End-of-chapter quiz

This chapter covered a lot of ground! Don't worry if you didn't grasp all the details; the next chapters will help you understand how things work under the hood.

Before moving on, though, let's test what you learned in this chapter.

1. The load_dataset() function in 🤗 Datasets allows you to load a dataset from which of the following locations?

- [x] Locally, e.g. on your laptop

Correct! Correct! You can pass the paths of local files to the `data_files` argument of `load_dataset()` to load local datasets.

- [x] The Hugging Face Hub

Correct! Correct! You can load datasets on the Hub by providing the dataset ID, e.g. `load_dataset('emotion')`.

- [x] A remote server

Correct! Correct! You can pass URLs to the `data_files` argument of `load_dataset()` to load remote files.

Submit **You got all the answers!**

2. Suppose you load one of the GLUE tasks as follows:

```
from datasets import load_dataset

dataset = load_dataset("glue", "mrpc", split="train")
```

Which of the following commands will produce a random sample of 50 elements from `dataset`?

- [ ] `dataset.sample(50)`
- [x] `dataset.shuffle().select(range(50))`

Correct! Correct! As you saw in this chapter, you first shuffle the dataset and then select the samples from it.

- [ ] `dataset.select(range(50)).shuffle()`

Submit **You got all the answers!**

3. Suppose you have a dataset about household pets called pets_dataset , which has a name column that denotes the name of each pet. Which of the following approaches would allow you to filter the dataset for all pets whose names start with the letter "L"?

- [x] `pets_dataset.filter(lambda x : x['name'].startswith('L'))`

Correct! Correct! Using a Python lambda function for these quick filters is a great idea. Can you think of another solution?

- [ ] `pets_dataset.filter(lambda x['name'].startswith('L'))`
- [x] Create a function like `def filter_names(x): return x['name'].startswith('L')` and run `pets_dataset.filter(filter_names)`.

Correct! Correct! Just like with `Dataset.map()`, you can pass explicit functions to `Dataset.filter()`. This is useful when you have some complex logic that isn't suitable for a short lambda function. Which of the other solutions would work?

Submit **You got all the answers!**

4. What is memory mapping?

- [ ] A mapping between CPU and GPU RAM
- [x] A mapping between RAM and filesystem storage

Correct! Correct! 🤗 Datasets treats each dataset as a memory-mapped file. This allows the library to access and operate on elements of the dataset without needing to fully load it into memory.

- [ ] A mapping between two files in the 🤗 Datasets cache

Submit **You got all the answers!**

5. Which of the following are the main benefits of memory mapping?

- [x] Accessing memory-mapped files is faster than reading from or writing to disk.

Correct! Correct! This allows 🤗 Datasets to be blazing fast. That's not the only benefit, though.

- [x] Applications can access segments of data in an extremely large file without having to read the whole file into RAM first.

Correct! Correct! This allows 🤗 Datasets to load multi-gigabyte datasets on your laptop without blowing up your CPU. What other advantage does memory mapping offer?

- [ ] It consumes less energy, so your battery lasts longer.

Submit **You got all the answers!**

6. Why does the following code fail?

```
from datasets import load_dataset

dataset = load_dataset("allocine", streaming=True, split="train")
dataset[0]
```

- [ ] It tries to stream a dataset that's too large to fit in RAM.
- [x] It tries to access an `IterableDataset`.

Correct! Correct! An `IterableDataset` is a generator, not a container, so you should access its elements using `next(iter(dataset))`.

- [ ] The `allocine` dataset doesn't have a `train` split.

Submit **You got all the answers!**

7. Which of the following are the main benefits of creating a dataset card?

- [x] It provides information about the intended use and supported tasks of the dataset so others in the community can make an informed decision about using it.

Correct! Correct! Undocumented datasets may be used to train models that may not reflect the intentions of the dataset creators, or may produce models whose legal status is murky if they're trained on data that violates privacy or licensing restrictions. This isn't the only benefit, though!

- [x] It helps draw attention to the biases that are present in a corpus.

Correct! Correct! Almost all datasets have some form of bias, which can produce negative consequences downstream. Being aware of them helps model builders understand how to address the inherent biases. What else do dataset cards help with?

- [x] It improves the chances that others in the community will use my dataset.

Correct! Correct! A well-written dataset card will tend to lead to higher usage of your precious dataset. What other benefits does it offer?

Submit **You got all the answers!**

8. What is semantic search?

- [ ] A way to search for exact matches between the words in a query and the documents in a corpus
- [x] A way to search for matching documents by understanding the contextual meaning of a query

Correct! Correct! Semantic search uses embedding vectors to represent queries and documents, and uses a similarity metric to measure the amount of overlap between them. How else might you describe it?

- [x] A way to improve search accuracy

Correct! Correct! Semantic search engines can capture the intent of a query much better than keyword matching and typically retrieve documents with higher precision. But this isn't the only right answer - what else does semantic search provide?

Submit **You got all the answers!**

9. For asymmetric semantic search, you usually have:

- [x] A short query and a longer paragraph that answers the query

Correct! Correct!

- [ ] Queries and paragraphs that are of about the same length
- [ ] A long query and a shorter paragraph that answers the query

Submit **You got all the answers!**

10. Can I use 🤗 Datasets to load data for use in other domains, like speech processing?

- [ ] No
- [x] Yes

Correct! Correct! Check out the exciting developments with speech and vision in the 🤗 Transformers library to see how 🤗 Datasets is used in these domains.

Submit **You got all the answers!**

</> Update on GitHub