МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ «КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ

імені ІГОРЯ СІКОРСЬКОГО»

Фізико-технічний інститут

КРИПТОГРАФІЯ

КОМП'ЮТЕРНИЙ ПРАКТИКУМ №3

Криптоаналіз афінної біграмної підстановки

Варіант №11

Виконали:

ФБ-32 Пінькас Б. О.

ФБ-32 Драчук О. І.

Київ 2025

Мета роботи

Набуття навичок частотного аналізу на прикладі розкриття моноалфавітної підстановки; опанування прийомами роботи в модулярній арифметиці.

Порядок виконання роботи

1. Реалізувати підпрограми із необхідними математичними операціями: обчисленням оберненого елементу за модулем із використанням розширеного алгоритму Евкліда, розв'язуванням лінійних порівнянь. При розв'язуванні порівнянь потрібно коректно обробляти випадок із декількома розв'язками, повертаючи їх усі.

2. За допомогою програми обчислення частот біграм, яка написана в ході виконання комп'ютерного практикуму №1, знайти 5 найчастіших біграм запропонованого шифртексту (за варіантом).

3. Перебрати можливі варіанти співставлення частих біграм мови та частих біграм шифртексту (розглядаючи пари біграм із п'яти найчастіших). Для кожного співставлення знайти можливі кандидати на ключ *(a,b)* шляхом розв'язання системи (1).

4. Для кожного кандидата на ключ дешифрувати шифротекст. Якщо шифртекст не є змістовним текстом російською мовою, відкинути цього кандидата.

5. Повторювати дії 3-4 доти, доки дешифрований текст не буде змістовним.

Хід роботи

Для знаходження ключа на основі двох з п'яти найчастіших біграм мови та двох з п'яти найчастіших біграм шифротексту, а також для дешифрування шифротексту за знайденими ключами та фільтрування розшифрованих текстів, які не відповідають змістовному тексту російською мовою, було написано скрипт lab3.py. Всі вхідні файли зберігаються в папці input (ШТ з варіанта), а вихідні файли - в папці output (весь вміст формується скриптом lab3.py). Перед безпосереднім запуском скрипта, рекомендується встановити необхідні бібліотеки (pip install -r requirements.txt)

На початку, скрипт фільтрує ШТ від зайвих символів. Всі символи, окрім літер російського алфавіту вилучено, а літери "ё", "ъ" замінено на "е" та "ь" відповідно. (оскільки ШТ вже відфільтрований, це необов'язково, але таким чином ми перестраховуємось від помилок)

Спочатку ми знайшли 5 найчастіших біграм шифротексту, скориставшись кодом програми, написаним у процесі виконання КП №1, фрагменти коду відображені у скрипт lab3.py.

```
ШТ: оквкпкящсройюфчвфбчллфэйлзщоыифуххгьижфбчбжройэжиавкхбоаэлбзьздблфюжвпыхожеуфыхь...

Bigram   Freq
    нк 0.0146
    юж 0.0136
    хб 0.0128
    шь 0.0128
    бй 0.0122
```

Повний текст збережено в файл FreqByBigram.xlsx

Також в методичних вказівках було надано 5 найчастіших біграм у мові:

“ст”, “но”, “то”, “на”, “ен”.

Перевіримо це за допомогою програми з КП №1, додамо в нього фрагмент коду для знаходження топ-7 біграм:

```
179     if n == 2 and spaces == False:
180         print("Біграма Частота")
181         sorted_freqs = sorted(freqs.items(), key=lambda x: x[1], reverse=True)
182         for word, count in sorted_freqs[:7]:
183             print(f"{word:<7} {count:.4f}")
184         print("\n")
```

```
Біграма Частота
то      0.0164
ст      0.0135
на      0.0127
но      0.0123
не      0.0119
ен      0.0110
по      0.0107
```

Дійсно, зазначені 5 найчастіших біграм у мові є в топ-6 біграм за частотою на обраному ВТ довжиною 3 Мб в КП №1

Для знаходження ключа, нам треба обрати 2 біграми з 5 найчастіших в ШТ і 2 біграми з 5 найчастіших в мові, зробити це можна відповідно

$C_5^2 * 2! * C_5^2 * 2! = 400$ способами, враховуючи перестановки в 2 обраних біграмах

$C_5^2 = \frac{5!}{2!*(5-2)!} = 10$

Переберемо всі можливі варіанти вибору біграм для аналізу та для кожного з варіантів розв'яжемо систему порівнянь:

$$\begin{cases} Y^* \equiv aX^* + b \pmod{m^2} \\ Y^{**} \equiv aX^{**} + b \pmod{m^2} \end{cases}, \quad (1)$$

З якої випливає порівняння:

$$Y^* - Y^{**} \equiv a(X^* - X^{**})(mod\, m^2) \quad (2)$$

Що являє собою метод частотного аналізу афінного шифру біграмної підстановки з порівнянням:

$$Y \equiv (aX + b)\, mod\, m^2, \quad (3)$$

де $(x_{2i-1}, x_{2i}) \leftrightarrow X_i = x_{2i-1}m + x_{2i}$, (x_{2i-1}, x_{2i}) - біграма ВТ

Аналогічно $Y_i = (y_{2i-1}, y_{2i}) = y_{2i-1}m + x_{2i}$

Відповідно **a** можемо знайти з порівняння (2) як

$a = (Y^* - Y^{**})^{-1}(X^* - X^{**})(mod\, m^2)$

Порівняння з оберненим розв'язуємо за розширеним алгоритмом Евкліда

b можемо знайти з системи порівнянь (1), підставивши **a** в будь-яке порівняння з системи, наприклад:

$b = (Y^* - aX^*)mod\, m^2$

Так як при обчисленні порівняння з оберненим в нас може бути декілька розв'язків при d>1, при чому цих розв'язків буде рівно d = gcd(a, m^2) , відповідно, виводимо ключі для всіх розв'язків **a** (в нас тут в кожному по одному розв'язку, тому що m^2 = 961 не взаємнопросте з **a** лише в тому випадку, якщо 31 кратне **a**, таких **a** небагато)

```
Ключі (X=ст, но | Y=нк, юж):
1: a =  807,  b =  736

Ключі (X=ст, но | Y=нк, хб):
1: a =  351,  b =  357

Ключі (X=ст, но | Y=нк, шь):
1: a =  275,  b =  454

Ключі (X=ст, но | Y=нк, бй):
1: a =  566,  b =  424

Ключі (X=ст, но | Y=юж, нк):
1: a =  154,  b =  582

Ключі (X=ст, но | Y=юж, хб):
1: a =  505,  b =  526
```

…

```
Ключі (X=ен, на | Y=шь, бй):
1: a =  917,  b =  474

Ключі (X=ен, на | Y=бй, нк):
1: a =  112,  b =  444

Ключі (X=ен, на | Y=бй, юж):
1: a =  842,  b =  812

Ключі (X=ен, на | Y=бй, хб):
1: a =  158,  b =  404

Ключі (X=ен, на | Y=бй, шь):
1: a =   44,  b =  336

Знайдено 178 ключів
```

Розшифруємо ШТ кожним з знайдених ключів. Для розшифрування використовується порівняння: $X \equiv a^{-1}(X - b) \bmod m^2$

Далі ми за допомогою автоматичного розпізнавача російської мови знаходимо змістовний ВТ:

```
Розшифрування за знайденими ключами та аналіз на відповідність текстів російській мові...
Виявлено потенційний текст: хорошосэрбиллнехотясунулденьгивкарманвотчтобиллвыпросто...

ШТ УСПІШНО РОЗШИФРОВАНО!
ВТ:хорошосэрбиллнехотясунулденьгивкарманвотчтобиллвыпростопосеетеэтуновуютравукогда...
Ключ: (703, 956)
```

Повний текст збережено в файл decrypted_11.txt

Оскільки маємо єдиний розв’язок, можемо вважати розшифрування і визначення змістовного тексту успішним, це і буде оригінальний текст

Однак, є момент, на який варто звернути увагу. При зашифруванні, можливо, випадково, в алфавіті переплутано місцями ы та ь

Правильний порядок: …щыьэ…

Порядок, при якому було отримано змістовний текст: …щьыэ…

Тому, для змістовного результату, було змінено порядок букв в алфавіті.

Така помилка була виявлена, на етапі коли автоматичний розпізнавач не був на стільки ретельним, для зручності було створено скрипт lab3_found_error.py, який дозволяє це побачити (прибрали критерій, який фільтрував цей текст), тоді маємо:

```
Знайдено 178 ключів

Розшифрування за знайденими ключами та аналіз на відповідність текстів російській мові...
Виявлено потенційний текст: хорошосэрбиллнехотясунулденыаивкарщгнвотчтобиллвьпросто...

ШТ УСПІШНО РОЗШИФРОВАНО!
ВТ:хорошосэрбиллнехотясунулденыаивкарщгнвотчтобиллвьпростопосеехеэтуновуюправукбасг...
Ключ: (703, 956)
```

Повний текст збережено в файл decrypted_with_error_11.txt

Співвідносимо помилкові символи з гіпотетично-правильними, розв’язуємо порівняння для того, щоб отримати помилковий символ, і для того, щоб отримати правильний, і бачимо, що ь та ы в алфавіті треба поміняти місцями

Автоматичний розпізнавач тексту, написаного російською мовою, працює на основі критеріїв, що спираються на властивості мови.

Так, в ньому реалізований критерій заборонених l-грам на основі біграм. Він шукає частоту появи кожної забороненої біграми, і якщо вона перевищує порогове значення, то це ознака незмістовного тексту. Таких біграм має бути знайдено більше двох, така реалізація припускає, що у тексті могло бути допущено описку, тому спрацьовує тільки після другої забороненої біграми.

Також було реалізовано критерій частих l-грам, за яким ми перевіряємо, дві з п’яти найчастіших біграм в тексті відповідають хоча б двом з п’яти найчастіших біграм в мові.

Останній схожий на попередній, але тут ми беремо сумарну частоту найчастіших літер в мові, і шукаємо сумарну частоту для цих же літер в тексті. Припускаємо, що значення сумарної частоти в тексті може коливатись в певному діапазону, і, відповідно, визначаємо діапазон коливання (в нашому випадку +-20% від сумарної частоти в мові)

Висновки:

У ході даного комп’ютерного практикуму, було експериментально досліджено розшифрування афінного шифру біграмної підстановки на ШТ. Для цього, ми шукали ключі для можливих пар біграм ШТ і біграм мови з п’яти найчастіших відповідно. Потім, розшифрували текст за цими ключами і перевіряли текст на валідність.

Валідність в нас визначається на основі того факту, що ВТ має бути змістовний, відносно мови, якою написаний. Тому, користуючись неоднорідністю мови, ми написали автоматичний розпізнавач мови, який працює за критеріями заборонених l-грам для біграм та частих l-грам для біграм та монограм (реалізація ідеї суттєво відрізняється).

Отже, змістовність тексту певної мови можна визначити автоматично на основі критеріїв, які відповідають властивостям цієї мови. Наприклад, на основі того, які l-грами частіше або рідше зустрічаються, або ж на основі статистичних характеристик (напр., ентропії тексту)

В ході виконання в нас виникла помилка з інтерпретацією алфавіту, змістовний розв’язок вдалось отримати при несуттєвій зміні порядку символів у алфавіті (поміняли місцями ы та ь), схоже на те, що цю помилку було допущено випадково при зашифруванні тексту за неправильним алфавітом