

МІНІСТЕРСТВО ОСВІТИ ТА НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ТАРАСА
ШЕВЧЕНКА

Факультет радіофізики, електроніки та комп'ютерних систем

Кафедра комп'ютерної інженерії

Лабораторна робота №1

за спеціальністю 123 Комп'ютерна інженерія

з предмету:

КОМП'ЮТЕРНІ СИСТЕМИ

Виконав студент 3-го курсу

Бількевич Борис Борисович

Науковий керівник:

кандидат технічних наук

Слюсар Євген Андрійович

КИЇВ 2020

«Дослідження кількості інформації при різних варіантах кодування»

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

План:

1. Дослідження кількості інформації в тексті
2. Дослідження способів кодування інформації на прикладі Base64

Хід роботи

1. Оберіть 3 текстових файла різного тематичного та лінгвістичного спрямування (наприклад, вірш Тараса Шевченка “Мені тринадцятий минало”, “Казка про репку” Леся Подерв'янського та специфікацію інтерфейсу PCI)

В процесі виконання лабораторно роботи, було обрано 2 статті та 1 вірш для аналізу:

- ❖ "І все на світі треба пережити" Ліна Костенко
- ❖ Фішинг
- ❖ Дихлордифенілтрихлоретан

2. Переконайтесь, що тексти, які ви використовуєте є унікальними і не повторюються у ваших колег! Використовуйте наявні електронні засоби зв'язку та документообігу, щоб уникнути дублювання! Вдруге аналіз того самого тексту не зараховується!

Для уникнення дублювання було використано спільний Excel документ:

- ❖ Комп'ютерні системи

3. Створіть програму (будь-якою зручною для вас мовою), яка в якості вхідних даних приймає текстовий файл, та аналізуючи його вміст:
- a) обраховує частоти (імовірності) появи символів в тексті
 - b) обраховує середню ентропію алфавіту для даного тексту
 - c) виходячи з ентропії визначає кількість інформації та порівнює її з розмірами файлів
 - d) виводить на екран значення частот, ентропії та кількості інформації

Дану програму вирішив реалізувати на мові Python:

Стаття про «Фішинг»

```
+++++
Фішинг – один з видів інтернет-шахрайства, коли «жертві» надсилаються повідомлення від імені відомих компаній або організацій (наприклад, банку, податкової служби, відомого інтернет-магазину), однак насправді вони не є справжніми. Мета фішингу – отримання доступу до конфіденційних даних користувачів (паролів, логінів, даних особових рахунків і банківських карт). Зазвичай використовується метод проведення масових розсилок від імені популярних компаній або організацій, які містять посилання на фейкові сайти, які важко зовні відрізнити від справжніх. У листах особу ввічливо просять оновити чи підтвердити правильність персональної інформації або інформують про які-небудь проблеми з даними, а після цього перенаправляють на підроблений сайт, де необхідно ввести облікові дані. Якщо «жертва» вводить свої дані на таких сайтах, то злочинцям стають відомі ці дані та вони можуть використати їх з метою крадіжки персональних даних, персональних коштів або іншого. Фішинг є одним з найпоширеніших видів кібератак.
+++++
```

Далі виводжу всю інформацію про даний уривок тексту, який я в процесі дослідив, а саме загальну кількість літер, середню ентропію, а також кількість інформації в бітах та байтах.

```
All characters in text = 847
Average entropy: 4.5195 bit 4.519490578634116
Count of information: 3828.0085 bit
Count of information: 478.5011 byte

File size: 1883 byte
File size > Count of information
```

Після чого проводжу перевірку на порівняння розмірів до стиснення та після.







Використав 5 варіантів стиснення (.rar, .zip, .gz, .bz2, .7z), які виконував за допомогою онлайн архіваторів.

```
Archive size .rar: 876
Archive size .rar > Count of information
Archive size .zip: 930
Archive size .zip > Count of information
Archive size .gz: 808
Archive size .gz > Count of information
Archive size .bz2: 675
Archive size .bz2 > Count of information
Archive size .7z: 924
Archive size .7z > Count of information
```

А також виводжу частоту появи кожної з букв:

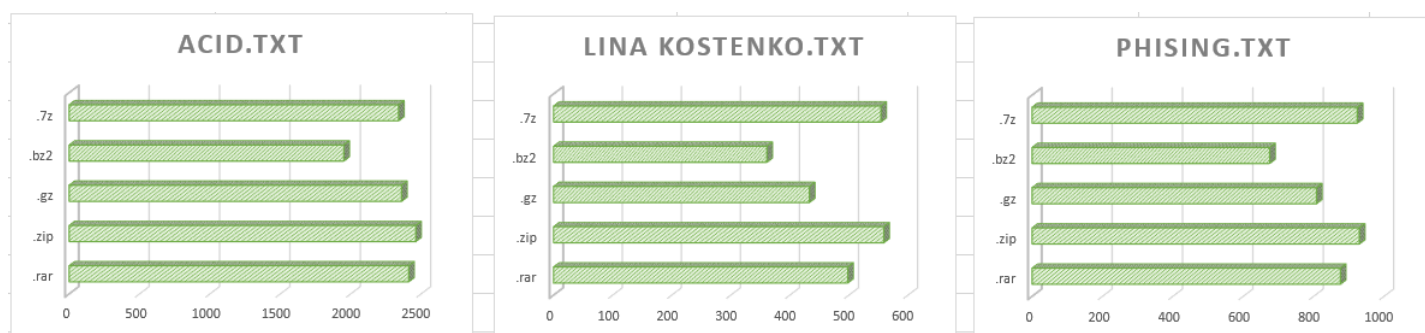
Letter	Counts	Frequency
А	70	8.2645
Б	15	1.771
В	51	6.0213
Г	10	1.1806
Ґ	0	0.0
Д	37	4.3684
Е	31	3.66
Є	3	0.3542
Ж	8	0.9445
З	13	1.5348
И	58	6.8477
І	66	7.7922
Ї	5	0.5903
Й	13	1.5348
К	29	3.4238
Л	22	2.5974
М	23	2.7155
Н	72	8.5006
О	79	9.327
П	27	3.1877
Р	38	4.4864
С	32	3.778
Т	44	5.1948
У	15	1.771
Ф	7	0.8264
Х	20	2.3613
Ц	7	0.8264
Ч	5	0.5903
Ш	8	0.9445
Щ	1	0.1181
Ь	17	2.0071
Ю	5	0.5903
Я	16	1.889

4. 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).

 acid.txt	03.02.2021 16:57	Текстовий докум...	7 КБ
 acid.txt.7z	03.02.2021 16:59	Архив WinRAR	3 КБ
 acid.txt.bz2	03.02.2021 16:59	Архив WinRAR	2 КБ
 acid.txt.gz	03.02.2021 16:58	Архив WinRAR	3 КБ
 acid.txt.rar	03.02.2021 16:58	Архив WinRAR	3 КБ
 acid.txt.zip	03.02.2021 16:58	Архив ZIP - WinR...	3 КБ

5. 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)

File	acid.txt	Lina Kostenko.txt	Phising.txt
Count of information	1663	228	478
Original	6441	981	1883
Archives			
.rar	2412	498	876
.zip	2462	559	930
.gz	2361	433	808
.bz2	1951	361	675
.7z	2340	554	924



Відповідно до графіків ми можемо побачити, що найкраще справився з стисненням bz2. А також на основі даних, можна зробити висновок, кількість інформації для всіх стиснених файлів менша, ніж фактичні розміри.

Як висновок, при ідеальному стисненні розмір файлу мав би бути рівним кількості інформації. Але насправді, ми можемо побачити з результатів таблиці у всіх випадках розміри архівованих файлів, дещо більші за кількість інформації. Це відбувається тому, що алгоритми архіваторів налаштовані таким чином, щоб використовувати повторювані частини тексту. Саме тому формула розрахунку кількості інформації, використана для програми, не є досконалою, бо вона не враховує даний уривок тексту.

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)

- а. перевірте коректність роботи програми, порівнявши результат з існуючими програмними засобами (наприклад, `openssl enc -base64`)

❖ **Acid.txt**

ДДТ має одну з найвищих суперечливих репутацій серед нещодавно винайдених хімічних препаратів. Він довів свою ефективність як інсектицид, проте його потужна токсичність впливає не тільки на комах.

У деяких країнах ДДТ все ще використовується, іноді - незаконно.

ДДТ, або дихлордифенілтрихлоретан, - це безбарвна тверда кристалічна речовина.

Вона належить до класу пестицидів органічолоридів і є повністю синтетичною - виробляється в лабораторіях і у природі не зустрічається.

ДДТ нерозчинний у воді, проте може розчинитися в органічних розчинниках, жирах і маслах.

Завдяки цій здатності ДДТ накопичується у жирових тканинах тварин, що піддавалися його впливу.

Внаслідок біокумуляції ДДТ залишається у харчовому ланцюгу, переходячи від риб, жаб і риб у тіла тварин, що ними харчуються.

Найбільш високий рівень ДДТ найчастіше виявляється в організмах тварин верхнього рівня харчового ланцюга, а особливо у таких хижих птахів, як орли, яструби, пелікани й кондори.

ДДТ згубно впливає і на здоров'я людини. Згідно ЕРА, ця речовина викликає пошкодження печінки (а також може призвести до розвитку раку печінки), впливає на нервову систему, викликає вроджені порушення та блокує роботу репродуктивної системи.

Коротка історія ДДТ

ДДТ було вперше синтезовано у 1874 році, проте можливість його застосування в якості універсального інсектициду було відкрито лише у 1939 році швейцарським біохіміком Полом Германом Мюллером , який отримав за своє відкриття Нобелівську премію.

До появи ДДТ маларія, висипного тифу, жовтої лихоманки та бубонної чуми загинули мільйони людей.

Під час Другої світової війни цей інсектицид став широко застосовуватися в американських військах - його використовували для боротьби з переносниками захворювань, відомі в Італії та у тропічних регіонах.

Після Другої світової війни фермери виявили, що ДДТ ефективно бореться зі шкідниками сільськогосподарських культур, і ДДТ став зброєю в боротьбі з маларією.

Однак слід зазначити, що деякі популяції комах розвинули стійкість до інсектициду.

ДДТ-це отрута

ДДТ ставав все більш популярним і вчені помітили, що його неконтрольоване застосування завдає шкоди природі.

У відомій книзі вченого та письменника Рейчел Карсон Мовчазна весна (Silent Spring) розповідається про наслідки забруднення довкілля ДДТ та іншими пестицидами. Назва книги говорить про те, що після використання ДДТ і подібних хімічних речовин у деяких регіонах зникли цілі популяції співочих птахів.

За повідомленнями вчених, у птахів, що піддаються впливу ДДТ, шаралупа ледь стає настільки тонкою, що тріскається до випулнення пташенят, що і призводить до їхньої загибелі.

«Мовчазна весна» стала однією з найпопулярніших книг в історії і їй приписують вплив на активізацію сучасного екологічного руху.

ДДТ заборонений у всьому світі

Після того, як шкоду ДДТ було доведено, уряди багатьох країн світу наклали заборону на його застосування або обмежили його використання. До 1978 року цей пестицид заборонили в Угорщині, Норвегії та Швеції; також, незважаючи на протистояння з боку хімічної промисловості, у США виробництво і використання ДДТ було заборонено у 1972 році.

У 2004 році 178 країн підписали договір, відомий як Стокгольмська конвенція про стійкі органічні забруднювачі (СОЗ), що обмежує використання ДДТ для боротьби з небезпечними комахами (наприклад, у разі нападу маларії).

Однак ДДТ все ще регулярно застосовується у сільському господарстві у деяких країнах (таких як Індія та країни Африки на південь від Сахари) для боротьби з москитами та іншими комахами.

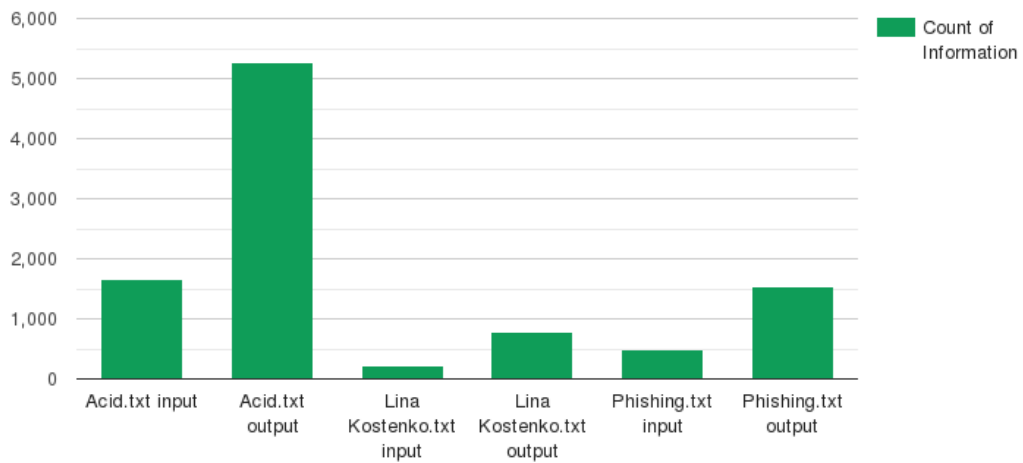
❖ **Phishing.txt**

Фішинг – один з видів інтернет-шахрайства, коли шахраї надсилають повідомлення від імені відомих компаній або організацій (наприклад, банку, податкової служби, відомого інтернет-магазину), щоб настраїли вони не є справжніми. Метою фішингу – отримання доступу до конфіденційних даних користувачів (паролів, логінів, даних особистих рахунків і банківських карт). Звичайно використовується метод проведення масових розсліду від імені популярних компаній або організацій, які містять посилання на фальшиві сайти, на яких шахраї підігрують відправників і листав особу ввічливо просити надати певні дані. Звичайно, якщо людина надасть потрібні дані, шахраї отримають доступ до інформації, а після цього перенаправляють на підбраний сайт, де необхідно ввести облікові дані. Шахраї починають вводити свої дані на таких сайтах, то злочинцям стають відомі ці дані та вони можуть використати їх з метою крадіжки персональних даних, персональних когтів або іншого. Фішинг є одним з найпоширеніших видів кібератак.

[illegible][illegible]

3. Закодуйте в Base64 обрані вами текстові файли
 - a. Обрахуйте кількість інформації в base64-закодованому варіанті файлу
 - b. Порівняйте отримане значення з кількістю інформації вихідного файлу
 - c. Зробіть висновки з отриманого результату

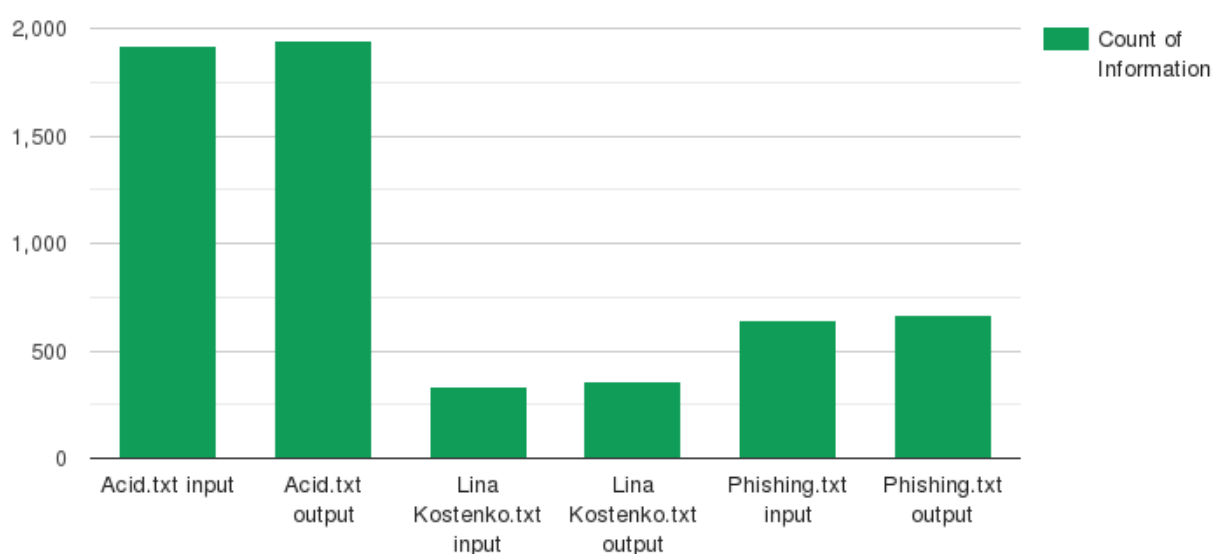
Archives (byte)	Acid.txt	Lina Kostenko.txt	Phishing.txt
Count of Information (in)	1663	228	478
Count of Information (out)	5275	787	1538
Percent of differences	3,171978352	3,451754386	3,217573222



Як висновок, можна сказати, що у закодованому файлі кількості інформації більше у 3.17 – 3.45 разів ніж у вихідному.

4. Закодуйте в Base64 стиснені кращим з алгоритмів текстові файли
 - a. Обрахуйте кількість інформації в base64-закодованому варіанті стисненого файлу
 - b. Порівняйте отримане значення з кількістю інформації вихідного файлу та base64-закодованого файлу
 - c. Зробіть висновки з отриманого результату

File (byte)	Acid.txt	Lina Kostenko.txt	Phishing.txt
Count of Information (in)	1919	331	644
Count of Information (out)	1947	356	668
Percent of differences	1,014590933	1,075528701	1,037267081



Як висновок, можна сказати, що у закодованому файлі кількості інформації більше у 1.015 – 1.075 разів, ніж у вихідному.

Висновок:

В процесі виконання лабораторної роботи, було опановано базові знання таких фундаментальних понять, як base64 та процес роботи, а також визначення ентропії інформації. Весь процес намагався побудувати на мові програмування Python, оскільки бачу перспективу даної мови, а також для розширення своїх вмінь та навичок. Протягом виконання роботи, було застосовано 5 варіантів стиснення текстового файлу, а саме ормати: .rar, .zip, .7z, .bz2 та .gz.

Розібрався з алгоритмом кодування base64. Всі додакові матеріали (файли, програми та архіви можна знайти в закріпленому нижче посиланні)

❖ Лабораторна робота №1 з Комп'ютерних систем



Laboratory work 1

Add files via upload

21 hours ago