Troubleshooting

This guide aims to provide you the tools and knowledge required to navigate some common issues. If the suggestions listed
in this guide do not cover your such situation, please refer to the Asking for Help section to learn where to
find help with your specific issue.

Issues when uploading datasets with `push_to_hub`

Authentication issues

If you are experiencing authentication issues when sharing a dataset on 🤗 Hub using Dataset.push_to_hub() and a Hugging Face
access token:

- Make sure that the Hugging Face token you're using to authenticate yourself is a token with **write** permission.
- On OSX, it may help to clean up all the huggingface.co passwords on your keychain access, as well as reconfigure `git config --global credential.helper osxkeychain`, before using `huggingface-cli login`.

Alternatively, you can use SSH keys to authenticate yourself - read more in the 🤗 Hub documentation.

Lost connection on large dataset upload

When uploading large datasets to Hub, if the number of dataset shards is large, it can create too many commits for the Hub in a
short period. This will result in a connection error.
The connection error can also be caused by a HTTP 500 error returned by AWS S3 bucket that Hub uses internally.
In either situation, you can re-run Dataset.push_to_hub() to proceed with the dataset upload. Hub will check the SHAs
of already uploaded shards to avoid reuploading them.

We are working on making upload process more robust to transient errors, so updating to the latest library version is always a good idea.

`Too Many Requests`

Uploading large datasets via `push_to_hub()` can result in an error:

```
HfHubHTTPError: 429 Client Error: Too Many Requests for url: ...
You have exceeded our hourly quotas for action: commit. We invite you to retry later.
```

If you encounter this issue, you need to upgrade the `datasets` library to the latest version (or at least `2.15.0`).

Issues when creating datasets from custom data

Loading images and audio from a folder

When creating a dataset from a folder, one of the most common issues is that the file structure does not follow the expected format, or there's an issue with the metadata file.

Learn more about required folder structure in corresponding documentation pages:

- AudioFolder
- ImageFolder

Pickling issues

Pickling issues when using `Dataset.from_generator`

When creating a dataset, IterableDataset.from_generator() and Dataset.from_generator() expect a "picklable" generator function.
This is required to hash the function using `pickle` to be able to cache the dataset on disk.

While generator functions are generally "picklable", note that generator objects are not. So if you're using a generator object, you will encounter a `TypeError` like this:

```
TypeError: cannot pickle 'generator' object
```

This error can also occur when using a generator function that uses a global object that is not "picklable", such as a DB connection, for example. If that's the case, you can initialize such object directly inside the generator function to avoid this error.

Pickling issues with `Dataset.map`

Pickling errors can also happen in the multiprocess Dataset.map() - objects are pickled to be passed to child processes.
If the objects used in the transformation are not picklable, it's not possible to cache the result of `map`, which leads to an error being raised.

Here are some ways to address this issue:

- A universal solution to pickle issues is to make sure the objects (or generator classes) are pickable manually by implementing `__getstate__` / `__setstate__` / `__reduce__`.
- You can also provide your own unique hash in `map` with the `new_fingerprint` argument.
- You can also disable caching by calling `datasets.disable_caching()`, however, this is undesirable - read more about importance of cache

Asking for help

If the above troubleshooting advice did not help you resolve your issue, reach out for help to the community and the team.

Forums

Ask for help on the Hugging Face forums - post your question in the 🤗Datasets category
Make sure to write a descriptive post with relevant context about your setup and reproducible code to maximize the likelihood that your problem is solved!

Discord

Post a question on Discord, and let the team and the community help you.

Community Discussions on 🤗 Hub

If you are facing issues creating a custom dataset on Hub, you can ask the Hugging Face team for help by opening a discussion in the Community tab of your dataset with this message:

```
# Dataset rewiew request for <Dataset name>

## Description

<brief description of the dataset>

## Files to review

- file1
- file2
- ...

cc @lhoestq @albertvillanova
```

GitHub Issues

Finally, if you suspect to have found a bug related to the library itself, create an Issue on the 🤗 Datasets GitHub repository. Include context regarding the bug: code snippet to reproduce, details about your environment and data, etc. to help us figure out what's wrong and how we can fix it.