CSSE 315 – Natural Language Processing
Rose-Hulman Institute of Technology

Worksheet 07

Name (Print):_____ Date:_____

1 Review Questions

1. Which LLM configuration setting controls the degree of **randomness** for the selection of the next predicted token?
   (a) Top-K
   (b) Top-P
   (c) Output Token Count
   (d) Temperature

2. What modalities can be converted into embeddings?
   (a) Image
   (b) Video
   (c) Text
   (d) All of the above

3. What of the following is a major advantage of ScaNN over other ANN search algorithms?
   (a) It is free
   (b) It returns exact matches
   (c) It is designed for high-dimensional data and has higher accuracy
   (d) It places high-dimensional data into buckets

4. What are some of the major weaknesses of bag-of-words (BoW) models for generating document embeddings?
   (a) The ignore word ordering and semantic meaning
   (b) They are computationally expensive and require large amounts of data
   (c) They cannot be used for topic discovery
   (d) Only effective for short documents and fails to capture long-range dependencies

5. Provide a model example for each transformer type
   (a) Decoder-only:
   (b) Encoder-only:
   (c) Endocer-Decoder:

6. You asked a language model to classify the sentiment of a review and you provided the following example: "I love this product - Positive" and "The service was terrible - Negative". This is an example of what type of task

   (a) Zero-Shot Prompting
   (b) Few-Shot Prompting
   (c) Fine-Tuning

7. To restrict the predicted next token, you can use several sampling settings, such as Top-K and Top-P. Which of the following statements correctly describes the difference between Top-K sampling and Top-P sampling in text generation?

   (a) Both Top-K and Top-P sampling always select the token with the highest probability
   (b) Top-K sampling selects tokens based on a fixed number of most probable tokens, while Top-P sampling selects tokens dynamically based on cumulative probability
   (c) Top-P sampling uses a fixed number of tokens, while Top-K sampling dynamically adjusts the number of tokens based on probability thresholds

2 Fine-Tuning

8. Supervised Fine-Tuning (SFT) involves training a pre-trained model on a _____ dataset to align the model's outputs with specific task requirements

9. _____ uses labeled data to train the model to perform a specific task.
   _____ uses human feedback to train a reward model.

10. What is quantization?

11. Calculate the memory size for 2X2 matrix with float32 and recalculate the memory after quantization to int8:

12. What is the primary goal of model distillation?

    (a) To increase the size of the model for higher accuracy
    (b) To train a smaller model to mimic a larger, pre-trained model
    (c) To fine-tune a model on a new dataset
    (d) To replace labeled data with unlabeled data for training