50 Theory Interview Questions on NLP for Data Scientists

1. Basics of NLP

1. What is Natural Language Processing, and why is it important?
2. Explain the difference between NLP, NLU (Natural Language Understanding), and NLG (Natural Language Generation).
3. What are the main challenges in processing natural language?
4. What is the difference between syntax and semantics in NLP?
5. What are stopwords, and why are they removed in text preprocessing?
6. Explain the difference between stemming and lemmatization.
7. What is tokenization, and what are its types?
8. How does POS (Part-of-Speech) tagging work, and why is it useful?
9. What is Named Entity Recognition (NER), and where is it applied?
10. What is n-gram, and how is it used in NLP tasks?

---

2. Text Preprocessing & Representation

11. What are word embeddings, and why are they useful in NLP?
12. Explain Bag of Words (BoW) and its limitations.
13. How does TF-IDF differ from BoW?
14. What are word2vec embeddings, and how do they work?
15. Explain CBOW vs Skip-gram models in word2vec.
16. What is GloVe embedding, and how is it different from word2vec?
17. What are contextual embeddings, and how do they improve over static embeddings?
18. Explain OOV (Out-of-Vocabulary) problem and possible solutions.
19. What are sentence embeddings, and where are they useful?
20. Explain subword tokenization and why it is important for modern NLP models.

---

3. Language Models & Deep Learning

21. What is a language model, and what are its applications?
22. Explain the difference between generative and discriminative models in NLP.
23. What is the intuition behind the Transformer architecture?
24. Explain the role of attention in NLP.
25. What is self-attention, and why is it important?
26. Explain BERT and its key innovations.
27. How does GPT differ from BERT?
28. What are encoder-decoder models, and where are they used?
29. Explain masked language modeling and next sentence prediction.
30. What are LSTMs, and how do they improve over vanilla RNNs in NLP?

4. Evaluation & Metrics

31. What is perplexity in NLP, and how is it used?
32. Explain BLEU score and its applications.
33. What are precision, recall, and F1 score in text classification tasks?
34. Explain ROUGE score and where it is applied.
35. What is word similarity, and how is it measured?
36. What are confusion matrices, and how are they useful in NLP model evaluation?
37. How do you evaluate a sentiment analysis model?
38. What challenges arise in evaluating machine translation models?
39. Explain cosine similarity in the context of word embeddings.
40. What are some common sources of bias in NLP models?

---

5. Applications of NLP

41. How does text classification work, and what are its real-world use cases?
42. Explain the working of sentiment analysis.
43. What is topic modeling, and how does LDA (Latent Dirichlet Allocation) work?
44. Explain how machine translation works.
45. How is NLP used in chatbots and virtual assistants?
46. What is information retrieval, and how is it applied in search engines?
47. Explain question answering systems in NLP.
48. What is text summarization, and what are the types (extractive vs abstractive)?
49. How is NLP applied in speech recognition and voice assistants?
50. What are ethical concerns and challenges in deploying NLP models in real-world applications?