Did you receive any help whatsoever from anyone in solving this assignment? No.
Did you give any help whatsoever to anyone in solving this assignment? No.
Did you find or come across code that implements any part of this assignment ? No.

1:

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
&= -\sum_x p(x)log(p(x)) + \sum_{x,y} p(x,y)log(p(x|y)) \\
&= -\sum_x p(x)log(p(x)) + \sum_{x,y} p(x,y)log(\frac{p(y|x)p(x)}{p(y)}) \\
&= -\sum_x p(x)log(p(x)) + \sum_{x,y} p(x,y)log(p(y|x)) + \sum_{x,y} p(x,y)log(p(x)) - \sum_{x,y} p(x,y)log(p(y)) \\
&= \sum_{x,y} p(x,y)log(p(y|x)) - \sum_{x,y} p(x,y)log(p(y)) \\
&= H(Y) - H(Y|X) \\
&= I(Y;X)
\end{aligned}
$$

3:

(a)

s	2	3	4	5	6	7	8	9	10	11	12
	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

$$
\begin{aligned}
H(S) &= \sum_s p(s)log(\frac{1}{p(s)}) \\
&= (\frac{1}{36} + \frac{1}{36})log(36) + (\frac{2}{36} + \frac{2}{36})log(\frac{36}{2}) + (\frac{3}{36} + \frac{3}{36})log(\frac{36}{3}) + \\
&(\frac{4}{36} + \frac{4}{36})log(\frac{36}{4}) + (\frac{5}{36} + \frac{5}{36})log(\frac{36}{5}) + \frac{6}{36}log(\frac{36}{6}) \\
&= 3.2744
\end{aligned}
$$

(b) Let W_i be the different choice (letter or digit) in position i,since all positions are independent, then the entropy of a sequence of choice is :

$$
\begin{aligned}
H(W_1...W_n) &= \sum_{i=1}^{n} H(W_i|W_1...W_{i-1}) \\
&= \sum_{i=1}^{n} H(W_i) = nH(W_i)
\end{aligned}
$$

For a single choice W_i, let A be the choice of letter or digit, p(A = letter) = λ, p(A = digit) = $1 - \lambda$, X be the choice within digit or letter then

$$\begin{aligned} H(W_i) &= H(X, A) \\ &= H(X|A) + H(A) \\ &= p(A = digit)H(X|A = digit) + p(A = letter)H(X|A = letter) + H(A) \end{aligned}$$

$$H(A) = H(\lambda), H(X|A = digit) = H_D, H(X|A = letter) = H_L$$

So

$$\begin{aligned} H(W_i) &= (1 - \lambda)H_D + \lambda H_L + H(\lambda) \\ H_p &= H(W_1...W_n) \\ &= n * H(W_i) \\ &= n * ((1 - \lambda)H_D + \lambda H_L + H(\lambda)) \end{aligned}$$

4:

X = 0,1,Y = 0,1 Z = 0,1

(0,0,0)	(0,0,1)	(0,1,0)	(0,1,1)	(1,0,0)	(1,0,1)	(1,1,0)	(1,1,1)
0.25	0	0	0.25	0	0.25	0.25	0

Here p(X = 0) = 0.5, p(X = 1) = 0.5, p(Y = 0) = 0.5, p(Y = 1) = 0.5, p(Z= 0) = 0.5, p(Z = 1) = 0.5
p(X = 0,Z = 0) = p(X = 0)p(Z = 0) = 0.25
p(X = 0,Z = 1) = p(X = 0)p(Z = 1) = 0.25
p(X = 1,Z = 0) = p(X = 1)p(Z = 0) = 0.25
p(X = 1,Z = 1) = p(X = 1)p(Z = 1) = 0.25
X Z are independent, (a) is satisfied
p(Y = 0,Z = 0) = p(Y = 0)p(Z = 0) = 0.25
p(Y = 0,Z = 1) = p(Y = 0)p(Z = 1) = 0.25
p(Y = 1,Z = 0) = p(Y = 1)p(Z = 0) = 0.25
p(Y = 1,Z = 1) = p(Y = 1)p(Z = 1) = 0.25
Y Z are independent, (b) is satisfied
H(Z) = -0.5 *log(0.5) -0.5*log(0.5) = 1
I(X,Y;Z) = 4* 0.25 * log(0.25/(0.25*0.5)) = 1
(c) is satisfied

5:

(a) Entropy = 8.75292253708

(b) The probability for word which is not in the sample, the Pml = 0 , which will run into problem when compute the cross entropy because CH requires take log to probability. To fix it, we can use the smoothing technique in the below question, add pseudo count or shrinkage for word not exist in the sample

(c) pseudo count(add 1 for not sampled word):8.77048363953
shrinkage:8.78714684234

(d)

The smoothing parameter in pseudo count is the count for word which does not exist in the sample. The larger pseudo count it is, the smaller the weight of the word count from the sample. In that case, the distribution after pseudo count smooth is less likely to reflect the word distribution from sample. So the cross entropy increases because the smoothed distribution has larger distance from the true probability. Similarly, the larger N it is, the less impact of the pseudo count it has because the term frequency from sample is large compared with pseudo count. Larger N also make sample more likely to approximate true probability. That's why larger N has small cross entropy and the change in cross entropy is smaller.
The smoothing parameter in shrinkage count is the weight for word which exist in the sample. The larger weight it is, the distribution after smoothing is less likely to take effect because the weight of unseen word is smaller. So the cross entropy decrease because the smoothed effect is smaller and unseen word tends to have smaller probability. Similarly, larger N also make sample more likely to approximate true probability. That's why larger N has small cross entropy.

6:

N is the sum of the total count

$$
\begin{aligned}
H &= \sum_{word_i} p(word_i) * log(\frac{1}{p(word_i)}) \\
&= \sum_{word_i} \frac{count(word_i)}{N} log(\frac{N}{count(word_i)}) \\
&= \frac{1}{N} \sum (count * log(N) - count * log(count)) \\
&= \frac{1}{N} log(N) \sum (count) - \frac{1}{N} \sum (count * log(count)) \\
&= log(N) - \frac{1}{N} \sum_{word_i} (count(word_i) * log(count(word_i)))
\end{aligned}
$$

So the input file can be treated as a data stream, we use a variable to keep track of the $\sum word_i(count(word_i) * log(count(word_i)))$, then every time a new count is read, the variable is updated, also the number of total count is updated too, after we read all the data , we can use the above formula to compute the entropy.