**Міністерство освіти і науки України**  
**Національний технічний університет України**  
**“Київський політехнічний інститут”**

Інститут Прикладного системного аналізу  
Кафедра Системного проектування

Лабораторна робота з дисципліни  
«Теорія інформації кодування»

на тему  
**«ВИЗНАЧЕННЯ КІЛЬКОСТІ ІНФОРМАЦІЇ В ДИСКРЕТНИХ ПОВІДОМЛЕННЯХ»**

Виконав:

студент 2 курсу

групи ДА-92

Насікан Д. Ю.

Київ 2021

**Мета роботи -** вивчення алгоритмів визначення кількості інформації в дискретних повідомленнях, оцінка надмірності повідомлень, розробка програми для дослідження ефективності кодування текстових даних у файлах довільної довжини.

1. **ЗАВДАННЯ**

1.1 Ознайомитися з методикою визначення кількості інформації в дискретних повідомленнях .

1.2 Використовуючи табличний процесор MS Excel визначити кількість інформації, що міститься в текстовому повідомленні, що складається з вашого прізвища, імені і по батькові. При цьому необхідно обчислити середню ентропію на символ кодованого повідомлення і середню довжину кодових комбінацій, необхідних для кодування кожного символу повідомлення. Знайти кількість інформації, що міститься в кодованому повідомленні, абсолютну і відносну надмірність повідомлення за умови кодування кожної букви повідомлення одним байтом.

1.3 Розробити програму на мові програмування високого рівня для обчислення ентропії та кількості інформації, що міститься в текстових файлах довільної довжини, максимально можливої ентропії за умови однакової імовірності символів алфавіту, абсолютної і відносної надмірності файлів. При вводі текстових даних передбачити два режими вводу кодів: введення з клавіатури, а також введення даних з текстового файлу з використанням однобайтових кодових таблиць. За допомогою розробленої програми перевірити дані, отримані при виконанні п. 1.2.

1.4 За допомогою розробленої при виконанні п.1.2 програми дослідити залежність ентропії від довжини тексту, а також залежність ентропіїї від мови. Для цього за допомогою доступних програм автоматичного перекладу підготувати однакові текстові файли на різних мовах (наприклад, англійська, українська, російська) різної довжини, для яких провести дослідження.

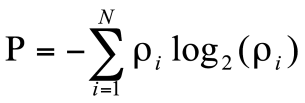
1. **ХІД РОБОТИ**
   1. Знайдемо імовірності появи кожного символу в рядку «насікан дмитро юрійович»:

|  |  |  |
| --- | --- | --- |
| **Символ** | **Кількість** | **Імовірність P[i]** |
| **\_** | **2** | **0.086956522** |
| **н** | **2** | **0.086956522** |
| **а** | **2** | **0.086956522** |
| **с** | **1** | **0.043478261** |
| **і** | **2** | **0.086956522** |
| **к** | **1** | **0.043478261** |
| **д** | **1** | **0.043478261** |
| **м** | **1** | **0.043478261** |
| **и** | **2** | **0.086956522** |
| **т** | **1** | **0.043478261** |
| **р** | **2** | **0.086956522** |
| **о** | **2** | **0.086956522** |
| **ю** | **1** | **0.043478261** |
| **й** | **1** | **0.043478261** |
| **в** | **1** | **0.043478261** |
| **ч** | **1** | **0.043478261** |
| **Сума** | **23** | **1** |

Як бачимо, сума ймовірностей появи кожного символа дорівнює одиниці, що свідчить про правильність їх знаходження.

* 1. Знайдемо кількість інформації повідомлення:

I = log N = n log m = 23 log 16 = 92

* 1. Розрахуємо ентропію повідомлення за формулою :

P = біт.

* 1. Визначимо середню довжину кодових комбінацій для, необхідних для кодування кожного символу повідомлення:

L^2 = 16

L = 4 біт

* 1. Знайдемо абсолютну і відносну надмірність повідомлення за умови кодування кожної букви повідомлення одним байтом:

Абсолютна надмірність повідомлення – 4 біт.

Відносна надмірність повідомлення – 8 біт

* 1. Розробимо програму на мові програмування пайтон для обчислення ентропії та кількості інформації, що міститься в текстових файлах довільної довжини, максимально можливої ентропії за умови однакової імовірності символів алфавіту, абсолютної і відносної надмірності файлів:

Код програми:

import math  
  
  
def main\_analyzer(message):  
 symbols = dict()  
  
 # рахуємо кількість символів у алфавіті та входження кожного символа  
 for sym in message:  
 symbols[sym] = symbols.get(sym, 0) + 1  
  
 # обчислюємо ентропію та кількість інформації  
 size = len(message)  
 entropy = 0  
 for num in symbols.values():  
 probability = num / size  
 entropy += probability \* math.log2(probability)  
  
 entropy = -entropy  
 information\_amount = size \* math.log2(len(symbols.keys()))  
  
 print(f'Довжина тексту - {size}')  
 print(f'Ентропія тексту - {entropy}')  
 print(f'Кількість інформації тексту - {information\_amount}')  
  
  
MODE = int(input('Виберіть режим роботи:\n1 - текстовий файл\n2 - введення повідомлення вручну\n\n'))  
  
if MODE == 1:  
 file\_name = input('Уведіть назву файлу без розширення: ') + '.txt'  
  
 # читаємо дані з файлу  
 with open(file\_name, 'r', encoding='utf-8') as file:  
 text\_message = file.read()  
 print(text\_message)  
 main\_analyzer(text\_message)  
  
elif MODE == 2:  
 text\_message = input('Введіть ваше текстове повідомлення: ')  
 main\_analyzer(text\_message)

* 1. Перевіримо розрахунки попередніх пунктів через програму:

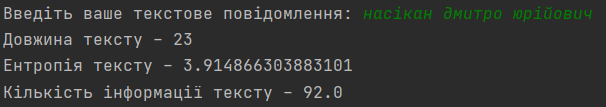


Рис 1.1 – вихідні дані програми

Як бачимо, розрахунки збігаються.

* 1. Користуючись розробленою програмою дослідимо залежність ентропії від довжини тексту та мови:

Після прогону підготовлених текстів українською, російською та англійської мови зведемо результати в порівняльну таблицю:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Мова** | **Довжина файлу** | | | | | | | |  |
| **489** | **850** | **1357** | **2451** | **3035** | **4532** | **5346** | **6876** | **Середнє значення ентропії** |
| **Тематика інформатика** | **Українська** | **4.485** | **4.8** | **4.926** | **4.83** | **4.717** | **4.653** | **4.953** | **4.7** | **4.758** |
| **Російська** | **4.38** | **4.642** | **4.874** | **4.721** | **4.588** | **4.623** | **4.84** | **4.532** | **4.65** |
| **Англійська** | **4.07** | **4.344** | **4.48** | **4.196** | **4.2595** | **4.316** | **4.415** | **4.371** | **4.3064375** |
|  | | **Довжина файлу** | | | | | | | |  |
| **339** | **796** | **1772** | **1615** | **2435** | **3545** | **4589** | **6438** |
| **Тематика загальна** | **Українська** | **4.6366** | **4.8675** | **5** | **4.7278** | **4.743** | **4.751** | **4.654** | **4.932** | **4.7889875** |
| **Російська** | **4.5447** | **4.6619** | **4.916** | **4.5693** | **4.43** | **4.61** | **4.438** | **4.742** | **4.6139875** |
| **Англійська** | **4.2686** | **4.4863** | **4.6819** | **4.4347** | **4.32** | **4.342** | **4.3** | **4.325** | **4.3948125** |

1. **ВИСНОВКИ**

У ході даної роботи було опрацьовано теоретичний матеріал з основ теорії інформації та визначено такі характеристики, як кількість інформації, ентропія та абсолютна й відносна надмірність для текстового рядка «насікан дмитро юрійович». Також, була розроблена програма на мові програмування python для, що може зчитувати текстові рядки з консолі або файлу та визначати ентропію й кількість інформації файлу.

З метою порівняння й дослідження залежностей ентропії від довжини та мови тексту з допомогою розробленої програми було визначено ентропію 48 текстових уривків українською, російською та англійською мовами різної довжини та тематики.

Після порівняння отриманих результатів можна зробити висновки, що:

1. Ентропія текстів ІТ тематики та загальної тематики суттєво не відрізняється.
2. Ентропія не залежить від довжини текстового повідомлення.
3. Ентропія тексту українською мовою є найвищою (приблизно 4.7 біт), трохи меншою (4.6 біт, що приблизно на 0.1 біт менше, ніж в української) є ентропія російського тексту, і найменшою є ентропія тексту на англійській мові (4.35 біт).

Така різниця ентропій текстів на різних мовах пояснюється різною кількістю символів у абетках різних мов, та різною частотою вживання певних символів й іншими лінгвістичними особливостями мов.