1.

1. 4.184582141829053
2. 7.7769156962060375
3. 10.857260892024893

We can see that the entropy is higher and higher

2.

I have done the same pre-processing as Q1. In my implementation of Q2 and Q3, all the punctuation and numbers have been excluded and all the characters are lowercase. In this question I assume the code of each character in the original file is ASCII, the length of each character is 8 bits.

Unigram\_Huffman:

compress 7.731055540705078e-05s

decompress 0.6217391447045734

compress radio 0.5264967113356406

sum code length 1283957

Bigram\_Huffman:

compress 0.00180585787913079

decompress 0.5210479810457858

compress radio 0.4877605917955615

sum code length 1189492

The optimal code length based one result of Q1 for unigram is 1275607.0972044594, for bigram is 1185341.9365843318, they are very near to the code length of result compressed by Huffman code.

3.

compress 0.24576405620769115

decompress 0.07969551310687417

compress radio 0.5873046074105663

sum code length 1432248

In Q3, I assume that each pair of output (index,lastch) is 24 bits long, 16 for the index, 8 for the character.

Then we can find that the sum length is larger than result from Q1. But it still have a good compression radio. In fact, we do not need to let every length of a [index] to be 16 bit, then the length of compressed code will be smaller and give a better compression radio.

4.

A)

They are 2 ways that the length are 3, each have possibility 1/8

2\*c(1,3) = 6 ways that have length 4, each have possibility 1/16

2\*c(2,4) = 12 ways that have length 5, each have possibility 1/32

H(X) = -2\*1/8log1/8 + -6\*1/16log1/16 + -12\*1/32log1/32 = 33/8

H(Y) = -1/4\*log1/4 + -2\*3/8log3/8 = 1.56

We know that H(Y|X) = 0

Then I(X;Y) = H(Y) = 1.56

H(X|Y) = H(X)-H(Y) = 2.56

We can easily know that D(q1||Y) is 0 because Y is independent of A.

Because pa(x) is p(X|A=wins) = 2\*p(X=x), D(pa||X) = 1

B)

With chain rule:

We have H(Y|Z) = H(Y ), then:

H(X,Y,Z) = H(X|Y,Z) + H(Y|Z) + H(Z) >= H(Y|Z) + H(Z) = H(Y ) + H(Z) = 2.

5.

The entropy is 9.71625847652071.

It is near the result of trigram in Q1.