Questionnaire (10th chapter)

Total de pontos 0/0

O e-mail do participante (**breno.xavier@ufpe.br**) foi registrado durante o envio deste formulário.

What is "self-supervised learning"? *

Is a training technique where a model learns from unlabeled data without requiring explicit supervision

What is a "language model"? *

It is achine learning model that is trained to predict the likelihood of words or sequences of words in a given text corpus

Why is a language model considered self-supervised? *

Considered self-supervised because it is trained on unlabeled data, without requiring explicit supervision from labeled data

What are self-supervised models usually used for? *

For pretraining tasks in natural language processing, such as language modeling, which can then be fine-tuned for downstream tasks like text classification or question answering

Why do we fine-tune language models? *

To adapt them to specific tasks and improve their performance on those tasks

What are the three steps to create a state-of-the-art text classifier? *

1- create a language model and fine-tune it on a large text corpus. 2 -use the fine-tuned model to generate embeddings for the labeled dataset. 3 - train a classifier on top of the embeddings

How do the 50,000 unlabeled movie reviews help us create a better text classifier for the IMDb dataset? *

Providing additional data for pretraining the language model, which can improve its ability to capture semantic relationships and general language patterns

What are the three steps to prepare your data for a language model? *

1- gather raw text data, 2 - tokenize and numericalize the text, and 3 - create batches of numericalized tokens for training

What is "tokenization"? Why do we need it? *

It is the process of splitting text into smaller units, usually words or subwords. To make it easier for a machine to process

Name three different approaches to tokenization. *

Word-based tokenization, Subword-based tokenization, and Character-based tokenization

What is `xxbos`? *

Is a special token that represents the beginning of a text sequence

List four rules that fastai applies to text during tokenization. *

1 - lowercase all text, 2 - handle punctuation, 3 - replace repeated characters, and 4 - replace unknown words with a special token

Why are repeated characters replaced with a token showing the number of repetitions and the character that's repeated? *

To help the model learn patterns in text that may not be immediately obvious

What is "numericalization"? *

Is the process of mapping tokens to unique integers to create a numerical representation of the text that can be fed into a machine learning model

Why might there be words that are replaced with the "unknown word" token? *

Words may be replaced with the "unknown word" token if they appear very infrequently in the training data, making it difficult for the model to learn anything useful about them

With a batch size of 64, the first row of the tensor representing the first batch contains the first 64 tokens for the dataset. What does the second row of that tensor contain? What does the first row of the second batch contain? *

With a batch size of 64, the second row of the tensor representing the first batch contains the 65th to 128th tokens of the dataset. The first row of the second batch contains the 129th to 192nd tokens of the dataset. This continues until all the tokens are processed

Why do we need padding for text classification? Why don't we need it for language modeling? *

Is needed for text classification to ensure that all sequences have the same length. It is not necessary for language modeling because the model can learn to process sequences of varying length

What does an embedding matrix for NLP contain? What is its shape? *

Contains a vector representation for each unique token in the vocabulary. Its shape is typically (vocabulary_size x embedding_dimension)

What is "perplexity"? *

A measure of how well a language model predicts a given sequence of text. It is calculated as the exponentiated average negative log-likelihood of the test set

Why do we have to pass the vocabulary of the language model to the classifier data block? *

Is passed to the classifier data block so that the classifier knows which tokens are valid inputs

What is "gradual unfreezing"? *

It is the process of gradually allowing the fine-tuning of deeper layers of a pre-trained model during training

Why is text generation always likely to be ahead of automatic identification of machine-generated texts? *

Because language models can be trained to generate text that closely mimics human writing, making it difficult to distinguish from human-generated text