CS4248
AY 2022/23 Semester 1
Tutorial 2

1. Text classification is the task of assigning a class c_i to a text t based on the content of t, where c_i is chosen from a predefined set of classes. The task can be formulated as a supervised learning task from labeled training texts, using Bayesian classification.

Suppose the following 4 labeled training texts have been collected for 2 classes:
c_1 : usa is the champion in tennis
c_1 : wimbledon is the ultimate in tennis
c_2 : brazil is the champion in soccer
c_2 : soccer is popular

(a) Compute the maximum likelihood estimate for $P(c_1)$ and $P(c_2)$, where $P(c_i)$ is the proportion of training texts in class c_i.

w	$P(w \mid c_1)$		$P(w \mid c_2)$	
	MLE	add-one	MLE	add-one
brazil				
champion				
in				
is				
popular				
soccer				
tennis				
the				
ultimate				
usa				
wimbledon				

(b) Let V be the set of words occurring in the training texts. For each word $w \in V$, compute the maximum likelihood estimate (MLE) for $P(w \mid c_1)$ and $P(w \mid c_2)$, where $P(w \mid c_i)$ is the probability of selecting word w from the bag of words that constitute all the training texts belonging to c_i. That is, provide the probabilities in the MLE columns in the above table.
(c) Use add-one smoothing to compute $P(w \mid c_1)$ and $P(w \mid c_2)$ such that $P(w \mid c_i) > 0$ for each word $w \in V$. That is, provide the probabilities in the add-one columns in the above table.
(d) For text classification, each text t can be represented as the bag of words occurring in t. The naïve Bayes assumption states that the words $w_1, \cdots, w_n$ in a text t are conditionally independent given the class c of t. That is,

$$P(w_1,\cdots,w_n \mid c) = P(w_1 \mid c)\cdot\ldots\cdot P(w_n \mid c)$$

Give the formula for determining the class that a new test text t belongs to, using Bayesian classification and making the naïve Bayes assumption.
(e) Given the following two texts:

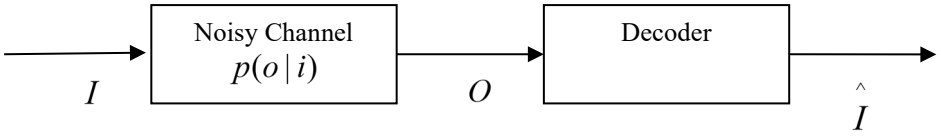
germany is the champion in soccer
wimbledon is played in the uk

determine the class to which each text is assigned. (Ignore any word in a test text that does not belong to V)

2. Give a trace of the minimum edit distance algorithm (a dynamic programming algorithm) to compute the minimum cost of transforming the string "cheap" to "help", by filling out every cell entry in the following table, where each cell entry denotes the minimum cost of transforming the associated substrings. Assume that the cost of inserting a character is 1, the cost of deleting a character is 1, and the cost of substituting a character by a different character is 2. Add appropriate backtrace pointer(s) to each cell entry, and trace the optimal paths in the table.

p	5				
a	4				
e	3				
h	2				
c	1				
	0	1	2	3	4
		h	e	l	p

3. The noisy channel model can be depicted as follows:



$$\hat{I} = \arg\max_i p(i \mid o) = \arg\max_i \frac{p(i)p(o \mid i)}{p(o)} = \arg\max_i p(i)p(o \mid i)$$

For each of the following applications:

(a) spelling error correction
(b) speech recognition

cast the application as a noisy channel model and specify the corresponding input to the noisy channel, the output of the noisy channel, $p(i)$, and $p(o \mid i)$.