

Introduction to Hugging Face

Estimated Time: 5 minutes

After completing this reading, you will be able to:

- Describe the purpose of the Hugging Face company
- Explain the libraries offered by Hugging Face
- Explain the purpose of the Hugging Face models you will use in this course

The Hugging Face open-source AI community provides developers and others interested in AI opportunities to collaborate through code development, chat, file sharing, and discussion of open-source libraries. The Hugging Face platform offers extensive open-source AI-related tools. The community and its platform allow users to build, train, and deploy AI models for a wide variety of applications.

You may hear the name Hugging Face to refer solely to its libraries, but this usage is not accurate. Hugging Face is a company that provides a platform for collaborating on AI-related libraries, models, and applications that use those models.

Hugging Face does offer extensive libraries that include Transformers, Datasets, and Tokenizers. Let's discuss each of these briefly.

Transformers library

The Transformers library provides pre-trained natural language processing (NLP) models that support text classification, translation, and summarization. This library offers a unified API for two popular deep-learning frameworks: TensorFlow and PyTorch. Developers can readily switch between the two as needed without tailoring their code to use the models in the Transformers library.

Datasets library

The Datasets library offers an expansive collection of data sets for training, testing, and evaluating models. For example, IMDb, the internet movie database, is available through the Hugging Face datasets library. IMDb contains movie reviews that you could use for sentiment analysis. Other data sets in the library include ones you can use for general language, translations, or questions and answers.

The library also contains tools and functions to manipulate and prepare data for processing. For example, with just a few lines of code, you can import a data set, ensure proper formatting, and split the data into training and testing sets.

Tokenizers library

NLP models and applications require text tokenization. Tokenization processes involve taking text and breaking it down into meaningful elements, such as prefixes or suffixes, to help identify linguistic properties and context. For example, tokenization can parse a sentence into words to identify and tag their parts of speech. The Hugging Face Tokenizers library offers a wide range of classes and functions to handle text tokenization.

Use of Hugging Face in this course

In the labs and projects in this course, you will use the Hugging Face HfInference module in the Hugging Face inference package. This module provides you access to several methods, including `featureExtraction()`. This method allows you to specify a model that you can then apply to given text inputs.

Specifically, you will use two models: a sentence-transformer model, `all-MiniLM-L6-v2`, and a model from the Facebook AI BART (Bidirectional and Auto-Regressive Transformers) family. The sentence transformer model generates embeddings while BART helps you classify text.

Conclusion

Hugging Face models are used for a wide variety of applications such as chatbots, sentiment analysis, text summaries, translation services, and for the purposes of this course, with vector data for embedding models.



Skills Network