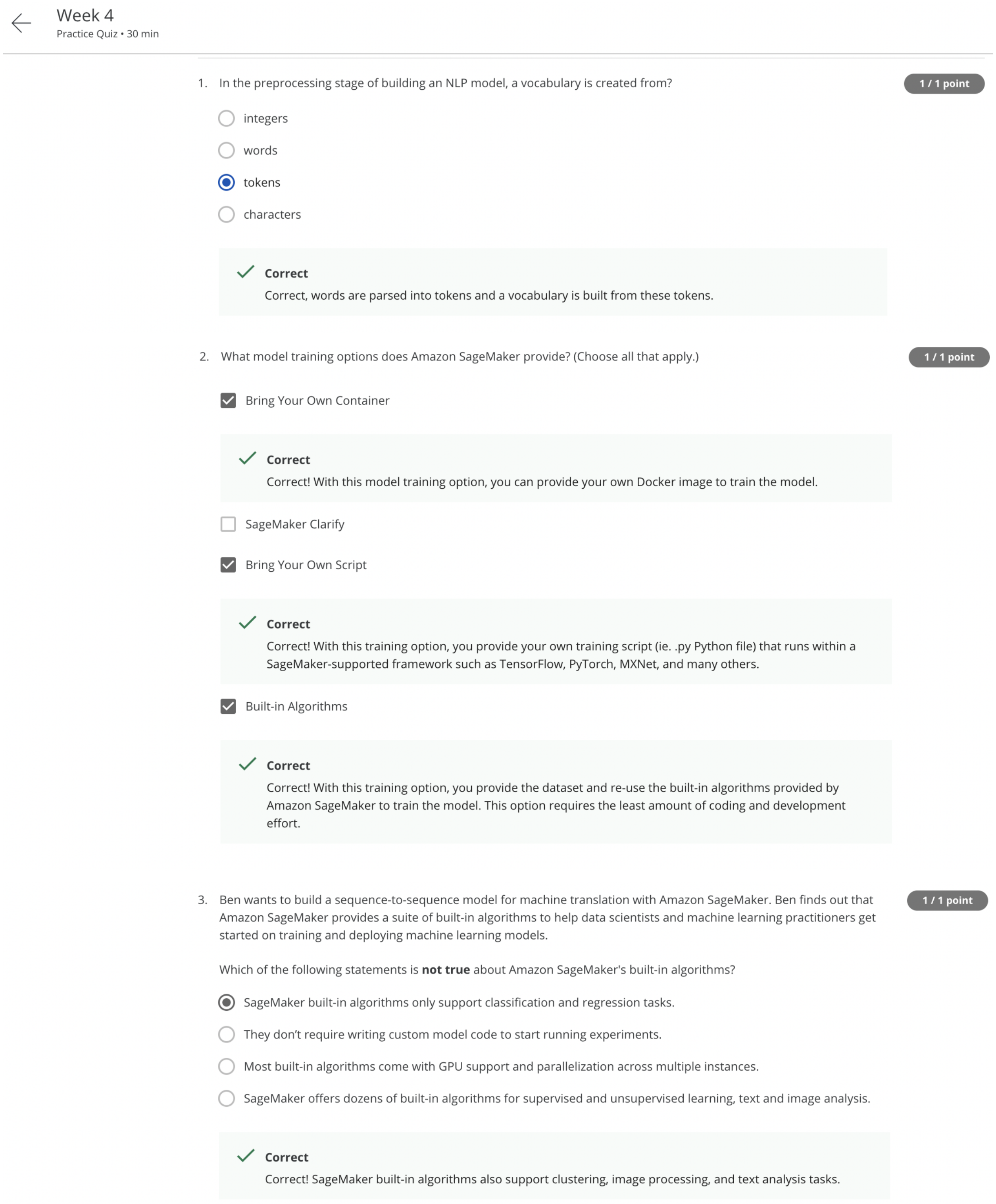
Hong Hanh

Practical Data Science Specialization



Week 4

Practice Quiz • 30 min

1. In the preprocessing stage of building an NLP model, a vocabulary is created from?

1 / 1 point

○ integers

○ words

● tokens

○ characters

✓ **Correct**

Correct, words are parsed into tokens and a vocabulary is built from these tokens.

2. What model training options does Amazon SageMaker provide? (Choose all that apply.)

1 / 1 point

☑ Bring Your Own Container

✓ **Correct**

Correct! With this model training option, you can provide your own Docker image to train the model.

□ SageMaker Clarify

☑ Bring Your Own Script

✓ **Correct**

Correct! With this training option, you provide your own training script (ie. .py Python file) that runs within a SageMaker-supported framework such as TensorFlow, PyTorch, MXNet, and many others.

☑ Built-in Algorithms

✓ **Correct**

Correct! With this training option, you provide the dataset and re-use the built-in algorithms provided by Amazon SageMaker to train the model. This option requires the least amount of coding and development effort.

3. Ben wants to build a sequence-to-sequence model for machine translation with Amazon SageMaker. Ben finds out that Amazon SageMaker provides a suite of built-in algorithms to help data scientists and machine learning practitioners get started on training and deploying machine learning models.

1 / 1 point

Which of the following statements is **not true** about Amazon SageMaker's built-in algorithms?

● SageMaker built-in algorithms only support classification and regression tasks.

○ They don't require writing custom model code to start running experiments.

○ Most built-in algorithms come with GPU support and parallelization across multiple instances.

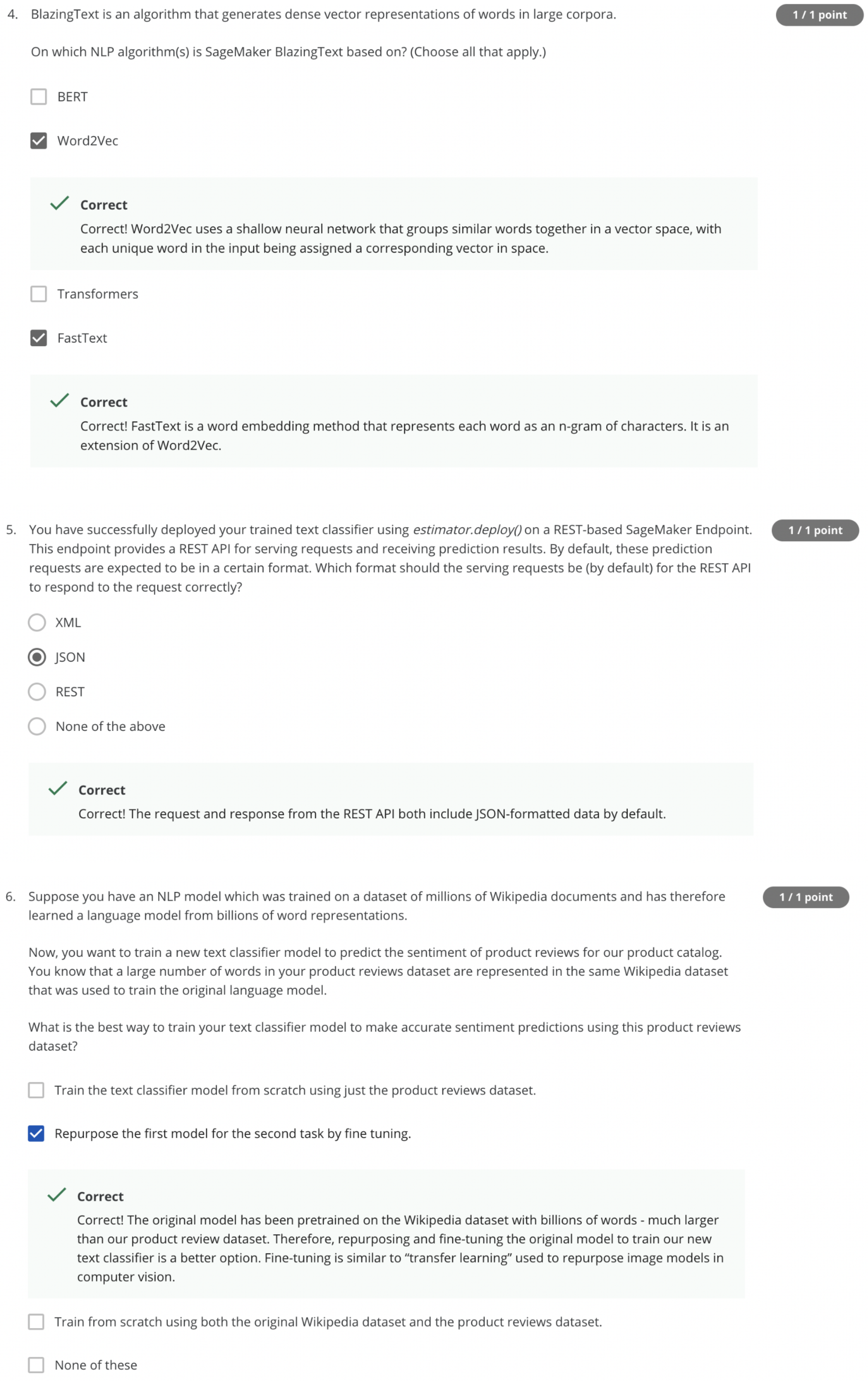
○ SageMaker offers dozens of built-in algorithms for supervised and unsupervised learning, text and image analysis.

✓ **Correct**

Correct! SageMaker built-in algorithms also support clustering, image processing, and text analysis tasks.

Hong Hanh

Practical Data Science Specialization



4. BlazingText is an algorithm that generates dense vector representations of words in large corpora.

1 / 1 point

On which NLP algorithm(s) is SageMaker BlazingText based on? (Choose all that apply.)

□ BERT

☑ Word2Vec

✓ **Correct**

Correct! Word2Vec uses a shallow neural network that groups similar words together in a vector space, with each unique word in the input being assigned a corresponding vector in space.

□ Transformers

☑ FastText

✓ **Correct**

Correct! FastText is a word embedding method that represents each word as an n-gram of characters. It is an extension of Word2Vec.

5. You have successfully deployed your trained text classifier using *estimator.deploy()* on a REST-based SageMaker Endpoint. This endpoint provides a REST API for serving requests and receiving prediction results. By default, these prediction requests are expected to be in a certain format. Which format should the serving requests be (by default) for the REST API to respond to the request correctly?

1 / 1 point

○ XML

● JSON

○ REST

○ None of the above

✓ **Correct**

Correct! The request and response from the REST API both include JSON-formatted data by default.

6. Suppose you have an NLP model which was trained on a dataset of millions of Wikipedia documents and has therefore learned a language model from billions of word representations.

1 / 1 point

Now, you want to train a new text classifier model to predict the sentiment of product reviews for our product catalog. You know that a large number of words in your product reviews dataset are represented in the same Wikipedia dataset that was used to train the original language model.

What is the best way to train your text classifier model to make accurate sentiment predictions using this product reviews dataset?

□ Train the text classifier model from scratch using just the product reviews dataset.

☑ Repurpose the first model for the second task by fine tuning.

✓ **Correct**

Correct! The original model has been pretrained on the Wikipedia dataset with billions of words - much larger than our product review dataset. Therefore, repurposing and fine-tuning the original model to train our new text classifier is a better option. Fine-tuning is similar to “transfer learning” used to repurpose image models in computer vision.

□ Train from scratch using both the original Wikipedia dataset and the product reviews dataset.

□ None of these