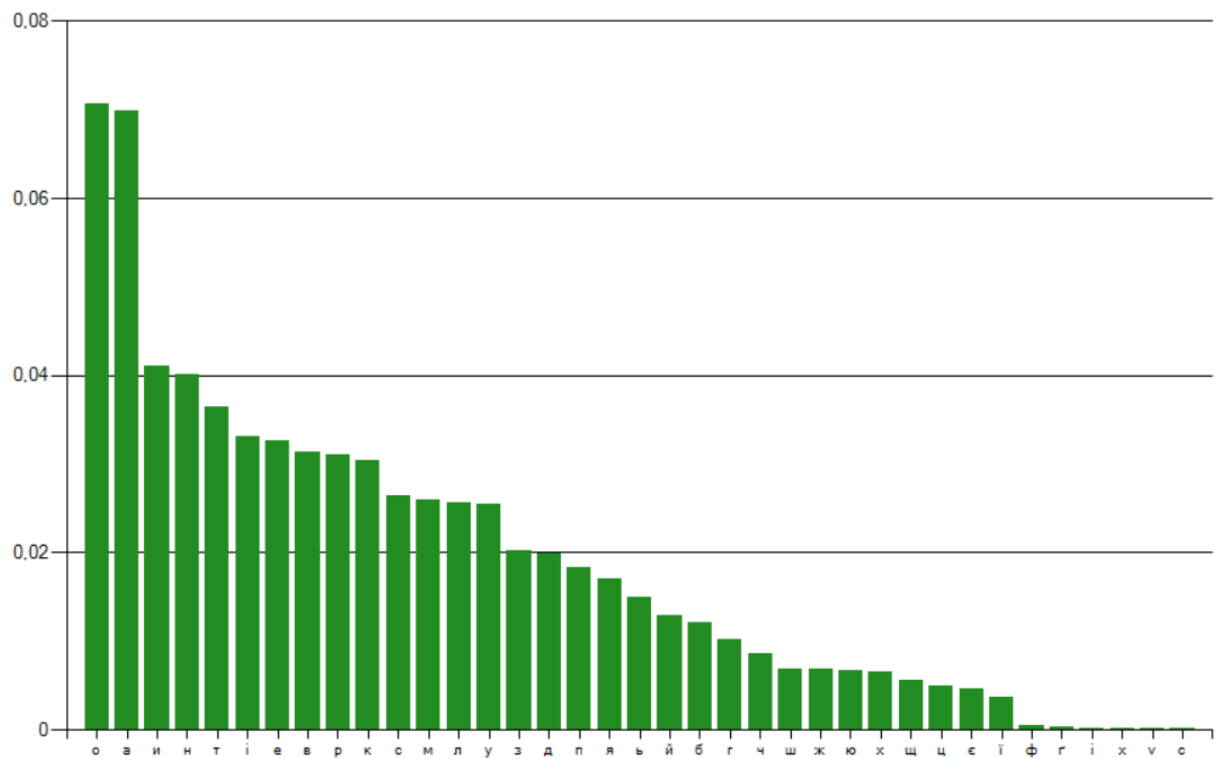
2. Для різних достатньо довгих текстів (*оберіть принаймні 2, причому тексти не повинні повторюватись в групі*), наприклад, художній твір, вірш, технічна документація, блог, тощо, проаналізуйте отримані результати та в звіті наведіть:

#### 1. Мина Мазайло, М Куліш

- а. діаграму відсортовану в алфавітному порядку



б. діаграму відсортовану по частотам появи літер

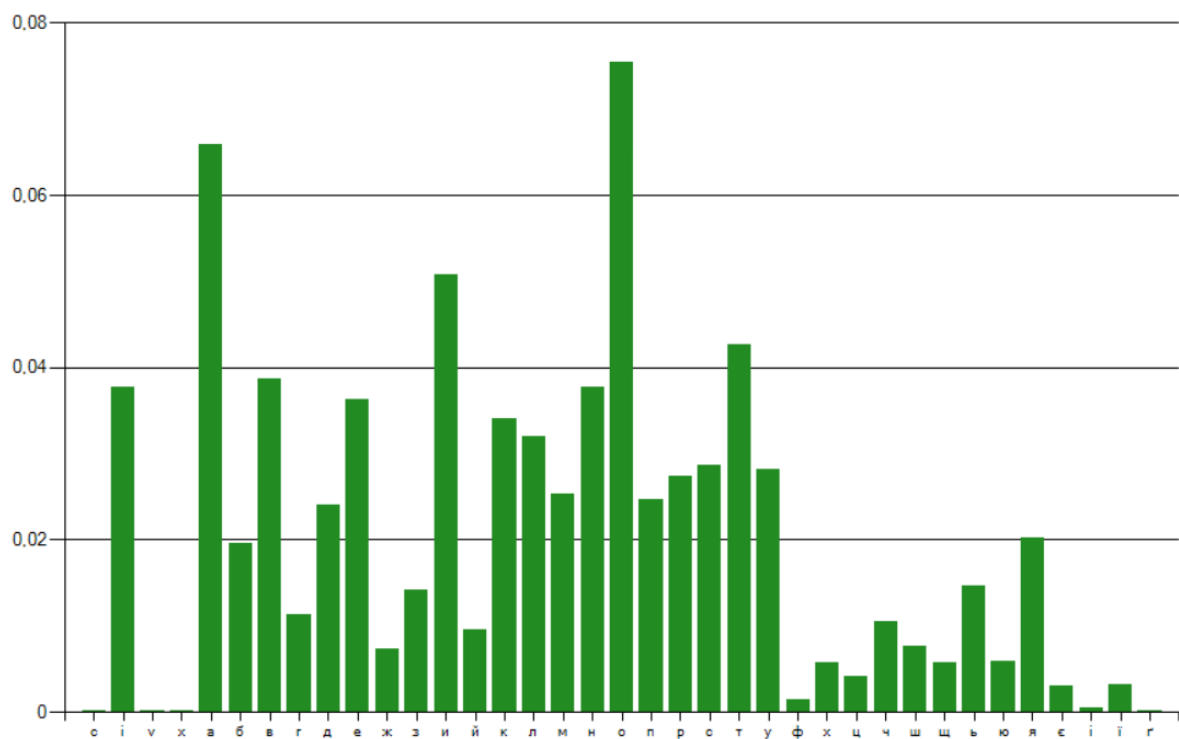


с. послідовність літер по мірі спадання частоти появи

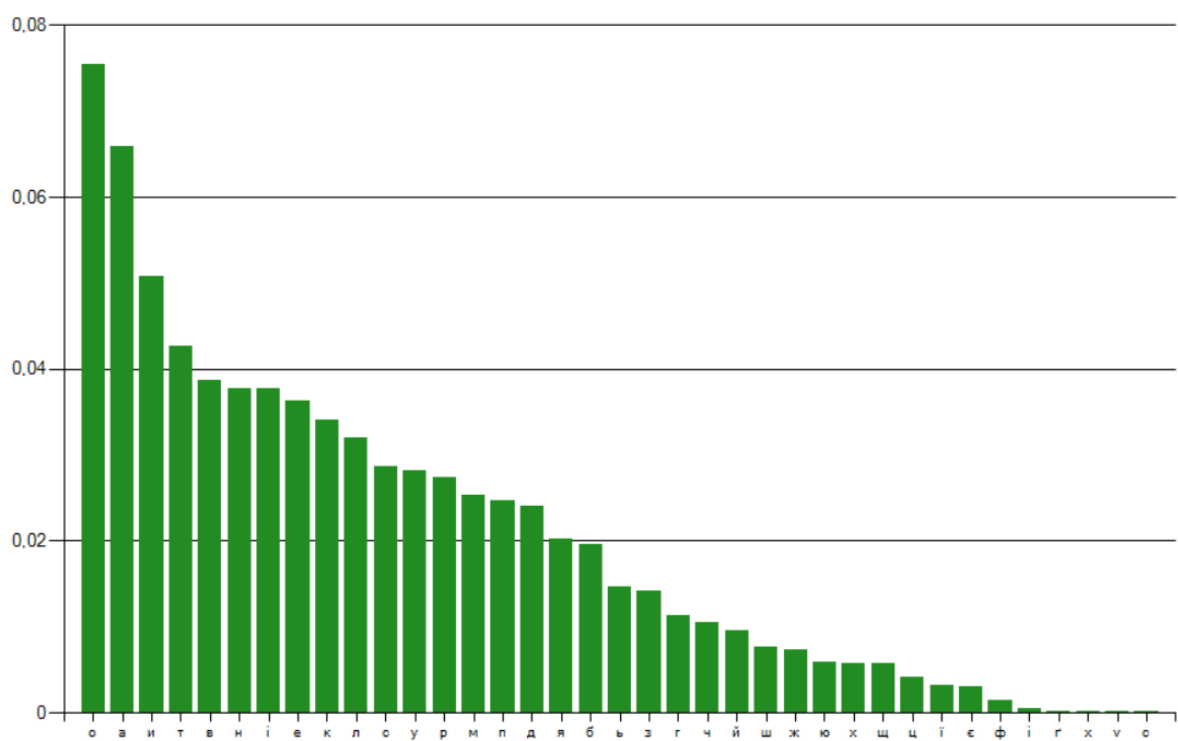


## 2. Моя автобіографія, Остап Вишня

а. діаграму відсортовану в алфавітному порядку



б. діаграму відсортовану по частотам появи літер



с. послідовність літер по мірі спадання частоти появи

Послідовність символів \ Таблица частот

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

## 2. Знаходження відносної частоти появи біграм українського алфавіту

1. Аналогічно до першого завдання, створіть програму яка підрахує відносні частоти появи біграм української мови аналізуючи вміст текстів, що задаються.
2. Для текстів, що були використані в першому завданні проаналізуйте отримані результати та в звіті наведіть:

### 1. Мина Мазайло, М Куліш

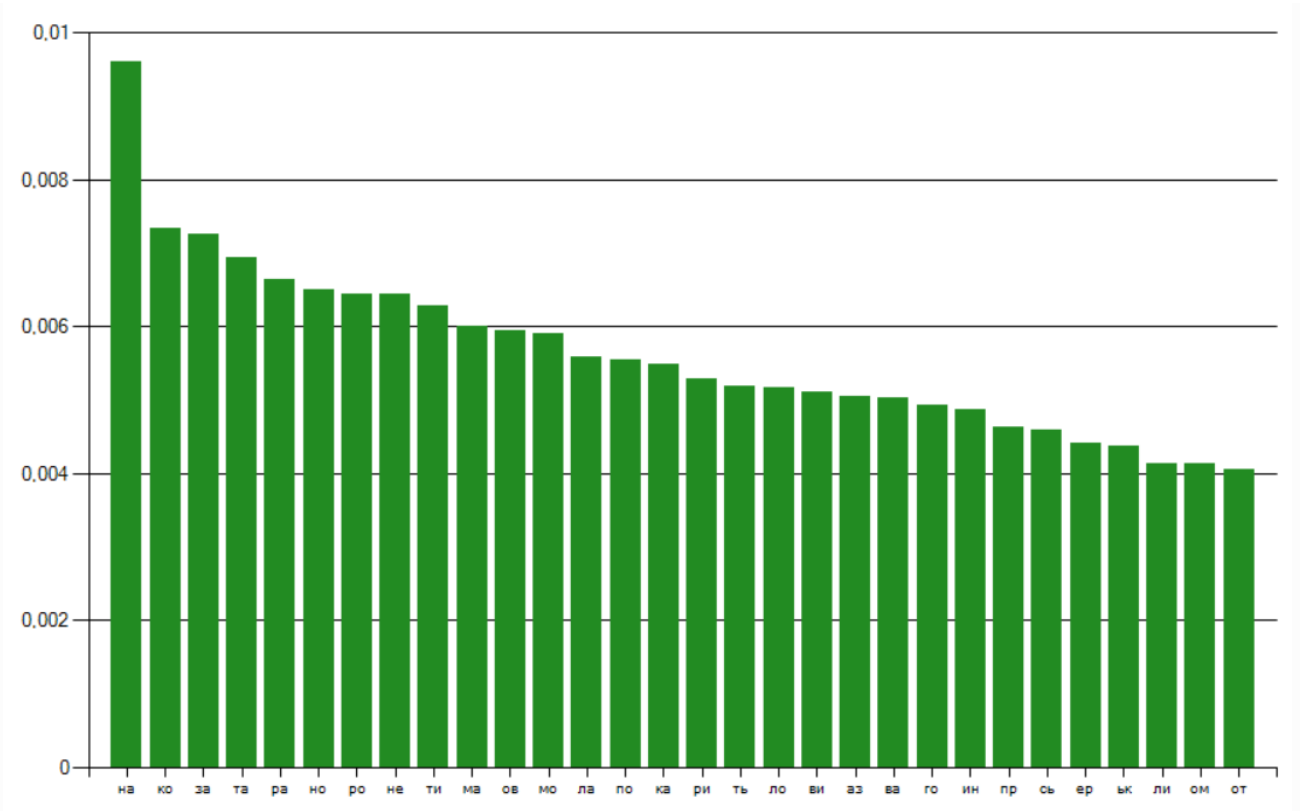
- а. таблицю з відносними частотами біграм, відсортовану за спаданням частоти (використовуйте декілька колонок, щоб результат був максимально компактним)

Послідовність символів \ Таблица частот	
на	0,00961
ко	0,00733
за	0,00726
та	0,00694
ра	0,00664
но	0,00651
ро	0,00644
не	0,00644
ти	0,0063
ма	0,00601
ов	0,00595
мо	0,00592
ла	0,00559
по	0,00555
ка	0,00549
ри	0,0053
ть	0,00518
ло	0,00516
ви	0,00511
аз	0,00505
ва	0,00503
го	0,00494
ин	0,00488
пр	0,00464
сь	0,00459
ер	0,00442
ьк	0,00438
ли	0,00415
ом	0,00414
от	0,00406

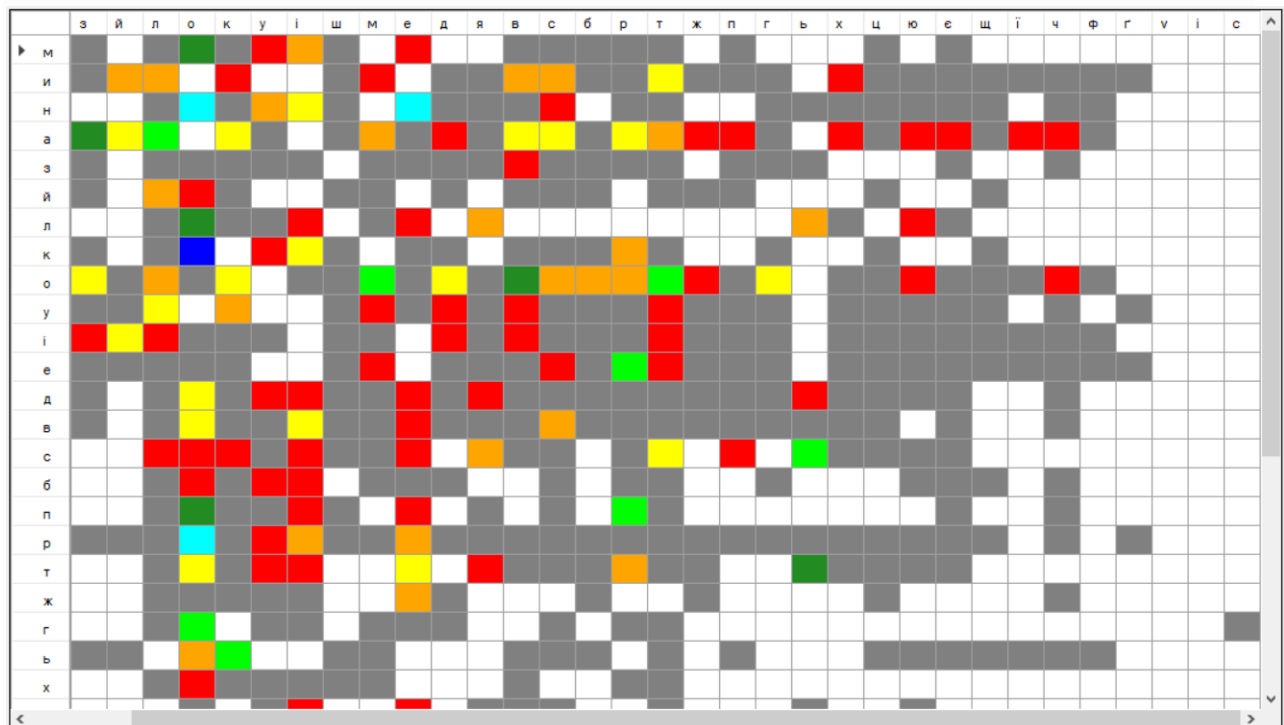
в. послідовність з 30-ти найбільш імовірних біграм

Послідовн
на
ко
за
та
ра
но
ро
не
ти
ма
ов
мо
ла
по
ка
ри
ть
ло
ви
аз
ва
го
ин
пр
сь
ер
ьк
ли
ом
от

с. діаграму відсортовану по частотам появи 30-ти найбільш імовірних біграм



d. матрицю частот появи біграм (імовірність позначена кольором)





## 2. Моя автобіографія, Остап Вишня

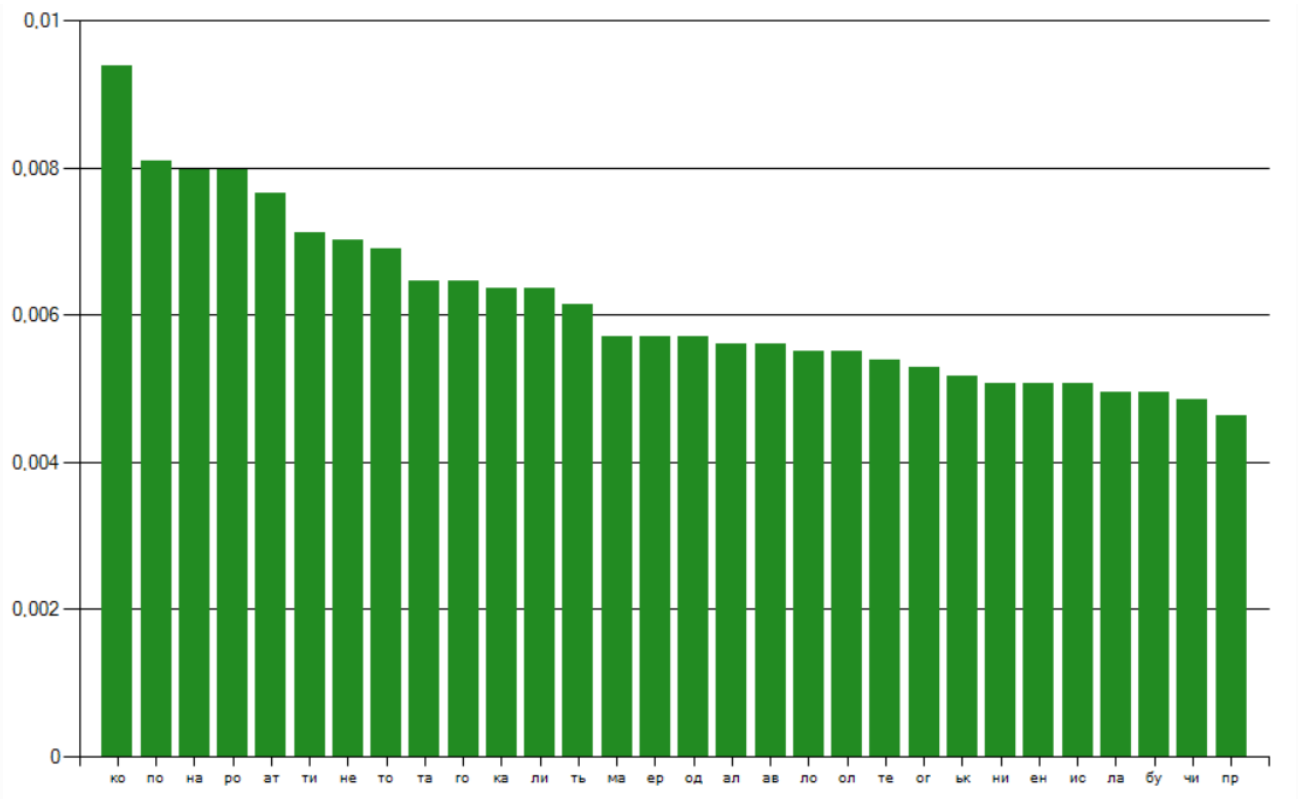
- а. таблицю з відносними частотами біграм, відсортовану за спаданням частоти  
(використовуйте декілька колонок, щоб результат був максимально компактним)

Послідовність символів \ Таблица частот	
ко	0,00939
по	0,00809
на	0,00798
ро	0,00798
ат	0,00766
ти	0,00712
не	0,00701
то	0,00691
та	0,00647
го	0,00647
ка	0,00637
ли	0,00637
ть	0,00615
ма	0,00572
ер	0,00572
од	0,00572
ал	0,00561
ав	0,00561
ло	0,0055
ол	0,0055
те	0,00539
ог	0,00529
ьк	0,00518
ни	0,00507
ен	0,00507
ис	0,00507
ла	0,00496
бу	0,00496
чи	0,00486
пр	0,00464

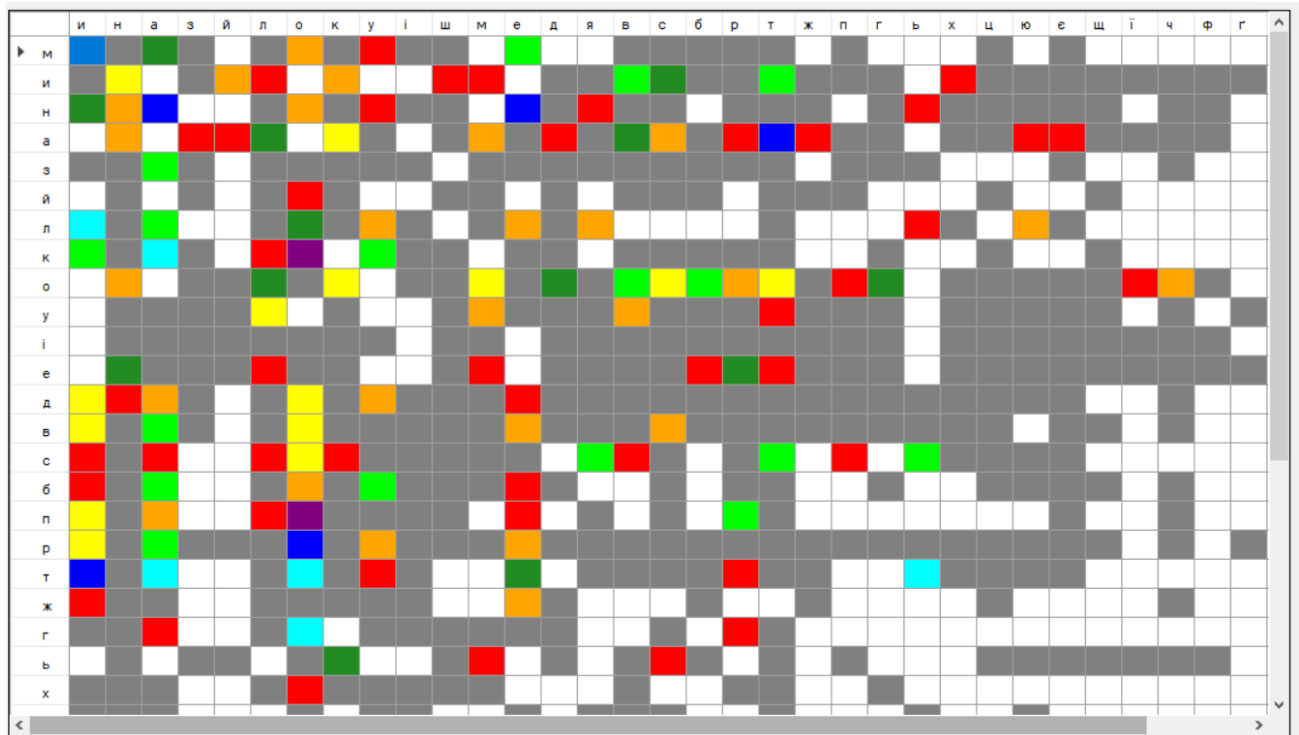
- б. послідовність з 30-ти найбільш імовірних біграм

Послідов	
ко	
по	
на	
ро	
ат	
ти	
не	
то	
та	
го	
ка	
ли	
ть	
ма	
ер	
од	
ал	
ав	
ло	
ол	
те	
ог	
ьк	
ни	
ен	
ис	
ла	
бу	
чи	
пр	

с.    діаграму відсортовану по частотам появи 30-ти найбільш імовірних біграм



d. матрицю частот появи біграм (імовірність позначена кольором)



### 3. Знаходження відносної частоти появи триграм українського алфавіту

1. Аналогічно до першого завдання, створіть програму яка підрахує відносні частоти появи триграм української мови аналізуючи вміст текстів, що задаються.
2. Для текстів, що були використані в першому завданні проаналізуйте отримані результати та в звіті наведіть:

#### 1. Мина Мазайло, М Куліш

- a. таблицю з відносними частотами триграм, відсортовану за спаданням частоти

Послідовність символів \ Таблица частот

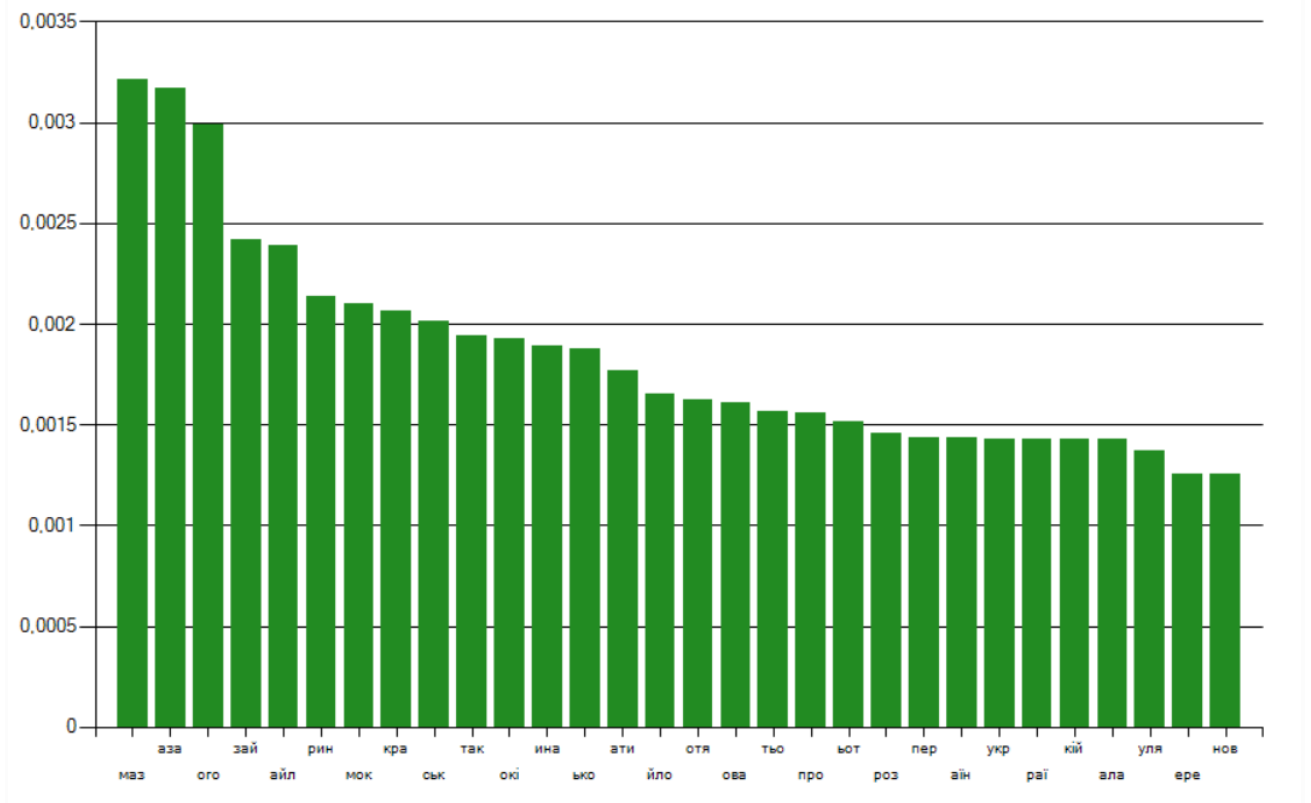
маз	0,00321
аза	0,00317
ого	0,00299
зай	0,00242
айл	0,00239
рин	0,00214
мок	0,0021
кра	0,00206
ськ	0,00202
так	0,00194
окі	0,00193
ина	0,00189
ько	0,00187
ати	0,00177
йло	0,00165
отя	0,00163
ова	0,00161
тьо	0,00157
про	0,00156
ьот	0,00151
роз	0,00146
пер	0,00144
аін	0,00144
укр	0,00143
раі	0,00143
кій	0,00143
ала	0,00143
уля	0,00137
ере	0,00126
нов	0,00126

в. послідовність з 30-ти найбільш імовірних триграм

Послідо

маз  
аза  
ого  
зай  
айл  
рин  
мок  
кра  
ськ  
так  
окі  
ина  
ько  
ати  
йло  
отя  
ова  
тьо  
про  
ьот  
роз  
пер  
аін  
укр  
раі  
кій  
ала  
уля  
ере  
нов

с. діаграму відсортовану по частотам появи 30-ти найбільш імовірних триграм



## 2. Моя автобіографія, Остап Вишня

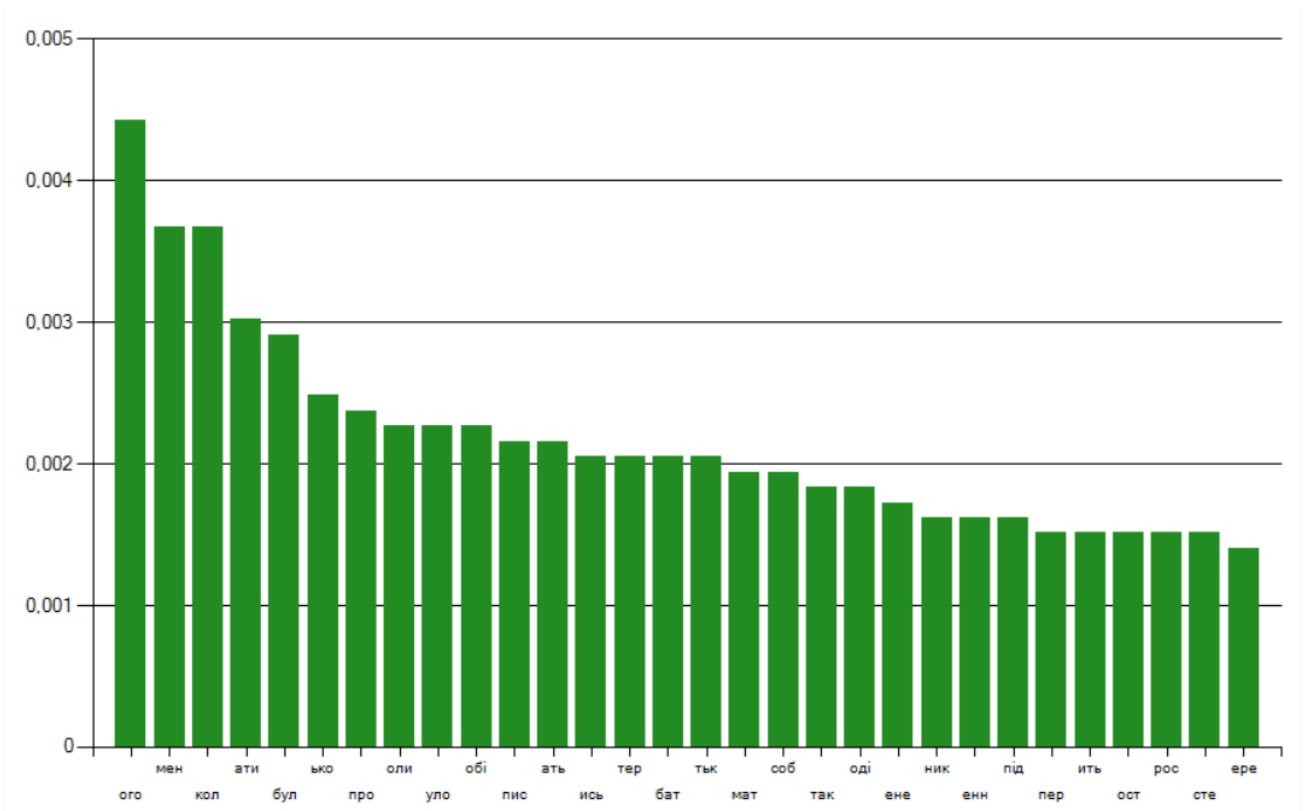
а. таблицю з відносними частотами триграм, відсортовану за спаданням частоти

Послідовність символів \ Таблица частот	
ого	0.00442
мен	0.00367
кол	0.00367
ати	0.00302
бул	0.00291
ько	0.00248
про	0.00237
оли	0.00227
уло	0.00227
обі	0.00227
пис	0.00216
ать	0.00216
ись	0.00205
тер	0.00205
бат	0.00205
тьк	0.00205
мат	0.00194
соб	0.00194
так	0.00183
оді	0.00183
ене	0.00173
ник	0.00162
енн	0.00162
під	0.00162
пер	0.00151
ить	0.00151
ост	0.00151
рос	0.00151
сте	0.00151
ере	0.0014

б. послідовність з 30-ти найбільш імовірних триграм



с. діаграму відсортовану по частотам появи 30-ти найбільш імовірних триграм



## 4. Криптоаналіз

1. Зашифруйте за допомогою *афінного шифру* (а та b оберіть випадковим чином) 3 фрагменти тексту та надайте отримані шифротексти колегам.

Текст №1:

[illegible]

Текст №2:

[illegible]



### Текст №3:

Тип операції:

☒ Шифрування

A:

7

B:

5

☐ Дешифрування

Часто вживана літера №1:

Часто вживана літера №2:

Виконати операцію

Пошук текст

Початковий текст:

Лівий, нижчий берег займала широка, рівна, вільна оболонь. Зелена таця, злегка пережилена до ріки. Раптово земля підносилася вгору. Наче кротовина, висаджена якимсь велетнем-кротом. Декілька городів, а за ними перші підміські вулички.

Гвоздики, георгіни, айстри, бруд, спокій і сентименталізм.

О меланхоліє малого містечка! О тишо синього неба, завішена над одноманітним щастям причинлених, згорблених домів! О солодко-млявий запах провинції!

Тут серця б'ють повільніше, й, мабуть, повільніше ходять годинники. Тут усе спізнається. Тут день є заспаний і ніч невиспана.[2] Тут сонце сонно позикає й під тихий вечір лічливо заплющує червоне око. Тут місяць забуває адресу останньої стації своєї подорожі і не знає, кудюо заранна вернути.[3] Вже світає, а він ще нудить світом над обрієм. Тут сон пахне сіном і стіни пахнуть сном. Тут весна є чутлива, як старосвітський роман, а дощиста оснь осоружна, мов дошкульний біль зубів. Тут вечір має колір какао.[4] а світанок барву двочих уст. Тут життя є банальне. Як рима: сонце – віконце. Птара, русяві коси, пропам'ятні віршики для девчат, Овідій, м'ятові цукерки, стара бритва, латані панчохи, вельшмерц, вишневий лікер, бажання направити світ, кминова юшка, широкі парасолі, мрійливі очі, кальоші, геніальні люди до маловажних справ, порошок на біль голови, приперчені дотепи, недіальні товариства, минулорічні журнали мод, злегка соціалізм, фіалка в петельці, світличні пристаркуватих фільмових авторок, кишені огрірки, огірчення на долю, огарки недокурених папіросок, осінній настрій, нокторн Шопена на розстроеному піаніно

Кінцевий текст:

омшешяодікйкцдозеізідйгдлмчйзімвнізізгізгчдікпкеікрпок  
йкклізіпміїзшмгйдеімеяйбгіяізігеізігдоццдегбеяпкйлхфібіме  
кйкгзеізізізілпзінігмдідобізізокеяйшхбейявймрчнпаязімг  
деізідхкпезххйквбцдхяояпшдрмояпшдрчдхбідпквідйпхмьей  
зхмхкібччохддйдрмрмйяяіймфпізпддшзрчмпяезпрзхсхкіяу  
угйбчдрчймшгдвдзхдхдрюпгмінгдлбвімпллйяехзозідоххдозія  
мьоддейзікйкшхкгдомоххоізігч'яфомізіхфкшддлмрг'япмоїб  
рімрячхооозрпкйіеймппкйчмідемевкохмшзізрчдйзілргкіншмвчп  
кйіезлгдоххфргчдлдізеймг'бпдчмхкікчзвгозлєяібпддчзлмейях  
звіпмвпбгзіквгякшхомятмібг'зоаяйзілбрпчзцідешбпдізбізр  
гкінхзлпмеіаогпмпейзедікхягббгдізххдмчмехзлвдгбіцкохол  
гдоохзлюібгбчкйізідлмвеххдхдхкєячяпшххкргіябіяярпнмш  
хяяхедг'япкмппмшчхавомоххлеяояйікохмоіяізішмймг'зогдрней  
кепмшдхобеймрчрчдйкш'яцш'яйдюгдрнеягмхдведобхдхдепр  
кейяозхдооххоог'зєрнгкемчмерпмеяяурчзлмуомрчбіцзп'язе  
пдікйкззоаяізеяедлдізпяхзр'яімеяіхмг'ячміяяркйкезхмвдсо  
хяркклпдчдізвмцкохзгдібгхдхцкпккчхорнг'зєкйрчюнія'яця  
єяімг'хоєяяпзхдвдг'яедхлпкоімімгдемєоярчоячйялзєйзрчд  
дідл'явцхмчп'яйзідзізіхбейюооярчмрбгдідмрчяпеземехярз  
ізрюпцйбвп'яєзхдхкоояв'язрчмо'яцбчхмібгзіяімеєейчз'яцр  
емйдіфєйр'ябіг'ячміяв'фчмеяпг'ядрчяп'бімппдшдлдізрчбдіяр  
юіміяфемх'яфхдхкозізхзпедінхмбшххопмізікшгмрххопмібінг  
зліфедізейзргдізрч'язідізг'ік'ячмізіяв'мшдізрюояп'яйв'орзедххд  
лмчкйейз'яймі'яй'ягхдх'япєяпкйчдіз'єюізізпхдєцхкчміярчдй  
злехмпц'ярчміг'яйфшмрхмшхмвбізп'іхмхмппмчк'юяпечдлг'зі  
дозрчмббдп'юярчб'обдхзш'япч'яєейялхбхдд'ег'з'єдйгкпч  
єз'єюізп'єйбшшюьпзч'йдл'єц'ярчдм'мхзг'язіхк'яізш'яйхміз  
емеп'єзхдл'йк'чмейз'ц'яйііхх'яц'як'ік'чдхдхцкдд'едрг'ц'ізіхд'йк  
оквхмоєбімемечоцдйкоп'явпмрчрч'ядхдхк'вдєяп'як'япдш'я  
одьшхххдл'зрх'ярг'їземізнювдоохх'єй'яг'юм'ячрчм'єйкувєзіз  
югхд'ьрг'яймг'к'вхзгдізп'ік'п'рчм'едік'яяг'юлб'омрчм'єєзлп  
ц'їз'яв'емінц'яф'рчйбгблпмрчк'іхх'япмлод'їзг'бізпд'єяпч'мі  
згдр'ючмінг'зп'мєа'омхшбіг'яч'ізіпзл'яірмх'о'япєбшк'едік'г'з'єйзі

- Отримайте 3 шифротексти від колег, виконайте їх криптоаналіз та проілюструйте в звіті процес його проведення за допомогою частотного аналізу.

### Тест №1:

Тип операції:

☐ Шифрування

A:

7

B:

2

☒ Дешифрування

Часто вживана літера №1:

o

Часто вживана літера №2:

i

Виконати операцію

Пошук текст

Початковий текст:

мйцтечймшжспцштїгпоіешфйжшїмігшзшсїміявгшнфюгеаншч  
мшжсмітхемімїтеїтмлітїгшїгшїмкбїрфшфїоїетфїсїлїтїгшїмк  
бїлюїевтїмїмкблїоафїшмєтїтїувїлфрєхвфшмтїтїагїштн  
кбїїлїнжшмунїзшцшїуєїтїаїжшїтїтїмїлїтїтїеїхемїшшїмїг'феншж  
йфїшшгемншфєвхмшшївїшїгшмемншїшмжвїтїчнейчїгїшшф  
юфкшшлїтїїоїнгїаефк'фюгегїгїїонжшмшлнежк'теллшмфїюзш  
слїжїшчтеяміяшзшслїжїшчтеявгшчюфкшзшїшшжшнїгїоцїофку  
шжшнїгїоцїофенбїйякшбїюїтїївфїлюхепкбїгшмїнфвфїїлї  
нтєпкбїгшлїнпшїтїгшжїчнжїйїгїфїжкбемїтїш'ягвнфкшїєїтїоф  
кнжкш'яенюлїєвншнєжшїг'єфкшнч'явм'яшфшїтїг'євфкшзш  
мвфкїшчнєтлшбтїшзїбївїеюгшжшмзшїшфк'зїмїцтечїмшжспч  
штїгїоїтїюбїгблбмшжшїтїонфюгегфїк'йгшлїгшжштїгвнфкшгюф  
фєнфкїгїшїш'ямвжєжн'к'яжкїюрн'г'єтєжшз'яувнєжлшмшцїтїшїагемгшс  
йфшнєжвєшзмізїюрн'г'єтєжшз'яувнєжлшмшцїтїшїагемгшс  
єечшїтїк

Кінцевий текст:

вічнийреволюціонердхшотїлорєдобоюреззопустшастяюволю  
вінживєвінєнєвмерніпопієскїїтортурінїпоремнісарськїмурианів  
їськамштрованїгїарматїлаштованїнішїонськеремєслєгрібїю  
гошенєзвєловїнєвмернінєжївєхочївїєшїаїтїродїєсїтаажвч  
рарозповїєсїяювласнїсїлїїдеїпростуєтьєсїаїмїнїєшїшїтїудїдєд  
нєслєвомсїлїнїмвотрубїомїлїонїзвєзсїомїлїонїлїонїрадоїдуть  
босєголєсдхачутьголєсдхачутїскрізьпокурїноххатахмужїчкїхп  
овєрстатхрємїснїхїхїпмїсїячнєдолїїслїзїдєтїлькївїнроздєст  
сїяєзнуьслєозїсумнєшастєсїлародїтєсїязавзятїтїярїдєтє  
добуватьхочїсїнаїмєнесобїрашудолєвоборьтєбїчїнїреволюцїю  
нєрдхнзукадумк'аволєнєуступїтьпїтїмїполїнадєстєспутатїсьтє  
пєррозвалїлєсьзларуїнапокотїлєсїявїнаїдєвсїїтїасїлашоб  
бїгїуїспїнїлашобзгїсїламогєнєррозвїднїяшїєсїядєнь



Тест №2:

Тип операції:

☐ Шифрування

A:

B:

☒ Дешифрування

Часто вживана літера №1:

Часто вживана літера №2:

Пошук текст

Початковий текст:

```

ейтмлїіжвюбйвертейнжмвймцжшеївлзівфшлщжщгйрюнлїїтір
кншгщжшяуетюфктваймщійхткбешибщнвтнцтпйявгтвоолїлез
ивффшонмщрчнфщицтпїгоаючфйжщлєгщжщхелавтаюммцщщц
пїойгщбхілещлєзівфшбщїявпкбщуищпогїоаючфйжщлєгщжщх
елавтаюммцщцпїойгщбхілещлєзівфшбщїявпкбщуищїю

```

Кінцевий текст:

```

шеневмерлаукраїниславаіолясенамбряттямолодіусміхеньсяд
олязгинутьнашіворіженькиаросанасонцізапануемімибраттвусво
їйсторонідшууйтіломиположимзанашувободупокажемщомибра
ттякозацькогородудушуйтіломиположимзанашувободупокажем
щомибраттякозацькогороду

```

Виконати операцію

Тест №3:

Тип операції:

☐ Шифрування

☒ Дешифрування

А:

В:

Пошук текст

Початковий текст:

Кінцевий текст:

Часто вживана літера N=1:

Часто вживана літера N=2:

Виконати операцію

## Висновки

У ході лабораторної роботи було досліджено вірогіднісні параметри появи літер української мови для різних типів текстів, а саме проведено аналіз найбільш імовірних літер, біграм та триграм для використання в частотному криптоаналізі. Продемонстровані навички моделювання діаграм, таблиць та матриць (за кольорами залежності частот) на основі отриманих даних. Також було зашифровано 3 фрагменти текстів та отримано 3 тексти від колег для розшифрування за допомогою афінного шифру.

Посилання на програму:

<https://github.com/VladyslavHolubtsov/ZIKS>