**Київський національний університет імені Тараса Шевченка**

**радіофізичний факультет**

Лабораторна робота № 1

## Тема: «Дослідження кількості інформації при різних варіантах кодування»

Роботу виконав

студент 3 курсу

Комп’ютерної інженерії

Науменко Павло Сергійович

Київ 2017

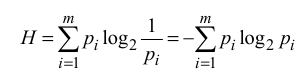
**Посилання на GitHub:** [**https://github.com/PashaNaumencko/ComputerSystemsLabs**](https://github.com/PashaNaumencko/ComputerSystemsLabs)

**Мета:** Дослідити імовірнісні параметри української мови для оцінки кількості інформації текстів. Дослідити вплив різних методів кодування інформації на її кількість.

# **Теоретичні відомості**

***Відносна частота появи символу*** - імовірність появи певного символу в певному місці тексту - відношення числа появи символу в тексті до загальної кількості символів.

***Середня ентропія нерівноймовірного алфавіту***:



де *m* - кількість символів алфавіту, *p* - імовірність появи символу

Ентропія вимірюється в **БІТАХ** (як представлення кількості можливих варіантів).

***Кількість інформації в тексті*** - середня ентропія вихідного алфавіту помножена на кількість символів тексту. (***HINT***: результат обрахунку для порівняння значення з розміром файлів треба перевести з бітів в байти)

# **1. Дослідження кількості інформації в тексті**

1. Оберіть 3 текстових файла різного тематичного та лінгвістичного спрямування (наприклад, вірш Тараса Шевченка “Мені тринадцятий минало”, “Казка про рєпку” Леся Подерв'янського та специфікацію інерфейсу PCI)
2. Створіть програму (будь-якою зручною для вас мовою), яка в якості вхідних даних приймає текстовий файл, та аналізуючи його вміст:
   1. обраховує частоти (імовірності) появи символів в тексті
   2. обраховує середню ентропію алфавіту для даного тексту
   3. виходячи з ентропії визначає кількість інформації та порівнює її з розмірами файлів
   4. виводить на екран значення частот, ентропії та кількості інформації

**Результати роботи:**

Trinadcyatui.txt

Frequency for each character:

"М":0.0050933786078098476

"е":0.04753820033955857

"н":0.04584040747028863

"і":0.031409168081494056

" ":0.166383701188455

"т":0.023769100169779286

"р":0.022071307300509338

"и":0.0365025466893039

"а":0.059422750424448216

"д":0.019524617996604415

"ц":0.0050933786078098476

"я":0.023769100169779286

"й":0.006791171477079796

"м":0.017826825127334467

"л":0.04838709677419355

"о":0.08573853989813243

".":0.030560271646859084

"\n":0.0432937181663837

"Я":0.001697792869269949

"п":0.01697792869269949

"с":0.030560271646859084

"г":0.021222410865874362

"з":0.010186757215619695

"Ч":0.0050933786078098476

"к":0.016129032258064516

"ч":0.012733446519524618

",":0.02631578947368421

"б":0.01697792869269949

"у":0.01697792869269949

"?":0.001697792869269949

"ю":0.010186757215619695

"Н":0.008488964346349746

"в":0.025466893039049237

"У":0.0025466893039049238

"ж":0.0050933786078098476

"А":0.001697792869269949

"'":0.0008488964346349745

"І":0.005942275042444821

"ь":0.010186757215619695

"Т":0.004244482173174873

"Г":0.0008488964346349745

"є":0.003395585738539898

"!":0.008488964346349746

"З":0.0008488964346349745

":":0.0008488964346349745

"С":0.0008488964346349745

"Б":0.001697792869269949

"П":0.003395585738539898

"ї":0.0025466893039049238

"О":0.0008488964346349745

"х":0.0025466893039049238

"—":0.0008488964346349745

"щ":0.0008488964346349745

"ш":0.0008488964346349745

"Л":0.0008488964346349745

Entropy for all text: 4.765494657749576

Amount of information in the text: 701.7190883536251 bytes

Amount of information/File size: 0.3373649463238582

PCI.txt

Frequency for each character:

"С":0.002074688796680498

"п":0.03319502074688797

"е":0.022821576763485476

"ц":0.016597510373443983

"и":0.03941908713692946

"ф":0.006224066390041493

"і":0.05186721991701245

"к":0.014522821576763486

"а":0.07261410788381743

"я":0.014522821576763486

" ":0.14107883817427386

"ш":0.01037344398340249

"н":0.05394190871369295

"P":0.002074688796680498

"C":0.002074688796680498

"I":0.002074688796680498

"\n":0.014522821576763486

"ч":0.006224066390041493

"с":0.035269709543568464

"т":0.04564315352697095

"о":0.04771784232365145

"—":0.01037344398340249

"3":0.03319502074688797

",":0.014522821576763486

"М":0.008298755186721992

"Г":0.006224066390041493

"б":0.02074688796680498

"6":0.012448132780082987

"р":0.04771784232365145

"д":0.03319502074688797

"х":0.002074688796680498

";":0.012448132780082987

"з":0.008298755186721992

"ь":0.012448132780082987

"2":0.01037344398340249

"4":0.006224066390041493

"м":0.004149377593360996

"у":0.02074688796680498

"л":0.012448132780082987

"в":0.02074688796680498

"(":0.008298755186721992

"ю":0.004149377593360996

"й":0.01867219917012448

")":0.008298755186721992

"-":0.004149377593360996

"г":0.006224066390041493

"щ":0.002074688796680498

"є":0.004149377593360996

"1":0.002074688796680498

"'":0.002074688796680498

"ї":0.004149377593360996

"5":0.004149377593360996

Entropy for all text: 4.942643765082409

Amount of information in the text: 297.7942868462152 bytes

Amount of information/File size: 0.3540954659289122

Poderev.txt

Frequency for each character:

"Л":6.622516556291391e-05

"е":0.026357615894039736

"с":0.029536423841059603

"ь":0.014635761589403974

" ":0.14701986754966886

"П":0.001390728476821192

"о":0.07443708609271524

"д":0.027350993377483444

"р":0.033973509933774834

"в":0.040264900662251656

"'":0.0007947019867549669

"я":0.01675496688741722

"н":0.04576158940397351

"к":0.033245033112582784

"и":0.057549668874172184

"й":0.007814569536423841

".":0.02258278145695364

"\n":0.010794701986754966

"К":0.0006622516556291391

"а":0.07688741721854304

"з":0.017483443708609273

"п":0.017549668874172187

"є":0.008741721854304637

"у":0.02456953642384106

",":0.021721854304635763

"б":0.013973509933774834

"Х":0.0010596026490066225

"л":0.02662251655629139

"і":0.033973509933774834

"?":0.002913907284768212

"Д":0.0015231788079470199

"С":0.010198675496688741

"О":0.0026490066225165563

"ч":0.016026490066225165

"м":0.02258278145695364

"В":0.004172185430463576

"ж":0.007019867549668875

"М":0.0027152317880794704

"т":0.03377483443708609

"Н":0.0017218543046357616

"ш":0.007748344370860927

"А":0.0038410596026490066

"ф":0.0028476821192052982

"i":6.622516556291391e-05

"-":0.0025827814569536426

"ц":0.004304635761589404

"З":0.0007947019867549669

"г":0.008410596026490066

"ю":0.006357615894039735

"щ":0.0016556291390728477

"І":0.0005960264900662252

"ї":0.0032450331125827817

"х":0.006688741721854304

"У":0.00013245033112582781

"(":0.0015231788079470199

")":0.0015231788079470199

"Ч":0.00019867549668874172

"!":0.0027152317880794704

"Б":0.0005298013245033113

"Щ":0.00026490066225165563

"Ш":0.00033112582781456954

"Р":0.00026490066225165563

"Я":0.0007947019867549669

"Ц":0.00026490066225165563

""":0.00026490066225165563

"Ж":6.622516556291391e-05

"Т":0.0017218543046357616

":":0.00033112582781456954

"Г":0.00019867549668874172

"5":0.00013245033112582781

"2":0.00013245033112582781

"Ї":0.00026490066225165563

"Й":6.622516556291391e-05

"Є":6.622516556291391e-05

"0":0.00013245033112582781

"Е":6.622516556291391e-05

Entropy for all text: 4.871128139532326

Amount of information in the text: 9194.254363367265 bytes

Amount of information/File size: 0.33910870664873916

1. Проведіть стиснення кожного вхідного файлу за допомогою 5 різних алгоритмів стиснення (zip, rar, gzip, bzip2, xz, або будь-які інші на ваш вибір, можна використовувати готові програмні засоби для стиснення).
2. Порівняйте результуючі обсяги архівів з обчисленою кількістю інформації та **наведіть у звіті висновки** щодо кореляції цих величин для обраних вами файлів (яка відмінність, що вийшло більше і чому)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | txt | zip | xz | gzip | bzip2 | rar | КІ |
| Trinadcyatui.txt | 2080 | 1 008 | 916 | 877 | 718 | 955 | 701 |
| PCI.txt | 841 | 542 | 472 | 420 | 376 | 487 | 297 |
| Poderev.txt | 27113 | 8279 | 7868 | 8153 | 6457 | 8385 | 9194 |

**Висновок:** після стиснення текстового файлу його зміст буде відрізнятися від оригінального файлу тому кількість інформації оригінального файлу та розмір стисненого файлу теж будуть відрізнятись.

# **2. Дослідження способів кодування інформації на прикладі Base64**

1. Ознайомтесь зі стандартом RFC4648
2. Для практичного засвоєння методу кодування, створіть програму, що кодує довільний файл в Base64 (шляхом реалізації алгоритму вручну, а не виклику бібліотечної функції)
   * 1. перевірте коректність роботи програми, порівнявши результат з існуючими програмними засобами (наприклад, openssl enc -base64)
3. Закодуйте в Base64 обрані вами текстові файли
   * 1. Обрахуйте кількість інформації в base64-закодованому варіанті файлу
     2. Порівняйте отримане значення з кількістю інформації вихідного файлу
     3. Зробіть висновки з отриманого результату
4. Закодуйте в Base64 стиснені кращим з алгоритмів текстові файли
   * 1. Обрахуйте кількість інформації в base64-закодованому варіанті стисненого файлу
     2. Порівняйте отримане значення з кількістю інформації вихідного файлу та base64-закодованого файлу
     3. Зробіть висновки з отриманого результату

|  |  |  |  |
| --- | --- | --- | --- |
|  | КІ txt | КІ Base64.txt | КІ Base64.bzip2 |
| Trinadcyatui.txt | 701 | 1706 | 709 |
| PCI.txt | 297 | 701 | 366 |
| Poderev.txt | 9194 | 22135 | 6450 |

**Висновок:** за результатами видно, що кількість інформації у Base64-закодованому файлі більша ніж в оригінальному файлі, тому що після кодування розмір файлу збільшується.