Part II: Short Answer

Q1. Understanding Model Behavior and Training in Machine Learning

(a) What is the main difference between generative and discriminative models? Provide one example of each.

The main difference between *generative* and *discriminative* models lies in what they learn:

- **Generative models** learn the **joint probability** P(x, y), i.e., how data and labels are generated. They can generate new samples.
  Example: Naïve Bayes, Gaussian Mixture Model.
- **Discriminative models** learn the **conditional probability** P(y | x) or directly a decision boundary between classes.
  Example: Logistic Regression, Support Vector Machine.

(b) What role does the cross-entropy loss function play during training?

The **cross-entropy loss** measures the distance between predicted probabilities and the true labels.
During training, minimizing cross-entropy pushes the model's predicted probability for the correct class closer to 1 and reduces it for incorrect classes.
It ensures faster convergence and stable gradients in probabilistic models like logistic regression or neural networks.

Q2. Generalization and Representation in NLP

(a) What is the difference between synonymy and word similarity? Provide an example.

Synonymy refers to words that have nearly the same meaning in context (e.g., "car" and "automobile").

Word similarity refers to words that are semantically related but not identical (e.g., "car" and "road").

(b) How does TF-IDF improve upon raw term frequency when representing document content?

TF-IDF down-weights very common words and up-weights rare but informative ones.

While raw term frequency counts how often a word appears, TF-IDF multiplies it by the inverse document frequency (IDF), giving higher weight to distinctive terms that help differentiate documents.
(e) Explain why PPMI values are clipped at 0. What does a negative PMI indicate?

PPMI (Positive PMI) clips negative PMI values at 0 because negative PMI means the word–context pair co-occurs *less often* than expected by chance.
Those pairs do not indicate semantic association, so they are set to 0 to keep only meaningful, positively associated pairs.

Q3. Neural Networks: Structure, Activation, and Representation

(a) Why are non-linear activation functions (like ReLU, sigmoid, tanh) necessary in neural networks?

Non-linear activations (ReLU, sigmoid, tanh) allow neural networks to model **non-linear relationships** between inputs and outputs.
Without them, multiple layers would collapse into a single linear transformation and could not capture complex patterns.

(b) Why are deep networks better at feature learning compared to shallow models like logistic regression?

Deep networks can learn **hierarchical features**: lower layers capture simple patterns (edges, n-grams), while higher layers combine them into more abstract representations (shapes, semantics).
Shallow models like logistic regression only learn a single linear boundary and cannot automatically compose complex features.