Вариант 1

Задание:

1. Подобрать англоязычный текстдлиной примерно 2 страницы (около 8000 символов), при этом для удобства все буквы переделать в строчные, оставив не более 32 различных символов.
2. Провести статистический анализ по частоте букв и по частоте пар букв
3. На основе этих статистик построить коды Хаффмана, закодировать текст этими кодами, сравнить количество бит закодированного текста с равномерными (5-ти битовыми) кодами, по формуле Шенона найти количество информации и сравнить с кодами Хаффмана.
4. Закодировать этот же текст по алгоритму LZW, сравнить количество бит закодированного текста с равномерными кодами и кодами Хаффмана.

Towards the end of the summer of 1969 – a few weeks after the moon landings, a few days after Woodstock, and a month before the first broadcast of Monty Python's Flying Circus – a large grey metal box was delivered to the office of Leonard Kleinrock, a professor at the University of California in Los Angeles. It was the same size and shape as a household refrigerator, and outwardly, at least, it had about as much charm. But Kleinrock was thrilled: a photograph from the time shows him standing beside it, in requisite late-60s brown tie and brown trousers, beaming like a proud father.

Had he tried to explain his excitement to anyone but his closest colleagues, they probably wouldn't have understood. The few outsiders who knew of the box's existence couldn't even get its name right: it was an IMP, or "interface message processor", but the year before, when a Boston company had won the contract to build it, its local senator, Ted Kennedy, sent a telegram praising its ecumenical spirit in creating the first "interfaith message processor". Needless to say, though, the box that arrived outside Kleinrock's office wasn't a machine capable of fostering understanding among the great religions of the world. It was much more important than that.

It's impossible to say for certain when the internet began, mainly because nobody can agree on what, precisely, the internet is. (This is only partly a philosophical question: it is also a matter of egos, since several of the people who made key contributions are anxious to claim the credit.) But 29 October 1969 – 40 years ago next week – has a strong claim for being, as Kleinrock puts it today, "the day the infant internet uttered its first words". At 10.30pm, as Kleinrock's fellow professors and students crowded around, a computer was connected to the IMP, which made contact with a second IMP, attached to a second computer, several hundred miles away at the Stanford Research Institute, and an undergraduate named Charley Kline tapped out a message. Samuel Morse, sending the first telegraph message 125 years previously, chose the portentous phrase: "What hath God wrought?" But Kline's task was to log in remotely from LA to the Stanford machine, and there was no opportunity for portentousness: his instructions were to type the command LOGIN.

To say that the rest is history is the emptiest of cliches – but trying to express the magnitude of what began that day, and what has happened in the decades since, is an undertaking that quickly exposes the limits of language. It's interesting to compare how much has changed in computing and the internet since 1969 with, say, how much has changed in world politics. Consider even the briefest summary of how much has happened on the global stage since 1969: the Vietnam war ended; the cold war escalated then declined; the Berlin Wall fell; communism collapsed; Islamic fundamentalism surged. And yet nothing has quite the power to make people in their 30s, 40s or 50s feel very old indeed as reflecting upon the growth of the internet and the world wide web. Twelve years after Charley Kline's first message on the Arpanet, as it was then known, there were still only 213 computers on the network; but 14 years after that, 16 million people were online, and email was beginning to change the world; the first really usable web browser wasn't launched until 1993, but by 1995 we had Amazon, by 1998 Google, and by 2001, Wikipedia, at which point there were 513 million people online. Today the figure is more like 1.7 billion.

Unless you are 15 years old or younger, you have lived through the dotcom bubble and bust, the birth of Friends Reunited and Craigslist and eBay and Facebook and Twitter, blogging, the browser wars, Google Earth, filesharing controversies, the transformation of the record industry, political campaigning, activism and campaigning, the media, publishing, consumer banking, the pornography industry, travel agencies, dating and retail; and unless you're a specialist, you've probably only been following the most attention-grabbing developments. Here's one of countless statistics that are liable to induce feelings akin to vertigo: on New Year's Day 1994 – only yesterday, in other words – there were an estimated 623 websites. In total. On the whole internet. "This isn't a matter of ego or crowing," says Steve Crocker, who was present that day at UCLA in 1969, "but there has not been, in the entire history of mankind, anything that has changed so dramatically as computer communications, in terms of the rate of change."

Looking back now, Kleinrock and Crocker are both struck by how, as young computer scientists, they were simultaneously aware that they were involved in something momentous and, at the same time, merely addressing a fairly mundane technical problem. On the one hand, they were there because of the Russian Sputnik satellite launch, in 1957, which panicked the American defence establishment, prompting Eisenhower to channel millions of dollars into scientific research, and establishing Arpa, the Advanced Research Projects Agency, to try to win the arms technology race. The idea was "that we would not get surprised again," said Robert Taylor, the Arpa scientist who secured the money for the Arpanet, persuading the agency's head to give him a million dollars that had been earmarked for ballistic missile research. With another pioneer of the early internet, JCR Licklider, Taylor co-wrote the paper, "The Computer As A Communication Device", which hinted at what was to come. "In a few years, men will be able to communicate more effectively through a machine than face to face," they declared. "That is rather a startling thing to say, but it is our conclusion."

On the other hand, the breakthrough accomplished that night in 1969 was a decidedly down-to-earth one. The Arpanet was not, in itself, intended as some kind of secret weapon to put the Soviets in their place: it was simply a way to enable researchers to access computers remotely, because computers were still vast and expensive, and the scientists needed a way to share resources. (The notion that the network was designed so that it would survive a nuclear attack is an urban myth, though some of those involved sometimes used that argument to obtain funding.) The technical problem solved by the IMPs wasn't very exciting, either. It was already possible to link computers by telephone lines, but it was glacially slow, and every computer in the network had to be connected, by a dedicated line, to every other computer, which meant you couldn't connect more than a handful of machines without everything becoming monstrously complex and costly. The solution, called "packet switching" – which owed its existence to the work of a British physicist, Donald Davies – involved breaking data down into blocks that could be routed around any part of the network that happened to be free, before getting reassembled at the other end.

"I thought this was important, but I didn't really think it was as challenging as what I thought of as the 'real research'," says Crocker, a genial Californian, now 65, who went on to play a key role in the expansion of the internet. "I was particularly fascinated, in those days, by artificial intelligence, and by trying to understand how people think. I thought that was a much more substantial and respectable research topic than merely connecting up a few machines. That was certainly useful, but it wasn't art."

Still, Kleinrock recalls a tangible sense of excitement that night as Kline sat down at the SDS Sigma 7 computer, connected to the IMP, and at the same time made telephone contact with his opposite number at Stanford. As his colleagues watched, he typed the letter L, to begin the word LOGIN.

"Have you got the L?" he asked, down the phone line. "Got the L," the voice at Stanford responded.

Kline typed an O. "Have you got the O?"

"Got the O," Stanford replied.

Kline typed a G, at which point the system crashed, and the connection was lost. The G didn't make it through, which meant that, quite by accident, the first message ever transmitted across the nascent internet turned out, after all, to be fittingly biblical:

"LO."

from collections import Counter  
import heapq  
  
with open("text.txt", 'r') as f:  
 text = f.read()  
text = text.lower()  
allowed\_chars = set("abcdefghijklmnopqrstuvwxyz ")  
filtered\_text = ''.join(filter(lambda x: x in allowed\_chars, text))  
  
unique\_chars = set(filtered\_text)  
print(f"Unique characters: {unique\_chars}")  
print(f"Total characters: {len(filtered\_text)}")  
  
letter\_freq = Counter(filtered\_text)  
pair\_freq = Counter(zip(filtered\_text[:-1], filtered\_text[1:]))  
  
print("Letter Frequencies:")  
for letter, freq in letter\_freq.items():  
 print(f"{letter}: {freq}")  
  
print("\nPair Frequencies:")  
for pair, freq in pair\_freq.items():  
 print(f"{pair}: {freq}")  
  
class Node:  
 def \_\_init\_\_(self, char, freq):  
 self.char = char  
 self.freq = freq  
 self.left = None  
 self.right = None  
  
 def \_\_lt\_\_(self, other):  
 return self.freq < other.freq  
  
def huffman\_coding(text):  
 frequency = Counter(text)  
 heap = [Node(char, freq) for char, freq in frequency.items()]  
 heapq.heapify(heap)  
  
 while len(heap) > 1:  
 left = heapq.heappop(heap)  
 right = heapq.heappop(heap)  
 merged = Node(None, left.freq + right.freq)  
 merged.left = left  
 merged.right = right  
 heapq.heappush(heap, merged)  
  
 root = heap[0]  
 codes = {}  
  
 def generate\_codes(node, current\_code):  
 if node.char is not None:  
 codes[node.char] = current\_code  
 return  
 generate\_codes(node.left, current\_code + "0")  
 generate\_codes(node.right, current\_code + "1")  
  
 generate\_codes(root, "")  
 return codes  
  
huffman\_codes = huffman\_coding(filtered\_text)  
encoded\_text\_huffman = ''.join(huffman\_codes[char] for char in filtered\_text)  
uniform\_bits = len(filtered\_text) \* 5  
huffman\_bits = len(encoded\_text\_huffman)  
  
print(f"Huffman encoded bits: {huffman\_bits}")  
print(f"Uniform encoded bits: {uniform\_bits}")  
  
entropy = -sum((freq / len(filtered\_text)) \* (freq / len(filtered\_text)).bit\_length() for freq in letter\_freq.values())  
print(f"Entropy (Shannon's information): {entropy}")  
  
def lzw\_encoding(text):  
 max\_table\_size = 256  
 table = {chr(i): i for i in range(max\_table\_size)}  
 code = max\_table\_size  
 p = ""  
 encoded\_output = []  
  
 for c in text:  
 pc = p + c  
 if pc in table:  
 p = pc  
 else:  
 encoded\_output.append(table[p])  
 table[pc] = code  
 code += 1  
 p = c  
  
 if p:  
 encoded\_output.append(table[p])  
  
 return encoded\_output  
  
lzw\_encoded = lzw\_encoding(filtered\_text)  
lzw\_bits = len(lzw\_encoded) \* (lzw\_encoded[-1].bit\_length() or 1)   
  
print(f"LZW encoded bits: {lzw\_bits}")

Total characters: 7834

Letter Frequencies:

t: 665

o: 462

w: 144

a: 537

r: 369

d: 236

s: 413

: 1362

h: 344

e: 804

n: 471

f: 108

u: 161

m: 161

k: 65

l: 267

i: 438

g: 143

y: 130

c: 240

b: 110

p: 133

x: 16

v: 46

z: 2

q: 5

j: 2

Pair Frequencies:

('t', 'o'): 62

('o', 'w'): 28

('w', 'a'): 40

('a', 'r'): 60

('r', 'd'): 13

('d', 's'): 6

('s', ' '): 166

(' ', 't'): 243

('t', 'h'): 193

('h', 'e'): 143

('e', ' '): 272

(' ', 'e'): 37

('e', 'n'): 56

('n', 'd'): 66

('d', ' '): 135

(' ', 'o'): 76

('o', 'f'): 37

('f', ' '): 34

(' ', 's'): 79

('s', 'u'): 8

('u', 'm'): 6

('m', 'm'): 7

('m', 'e'): 38

('e', 'r'): 106

('r', ' '): 65

(' ', ' '): 33

(' ', 'a'): 161

('a', ' '): 49

(' ', 'f'): 39

('f', 'e'): 14

('e', 'w'): 7

('w', ' '): 16

(' ', 'w'): 99

('w', 'e'): 22

('e', 'e'): 13

('e', 'k'): 2

('k', 's'): 4

('a', 'f'): 5

('f', 't'): 5

('t', 'e'): 82

(' ', 'm'): 55

('m', 'o'): 16

('o', 'o'): 7

('o', 'n'): 73

('n', ' '): 88

(' ', 'l'): 29

('l', 'a'): 21

('a', 'n'): 95

('d', 'i'): 13

('i', 'n'): 146

('n', 'g'): 62

('g', 's'): 3

(' ', 'd'): 30

('d', 'a'): 14

('a', 'y'): 23

('y', 's'): 7

('w', 'o'): 17

('o', 'd'): 6

('s', 't'): 65

('o', 'c'): 16

('c', 'k'): 19

('k', ' '): 21

('n', 't'): 67

('h', ' '): 43

(' ', 'b'): 71

('b', 'e'): 26

('e', 'f'): 9

('f', 'o'): 18

('o', 'r'): 55

('r', 'e'): 83

('f', 'i'): 14

('i', 'r'): 14

('r', 's'): 32

('t', ' '): 166

('b', 'r'): 9

('r', 'o'): 47

('o', 'a'): 1

('a', 'd'): 17

('d', 'c'): 1

('c', 'a'): 27

('a', 's'): 62

('t', 'y'): 7

('y', ' '): 90

(' ', 'p'): 49

('p', 'y'): 1

('y', 't'): 4

('h', 'o'): 29

('n', 's'): 16

('f', 'l'): 2

('l', 'y'): 32

('y', 'i'): 3

('g', ' '): 51

(' ', 'c'): 80

('c', 'i'): 17

('r', 'c'): 9

('c', 'u'): 4

('u', 's'): 21

('r', 'g'): 4

('g', 'e'): 26

(' ', 'g'): 18

('g', 'r'): 10

('e', 'y'): 11

('e', 't'): 29

('t', 'a'): 36

('a', 'l'): 34

('l', ' '): 34

('b', 'o'): 8

('o', 'x'): 3

('x', ' '): 3

('d', 'e'): 36

('e', 'l'): 26

('l', 'i'): 47

('i', 'v'): 9

('v', 'e'): 33

('e', 'd'): 70

('o', ' '): 58

('f', 'f'): 3

('i', 'c'): 37

('c', 'e'): 27

('l', 'e'): 54

('e', 'o'): 7

('n', 'a'): 9

(' ', 'k'): 17

('k', 'l'): 15

('e', 'i'): 12

('n', 'r'): 7

('p', 'r'): 17

('e', 's'): 63

('s', 's'): 24

('s', 'o'): 16

('a', 't'): 73

(' ', 'u'): 16

('u', 'n'): 26

('n', 'i'): 21

('s', 'i'): 28

('i', 't'): 56

('i', 'f'): 4

('r', 'n'): 13

('i', 'a'): 11

(' ', 'i'): 97

('l', 'o'): 20

('o', 's'): 17

('s', 'a'): 21

('a', 'm'): 17

('i', 'z'): 1

('z', 'e'): 1

('s', 'h'): 10

('h', 'a'): 71

('a', 'p'): 12

('p', 'e'): 20

(' ', 'h'): 45

('o', 'u'): 48

('s', 'e'): 41

('e', 'h'): 1

('o', 'l'): 20

('l', 'd'): 16

(' ', 'r'): 34

('f', 'r'): 5

('r', 'i'): 14

('i', 'g'): 13

('r', 'a'): 21

('u', 't'): 41

('t', 'w'): 7

('d', 'l'): 3

('e', 'a'): 38

('a', 'b'): 12

('m', 'u'): 13

('u', 'c'): 11

('c', 'h'): 46

('r', 'm'): 5

('m', ' '): 17

('b', 'u'): 17

('h', 'r'): 6

('i', 'l'): 18

('l', 'l'): 31

('p', 'h'): 11

('o', 't'): 21

('o', 'g'): 10

('o', 'm'): 33

('t', 'i'): 51

('i', 'm'): 20

('w', 's'): 3

('h', 'i'): 38

('i', 'd'): 12

('e', 'q'): 1

('q', 'u'): 5

('u', 'i'): 5

('i', 's'): 46

('w', 'n'): 7

('i', 'e'): 15

('t', 'r'): 17

('m', 'i'): 11

('i', 'k'): 4

('k', 'e'): 14

('u', 'd'): 3

('f', 'a'): 9

('r', 'h'): 1

('e', 'x'): 12

('x', 'p'): 5

('p', 'l'): 12

('a', 'i'): 17

('x', 'c'): 3

('e', 'm'): 10

('n', 'y'): 4

('y', 'o'): 10

('n', 'e'): 63

('c', 'l'): 9

('c', 'o'): 52

('a', 'g'): 17

('g', 'u'): 5

('u', 'e'): 4

('o', 'b'): 9

('b', 'a'): 8

('b', 'l'): 21

('u', 'l'): 10

('d', 'n'): 5

('a', 'v'): 6

('t', 's'): 19

('w', 'h'): 22

('k', 'n'): 2

('x', 's'): 1

('x', 'i'): 3

('n', 'c'): 15

('e', 'v'): 12

(' ', 'n'): 24

('g', 'h'): 13

('h', 't'): 7

('m', 'p'): 30

('p', ' '): 5

('r', 'f'): 2

('a', 'c'): 24

(' ', 'y'): 19

('y', 'e'): 10

('p', 'a'): 17

('c', 't'): 16

('n', 'n'): 9

('d', 'y'): 3

('e', 'g'): 8

('e', 'c'): 30

('s', 'p'): 5

('p', 'i'): 3

('c', 'r'): 12

('u', 'g'): 10

('r', 'r'): 1

('s', 'n'): 6

('m', 'a'): 24

('g', 'i'): 9

('i', 'o'): 21

('r', 'l'): 10

('p', 'o'): 18

('r', 't'): 19

('i', 'b'): 5

('g', 'a'): 3

('n', 'l'): 10

('a', 'u'): 5

('n', 'o'): 14

('t', 'l'): 4

('o', 'p'): 10

(' ', 'q'): 4

('l', 's'): 2

('t', 't'): 11

('g', 'o'): 11

('n', 'x'): 1

('x', 't'): 1

('p', 'u'): 17

('n', 'f'): 6

('p', 'm'): 2

('t', 'u'): 5

('w', 'd'): 1

('w', 'i'): 13

('h', 'u'): 1

('d', 'r'): 3

('a', 'w'): 2

('d', 'u'): 4

('u', 'a'): 3

('p', 'p'): 6

('v', 'i'): 7

('s', 'l'): 6

('w', 'r'): 2

('s', 'k'): 2

('r', 'u'): 3

('y', 'p'): 4

('r', 'y'): 13

('p', 't'): 2

('g', 'n'): 4

('a', 'k'): 6

('k', 'i'): 8

('c', 's'): 2

('g', 'l'): 5

(' ', 'v'): 6

('t', 'n'): 2

('s', 'c'): 7

('s', 'm'): 4

('p', 's'): 2

('c', ' '): 4

('f', 'u'): 4

('u', 'r'): 10

('u', 'p'): 2

('w', 't'): 1

('e', 'b'): 5

('b', ' '): 2

('l', 'v'): 5

('r', 'p'): 6

('r', 'k'): 6

('b', 'y'): 10

('a', 'z'): 1

('z', 'o'): 1

('i', 'p'): 1

('o', 'i'): 3

('b', 'i'): 4

('n', 'u'): 3

('u', ' '): 5

('d', 'o'): 8

('t', 'c'): 3

('u', 'b'): 3

('b', 'b'): 2

('e', 'u'): 1

('o', 'k'): 2

('g', 'g'): 1

('o', 'v'): 2

('s', 'f'): 1

('n', 'k'): 5

('h', 'y'): 2

('u', 'v'): 1

('b', 's'): 2

('m', 's'): 2

('l', 't'): 1

('n', 'v'): 3

('v', 'o'): 4

('d', 'd'): 1

('h', 'n'): 3

('h', 'm'): 1

('n', 'h'): 2

('d', 'v'): 1

('v', 'a'): 2

('o', 'j'): 1

('j', 'e'): 1

('c', 'y'): 2

('g', 'y'): 1

('y', 'l'): 2

(' ', 'j'): 1

('j', 'c'): 1

('l', 'u'): 2

('k', 't'): 1

('c', 'c'): 3

('o', 'e'): 1

('l', 'f'): 1

('r', 'v'): 1

('r', 'b'): 1

('m', 'y'): 1

('b', 't'): 1

('l', 'r'): 1

('e', 'p'): 3

('d', 'f'): 1

('s', 'w'): 1

('m', 'b'): 2

('i', ' '): 5

('s', 'd'): 1

('g', 'm'): 1

('d', 'k'): 2

('s', 'y'): 1

Huffman encoded bits: 32449

Uniform encoded bits: 39170

Entropy (Shannon's information): -0.07539069076234602

LZW encoded bits: 20475