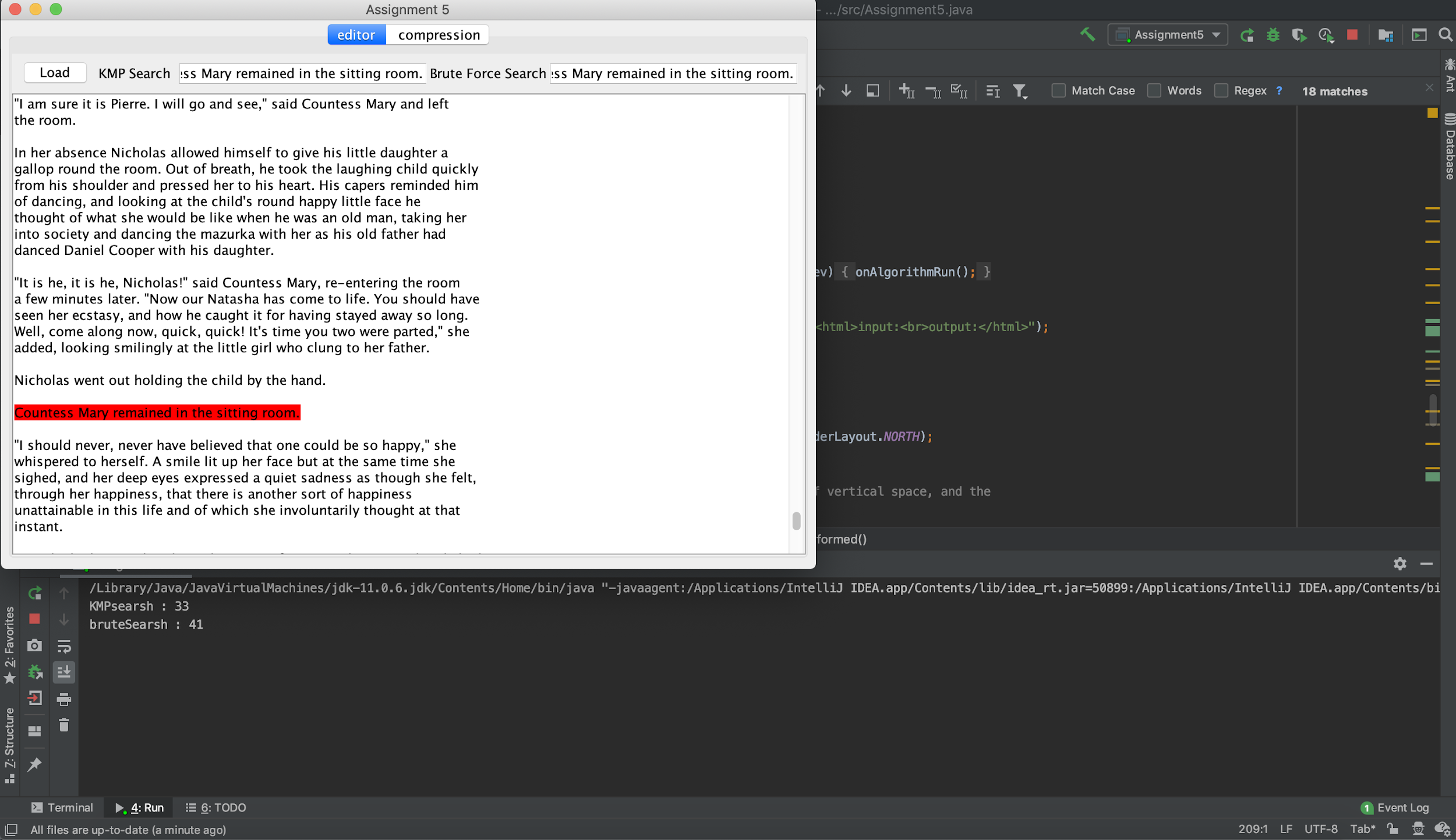
**Question 1 :**

I used System.*currentTimeMillis*() to measure the time taken . And the KMP algorithm is taking less time to search than the brute force algorithm .



**Question 2 :**

**binary tree:**

:111010

:111001

:110

!:1110000111

":11111010

':111000010

(:1111101111111

):011000111000

\*:11111011010010

,:1111111

-:100101001

.:1110001

/:011000111001010111110

0:111110110100001

1:11111011010001

2:111110110100000

3:0110001110010111

4:01100011100101010

5:0110001110010100

6:0110001110010110

7:01100011100111110

8:01100011100100

9:01100011100111101

::111000001001

;:111110110101

=:011000111001010111111

?:1001010100

A:011000110

B:1110000001

C:01100010000

D:11111011000

E:01100010001

F:11100000101

G:111110111101

H:1110000011

I:100101011

J:11111011010011

K:111110111100

L:1111101111110

M:1001010101

N:1110000000

O:01100011101

P:011000101

Q:01100011100111111

R:11111011011

S:0110001111

T:100101000

U:01100011100110

V:111000001000

W:0110001001

X:01100011100111100

Y:111110111110

Z:011000111001110

à:0110001110010101110

a:1000

b:1111100

c:101111

d:10110

ä:0110001110010101111010

e:000

f:100110

g:100100

h:0011

é:0110001110010101111011

i:0100

j:11111011001

ê:011000111001010110

k:0110000

l:01101

m:101110

n:0101

o:0111

p:1111110

q:11111011101

r:11110

s:0010

t:1010

u:111011

v:1001011

w:100111

x:1110000110

y:011001

z:11111011100

﻿:011000111001010111100

The size after huffman coding is :1848598 bytes

**Question 3:**

for war\_and\_peace.txt:

input length: 3258246 bytes

output length: 1848598 bytes

there are 1409648 bytes has been reducted which is 43.26% of the original input

for taisho.txt :

input length: 3649944 bytes

output length: 1542656 bytes

there are 2107288 bytes has been reducted which is 57.73% of the original input

for Pi.txt :

input length: 1010003 bytes

output length: 443632 bytes

there are 566371 bytes has been reducted which is 56.07% of the original input

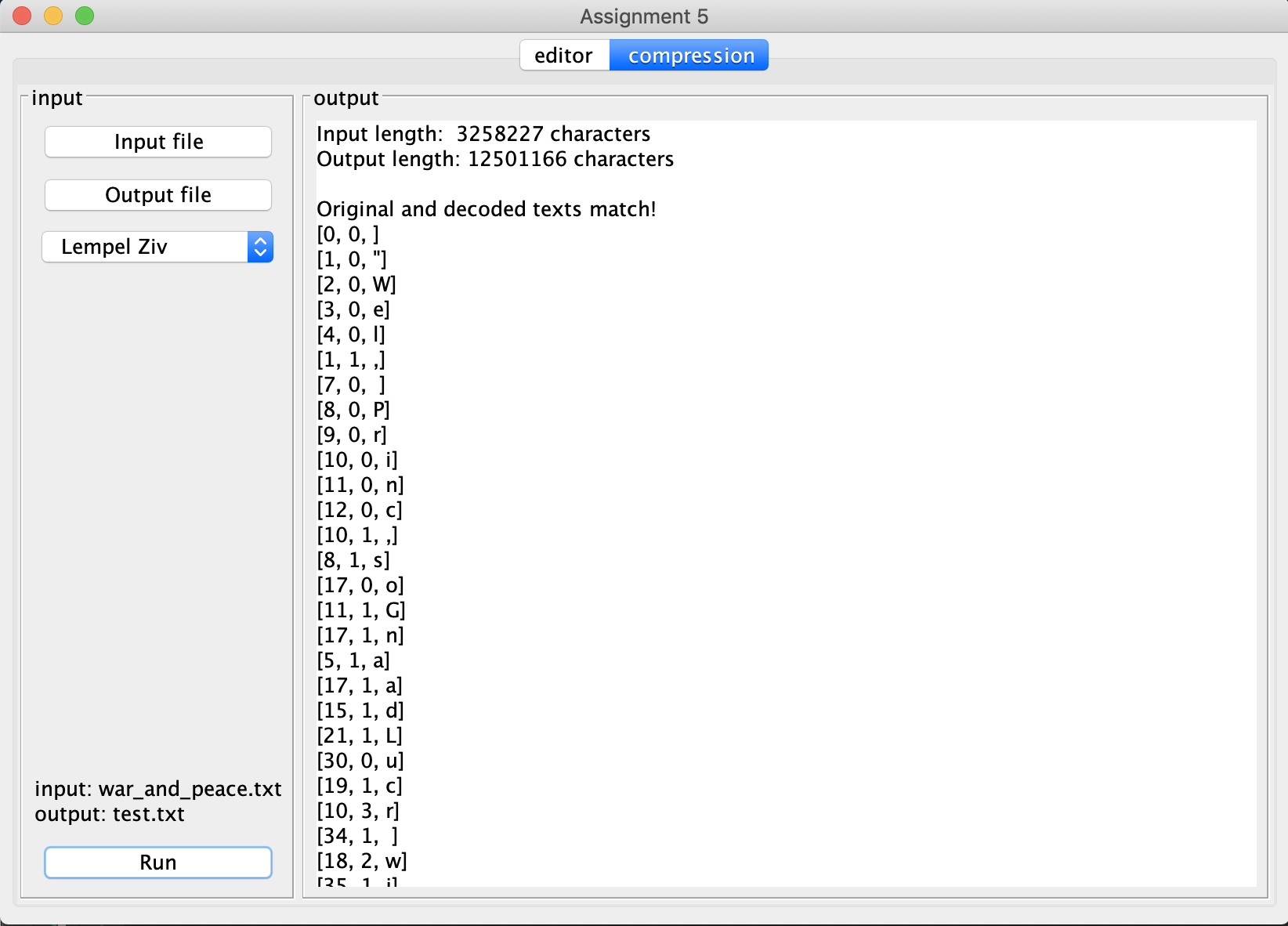
the taisho.txt has the best compression, which is 57 percent less memory taking before compression , I think it is because the Chinese characters can be taking more memory occupy. Pi is the second best,this is because the Pi only contains number so the binary tree is smaller because no letter have to be compress.

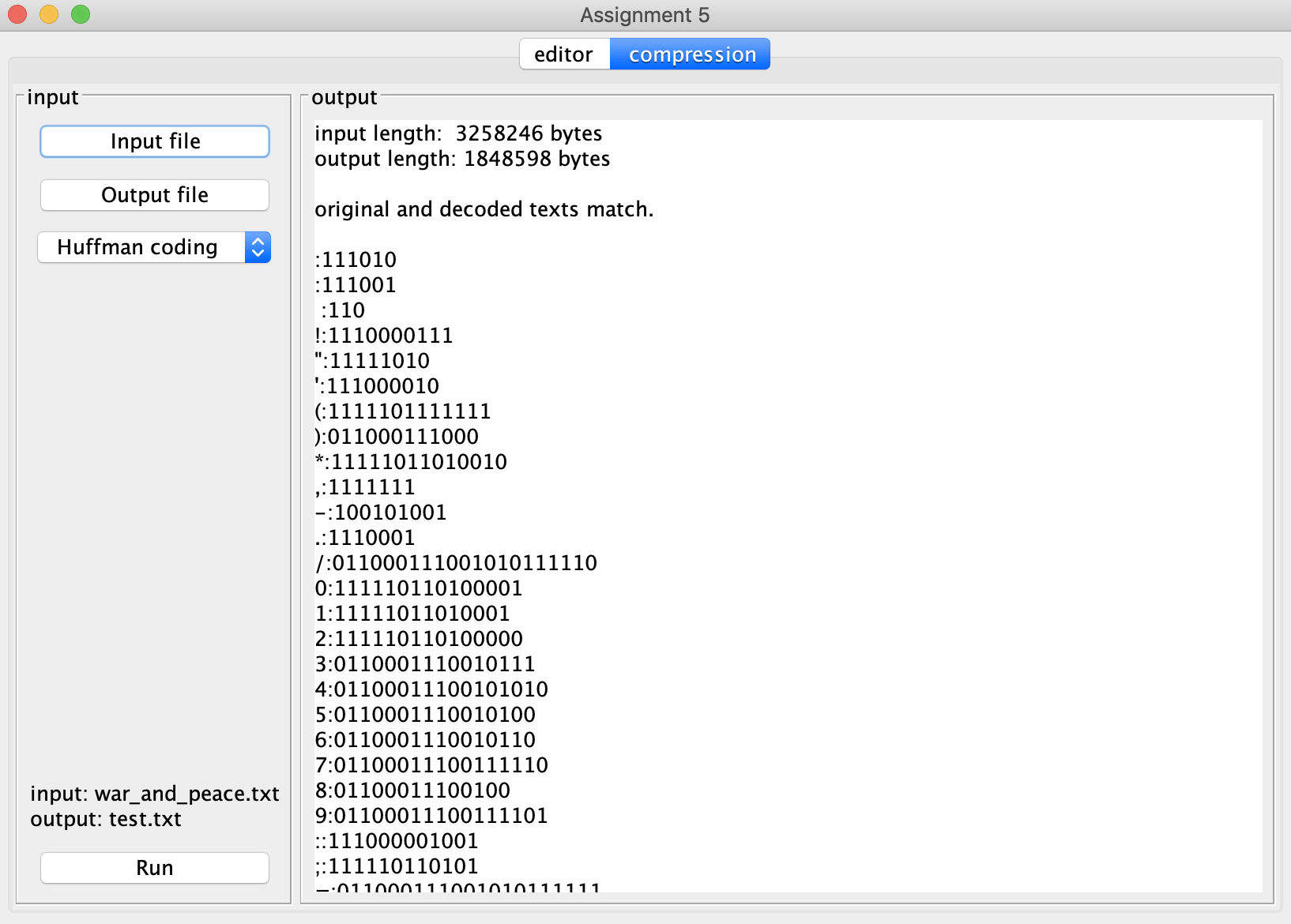
**Question 4:**

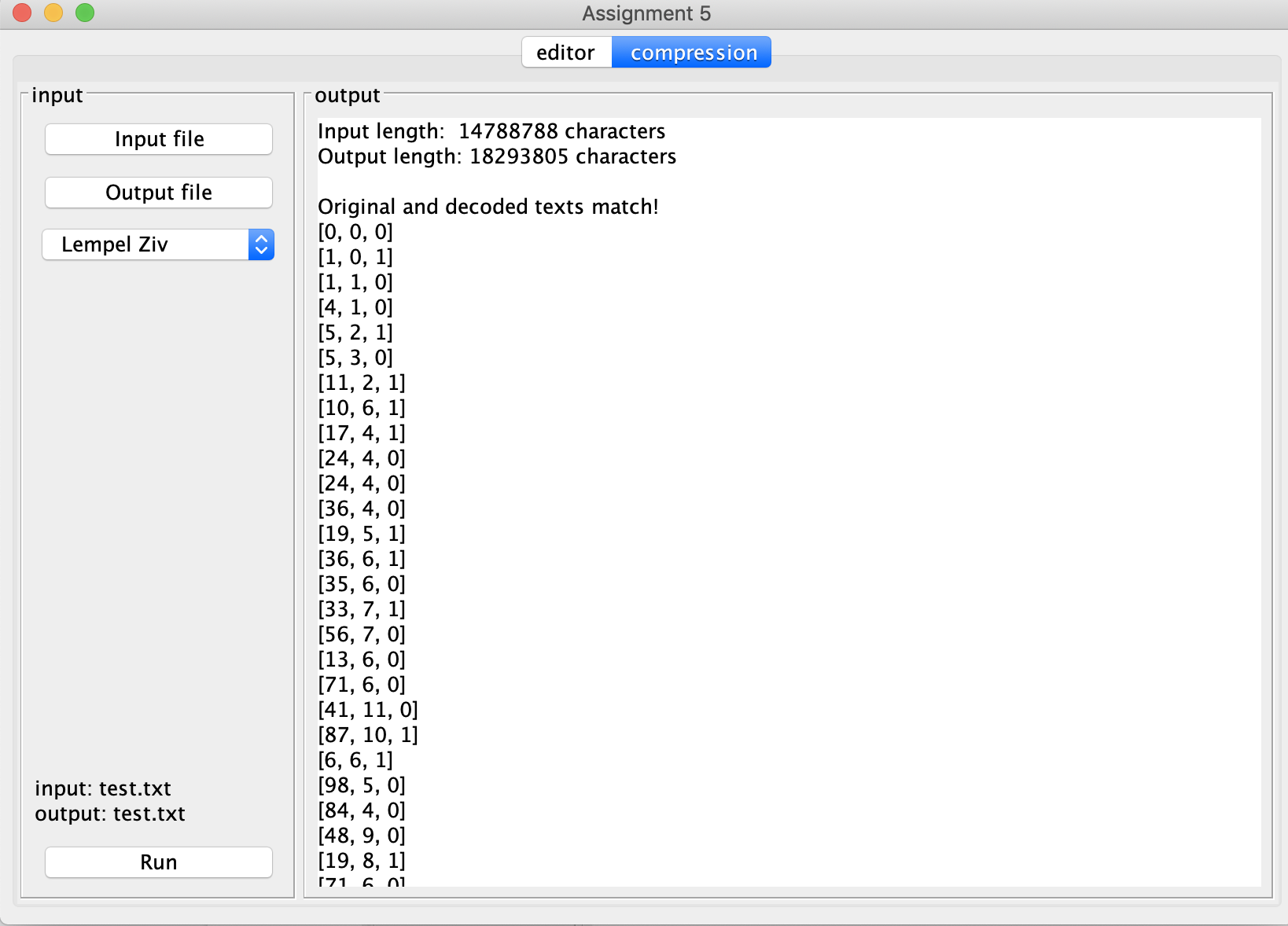
When we use a bigger window size, the quality of compression will become better, size of tuples and bytes using after compression will be reduced . If the window size get lower, the quality of compression will decrease as well.

**Question 5:**

if I Huffman encode War and Peace *before* applying Lempel-Ziv compression to it, using Huffman compression will encode War\_and\_Peace.txt into codes contains only 0 and 1 to represent each chars, so , if I use Lempel-Ziv after it , it will use tuple to store same repeated order of binary code . The size of file will not be smaller because after Huffman compression , there are more characters overall needed to be compress.

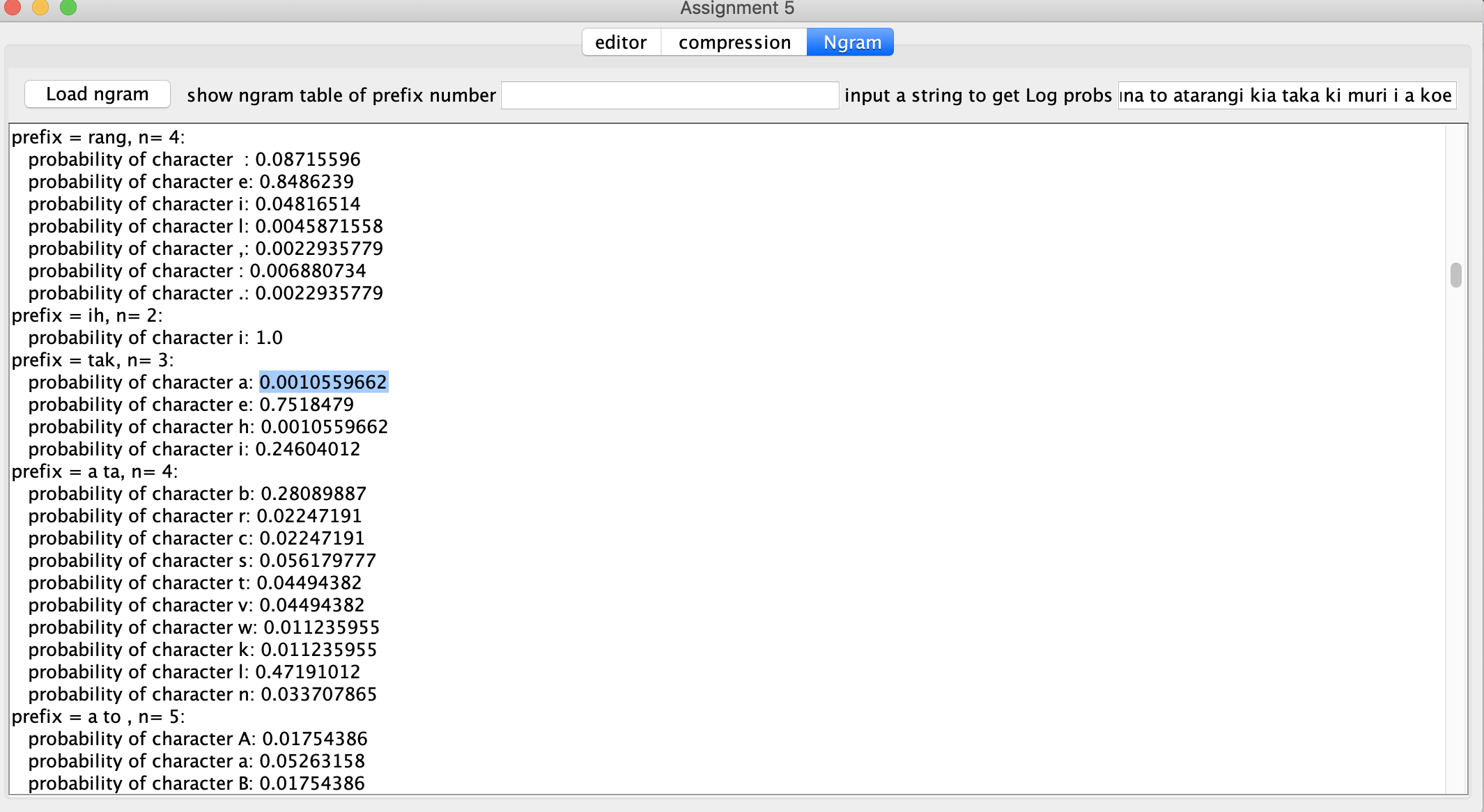






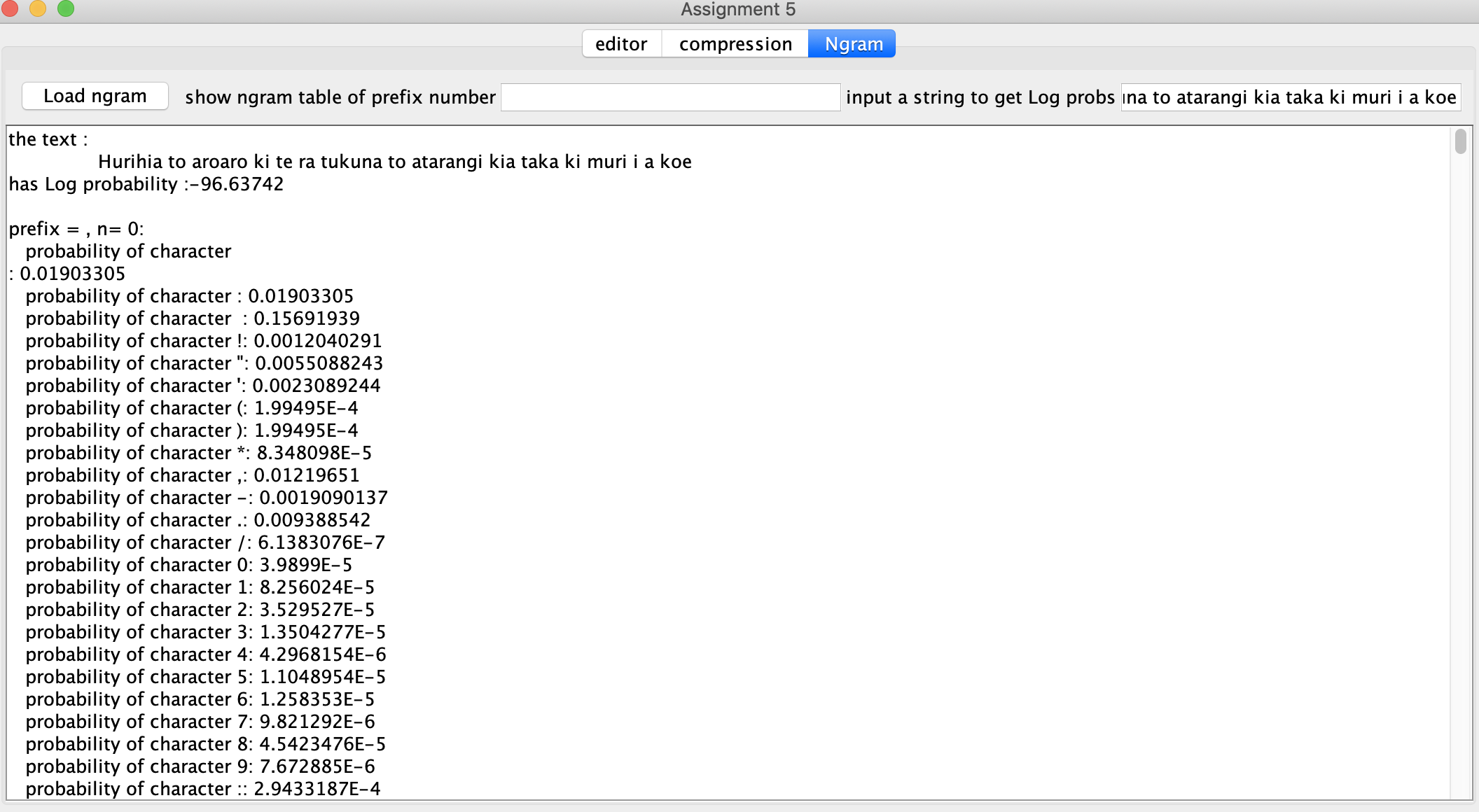
* The whakatauki contains the substring "kia taka". Given the prefix "a tak", what is the probability of "a", and the value of n used, based on Ngrams trained on the text of war\_and\_peace.txt?

The probability of a is 0.00105596624 , and the value of n used is 3.

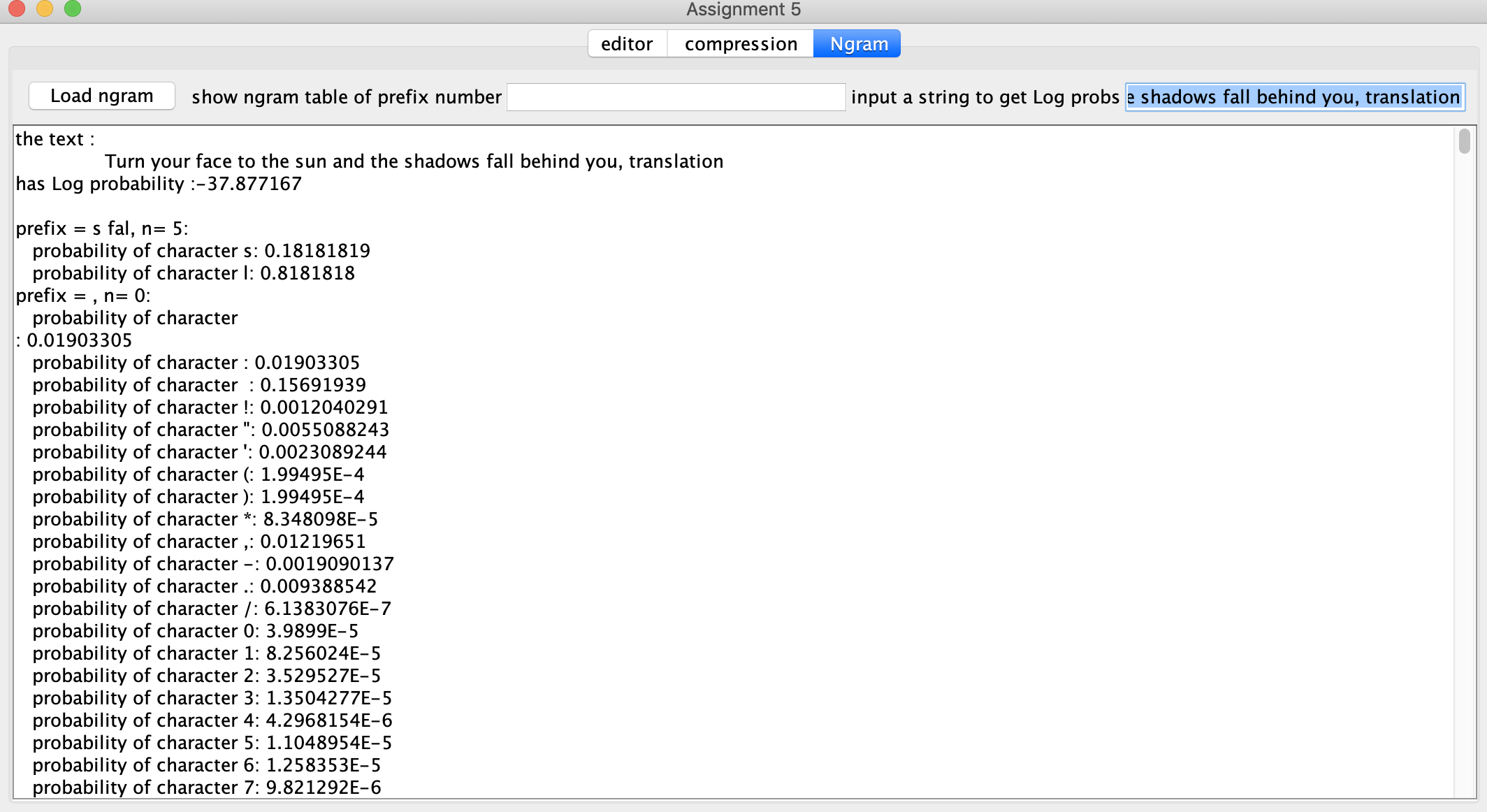


* What is the log probability of the Te Reo version, and the English version, of the whakatauki?

The Log probability of Tr Reo Version is -96.63742



The Log probability of English Version is -37.877167



**Question 6:**

The reason that the log probability of English version and Te Reo version is that they are using different languages ,different languages have very different grammar and habit so the character probability of after a prefix is very different. Also, the ngram table that we created base on War\_and\_Peace.txt is in English, so the English version is more likely to find a same prefix in the main map that Te Reo version and the English version will also have a higher n value; So the log probability of them are very different.