

Звіт

З Лабораторної роботи N1

З курсу «Комп'ютерні системи»

Студента 3-го курсу КІ ФРЕКС

Корнійчук Олександр

<https://github.com/Gastonovich/labs>

1 Частина

1. Оберіть 3 текстових файла різного тематичного та лінгвістичного спрямування

Я обрав океан ельзи на небі, <https://www.bbc.com/ukrainian/news-47608602>,
<https://www.bbc.com/ukrainian/news-47608600>

2. 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

```
o: 69
r: 3
c: 13
ь: 5
т: 31
і: 49
и: 37
ю: 22
й: 7
з: 23
в: 41
н: 70
.: 9
H: 1
I: 2
y: 11
ц: 1
x: 6
.: 30
I: 11
x: 3
п: 3
e: 3
Б: 2
K: 2
б: 24
l: 3
M: 6
З: 4
р: 3
ш: 4
ч: 3
A: 1
i: 1
л: 1
freq: {'\ufeff': 0.0010384215991692627, '\n': 0.029875804776739357, 'T': 0.0020768431983385254, 'a': 0.06230529595015576, 'm': 0.035306334371754934, ',': 0.023883696780893044,
' ': 0.18691588785046728, 'д': 0.012461059190031152, 'e': 0.04257528556593977, 'л': 0.034267912727258567, 'я': 0.023883696780893044, 'к': 0.02284527518172378, 'o': 0.071651090342
67912, 'r': 0.003115264797507788, 'c': 0.013499480789200415, 'ь': 0.005192107995846314, 'т': 0.032191069574247146, 'i': 0.05088265835929388, 'и': 0.03842159916926272, 'ю': 0.022
84527518172378, 'й': 0.007268951194184839, 'з': 0.023883696780893044, 'в': 0.04257528556593977, 'н': 0.07268951194184839, ':': 0.009345794392523364, 'H': 0.0010384215991692627,
'I': 0.0020768431983385254, 'y': 0.01142263759086189, 'ц': 0.0010384215991692627, 'x': 0.006230529595015576, '.:': 0.03115264797507788, 'I': 0.01142263759086189, 'x': 0.003115264
797507788, 'п': 0.003115264797507788, 'e': 0.003115264797507788, 'Б': 0.0020768431983385254, 'K': 0.0020768431983385254, 'б': 0.024922110380062305, 'l': 0.003115264797507788, 'M':
0.006230529595015576, 'З': 0.004153686396677051, 'р': 0.003115264797507788, 'ш': 0.004153686396677051, 'ч': 0.003115264797507788, 'A': 0.0010384215991692627, 'i': 0.001038421
5991692627, 'л': 0.0010384215991692627}
Entropy: 4.54356941513016
```

Для другого файлу

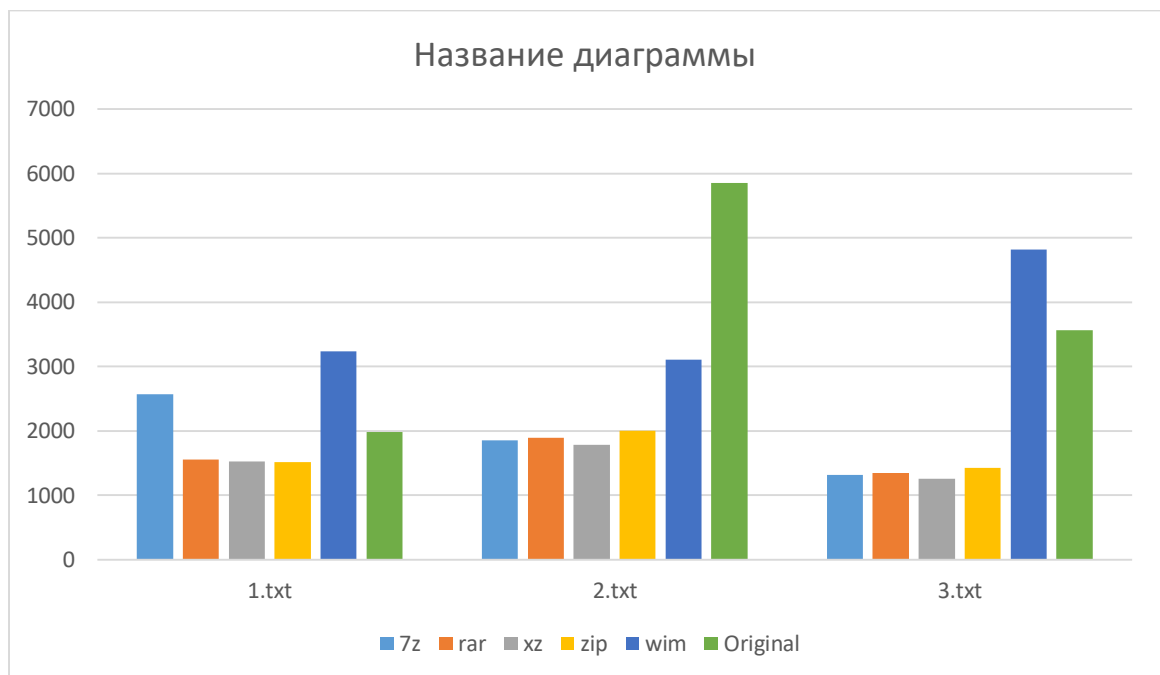
Ta

3. 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).

4. 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)

	1.txt	2.txt	3.txt
7z	3 566 байт	1 849 байт	1 318 байт
rar	559 байт	1 894 байт	1 346 байт
xz	528 байт	1 788 байт	1 256 байт

zip	515 байт	2 005 байт	1 427 байт
wim	3 236 байт	7 106 байт	4 820 байт
Original	1 982 байт	5 852 байт	3 566 байт
Information Size	785	2120	1340



Як видно з графіку розміри файлів відрізняються від початкового розміру, що пояснюється використанням алгоритмів архіваторів повторюваного тексту та через те, що формула для підрахунку кількості інформації не враховує частоту появи символів та передбачення наступної частини тексту. З статистики можемо побачити що одним з оптимальних архіваторів є xz

2 Частина

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

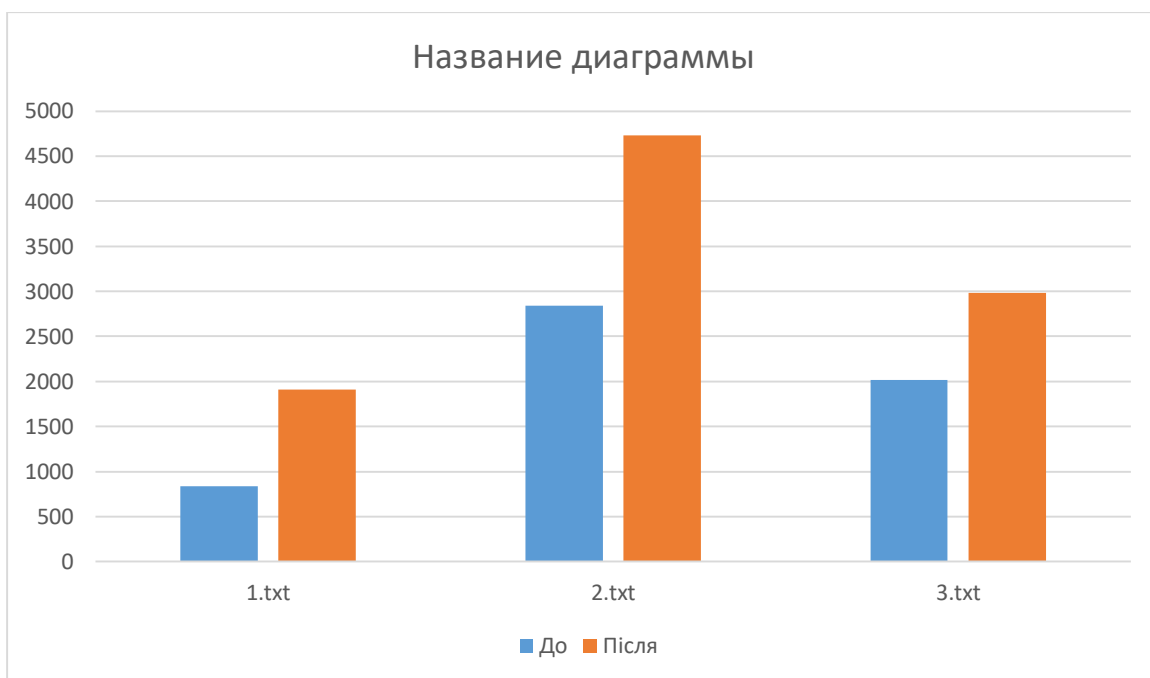
b. Порівняйте отримане значення з кількістю інформації вихідного файлу та base64-закодованого файлу

c. Зробіть висновки з отриманого результату

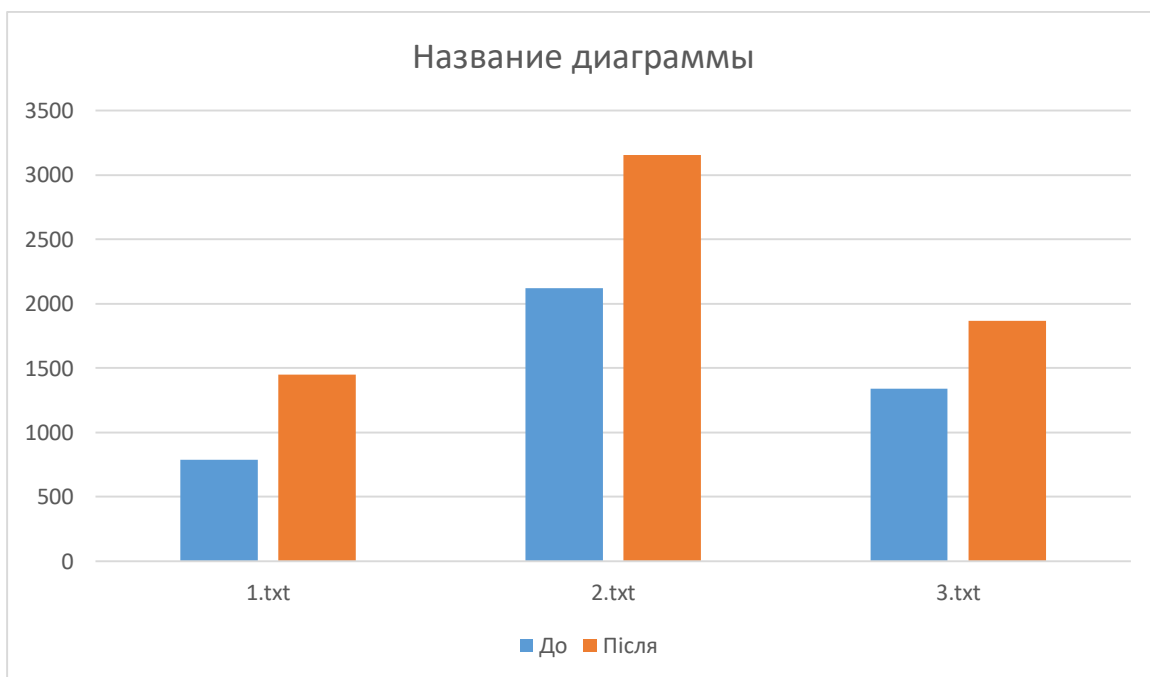
```
180: 1
250: 1
173: 1
114: 1
180: 1
18: 1
16: 1
143: 1
129: 1
13: 2
freq: {253: 0.0036231884057971015, 55: 0.009057971014492754, 122: 0.005434782608695652, 88: 0.009057971014492754, 90: 0.007246376811594203, 0: 0.034420289855072464, 1: 0.012681159420289856, 105: 0.009
057971014492754, 34: 0.009057971014492754, 222: 0.0036231884057971015, 54: 0.009057971014492754, 2: 0.016811594202898557, 39: 0.005434782608695652, 33: 0.005434782608695652, 151: 0.0036231884057971015,
214: 0.0036231884057971015, 221: 0.000957971014492754, 61: 0.0036231884057971015, 144: 0.009057971014492754, 233: 0.014420289855072464, 63: 0.0036231884057971015, 119: 0.007246376811594203, 174: 0.0009
57971014492754, 211: 0.0036231884057971015, 201: 0.0036231884057971015, 209: 0.001811594202898557, 106: 0.005434782608695652, 160: 0.012681159420289856, 30: 0.001811594202898557, 96: 0.00362318840579
71015, 240: 0.005434782608695652, 17: 0.001811594202898557, 238: 0.007246376811594203, 38: 0.0036231884057971015, 220: 0.005434782608695652, 78: 0.001811594202898557, 148: 0.010869565217391304, 89: 0
.005434782608695652, 193: 0.005434782608695652, 59: 0.007246376811594203, 61: 0.001811594202898557, 234: 0.010869565217391304, 239: 0.007246376811594203, 152: 0.0036231884057971015, 203: 0.00362318840
57971015, 108: 0.0036231884057971015, 198: 0.005434782608695652, 147: 0.005434782608695652, 11: 0.007246376811594203, 86: 0.0036231884057971015, 58: 0.001811594202898557, 49: 0.005434782608695652, 74
: 0.007246376811594203, 133: 0.0036231884057971015, 238: 0.0036231884057971015, 235: 0.0036231884057971015, 60: 0.009057971014492754, 168: 0.0036231884057971015, 21: 0.0018
11594202898557, 172: 0.0036231884057971015, 185: 0.009057971014492754, 67: 0.007246376811594203, 38: 0.0036231884057971015, 104: 0.005434782608695652, 104: 0.0036231884057971015, 170: 0.001811594202898557, 50: 0.00362318840579
71015, 255: 0.005434782608695652, 156: 0.005434782608695652, 9: 0.007246376811594203, 208: 0.0036231884057971015, 46: 0.007246376811594203, 246: 0.0036231884057971015, 242: 0.0036231884057971015, 76: 0.007246376811594203, 36: 0.0036231884057971015, 115: 0.0036231884057971015, 99: 0.009057971014492754, 37: 0.005434782608695652, 164: 0.005434782608695652, 56: 0.007246376811594203, 70: 0.00362318840579
71015, 165: 0.0036231884057971015, 28: 0.007246376811594203, 117: 0.005434782608695652, 221: 0.007246376811594203, 235: 0.001811594202898557, 72: 0.001811594202898557, 98: 0.0036231884057971015, 142: 0
.0036231884057971015, 216: 0.007246376811594203, 216: 0.0036231884057971015, 177: 0.012681159420289856, 35: 0.0036231884057971015, 154: 0.0036231884057971015, 158: 0.0036231884057971015, 199: 0.0054347
82608695652, 181: 0.001811594202898557, 63: 0.0036231884057971015, 145: 0.007246376811594203, 31: 0.001811594202898557, 72: 0.005434782608695652, 166: 0.005434782608695652, 91: 0.001811594202898557, 166: 0.005434782608695652, 102: 0.007246376811594203, 229: 0.007246376811594203, 102: 0.001811594202898557, 49: 0.007246376811594203, 239: 0.0036231884057971015, 189: 0.003
6231884057971015, 40: 0.001811594202898557, 45: 0.005434782608695652, 3: 0.0036231884057971015, 169: 0.001811594202898557, 132: 0.005434782608695652, 159: 0.009057971014492754, 22: 0.005434782608695652, 157: 0.009057971014492754, 44: 0.005434782608695652, 8: 0.0036231884057971015, 51: 0.007246376811594203, 192: 0.005434782608695652, 138: 0.001811594202898557, 135: 0.0036231884057971015, 128: 0.0036231884057971015, 238: 0.0036231884057971015, 171: 0.005434782608695652, 215: 0.001811594202898557, 210: 0.007246376811594203, 176: 0.005434782608695652, 9
7: 0.005434782608695652, 247: 0.009057971014492754, 101: 0.001811594202898557, 107: 0.0036231884057971015, 195: 0.0036231884057971015, 140: 0.001811594202898557, 140: 0.001811594202898557, 68: 0.003
6231884057971015, 197: 0.005434782608695652, 161: 0.001811594202898557, 23: 0.0036231884057971015, 71: 0.005434782608695652, 202: 0.0036231884057971015, 271: 0.0036231884057971015, 82: 0.0036231884057
971015, 12: 0.001811594202898557, 41: 0.0036231884057971015, 180: 0.0036231884057971015, 169: 0.001811594202898557, 83: 0.001811594202898557, 7: 0.001811594202898557, 251: 0.001811594202898557, 25
: 0.0036231884057971015, 227: 0.0036231884057971015, 175: 0.005434782608695652, 118: 0.0036231884057971015, 112: 0.005434782608695652, 245: 0.001811594202898557, 146: 0.0036231884057971015, 124: 0.0036231884057971015, 81: 0.001811594202898557, 57: 0.001811594202898557, 62: 0.001811594202898557, 151: 0.001811594202898557, 212: 0.001811594202898557, 75: 0.0036231884057971015, 184: 0.0036231884057971015, 81: 0.001811594202898557, 32: 0.005434782608695652, 134: 0.001811594202898557, 208: 0.001811594202898557, 186: 0.0036231884057971015, 26: 0.0036231884057971015, 121: 0.001811594202898557, 64: 0.001811594202898557, 64: 0.001811594202898557, 179: 0.003623188405
7971015, 87: 0.001811594202898557, 111: 0.001811594202898557, 153: 0.0036231884057971015, 197: 0.0036231884057971015, 120: 0.001811594202898557, 120: 0.001811594202898557, 236: 0.001811594202898557, 126: 0.001811594202898557, 126: 0.001811594202898557, 173: 0.005434782608695652, 157: 0.005434782608695652, 123: 0.001811594202898557, 252: 0.001811594202898557, 170: 0.001811594202898557, 108: 0.001811594202898557, 250: 0.001811594202898557, 173:
```

```
149: 5
129: 4
109: 4
86: 4
52: 4
87: 1
freq: {253: 0.004807692307692308, 55: 0.00541880341880342, 122: 0.00541880341880342, 88: 0.004807692307692308, 90: 0.0058766837668376, 0: 0.011217948717948718, 1: 0.0080128051280512, 105: 0.0042
73504273504274, 34: 0.00541880341880342, 222: 0.0037393162393162395, 54: 0.00267904170904171, 2: 0.0037393162393162395, 33: 0.0080128051280512, 39: 0.003205128051280512, 44: 0.003205128051280512, 45: 0.003205128051280512, 63: 0.003205128051280512, 67: 0.003205128051280512, 68: 0.003205128051280512, 72: 0.003205128051280512, 73: 0.003205128051280512, 74: 0.003205128051280512, 75: 0.003205128051280512, 76: 0.003205128051280512, 77: 0.003205128051280512, 78: 0.003205128051280512, 79: 0.003205128051280512, 80: 0.003205128051280512, 81: 0.003205128051280512, 82: 0.003205128051280512, 83: 0.003205128051280512, 84: 0.003205128051280512, 85: 0.003205128051280512, 86: 0.003205128051280512, 87: 0.003205128051280512, 88: 0.003205128051280512, 89: 0.003205128051280512, 90: 0.003205128051280512, 91: 0.003205128051280512, 92: 0.003205128051280512, 93: 0.003205128051280512, 94: 0.003205128051280512, 95: 0.003205128051280512, 96: 0.003205128051280512, 97: 0.003205128051280512, 98: 0.003205128051280512, 99: 0.003205128051280512, 100: 0.003205128051280512, 101: 0.003205128051280512, 102: 0.003205128051280512, 103: 0.003205128051280512, 104: 0.003205128051280512, 105: 0.003205128051280512, 106: 0.003205128051280512, 107: 0.003205128051280512, 108: 0.003205128051280512, 109: 0.003205128051280512, 110: 0.003205128051280512, 111: 0.003205128051280512, 112: 0.003205128051280512, 113: 0.003205128051280512, 114: 0.003205128051280512, 115: 0.003205128051280512, 116: 0.003205128051280512, 117: 0.003205128051280512, 118: 0.003205128051280512, 119: 0.003205128051280512, 120: 0.003205128051280512, 121: 0.003205128051280512, 122: 0.003205128051280512, 123: 0.003205128051280512, 124: 0.003205128051280512, 125: 0.003205128051280512, 126: 0.003205128051280512, 127: 0.003205128051280512, 128: 0.003205128051280512, 129: 0.003205128051280512, 130: 0.003205128051280512, 131: 0.003205128051280512, 132: 0.003205128051280512, 133: 0.003205128051280512, 134: 0.003205128051280512, 135: 0.003205128051280512, 136: 0.003205128051280512, 137: 0.003205128051280512, 138: 0.003205128051280512, 139: 0.003205128051280512, 140: 0.003205128051280512, 141: 0.003205128051280512, 142: 0.003205128051280512, 143: 0.003205128051280512, 144: 0.003205128051280512, 145: 0.003205128051280512, 146: 0.003205128051280512, 147: 0.003205128051280512, 148: 0.003205128051280512, 149: 0.003205128051280512, 150: 0.003205128051280512, 151: 0.003205128051280512, 152: 0.003205128051280512, 153: 0.003205128051280512, 154: 0.003205128051280512, 155: 0.003205128051280512, 156: 0.003205128051280512, 157: 0.003205128051280512, 158: 0.003205128051280512, 159: 0.003205128051280512, 160: 0.003205128051280512, 161: 0.003205128051280512, 162: 0.003205128051280512, 163: 0.003205128051280512, 164: 0.003205128051280512, 165: 0.003205128051280512, 166: 0.003205128051280512, 167: 0.003205128051280512, 168: 0.003205128051280512, 169: 0.003205128051280512, 170: 0.003205128051280512, 171: 0.003205128051280512, 172: 0.003205128051280512, 173: 0.003205128051280512, 174: 0.003205128051280512, 175: 0.003205128051280512, 176: 0.003205128051280512, 177: 0.003205128051280512, 178: 0.003205128051280512, 179: 0.003205128051280512, 180: 0.003205128051280512, 181: 0.003205128051280512, 182: 0.003205128051280512, 183: 0.003205128051280512, 184: 0.003205128051280512, 185: 0.003205128051280512, 186: 0.003205128051280512, 187: 0.003205128051280512, 188: 0.003205128051280512, 189: 0.003205128051280512, 190: 0.003205128051280512, 191: 0.003205128051280512, 192: 0.003205128051280512, 193: 0.003205128051280512, 194: 0.003205128051280512, 195: 0.003205128051280512, 196: 0.003205128051280512, 197: 0.003205128051280512, 198: 0.003205128051280512, 199: 0.003205128051280512, 200: 0.003205128051280512, 201: 0.003205128051280512, 202: 0.003205128051280512, 203: 0.003205128051280512, 204: 0.003205128051280512, 205: 0.003205128051280512, 206: 0.003205128051280512, 207: 0.003205128051280512, 208: 0.003205128051280512, 209: 0.003205128051280512, 210: 0.003205128051280512, 211: 0.003205128051280512, 212: 0.003205128051280512, 213: 0.003205128051280512, 214: 0.003205128051280512, 215: 0.003205128051280512, 216: 0.003205128051280512, 217: 0.003205128051280512, 218: 0.003205128051280512, 219: 0.003205128051280512, 220: 0.003205128051280512, 221: 0.003205128051280512, 222: 0.003205128051280512, 223: 0.003205128051280512, 224: 0.003205128051280512, 225: 0.003205128051280512, 226: 0.003205128051280512, 227: 0.003205128051280512, 228: 0.003205128051280512, 229: 0.003205128051280512, 230: 0.003205128051280512, 231: 0.003205128051280512, 232: 0.003205128051280512, 233: 0.003205128051280512, 234: 0.003205128051280512, 235: 0.003205128051280512, 236: 0.003205128051280512, 237: 0.003205128051280512, 238: 0.003205128051280512, 239: 0.003205128051280512, 240: 0.003205128051280512, 241: 0.003205128051280512, 242: 0.003205128051280512, 243: 0.003205128051280512, 244: 0.003205128051280512, 245: 0.003205128051280512, 246: 0.003205128051280512, 247: 0.003205128051280512, 248: 0.003205128051280512, 249: 0.003205128051280512, 250: 0.003205128051280512, 251: 0.003205128051280512, 252: 0.003205128051280512, 253: 0.003205128051280512, 254: 0.003205128051280512, 255: 0.003205128051280512, 256: 0.003205128051280512, 257: 0.003205128051280512, 258: 0.003205128051280512, 259: 0.003205128051280512, 260: 0.003205128051280512, 261: 0.003205128051280512, 262: 0.003205128051280512, 263: 0.003205128051280512, 264: 0.003205128051280512, 265: 0.003205128051280512, 266: 0.003205128051280512, 267: 0.003205128051280512, 268: 0.003205128051280512, 269: 0.003205128051280512, 270: 0.003205128051280512, 271: 0.003205128051280512, 272: 0.003205128051280512, 273: 0.003205128051280512, 274: 0.003205128051280512, 275: 0.003205128051280512, 276: 0.003205128051280512, 277: 0.003205128051280512, 278: 0.003205128051280512, 279: 0.003205128051280512, 280: 0.003205128051280512, 281: 0.003205128051280512, 282: 0.003205128051280512, 283: 0.003205128051280512, 284: 0.003205128051280512, 285: 0.003205128051280512, 286: 0.003205128051280512, 287: 0.003205128051280512, 288: 0.003205128051280512, 289: 0.003205128051280512, 290: 0.003205128051280512, 291: 0.003205128051280512, 292: 0.003205128051280512, 293: 0.003205128051280512, 294: 0.003205128051280512, 295: 0.003205128051280512, 296: 0.003205128051280512, 297: 0.003205128051280512, 298: 0.003205128051280512, 299: 0.003205128051280512, 300: 0.003205128051280512, 301: 0.003205128051280512, 302: 0.003205128051280512, 303: 0.003205128051280512, 304: 0.003205128051280512, 305: 0.003205128051280512, 306: 0.003205128051280512, 307: 0.003205128051280512, 308: 0.00320512805128051
```

Розмір файлу до і після кодування

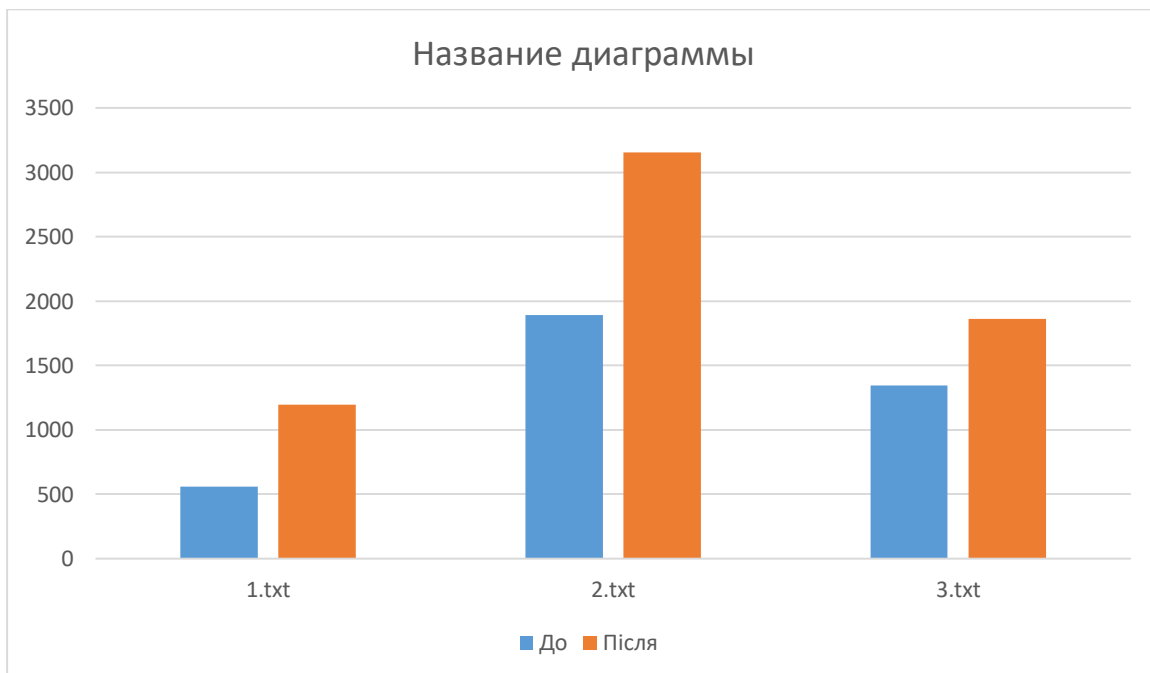


Кількість інформації у файлах до і після кодування



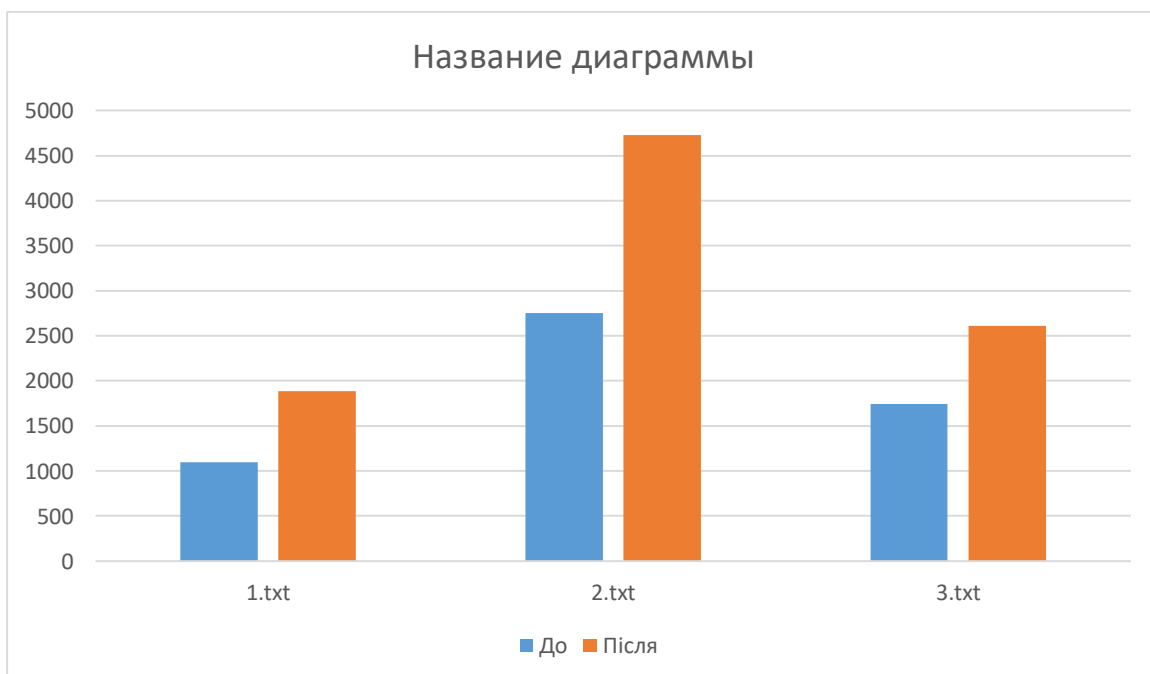
Найкращим виявився архіватор xz

Розмір файлів до і після архівування



Після кодування розмір файлів зростає приблизно у 1.43 рази

Кількість інформації у файлах до і після архівування:



Кількість інформації зросла приблизно в 1.33 рази

Висновок:

Файли обрані дуже майстерно і я зрозумів, що саме про ентропію мені казала вчителька молодших класів коли говорила про багатобарвність української мови

Під час виконання цієї лабораторної роботи було досліджено імовірнісні параметри української мови для оцінки кількості інформації текстів, вплив різних методів кодування інформації на її кількість; було ознайомлено з деякими термінами (такими як середня ентропія нерівноймовірного алфавіту); розглянуто алгоритм кодування інформації Base64; закріплено навички роботи з мовою програмування Python завдяки створенню необхідної програми