📚 Text and Language Datasets

- **The Pile**: An 800GB diverse, open-source dataset of English text, designed for training large language models.
- **Harvard Public Domain Books**: A collection of nearly one million public-domain books, offering a rich resource for training AI models.
- **IIT-Bombay AIKosh Dataset**: Comprises approximately 218,000 sentences and 1.5 million words from ancient Indian texts, aiding in AI research related to Indian knowledge systems.

🎁 1. The Pile (800GB diverse English corpus)

To work with this dataset:

Step 1: Install and load from Hugging Face (subset only recommended)

```
bash
CopyEdit
pip install datasets
python
CopyEdit
from datasets import load_dataset

# Load a small slice of The Pile
dataset = load_dataset("the_pile", split="train[:1000]")  # Only 1K records

# Preview
print(dataset[0]['text'][:500])
```

⚠️ **Note**: Full dataset is ~800GB. Use cloud/GPU environments for full-scale training.

📚 2. Harvard Public Domain Books

Harvard has partnered with Internet Archive. You can access books using the `internetarchive` Python package:

Step 1: Install the library

```
bash
CopyEdit
pip install internetarchive
```

Step 2: Download a public domain book (e.g., Shakespeare)

```
python
CopyEdit
import internetarchive

# Search for books from Harvard's public domain collection
results = internetarchive.search_items('collection:harvardlibrary AND
mediatype:texts')
```

```
for item in results:
    print(item['identifier'])
    break  # Just show one ID

# Download a specific item by ID
item = internetarchive.get_item('shakespearefoli00shak')  # Example
item.download(files=['shakespearefoli00shak_djvu.txt'])
```

Step 3: Read and use text

```
python
CopyEdit
with open('shakespearefoli00shak_djvu.txt', 'r', encoding='utf-8') as f:
    text = f.read()

print(text[:1000])
```

🕮 3. IIT-Bombay AIKosh Dataset

Currently hosted on **AI4Bharat** or similar institutional platforms. It may not be hosted via APIs, but you can work with any downloaded `.csv` or `.json` file.

Sample usage after downloading (assuming CSV format):

```
python
CopyEdit
import pandas as pd

# Load dataset
df = pd.read_csv("aikosh_dataset.csv")  # Update with the real path

# Preview
print(df.head())

# Analyze ancient sentence structure
print(df['sentence'].head())
```