data source 1

data source 2

data source N

CompositeDatasetDriver

- collect file names
- shuffle (seed)
- split by rank
- read file contents
- shuffle (seed)
- split by worker

CompositeDatasetDriver

- collect file names
- shuffle (seed)
- split by rank
- read file contents
- shuffle (seed)
- split by worker

CompositeDatasetDriver

- collect file names
- shuffle (seed)
- split by rank
- read file contents
- shuffle (seed)
- split by worker

normalize data dicts

normalize data dicts

normalize data dicts

multiplexing

- Tokenize
- Pad or stuff content
- log meta data
- filter keys
- batch
- to tensor
- to torch iterable

torch data loader