**Київський національний університет імені Тараса Шевченка**

**Факультет радіофізики, електроніки та комп’ютерних систем**

**Лабораторна робота № 4**

на тему «Дослідження частотних характеристик української мови»

з предмету “Захист інформації в комп’ютерних системах”

Роботу виконав

студент 4 курсу

напряму підготовки «Комп’ютерна інженерія»

Ліщенко Юрій Вікторович

Київ 2024

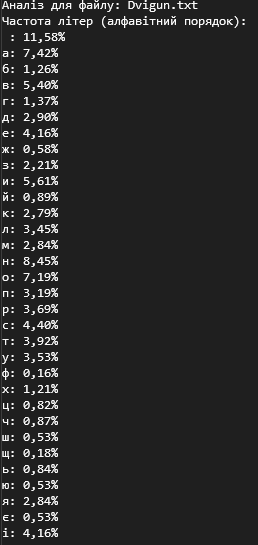
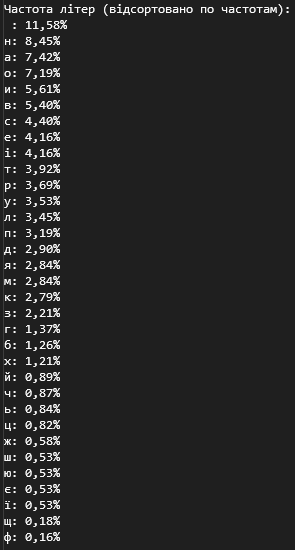
**Мета:** Дослідити вірогіднісні параметри появи літер української мови для різних типів текстів. Аналіз найбільш імовірних літер, біграм та триграм для використання в частотному криптоаналізі.

**Хід роботи**

**Посилання на репозиторій https://github.com/YuriiLishchenko1/ZICS**

**1. Знаходження відносної частоти появи літер українського алфавіту**

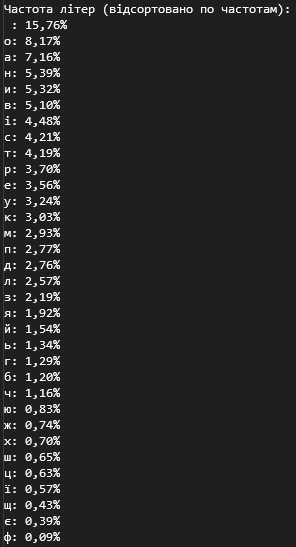
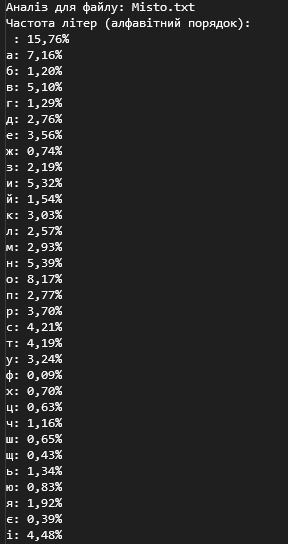
1. Створіть програму (будь-якою зручною для вас мовою), яка в якості аргументів приймає перелік текстових файлів, та аналізуючи їх вміст обраховує частоти появи літер українського алфавіту. Для спрощення аналізу з тексту виключаються всі знаки пунктуації окрім “пробілу”.
2. Для різних достатньо довгих 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. діаграму відсортовану в алфавітному порядку
   2. діаграму відсортовану по частотам появи літер
   3. послідовність літер по мірі спадання частоти появи

  
Для файлу Dvigun.txt:

Послідовність літер по мірі спадання частоти:

наоивсеітрулпдямкзгбхйчьцжшюєїщф

Відповідно до даних бачимо що літера «Н» найчастіше з’являється у тексті, а літера «ф» найрідше (9,56% та 0,18% відповідно).

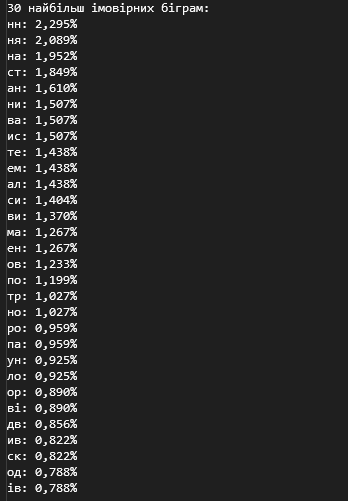
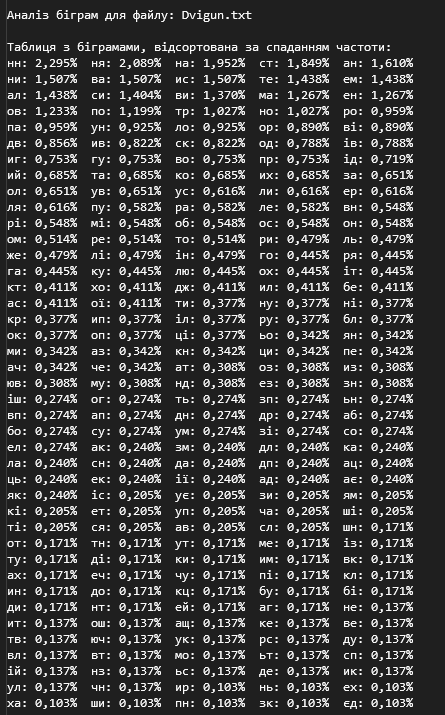
Для файлу Misto.txt  


Послідовність літер по мірі спадання частоти:

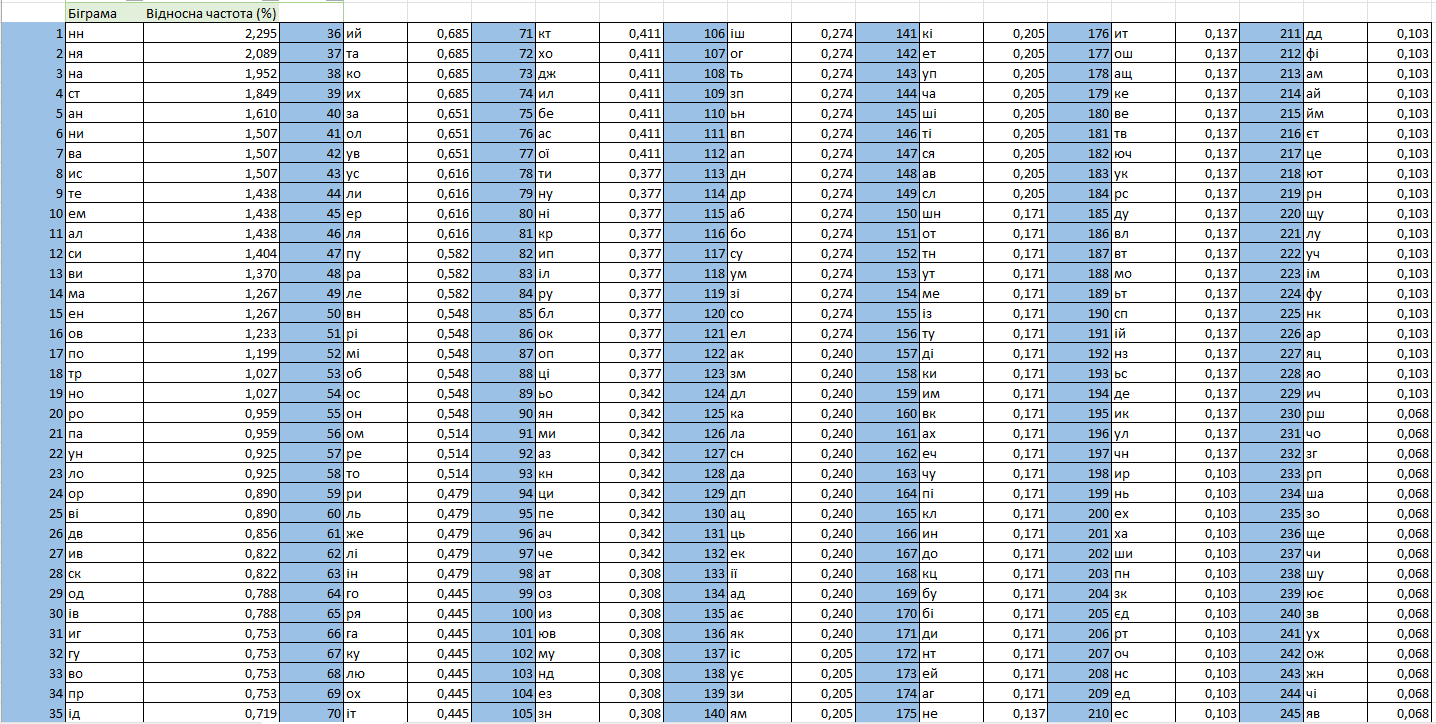
оанивістреукмпдлзяйьгбчюжхшцїщєф

**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 результаті та в звіті наведіть:
   1. таблицю з відносними частотами біграм, відсортовану за спаданням частоти (використовуйте декілька колонок, щоб результат був максимально компактним)
   2. послідовність з 30-ти найбільш імовірних біграм
   3. діаграму відсортовану по частотам появи 30-ти найбільш імовірних біграм
   4. матрицю частот появи біграм (імовірність позначена кольором)

Dvigun.txt:

Таблиця

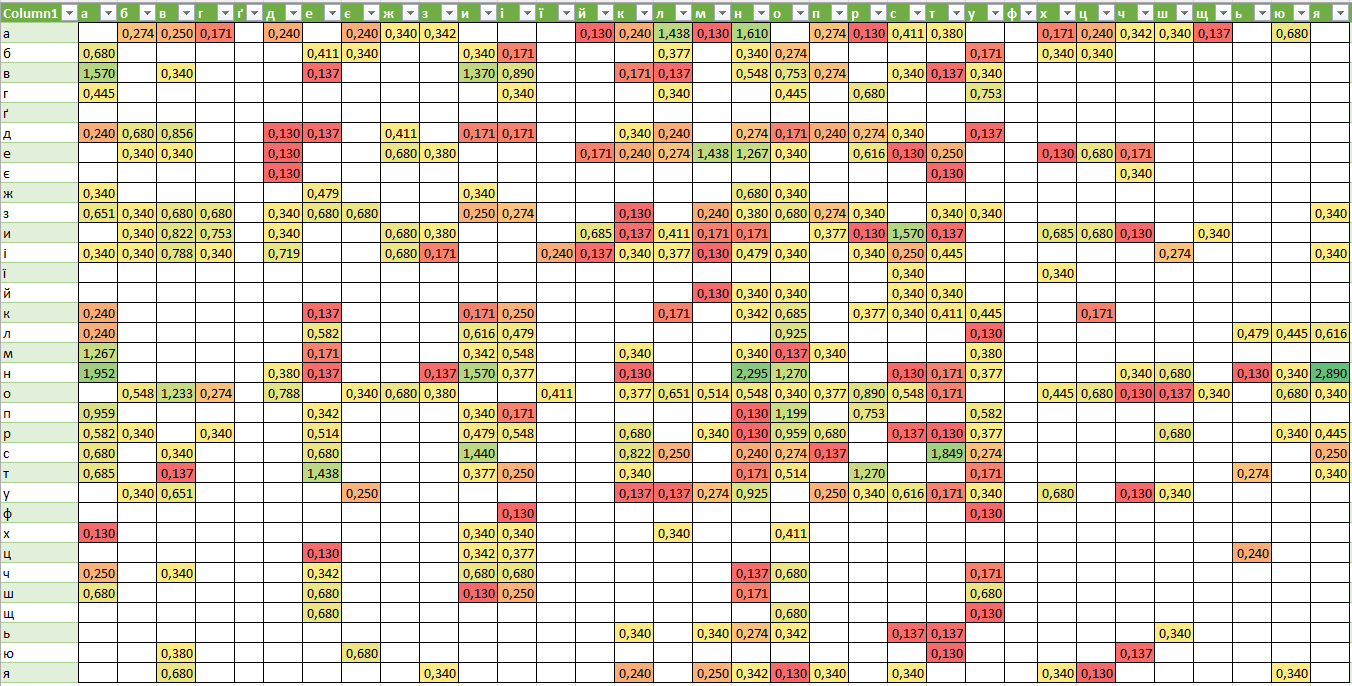


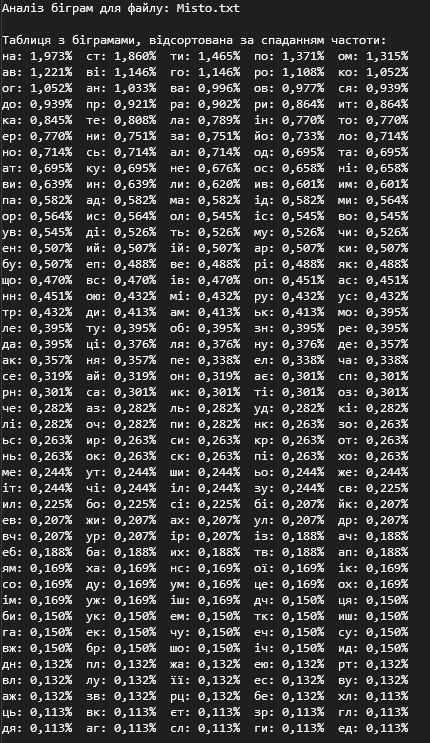
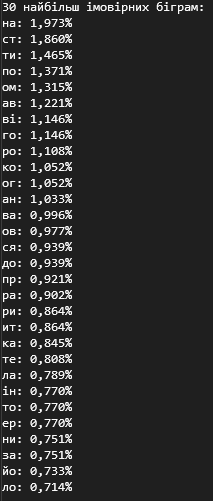
Загальна кількість біграм 343.

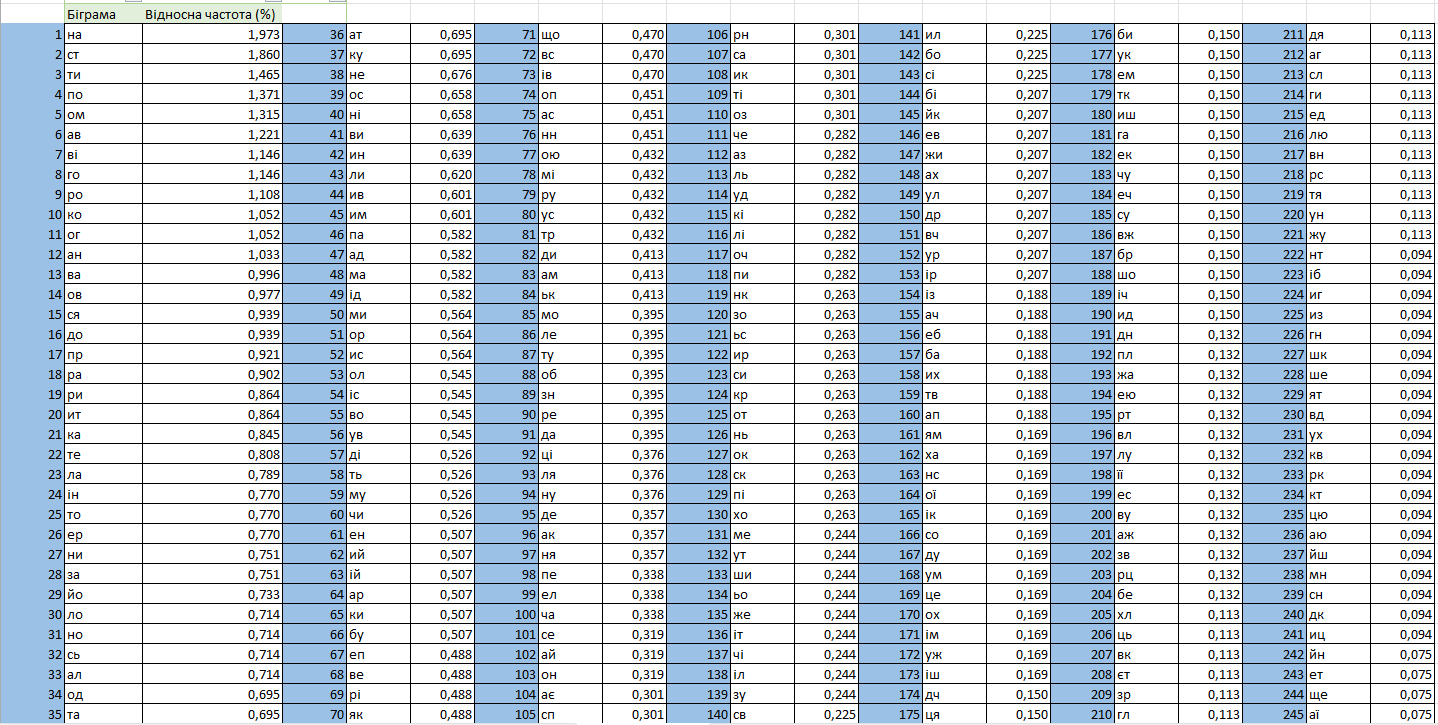
30 найбільш імовірних біграм:

нн ня на ст ан ни ва ис те ем ал си ви ма ен ов по тр но ро па ун ло ор ві дв ив ск од ів

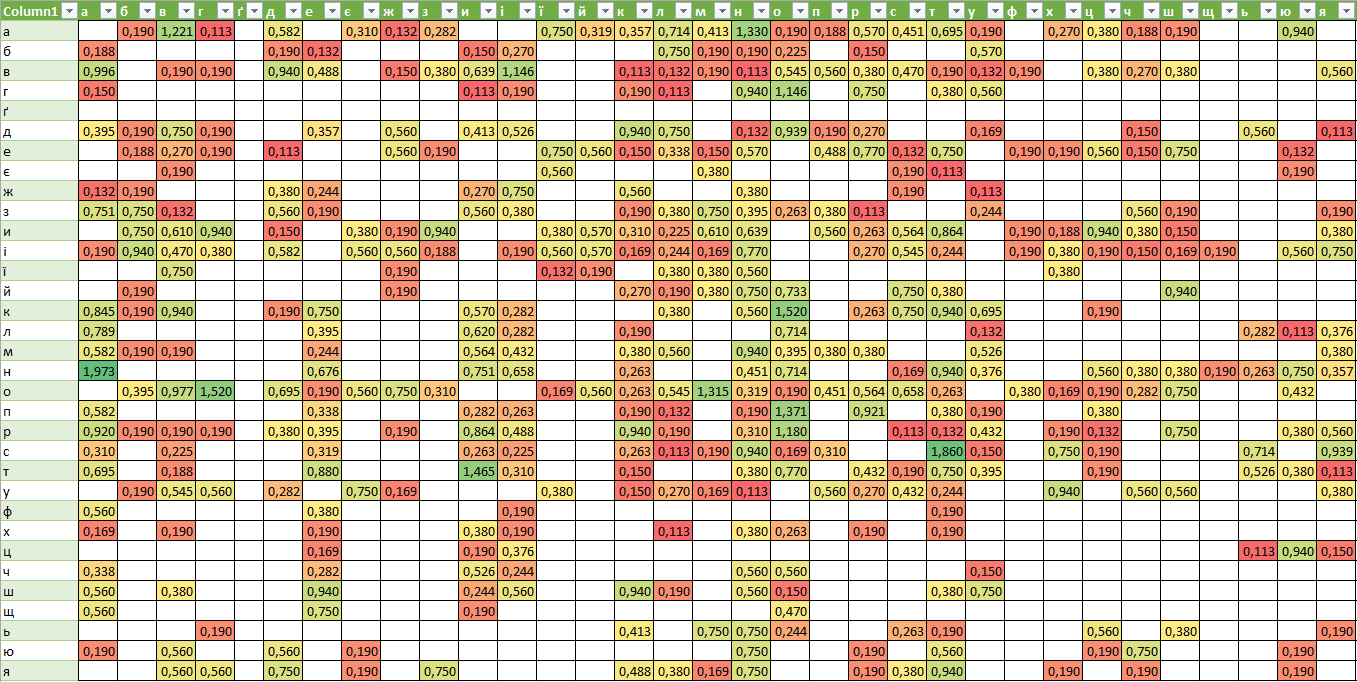
Діаграма:

Матриця частот появи біграм: (якщо клітинка пуста – частота = 0)   


Для файлу Misto.txt:

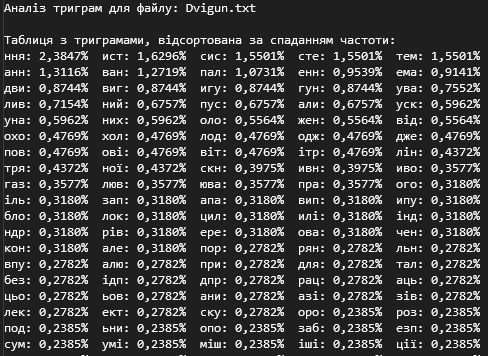
Таблиця:  


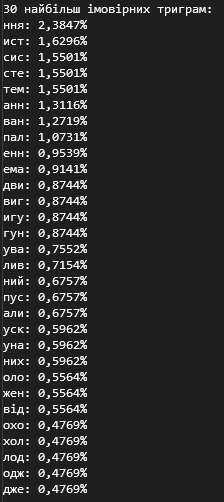
Загальна кілкість біграм 453.  
  
Діаграма:

Матриця частот появи біграм: (якщо клітинка пуста – частота = 0)   


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

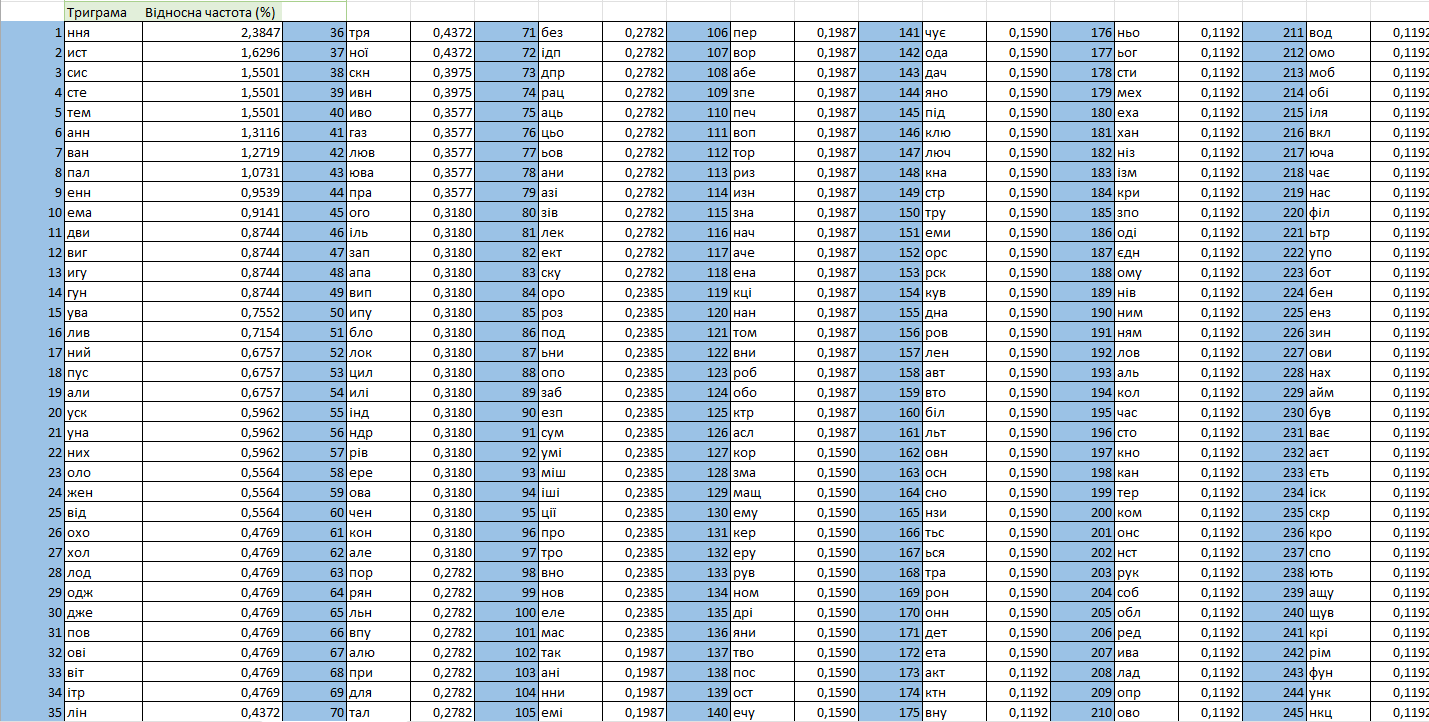
1. Аналогічно до першого завдання, створіть програму яка підрахує відносні частоти появи триграм української мови аналізуючи вміст текстів, що задаються.
2. Для тексті, що були використані в першому завданні проаналізуйте отримані результаті та в звіті наведіть:
   1. таблицю з відносними частотами триграм, відсортовану за спаданням частоти
   2. послідовність з 30-ти найбільш імовірних триграм
   3. діаграму відсортовану по частотам появи 30-ти найбільш імовірних триграм

Для Dvigun.txt:

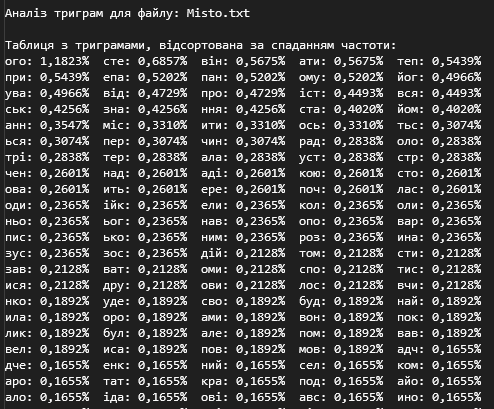


30 найбільш імовірних триграм:

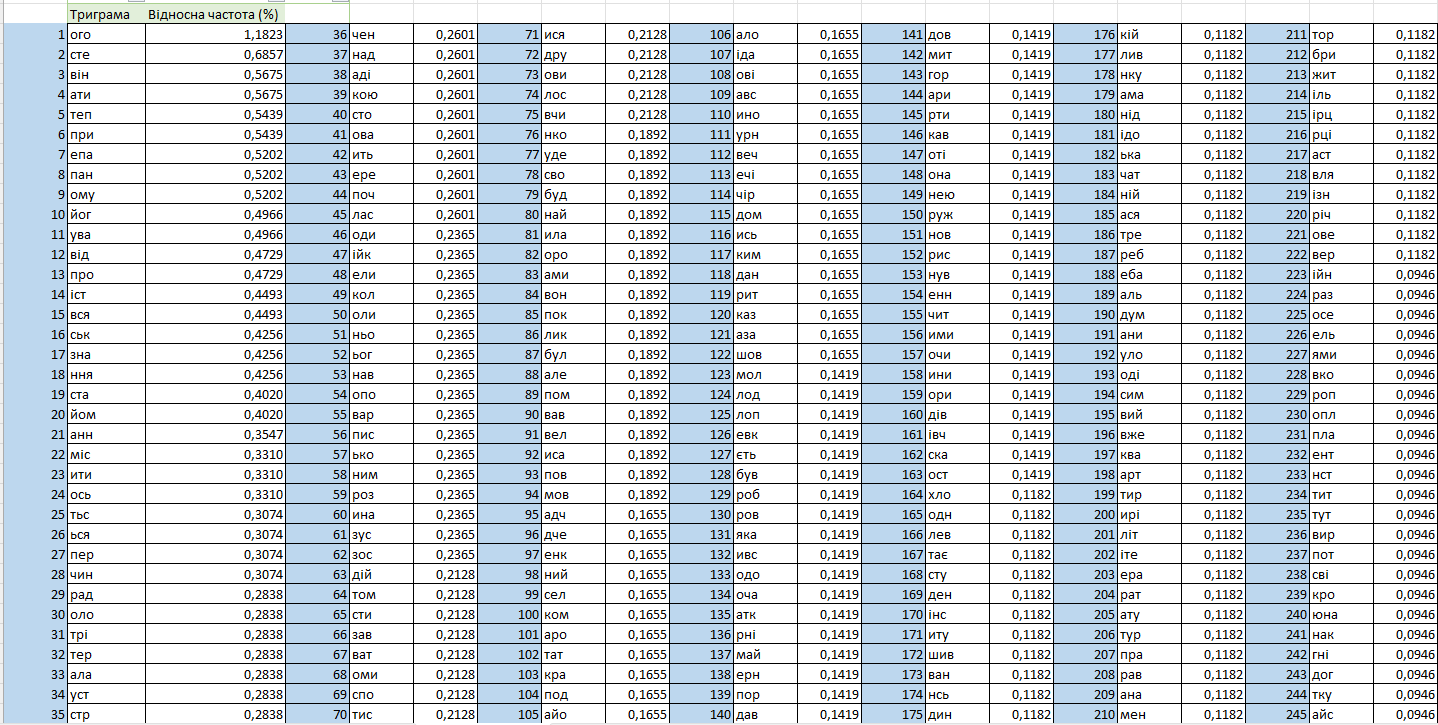
ння ист сис сте тем анн ван пал енн ема дви виг игу гун ува лив ний пус али уск уна них оло жен від охо хол лод одж дже



Діаграма

Для Misto.txt

30 найбільш імовірних триграм:

  
  
30 найбільш імовірних триграм:  
ого сте він ати теп при епа пан ому йог ува від про іст вся ськ зна ння ста йом анн міс ити ось тьс ься пер чин рад оло

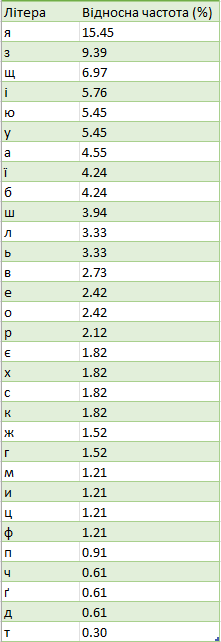
Діаграма:

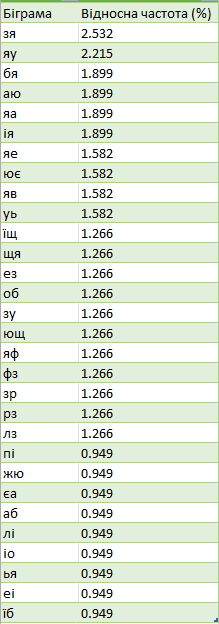
**4. Криптоаналіз**

Зашфрували фрагменти тексту: а = 33 b = 21

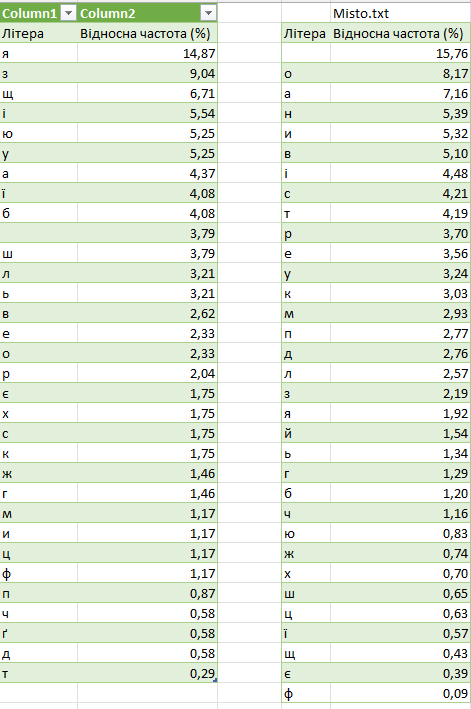
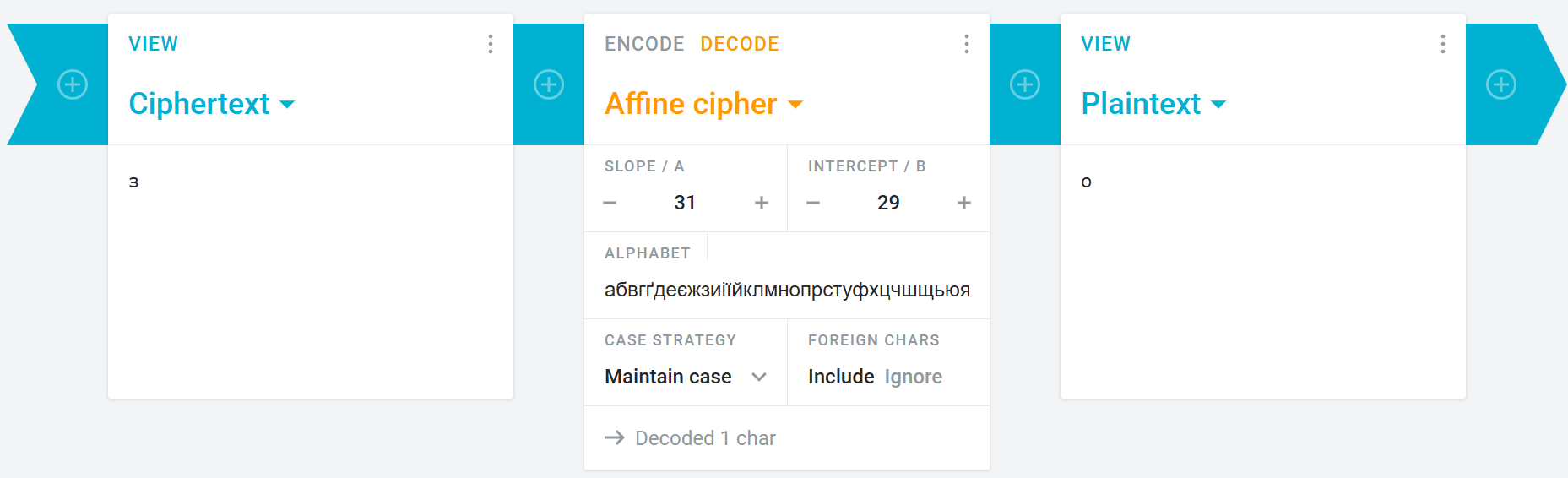
Вбіьгмі хТїпиа єсТвбгТплеієіжТогґгбсбТїсТїрибєяТгвгпимсґх.ТЩлТмяйлТ ичі хТА лвсґс.ТСТґлпмгпїиТпиґТїяа бишскТапгогТїлдеуєсТРгбіас.Т гжТвгмиеіпауТбсмиа фТ–ТпиґТгмбяйяк хау.ТЇсТгвіагдТдсжря ґхгзТмбяйіґіТ гпсбічс,ТА лвсґТпвиїґскТҐсмижєя.ТПимТ гогТжгдяТшгдяахТа скТґлвбікдґгТжТвбієбг:Т«ОімєідТїегшіґгдТяупеуегауТжгдяТгрлбґя іТресєі ґггєяТҐсмижєяТпТєяьгпсбєя,Твбірібсехґіщф,ТпТгьгбгґщуТвиаґгогТмгрбгря яТдгегмгогТдицсґіґс».

Отриманні шифротексти від колеги:  
АпіїщЯезм їщжюєабЯвЯезбу ЯЛіхьаюзхіобЯшЯаубюсзузлшЯзкбвь.ЯУьїЯвщеіуїбжЯХщшающ,ЯизЯющсіЯуцгщїїбЯгщвзлЯвЯуіаіозищл ЯезвцщуобюєЯфзрзЯуьмїзчЯюшр .ЯЩоіЯХщшаюЯуьклзуобжюєаб,Явщеіуїбґм ЯЛіхьаюзхіоб,ЯизЯпіЯїіЯу оьсшжЯфзрзЯуькЯдшгц ,ЯбсщЯающїіЯфзрзЯашешюї піґЯкзЯащлзчЯалігюь.ЯКштЯїщлщрщжюєабЯеігісзїщю Ямзозуьсщ,ЯизЯвлздіЯфзрзЯезвцщу ю Яїшкєр ЯьЯающу юєЯшлзуш.

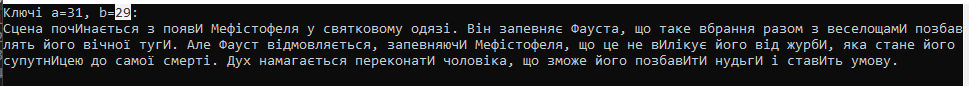




Порівняємо частоту появу літер нашого тексту та шифротексту.

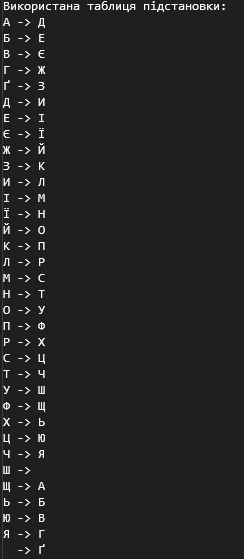
  
Припустимо що я = «пробіл», з = «о»  


Запам’ятаємо ключі.

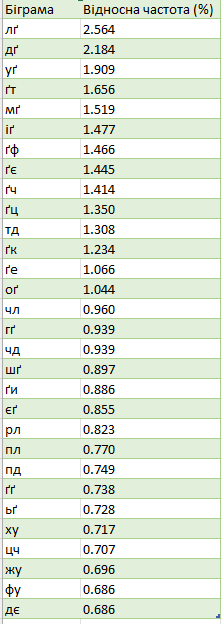
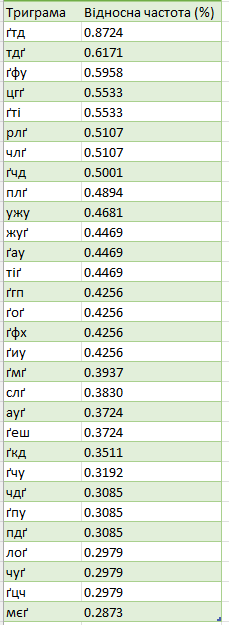
Була зроблена програма, що просто перебирає всі a та b для поданого тексту на вхід.   
  
  
Текст розшифровано.

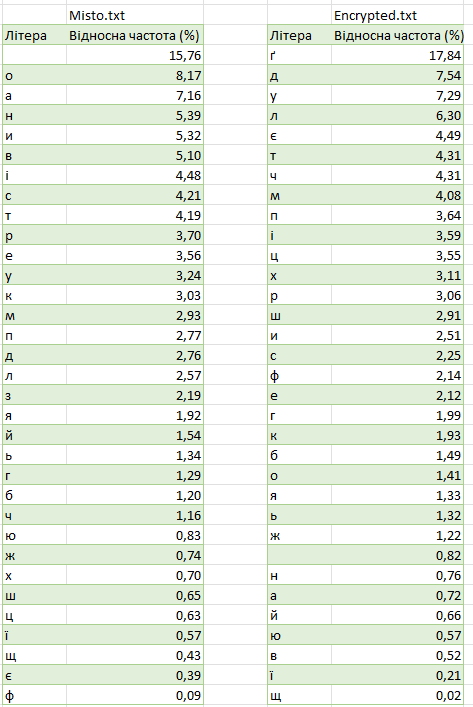
**5. Криптоаналіз довільного моноалфавітного шифру (додаткове завдання на +5 балів)**

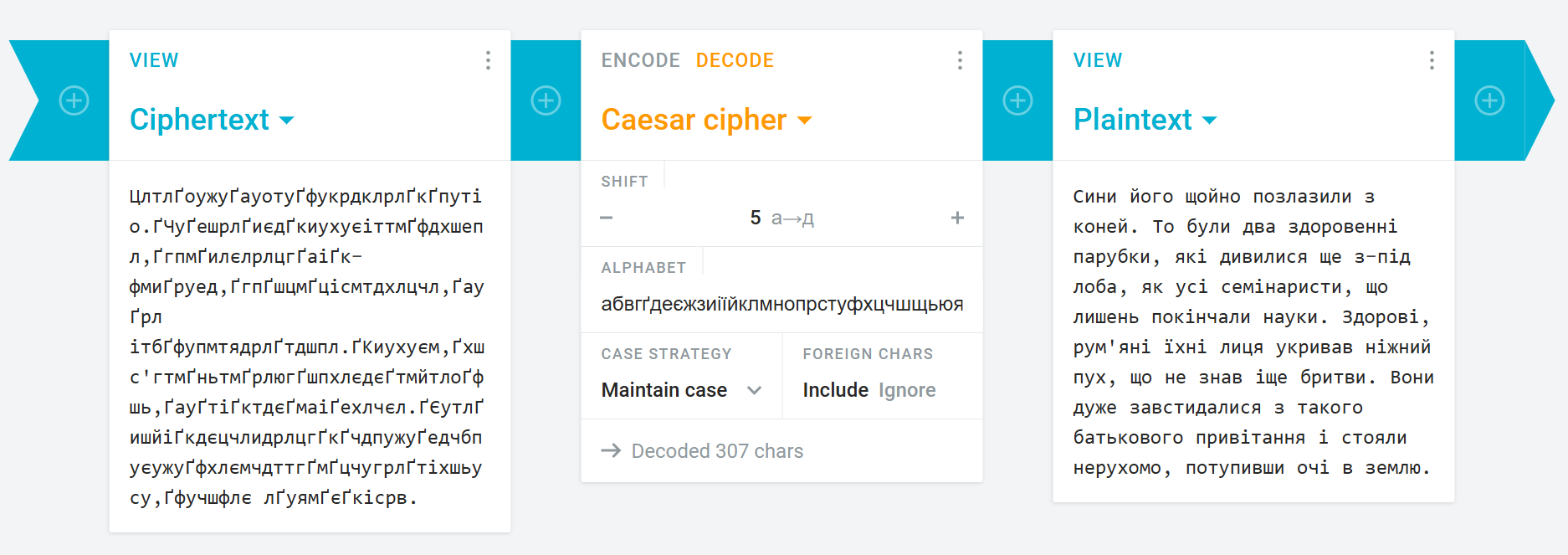
1. Згенеруйте випадкову моноалфавітну підстановку (таблицю відповідності) та зашифруте за її допомогою достатньо довгий фрагмент тексту.
2. Виконайте криптоаналіз отриманого шифротексту, використовуючи частотний аналіз (в т.ч. частоти біграм, триграм та припущень).



Вигляд зашифрованого тексту:  
ЦлтлҐоужуҐауотуҐфукрдклрлҐкҐпутіо.ҐЧуҐешрлҐиєдҐкиухуєіттмҐфдхшепл,ҐгпмҐилєлрлцгҐаіҐк-фмиҐруед,ҐгпҐшцмҐцісмтдхлцчл,ҐауҐрл ітбҐфупмтядрлҐтдшпл.ҐКиухуєм,Ґхшс'гтмҐньтмҐрлюгҐшпхлєдєҐтмйтлоҐфшь,ҐауҐтіҐктдєҐмаіҐехлчєл.ҐЄутлҐишйіҐкдєцчлидрлцгҐкҐчдпужуҐедчбпуєужуҐфхлємчдттгҐмҐцчугрлҐтіхшьусу,Ґфучшфлє лҐуямҐєҐкісрв.

   
  
Співставимо відносні частоти літер до попередніх даних з файлу Misto.txt



Можна помітити що «пробіл» відповідає літері «Ґ»  
Припустимо що пробіл буде 34 елементом у алфавіті.  
Таким чином 34 + 5 = 5 літера = Ґ.  
Відповідно шифр розгадано і якщо ми виконаємо зсув на 5 позицій вправо отримає розшифрований текст.  
Це шифр Цезаря.  
  
Як бачимо припущення було правильним.  
Текст розшифровано

**Висновок:** у процесі виконання лабораторної роботи було проаналізовано ймовірнісні параметри появи літер української абетки в різних українських текстах. Досліджено частотність літер, біграм і триграм. Розглянуто роботу афінного шифру та здійснено його розшифрування за допомогою методів криптоаналізу. Для розробки програмного забезпечення використовувалась мова програмування С#. Було створено моноалфавітний шифр та виконано криптоаналіз шифротексту.