100 NLP interview questions

Milana Shkhanukova · Follow

6 min read · Dec 31, 2023

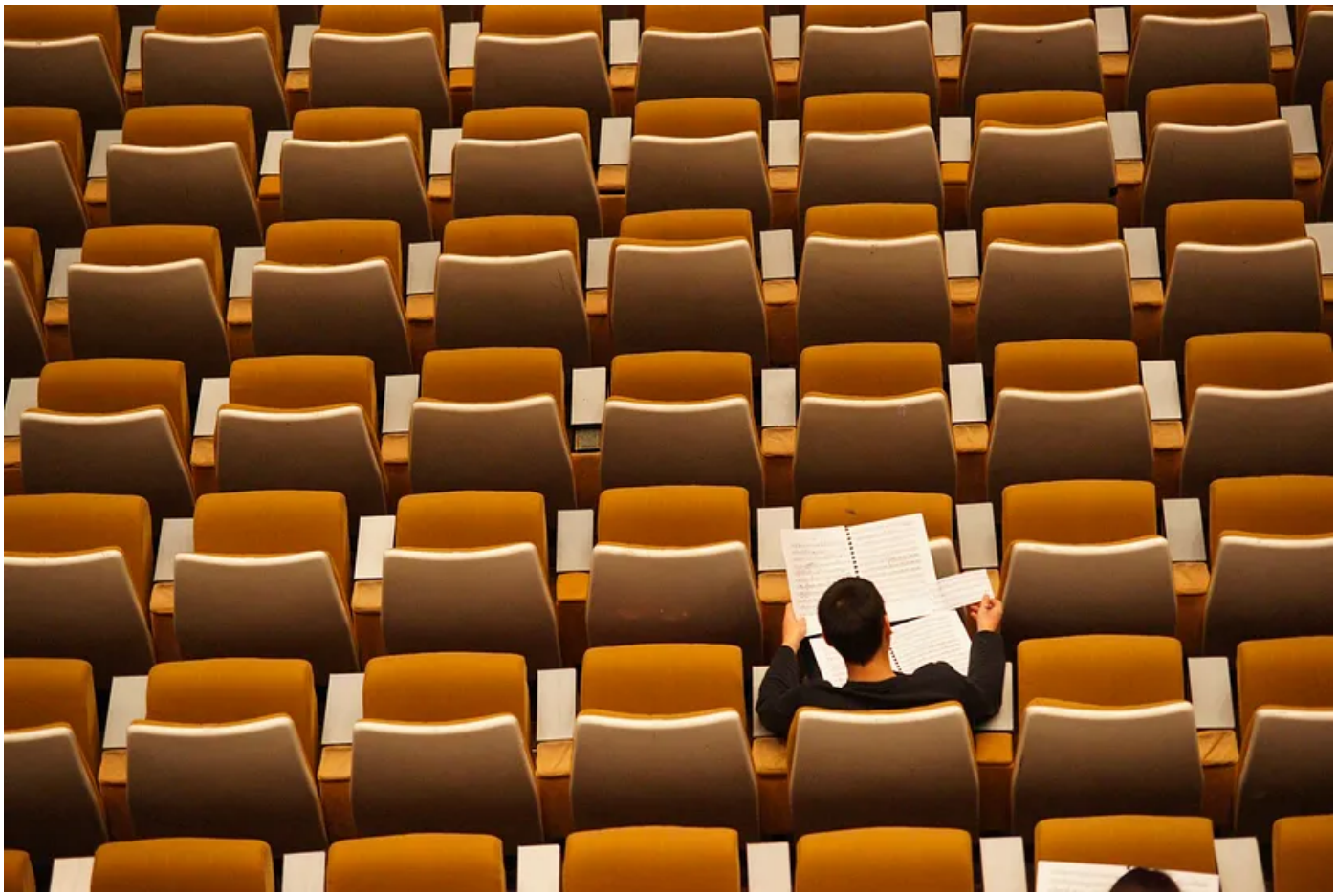
*One of the cool honed skills is **asking questions**. Not knowing the answer is not bad; it's bad not even to Google it. Before the interview, copy this document to yourself.*

Check out the notion version in english, in russian.

The questions were created with the help of:

ds girl
Alexander Babiy
канал что-то на DL-ском
канал Dealer.AI

канал Плюшевый Питон



CLASSIC NLP

TF-IDF & ML (8)

1. Write TF-IDF from scratch.

2. What is normalization in TF-IDF ?

3. Why do you need to know about TF-IDF in our time, and how can you use it in complex models?

4. Explain how Naive Bayes works. What can you use it for?

5. How can SVM be prone to overfitting?

6. Explain possible methods for text preprocessing (lemmatization and stemming). What algorithms do you know for this, and in what cases would you use them?

7. What metrics for text similarity do you know?

8. Explain the difference between cosine similarity and cosine distance. Which of

these values can be negative? How would you use them?

METRICS (7)

9. Explain precision and recall in simple words and what you would look at in the absence of F1 score?

10. In what case would you observe changes in specificity ?

11. When would you look at macro, and when at micro metrics? Why does the weighted metric exist?

12. What is perplexity? What can we consider it with?

13. What is the BLEU metric?

14. Explain the difference between different types of ROUGE metrics?

15. What is the difference between BLUE and ROUGE?

WORD2VEC(9)

16. Explain how Word2Vec learns? What is the loss function? What is maximized?

17. What methods of obtaining embeddings do you know? When will each be better?

18. What is the difference between static and contextual embeddings?

19. What are the two main architectures you know, and which one learns faster?

20. What is the difference between Glove, ELMO, FastText, and Word2Vec ?

21. What is negative sampling and why is it needed? What other tricks for Word2Vec do you know, and how can you apply them?

22. What are dense and sparse embeddings? Provide examples.

23. Why might the dimensionality of embeddings be important?

24. What problems can arise when training Word2Vec on short textual data, and how can you deal with them?

RNN & CNN(7)

25. How many training parameters are there in a simple 1-layer RNN ?

26. How does RNN training occur?

27. What problems exist in RNN?

28. What types of RNN networks do you know? Explain the difference between GRU and LSTM?

29. What parameters can we tune in such networks? (Stacking, number of layers)

30. What are vanishing gradients for RNN? How do you solve this problem?

31. Why use a Convolutional Neural Network in NLP, and how can you use it? How can you compare CNN within the attention paradigm?

NLP and TRANSFORMERS

ATTENTION AND TRANSFORMER ARCHITECTURE (15)

32. How do you compute attention ? (additional: for what task was it proposed, and why?)

33. Complexity of attention? Compare it with the complexity in RNN.

34. Compare RNN and attention . In what cases would you use attention, and when RNN?

35. Write attention from scratch.

36. Explain masking in attention.

37. What is the dimensionality of the self-attention matrix?

38. What is the difference between BERT and GPT in terms of attention calculation ?

39. What is the dimensionality of the embedding layer in the transformer?

40. Why are embeddings called contextual? How does it work?

41. What is used in transformers, layer norm or batch norm , and why?

42. Why do transformers have PreNorm and PostNorm ?

43. Explain the difference between soft and hard (local/global) attention?

44. Explain multihead attention.

45. What other types of attention mechanisms do you know? What are the purposes of these modifications?

46. How does self-attention become more complex with an increase in the number of heads?

TRANSFORMER MODEL TYPES (7)

47. Why does BERT largely lag behind RoBERTa , and what can you take from RoBERTa?

48. What are T5 and BART models? How do they differ?

49. What are task-agnostic models? Provide examples.

50. Explain transformer models by comparing BERT, GPT, and T5.

51. What major problem exists in BERT, GPT, etc., regarding model knowledge? How can this be addressed?

52. How does a decoder-like GPT work during training and inference? What is the difference?

53. Explain the difference between heads and layers in transformer models.

POSITIONAL ENCODING (6)

54. Why is information about positions lost in embeddings of transformer models with attention?

55. Explain approaches to positional embeddings and their pros and cons.

56. Why can't we simply add an embedding with the token index?

57. Why don't we train positional embeddings?

58. What is relative and absolute positional encoding?

59. Explain in detail the working principle of rotary positional embeddings.

PRETRAINING (4)

60. How does causal language modeling work?

61. When do we use a pretrained model?

62. How to train a transformer from scratch? Explain your pipeline, and in what cases would you do this?

63. What models, besides BERT and GPT, do you know for various pretraining tasks?

TOKENIZERS (9)

64. What types of tokenizers do you know? Compare them.

65. Can you extend a tokenizer? If yes, in what case would you do this? When would you retrain a tokenizer? What needs to be done when adding new tokens?

66. How do regular tokens differ from special tokens?

67. Why is lemmatization not used in transformers? And why do we need tokens?

68. How is a tokenizer trained? Explain with examples of WordPiece and BPE .

69. What position does the CLS vector occupy? Why?

70. What tokenizer is used in BERT, and which one in GPT?

71. Explain how modern tokenizers handle out-of-vocabulary words?

72. What does the tokenizer vocab size affect? How will you choose it in the case of new training?

TRAINING (8)

73. What is class imbalance? How can it be identified? Name all approaches to solving this problem.

74. Can dropout be used during inference, and why?

75. What is the difference between the Adam optimizer and AdamW?

76. How do consumed resources change with gradient accumulation?

77. How to optimize resource consumption during training?

78. What ways of distributed training do you know?

79. What is textual augmentation? Name all methods you know.

80. Why is padding less frequently used? What is done instead?

81. Explain how warm-up works.

82. Explain the concept of gradient clipping?

83. How does teacher forcing work, provide examples?

84. Why and how should skip connections be used?

85. What are adapters? Where and how can we use them?

86. Explain the concepts of metric learning. What approaches do you know?

INFERENCE (4)

87. What does the temperature in softmax control? What value would you set?

88. Explain types of sampling in generation? top-k, top-p, nucleus sampling?

89. What is the complexity of beam search, and how does it work?

90. What is sentence embedding? What are the ways you can obtain it?

LLM (13)

91. How does LoRA work? How would you choose parameters? Imagine that we want to fine-tune a large language model, apply LORA with a small R, but the model still doesn't fit in memory. What else can be done?

92. What is the difference between prefix tuning , p-tuning , and prompt tuning ?

93. Explain the scaling law .

94. Explain all stages of LLM training. From which stages can we abstain, and in what cases?

95. How does RAG work? How does it differ from few-shot KNN?

96. What quantization methods do you know? Can we fine-tune quantized models?

97. How can you prevent catastrophic forgetting in LLM?

98. Explain the working principle of KV cache , Grouped-Query Attention , and MultiQuery Attention .

99. Explain the technology behind MixTral, what are its pros and cons?

100. How are you? How are things going?

If you found the information helpful and want to thank me in other ways, you can buy me coffee.