[redacted]

Question 1. Data preprocessing in NLP.

a) It is possible to tokenize based on space alone?

1. True for Vietnamese, False for Chinese
2. False for Vietnamese, True for Chinese
2. True for Vietnamese, True for Chinese
2. False for Vietnamese, False for Chinese

b) As preprocessing, one should always tokenize and lemmatize the data:

1. True
2. False

c) What is the tokenization of the following sentence?

There aren't any mistakes!

1. `[There] [are] [not] [any] [mistakes] [!]`
2. `[There] [aren] ['t] [any] [mistakes] [!]`
3. `[There] [aren't] [any] [mistakes!]`
4. it depends on the tokenizer

Question 2. What is the expression of $p(Y = y \mid X = \mathbf{x})$ for the binary logistic regression, as a function of the parameters $\mathbf{w} \in \mathbb{R}^d$:

Question 3. Let V be the size of the vocabulary, $\mathbf{p} \in \mathbb{R}^V$ and $b \in \mathbb{R}$ such that:

$$\mathbf{p}_i = \log\left(P(W = i|Y = 1)\right) - \log\left(P(W = i|Y = 0)\right),$$
$$b = \log\left(P(Y = 1)\right) - \log\left(P(Y = 0)\right).$$

Let $\mathbf{x} \in \mathbb{R}^V$ a vector representing the bag of words of a document D. What is the expression of the log probability ratio $\dfrac{P(Y = 1, D)}{P(Y = 0, D)}$ in the case of categorical naive Bayes?

Question 4. Let V be the size of the vocabulary, $c(y)$ the count of class y and $c(y, w)$ the count of class y and word w. What is the expression of $P(W = w \mid Y = y)$ corresponding to categorical naive Bayes with Laplacian smoothing?

Question 5. Given word vectors $\mathbf{u}$ and $\mathbf{v}$, what is the cosine similarity between $\mathbf{u}$ and $\mathbf{v}$?

Question 6. Given word vectors for the words *strong*, *stronger* and *tall*, how can we find the word vector corresponding to the superlative of *tall* (ie. *taller*)?

Question 7. What is the difference between a bag of words and a set of words?

Question 8. Given a corpus of N tokens, let's denote by $c(a)$ the count of word a and $c(a, b)$ the count of bigram ab (i.e. the number of times the word b appears after the word a). What is the probability $p(w_t \mid w_{t-1})$ according to a simple linear interpolation model?

Question 9. Let $c \in \mathbb{R}$ and $\mathbf{s} \in \mathbb{R}^d$. Let note f the softmax operator, such that $f_i(\mathbf{s}) = \dfrac{\exp(s_i)}{\sum_{j=1}^{d} \exp(s_j)}$.

We have:

1. $f_i(\mathbf{s}) < f_i(\mathbf{s} + c)$
2. $f_i(\mathbf{s}) > f_i(\mathbf{s} + c)$
3. $f_i(\mathbf{s}) = f_i(\mathbf{s} + c)$
4. It depends on the sign of c.

Question 10. Given the confusion matrix, where y is the ground truth and $\hat{y}$ is the prediction of the model:

	$y = 1$	$y = -1$
$\hat{y} = 1$	TP	FP
$\hat{y} = -1$	FN	TN

What is the expression of the recall:

1. $\text{RE} = \dfrac{\texttt{TP}}{\texttt{FP} + \texttt{FN}}$
2. $\text{RE} = \dfrac{\texttt{TP}}{\texttt{TP} + \texttt{FP}}$
3. $\text{RE} = \dfrac{\texttt{TP}}{\texttt{TP} + \texttt{FN}}$
4. $\text{RE} = \dfrac{\texttt{TP}}{\texttt{TP} + \texttt{TN}}$

Question 11. Given a query $\mathbf{q} \in \mathbb{R}^d$ and a set of keys and values $\mathbf{k}_i \in \mathbb{R}^d, \mathbf{v}_i \in \mathbb{R}^d$ for $i \in \{1, ..., N\}$, what is the expression of the output of the attention mechanism?

Question 12. How is the word order captured in transformer networks?

Question 13. What is the perplexity of a sentence $W = (w_1, \dots, w_T)$:

1. $PP(W) = \left(\prod_{t=1}^{T} P(w_t \mid w_{t-1}, \dots, w_1)\right)^{-T}$
2. $PP(W) = \left(\prod_{t=1}^{T} P(w_t \mid w_{t-1}, \dots, w_1)\right)^{-\frac{1}{T}}$
3. $PP(W) = \left(\sum_{t=1}^{T} P(w_t \mid w_{t-1}, \dots, w_1)\right)^{-T}$
4. $PP(W) = \sum_{t=1}^{T} P(w_t \mid w_{t-1}, \dots, w_1)^{-\frac{1}{T}}$

Question 14. In `word2vec`, negative sampling is used

1. To make the training more efficient
2. To regularize the model
3. To avoid overflow in the softmax
4. To avoid exploding gradient

Question 15. Gradient clipping in a recurrent neural network:

1. prevents gradient explosion
2. prevents vanishing gradient
3. regularizes the model to avoid overfitting
4. approximates the gradient with truncation in time

Question 16. We have a dataset split for positive ($+$) and negative ($-$) movie reviews:

$+$	the story was amazing
$+$	movie was super good
$-$	not a good story
$-$	movie was boring
$-$	story is bad

We have a new review:

story was good

We want to know what is the most likely label and the associated joint probability given by a Naive Bayes model?

1. The most likely label is $+$ and $P(+, \text{ story was good}) = \frac{1}{540}$
2. The most likely label is $+$ and $P(+, \text{ story was good}) = \frac{1}{640}$
3. The most likely label is $-$ and $P(-, \text{ story was good}) = \frac{3}{2500}$
4. The most likely label is $-$ and $P(-, \text{ story was good}) = \frac{1}{2500}$

Question 17. We have a dataset containing $N = 600$ words in total with a vocabulary of 25 unique words. We want to estimate the probability of the following sentence:

$$<s>\ i\ study\ machine\ learning$$

with a counted based model. To do so, we provide the unigrams counts for 8 tokens:

i	want	study	math	<s>	machine	learning	like
30	132	25	174	50	36	60	64

as well as their bigram counts:

$w_1 \backslash w_2$	<s>	i	want	study	math	machine	learning	like
<s>	0	20	0	0	0	1	5	0
i	0	0	4	9	3	0	0	10
want	0	1	0	1	2	6	3	0
study	0	0	0	0	11	7	4	4
math	0	0	0	5	0	0	0	0
machine	0	0	0	6	0	0	6	5
learning	0	0	2	0	0	3	0	22
like	0	8	0	1	0	2	8	0

(For example $c(\text{i want}) = 4$ and $c(\text{want i}) = 1$)

We remind that, by convention, $P(<\text{s}>) = 1$.

a. What is the probability of the sentence given by a unigram model?

1. $\frac{1}{40000}$
2. $\frac{1}{60000}$
3. $\frac{1}{80000}$
4. $\frac{1}{120000}$

b. What is the probability of the sentence given by a bigram model?

1. $\frac{1}{60}$
2. $\frac{1}{120}$
3. $\frac{1}{250}$
4. $\frac{1}{400}$