CS4248
AY 2022/23 Semester 1
Tutorial 3

1. Consider a language with only two symbols I and O. The probability p_I of the symbol I occurring in the language is unknown. Suppose one particular sample text consists of a sequence of N symbols drawn randomly and independently from the language, out of which the symbol I occurs N_I times and the symbol O occurs $N - N_I$ times in this particular sample text.
(a) Write an expression for the probability of the sample text in terms of p_I, N_I, and N.
(b) Derive mathematically the value of p_I that maximizes the probability of the sample text, in terms of N_I and N. This value serves as the maximum likelihood estimate of p_I.

2. The following bigram counts were collected from a corpus:

	w_1	w_2	w_3
w_1	100	0	30
w_2	0	50	0

That is, the bigram $w_1 w_1$ occurred 100 times, the bigram $w_1 w_2$ occurred 0 times, the bigram $w_1 w_3$ occurred 30 times, etc. The frequency of each word w_i in the corpus is tabulated as follows:

w_1	4,000
w_2	2,500

The number of word types following each word w_i is tabulated as follows:

w_1	150
w_2	80

The vocabulary size of this corpus is 15,000.

Compute the Witten-Bell smoothed bigram counts and tabulate the smoothed bigram counts in a table as follows. Show clearly the steps of your computation in deriving the smoothed bigram counts.

	w_1	w_2	w_3
w_1			
w_2			

3. Let *H(X)*, *H(X,Y)*, and *H(Y|X)* denote the entropy, joint entropy, and conditional entropy of discrete random variables *X* and *Y*, defined as follows:

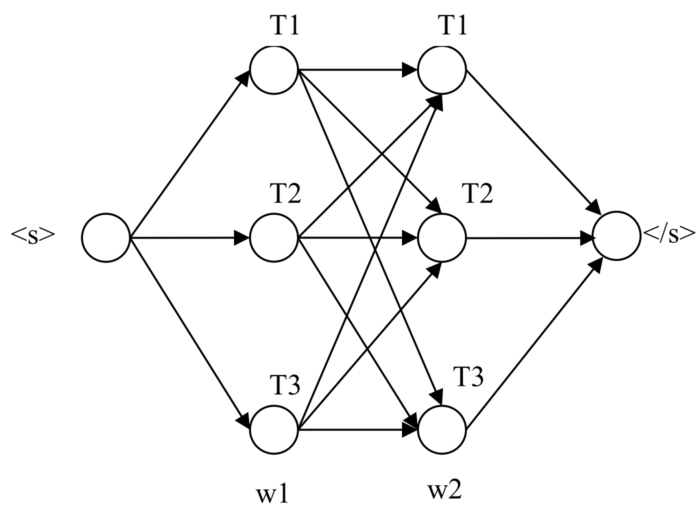
$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$
$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x,y)$$
$$H(Y \mid X) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(y \mid x)$$

Consider the following statement:

$$H(X,Y) = H(X) + H(Y \mid X)$$

Is the statement true? If it is true, give a formal proof and clearly justify each step of your proof. If it is false, provide a concrete counter-example.

4. Consider the following HMM:



Suppose this HMM has the following set of parameters:

P(T1|<s>) = 1/5	P(T1|T1) = 0	P(T1|T2) = 1/8	P(T1|T3) = 3/5
P(T2|<s>) = 0	P(T2|T1) = 5/6	P(T2|T2) = 1/2	P(T2|T3) = 1/5
P(T3|<s>) = 4/5	P(T3|T1) = 0	P(T3|T2) = 1/4	P(T3|T3) = 1/5
	P(</s>|T1) = 1/6	P(</s>|T2) = 1/8	P(</s>|T3) = 0
P(w1|T1) = 1/20	P(w2|T1) = 1/10		
P(w1|T2) = 1/5	P(w2|T2) = 1/10		
P(w1|T3) = 1/10	P(w2|T3) = 1/10		

T1, T2, T3 are part-of-speech tags.
Consider the input sentence "w1 w2", where w1 and w2 are words. Trace the Viterbi algorithm, by providing the values of the cells v(T, w) where T ∈ { T1, T2, T3 }, and w ∈ { w1, w2 }, and determine the optimal sequence of part-of-speech tags.