1. What is the purpose of tokenizing text into sentences and words?
   Given a document, which in essence, is a string of characters, tokenizing is a means to subdividing such a collection into its constituent parts either based on sentences, whose ends are marked by full-stop (.), question marks (?) or exclamation mark (!), or words. Through this process, textual data is made suitable for tasks like information retrieval, textual analysis, etc, as it is available in its primitive form, providing room to create associations adhering to the rules of English grammar.

2. What are n-gram language models and how are they useful in NLP?
   Sequences of words taken n at a time from a text corpus. Simple bi-gram and tri-gram models can help identifying commonly occurring phrases and clauses. In term of text generation, n-gram indicates that $n - 1$ number of words from the end are taken into consideration for contextually predicting the next word.

3. What is the naive Bayes assumption and how does it relate to text classification?
   Given a set of features and a target variable, $\hat{c}$, the algorithm assumes that features are independent of each other, while conditionally being dependent on their existence with respect to the document d. For any document d, it calculates the posterior probability $P(c|d)$ for every class defined in d, selecting the maximum probability (**argmax**) to classify the text.

4. What are some of the advantages and disadvantages of naive Bayes classifiers compared to logistic regression?
   Taking the points discussed in 3, given its high bias, Naive Bayes serves as a good base estimator while ignoring the sequence of words, leading to low variance. Nevertheless, it is simple, efficient and can handle sparse data. Logistic Regression assumes a linear relationship of the features, calculating a logistic function to classify the text. Given its nature, it can handle correlated texts and ensures higher variance while being computationally expensive than the former.

5. What do we mean by ”features” in the context of text classification? Give some examples of features that might be useful for distinguishing different newsgroup topics.
   For a given document d, a set of features f_i, where $i \in [1, n]$, can be defined in the following ways:
   - PoS tags: Dictionary of Parts of Speech, abiding by the rules of English Grammar to create a counter for each category.
   - Keyword based searching.
   - Named entities.
   - Dates.
   - BoW (*discussed in 9*). The above ways of discribing features could be applicable in the problem statement mentioned in the question.

6. What is the purpose of a test set in machine learning? Why do we need separate training and test sets?
   As the name suggests, a test set is primarily used to evaluate our models. The whole ideology behind ML is to enable our selected algorithms to learn the pattern in our *train* set. While doing so, certain transformations and scaling techniques can be applied to the train set to optimize learning. Equations involved here only consider the statistical metrics of the train set only, which are not available when the model is deployed. Hence, a test set is separated out as a means to evaluate and iteratively improve the model before deploying it. Note that the performance when deployed is not guaranteed to be the same as that on test data.

7. What metrics could you use to evaluate the performance of a text classification model? Define accuracy and any other relevant metrics.
   T - True, F - False, P - Positive, N - Negative

- Accuracy score = $\frac{TP+TN}{TP+TN+FP+FN}$
- Precision = $\frac{TP}{TP+FP}$
- Recall = $\frac{TP}{TP+FN}$
- Specificity = $\frac{TN}{TN+FP}$

8. How could you determine which features are most important or indicative for a logistic regression text classification model?
   We could measure the importance of each feature based on the absolute values of the coefficients learned by the model on the training data. We could also use *Shapely values* to measure the impact of a feature on each data point of the target variable.

9. What is the bag-of-words representation and what are some of its limitations for text classification?
   The entire document is divided into a set of words (*key*) while storing respective frequencies (*value*) like a dictionary. Its limitations are as follows:
   - High dimensional feature space.
   - Loss of word order.
   - Difficult to capture semantic relationships, leading to redundancy at times.
   - Highly sensitive to stop words like articles, conjunctions, etc.

10. What is overfitting in machine learning? How could you tell if your text classification model is overfitting the training data? Describe two ways to reduce overfitting.
    Considering the test and train sets, we observe an instance of overfitting when our model performs far better on the train set than the test set. It can also be an instance of high variance and low bias in our model. Overfitting can be reduced by:
    - Regularization : Adds a penalty term to the loss function that reduces the large magnitude of coefficients (L2) or even nullifying them (L1).
    - k-Fold Cross-validation : The train set is divided into k folds, trained on k-1 folds and evaluated on the last one, repeated k times, ensuring that each fold is used for evaluating the final average performance.