**Short Answer Questions:**

1. What is the difference between stemming and lemmatization? Provide examples with the word “running.”

* A) **Stemming** cuts words to their root form without understanding context.

Example: "running" → "run"

* **Lemmatization** returns the base word using vocabulary and grammar rules.

Example: "running" → "run" (as a verb)

1. Why might removing stop words be useful in some NLP tasks, and when might it actually be harmful?

* A) **Helpful** in tasks like topic modeling or document classification to remove common, non-informative words.
* **Harmful** in tasks like sentiment analysis or translation where words like **"not"** change meaning.

1. How does NER differ from POS tagging in NLP?

* A) **NER (Named Entity Recognition)** identifies **real-world entities** like names, locations, and dates (e.g., "Barack Obama" → PERSON).
* **POS tagging (Part-of-Speech tagging)** labels words with their **grammatical role** (e.g., "run" → verb or noun).

1. Describe two applications that use NER in the real world (e.g., financial news, search engines).

* A) **Financial News Analysis**: Extracts company names, stock symbols, or monetary values from news for automated trading.
* **Search Engines**: Helps identify entities in queries to improve search relevance (e.g., distinguishing between "Apple" as a fruit vs. a company).

1. Why do we divide the attention score by √d in the scaled dot-product attention formula?

* A) To **prevent extremely large values** in the dot product when the dimension **d** is large, which could make the softmax output very small and unstable. Dividing by √d **stabilizes training**.

1. How does self-attention help the model understand relationships between words in a sentence?

* A) Self-attention allows the model to **weigh the importance of other words** in a sentence for each word, helping it capture **context and dependencies**, even from distant words.

1. What is the main architectural difference between BERT and GPT? Which uses an encoder and which uses a decoder?

* A) **BERT** uses only the **encoder** part of the Transformer (good for understanding).
* **GPT** uses only the **decoder** part (good for generating text).

1. Explain why using pre-trained models (like BERT or GPT) is beneficial for NLP applications instead of training from scratch.

* A) They save time and resources by learning from massive datasets.
* They provide **strong performance out of the box** and can be fine-tuned easily for specific tasks.