**Natural Language Processing (NLP) - 20 Questions**

1. What is tokenization in NLP?

* *is the process of breaking down text into smaller units called tokens—usually words, phrases, or sentences.*

1. Explain stemming vs lemmatization.

* ***Stemming*** *cuts words down to their root form by chopping off endings (e.g., “running” → “run”, “flies” → “fli”).*
* ***Lemmatization*** *uses grammar rules and dictionaries to get real words (e.g., “running” → “run”, “flies” → “fly”).*

1. What are stopwords, and why are they removed?

* *Stopwords are common words like “the,” “is,” “and” etc., which usually don’t carry important meaning. They’re removed to reduce noise and improve efficiency in text analysis.*

1. What is part-of-speech tagging?

* *assigns labels like noun, verb, adjective, etc., to each word in a sentence to help understand its grammatical structure.*

1. How does Named Entity Recognition (NER) work?

* *NER identifies and classifies entities in text—like names of people, organizations, locations, dates, etc.—using machine learning or rule-based approaches.*

1. What is the bag-of-words model?

* *It’s a way to represent text by counting how often each word appears, ignoring grammar and word order. Each document becomes a vector of word frequencies.*

1. Explain TF-IDF and its importance.

* *TF-IDF (Term Frequency–Inverse Document Frequency) measures how important a word is in a document relative to all documents. It reduces the weight of common words and highlights rare but significant terms.*

1. What is word embedding?

* *Word embeddings are dense vector representations of words that capture their meanings based on context. Words with similar meanings have similar vectors (e.g., “king” and “queen” will be close in vector space).*

1. How is Word2Vec different from GloVe?

* *Word2Vec learns word relationships by predicting surrounding words (context).*
* *GloVe, on the other hand, looks at how often words appear together in the whole dataset to find patterns.*

1. What is a language model in NLP?

* *A language model predicts the next word in a sequence. It helps with tasks like text generation, autocomplete, and speech recognition. Examples: GPT, BERT.*

1. What is syntactic vs sematic analysis?

* *Syntactic analysis looks at grammar and sentence structure (e.g., parsing).*
* *Semantic analysis focuses on meaning and understanding context, like detecting intent or relationships between words.*

1. Describe how sentiments analysis works.

* *Sentiment analysis classifies text as positive, negative, or neutral by analyzing word choices, tone, and context—often using machine learning models or rule-based systems.*

1. What are n-grams in NLP?

*N-grams are sequences of N words or characters that appear together. For example:*

* *Unigram: “I”,*
* *Bigram: “I love”,*
* *Trigram: “I love NLP.”  
  They help analyze word patterns and predict what words come next*

1. What is the role of attention mechanisms in NLP?

* *Attention helps the model focus on the most relevant words in a sentence when making predictions or understanding context.*

1. Describe how a transformer model processes text.

* *A transformer reads all words at once and uses attention to understand how they relate to each other.*

1. What is the difference between an encoder and decoder in transformers?

* *The encoder understands the input, while the decoder uses that understanding to generate output (like a translation).*

1. What is BERT, and how is it pre-trained?

* *BERT is a language model trained to understand context by predicting missing words and sentence relationships.*

1. How does NLP handle ambiguity in language (like polysemy)?

* *It uses context to figure out the correct meaning of words that have more than one definition.*

1. What are some key challenges in NLP for low-resource languages?

* *There’s not enough data or tools for training models, making it hard to understand or generate text in those languages.*

1. How is NLP applied in real-world tasks like chatbots or search engines?

* NLP helps chatbots understand questions and give answers and helps search engines show the most relevant results.

**Language Modeling & Machine Learning (LMM) – 20 Questions**

1. What is a language model?

* *A language model predicts the next word in a sentence based on the words that came before. It helps machines understand and generate human language.*

1. Explain the difference between generative and discriminative models.

* *Generative models learn how data is created and can generate new data. Discriminative models only learn to tell the difference between classes or labels*.

1. How do Markov models work in language modeling?

* *Markov models predict the next word using only a limited number of previous words, assuming the future depends only on the recent past.*

1. What is perplexity in the context of language models?

* *Perplexity measures how well a language model predicts text—a lower score means better predictions.*

1. What is the SoftMax function, and why is it used?

* *SoftMax turns numbers (logits) into probabilities that add up to 1, helping models pick the most likely word or class.*

1. How does overfitting affect machine learning models?

* *Overfitting happens when a model learns the training data too well, including noise, and performs poorly on new data.*

1. What are the common regularization techniques?

* *Common techniques include L1 and L2 regularization, dropout, and early stopping—they help prevent overfitting.*

1. Compare supervised, unsupervised, and reinforcement learning.

* *Supervised learning uses labeled data.*
* *Unsupervised learning finds patterns in unlabeled data.*
* *Reinforcement learning learns through trial and error using rewards.*

1. What is the difference between classification and regression?

* *Classification predicts categories (like spam or not), while regression predicts continuous values (like house prices).*

1. Explain how decision trees work.

* *Decision trees split data into branches based on features, leading to a final decision at the leaves.*