Lesson   Downloads

Show TranscriptSummarize Video

HuggingFace Datasets library is a powerful tool for managing a variety of data types, like text and images, efficiently and easily. This resource is incredibly fast and doesn't use a lot of computer memory, making it great for handling big projects without any hassle.

Technical Terms Explained:

IMDb dataset: A dataset of movie reviews that can be used to train a machine learning model to understand human sentiments.

Apache Arrow: A software framework that allows for fast data processing

Code Example

Note that this code uses IPython functions (display and HTML) so it should be run in an IPython environment (e.g. Jupyter Notebook).

```python
from datasets import load_dataset
from IPython.display import HTML, display


# Load the IMDB dataset, which contains movie reviews
# and sentiment labels (positive or negative)
dataset = load_dataset("imdb")


# Fetch a revie from the training set
review_number = 42
sample_review = dataset["train"][review_number]


display(HTML(sample_review["text"][:450] + "..."))
# WARNING: This review contains SPOILERS. Do not read if you don't want some p
# film.
#
# With a cast like this, you wonder whether or not the actors and actresses kn
# see the script and say, `Hey, Close Encounters of the Third Kind was such a
# Did they even think to check on the director's credentials...


if sample_review["label"] == 1:
    print("Sentiment: Positive")
else:
    print("Sentiment: Negative")
# Sentiment: Negative
```

Resources

Hugging Face Datasets documentation index
Hugging Face Datasets search
Hugging Face load_datasets documentation
imdb Dataset documentation on Hugging Face