Sandbox for Beginners

In this notebook, we will dive into Sandbox, a comprehensive suite in Data-Juicer for data-model co-development. With Sandbox, users can experiment, iterate and refine data recipes and datasets based on some small-scale models and datasets to obtain some data processing inghts with low overhead. These insights then can be transfered to large scale to produce high-quality and high-performance models and datsets.

In Sandbox, users can use various configurable components from Data-Juicer, such as analysis and probe tools and data processing pipelines, and other open-source tools or architectures, such as model training and evalution metrics, to construct a data-model feedback loop for one-stop data recipe refinement.

For more details, please refer to the [Sandbox doc]https://github.com/modelscope/data-juicer/blob/main/docs/Sandbox.md().

Now let's begin to use sandbox to refine a initial data recipe in a simple example.

Prepare Config Files for Sandbox

The default one-trial pipeline of Sandbox including 4 types of jobs. Each type of jobs corresponds to a config group with one or more hooks and their config files. So there are also 4 types of config groups:

- Data/Model Probe -- `probe_job_configs`
- Iterative Recipe Refinement based on Probe Results -- `refine_recipe_job_configs`
- Dataset Processing and Model Training -- `execution_job_configs`
- Data/Model Evaluation -- `evaluation_job_configs`

In each config group, various hooks can be mounted and are organized as configurable job list. For each hook, we need to specify:

- hook name -- `hook`
- tag name for recording intermediate results -- `meta_name`
- Data-Juicer recipe config for some Data-Juicer components -- `dj_config`
- some specific parameters for this hook -- `extra_configs`

Here we take a typical data recipe refinement using the k-sigma method pipeline as the example. The sandbox config file could be:

```
# global parameters
project_name: 'sandbox-recipe-refinement'
experiment_name: 'sandbox-recipe-refinement-run0' # for wandb
tracer name
work_dir: './outputs/sandbox-process/'
```

```yaml
hpo_config: null                                    # path to a
configuration file when using auto-HPO tool.

# configs for each job, the jobs will be executed according to
the order in the list
probe_job_configs:
  - hook: 'ProbeViaAnalyzerHook'
    meta_name: 'analysis_ori_data'
    dj_configs: 'dj_process.yaml'
    extra_configs:

refine_recipe_job_configs:
  - hook: 'RefineRecipeViaKSigmaHook'
    meta_name: 'analysis_ori_data'
    dj_configs: 'dj_process.yaml'
    extra_configs:
      path_k_sigma_recipe: './outputs/sandbox-
process/k_sigma_new_recipe.yaml'

execution_job_configs:
  - hook: 'ProcessDataHook'
    meta_name:
    dj_configs: './outputs/sandbox-
process/k_sigma_new_recipe.yaml'
    extra_configs:
  - hook: 'TrainModelHook'
    meta_name:
    dj_configs:
    extra_configs: 'gpt3_extra_train_config.yaml'

evaluation_job_configs:
  - hook: 'ProbeViaAnalyzerHook'
    meta_name: 'analysis_processed_data'
    dj_configs: 'dj_process.yaml'
    extra_configs:
  - hook: 'EvaluateDataHook'
    meta_name: 'eval_data'
    dj_configs:
    extra_configs: 'gpt3_data_quality_eval_config.yaml'
```

All 4 steps and config groups are activated in this sandbox config:

1. In probe jobs, sandbox would analyze the original dataset with the `ProbeViaAnalyzerHook` using a Data-Juicer data recipe.
2. In iterative recipe refinement jobs, sandbox would refine the original recipe and generate a new refined recipe with the `RefineRecipeViaKSigmaHook` using the k-sigma method.
3. In execution jobs, sandbox processes the original dataset first using the refined data recipe with the `ProcessDataHook` and then train a GPT-3 model using the processed dataset with the `TrainModelHook`.
4. In evaluation jobs, sandbox analyze the processed dataset with the `ProbeViaAnalyzerHook` again to cehck the data quality from a data perspective and evalute the quality score of processed dataset with the `EvaluateDataHook` to check the data quality form a model perspective.

Now we write this sandbox config to a config file.

In [1]:
```
sandbox_config = '''
# global parameters
project_name: 'sandbox-recipe-refinement'
experiment_name: 'sandbox-recipe-refinement-run0' # for wandb tracer name
work_dir: './outputs/sandbox-process/'
hpo_config: null                                  # path to a configurati

# configs for each job, the jobs will be executed according to the order
probe_job_configs:
  - hook: 'ProbeViaAnalyzerHook'
    meta_name: 'analysis_ori_data'
    dj_configs: 'dj_process.yaml'
    extra_configs:

refine_recipe_job_configs:
  - hook: 'RefineRecipeViaKSigmaHook'
    meta_name: 'analysis_ori_data'
    dj_configs: 'dj_process.yaml'
    extra_configs:
      path_k_sigma_recipe: './outputs/sandbox-process/k_sigma_new_recipe.

execution_job_configs:
  - hook: 'ProcessDataHook'
    meta_name:
    dj_configs: './outputs/sandbox-process/k_sigma_new_recipe.yaml'
    extra_configs:
  - hook: 'TrainModelHook'
    meta_name:
    dj_configs:
    extra_configs: 'gpt3_extra_train_config.yaml'

evaluation_job_configs:
  - hook: 'ProbeViaAnalyzerHook'
    meta_name: 'analysis_processed_data'
    dj_configs: 'dj_process.yaml'
    extra_configs:
  - hook: 'EvaluateDataHook'
    meta_name: 'eval_data'
    dj_configs:
    extra_configs: 'gpt3_data_quality_eval_config.yaml'
'''

with open('sandbox_pipeline.yaml', 'w') as fout:
    fout.write(sandbox_config)
```

At the same time, we need to prepare config files for each hook as well.

`ProbeViaAnalyzerHook`

We need a initial Data-Juicer data recipe. Like previous notebooks, we use the demo processing recipe as an exmpale as well.

In [2]:
```
dj_process_config = '''
# global parameters
project_name: 'demo-process'
```

```
dataset_path: '../demos/data/demo-dataset.jsonl'  # path to your dataset
np: 4  # number of subprocess to process your dataset

export_path: './outputs/sandbox-process/demo-processed.jsonl'

# process schedule
# a list of several process operators with their arguments
process:
  - language_id_score_filter:
      lang: 'zh'
      min_score: 0.8
'''

with open('dj_process.yaml', 'w') as fout:
    fout.write(dj_process_config)
```

`RefineRecipeViaKSigmaHook`

We refine the initial data recipe above in this hook and generate a refined recipe to the outputs directory. So we only need to specify the path to store the generated recipe for this hook.

`ProcessDataHook`

In this hook, we use the generated refined recipe to process the original dataset, which has the same path to the previous refined recipe.

`TrainModelHook`

Training a model requires lots of parameters. In this hook, we train a GPT-3 model based on ModelScope, so we need training configs following the documents of ModelScope, which could be a JSON file or a YAML file. Here we take the YAML format as an example.

In [3]:
```
training_config = '''
type: modelscope
train_dataset: './outputs/sandbox-process/demo-processed.jsonl'
work_dir: './outputs/sandbox-process/'
model_name: "iic/nlp_gpt3_text-generation_chinese-base"
trainer_name: "nlp-base-trainer"
key_remapping:
  text: "src_txt"
train:
  max_epochs: 2
  lr_scheduler:
    type: "StepLR"
    step_size: 2
    options:
      by_epoch: false
  optimizer:
    type: "AdamW"
    lr: 0.0003
  dataloader:
    batch_size_per_gpu: 2
```

```
    workers_per_gpu: 0
'''

with open('gpt3_extra_train_config.yaml', 'w') as fout:
  fout.write(training_config)
```

`ProbeViaAnalyzerHook`

For this hook, we use the same data recipe to analyze the processed dataset, so the config is the same as the previous `ProbeViaAnalyzerHook`.

`EvaluateDataHook`

In this hook, we use the quality classifier to score for the samples in the processed dataset. So we need a config file for it and it's quite simple.

In [4]:
```
data_eval_config = '''
type: dj_text_quality_classifier
dataset_path: './outputs/sandbox-process/demo-processed.jsonl'
'''

with open('gpt3_data_quality_eval_config.yaml', 'w') as fout:
  fout.write(data_eval_config)
```

Start Sandbox

After all these config files are ready, we can start the sandbox by running the sandbox entry in Data-Juicer: `tools/sandbox_starter.py`. The usage is similar to the data processing and analysis tool:

In [5]:
```
!python ../tools/sandbox_starter.py --config sandbox_pipeline.yaml
```

```
wandb: Currently logged in as: hyl1024. Use `wandb login --relogin` to for
ce relogin
wandb: wandb version 0.17.6 is available!  To upgrade, please run:
wandb:  $ pip install wandb --upgrade
wandb: Tracking run with wandb version 0.17.4
wandb: Run data is saved locally in /root/projects/kdd_tutorial_notebooks/
wandb/run-20240809_103105-3l1fauog
wandb: Run `wandb offline` to turn off syncing.
wandb: Syncing run sandbox-recipe-refinement-run0
wandb: ⭐️ View project at https://wandb.ai/hyl1024/sandbox-recipe-refineme
nt
wandb: 🚀 View run at https://wandb.ai/hyl1024/sandbox-recipe-refinement/r
uns/3l1fauog
2024-08-09 10:31:05.978 | INFO     | data_juicer.core.sandbox.hooks:specif
y_dj_and_extra_configs:33 - Parsing Data-Juicer configs in the job.
2024-08-09 10:31:08 | INFO     | data_juicer.config.config:618 - Back up t
he input config file [/tmp/job_dj_config.json] into the work_dir [/root/pr
ojects/kdd_tutorial_notebooks/outputs/sandbox-process]
2024-08-09 10:31:08 | INFO     | data_juicer.config.config:640 - Configura
tion table:
╒═══════════════════════╤═══════════════════════════════════════════════════
═══════════════════════════════════════╕
│ key                   │ values
│
╞═══════════════════════╪═══════════════════════════════════════════════════
═══════════════════════════════════════╡
│ config                │ [Path_fr(/tmp/job_dj_config.json)]
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────────────────────┤
│ hpo_config            │ None
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────────────────────┤
│ data_probe_algo       │ 'uniform'
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────────────────────┤
│ data_probe_ratio      │ 1.0
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────────────────────┤
│ project_name          │ 'demo-process'
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────────────────────┤
│ executor_type         │ 'default'
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────────────────────┤
│ dataset_path          │ '/root/projects/data-juicer/demos/data/demo-da
taset.jsonl'                          │
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────────────────────┤
│ export_path           │ '/root/projects/kdd_tutorial_notebooks/output
s/sandbox-process/demo-processed.jsonl' │
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────────────────────┤
│ export_shard_size     │ 0
│
```

```
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ export_in_parallel      │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ keep_stats_in_res_ds    │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ keep_hashes_in_res_ds   │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ np                      │ 4
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ text_keys               │ 'text'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ image_key               │ 'images'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ image_special_token     │ '<__dj__image>'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ audio_key               │ 'audios'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ audio_special_token     │ '<__dj__audio>'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ video_key               │ 'videos'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ video_special_token     │ '<__dj__video>'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ eoc_special_token       │ '<|__dj__eoc|>'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ suffixes                │ []
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ use_cache               │ True
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ ds_cache_dir            │ '/root/.cache/huggingface/datasets'
│
```

```
├─────────────────────────┼──────────────────────────────────────────────
───────────────────────────────────────┤
│ cache_compress          │ None
│
├─────────────────────────┼──────────────────────────────────────────────
───────────────────────────────────────┤
│ use_checkpoint          │ False
│
├─────────────────────────┼──────────────────────────────────────────────
───────────────────────────────────────┤
│ temp_dir                │ None
│
├─────────────────────────┼──────────────────────────────────────────────
───────────────────────────────────────┤
│ open_tracer             │ False
│
├─────────────────────────┼──────────────────────────────────────────────
───────────────────────────────────────┤
│ op_list_to_trace        │ []
│
├─────────────────────────┼──────────────────────────────────────────────
───────────────────────────────────────┤
│ trace_num               │ 10
│
├─────────────────────────┼──────────────────────────────────────────────
───────────────────────────────────────┤
│ op_fusion               │ False
│
├─────────────────────────┼──────────────────────────────────────────────
───────────────────────────────────────┤
│ process                 │ [{'language_id_score_filter': {'accelerator':
None,                                  │
│                         │                               'audio_key': 'a
udios',                                │
│                         │                               'cpu_required':
1,                                    │
│                         │                               'image_key': 'i
mages',                                │
│                         │                               'lang': 'zh',
│
│                         │                               'mem_required':
0,                                    │
│                         │                               'min_score': 0.
8,                                     │
│                         │                               'num_proc': 4,
│
│                         │                               'stats_export_p
ath': None,                            │
│                         │                               'text_key': 'te
xt',                                   │
│                         │                               'video_key': 'v
ideos'}}]                              │
├─────────────────────────┼──────────────────────────────────────────────
───────────────────────────────────────┤
│ percentiles             │ []
│
├─────────────────────────┼──────────────────────────────────────────────
───────────────────────────────────────┤
│ export_original_dataset │ False
│
```

```
├─────────────────────────┼───────────────────────────────────────────────
──────────────────────────────────────┤
│ save_stats_in_one_file  │ False
│
├─────────────────────────┼───────────────────────────────────────────────
──────────────────────────────────────┤
│ ray_address             │ 'auto'
│
├─────────────────────────┼───────────────────────────────────────────────
──────────────────────────────────────┤
│ debug                   │ False
│
├─────────────────────────┼───────────────────────────────────────────────
──────────────────────────────────────┤
│ work_dir                │ '/root/projects/kdd_tutorial_notebooks/output
s/sandbox-process'                     │
├─────────────────────────┼───────────────────────────────────────────────
──────────────────────────────────────┤
│ timestamp               │ '20240809103107'
│
├─────────────────────────┼───────────────────────────────────────────────
──────────────────────────────────────┤
│ dataset_dir             │ '/root/projects/data-juicer/demos/data'
│
├─────────────────────────┼───────────────────────────────────────────────
──────────────────────────────────────┤
│ add_suffix              │ False
│
╘═════════════════════════╧═══════════════════════════════════════════════
══════════════════════════════════════╛
2024-08-09 10:31:08 | INFO     | data_juicer.core.analyzer:37 - Using cach
e compression method: [None]
2024-08-09 10:31:08 | INFO     | data_juicer.core.analyzer:42 - Setting up
data formatter...
2024-08-09 10:31:08 | INFO     | data_juicer.core.analyzer:51 - Preparing
exporter...
2024-08-09 10:31:08 | INFO     | data_juicer.core.sandbox.hooks:63 - Begin
to analyze data
2024-08-09 10:31:08 | INFO     | data_juicer.core.analyzer:75 - Loading da
taset from data formatter...
2024-08-09 10:31:09 | INFO     | data_juicer.format.formatter:185 - Unifyi
ng the input dataset formats...
2024-08-09 10:31:09 | INFO     | data_juicer.format.formatter:200 - There
are 6 sample(s) in the original dataset.
Filter (num_proc=4): 100%|##########| 6/6 [00:00<00:00, 62.62 examples/s]
2024-08-09 10:31:09 | INFO     | data_juicer.format.formatter:214 - 6 samp
les left after filtering empty text.
2024-08-09 10:31:09 | INFO     | data_juicer.format.mixture_formatter:137
- sampled 6 from 6
2024-08-09 10:31:09 | INFO     | data_juicer.format.mixture_formatter:143
- There are 6 in final dataset
2024-08-09 10:31:09 | INFO     | data_juicer.core.analyzer:81 - Preparing
process operators...
2024-08-09 10:31:09 | INFO     | data_juicer.utils.model_utils:103 - Loadi
ng fasttext language identification model...
Warning : `load_model` does not return WordVectorModel or SupervisedModel
any more, but a `FastText` object which is very similar.
2024-08-09 10:31:09 | INFO     | data_juicer.core.analyzer:86 - Computing
the stats of dataset...
Adding new column for stats (num_proc=4): 100%|##########| 6/6 [00:00<00:0
```

```
0, 68.28 examples/s]
language_id_score_filter_compute_stats (num_proc=4):   0%|          | 0/6
[00:00<?, ? examples/s]2024-08-09 10:31:09 | INFO     | data_juicer.utils.
model_utils:103 - Loading fasttext language identification model...
Warning : `load_model` does not return WordVectorModel or SupervisedModel
any more, but a `FastText` object which is very similar.
2024-08-09 10:31:09 | INFO     | data_juicer.utils.model_utils:103 - Loadi
ng fasttext language identification model...
Warning : `load_model` does not return WordVectorModel or SupervisedModel
any more, but a `FastText` object which is very similar.
2024-08-09 10:31:09 | INFO     | data_juicer.utils.model_utils:103 - Loadi
ng fasttext language identification model...
Warning : `load_model` does not return WordVectorModel or SupervisedModel
any more, but a `FastText` object which is very similar.
2024-08-09 10:31:09 | INFO     | data_juicer.utils.model_utils:103 - Loadi
ng fasttext language identification model...
Warning : `load_model` does not return WordVectorModel or SupervisedModel
any more, but a `FastText` object which is very similar.
language_id_score_filter_compute_stats (num_proc=4): 100%|##########| 6/6
[00:00<00:00, 36.36 examples/s]
language_id_score_filter_process (num_proc=4): 100%|##########| 6/6 [00:00
<00:00, 66.52 examples/s]
2024-08-09 10:31:09 | INFO     | data_juicer.core.data:193 - OP [language_
id_score_filter] Done in 0.484s. Left 6 samples.
2024-08-09 10:31:09 | INFO     | data_juicer.core.analyzer:101 - Exporting
dataset to disk...
2024-08-09 10:31:09 | INFO     | data_juicer.core.exporter:111 - Exporting
computed stats into a single file...
Creating json from Arrow format: 100%|##########| 1/1 [00:00<00:00, 205.87
ba/s]
2024-08-09 10:31:09 | INFO     | data_juicer.core.analyzer:113 - Applying
overall analysis on stats...
100%|##########| 2/2 [00:00<00:00, 19065.02it/s]
2024-08-09 10:31:10 | INFO     | data_juicer.core.analyzer:120 - The overa
ll analysis results are:        lang lang_score
count   6.0        6.0
unique  3.0        NaN
top      en        NaN
freq    3.0        NaN
mean    NaN    0.97077
std     NaN   0.021711
min     NaN   0.945098
25%     NaN   0.958538
50%     NaN   0.964044
75%     NaN   0.987662
max     NaN   0.999194
2024-08-09 10:31:10 | INFO     | data_juicer.core.analyzer:122 - Applying
column-wise analysis on stats...
Column: 100%|##########| 2/2 [00:00<00:00,  6.50it/s]
2024-08-09 10:31:10 | INFO     | data_juicer.core.sandbox.hooks:33 - Parsi
ng Data-Juicer configs in the job.
2024-08-09 10:31:13 | INFO     | data_juicer.config.config:618 - Back up t
he input config file [/tmp/job_dj_config.json] into the work_dir [/root/pr
ojects/kdd_tutorial_notebooks/outputs/sandbox-process]
2024-08-09 10:31:13 | INFO     | data_juicer.config.config:640 - Configura
tion table:
╒═══════════════════════╤═══════════════════════════════════════════════════
═══════════════════════════════════╕
│ key                   │ values
│
```

```
╞═════════════════════════╪═══════════════════════════════════════════════
═══════════════════════════════════════╡
│ config                  │ [Path_fr(/tmp/job_dj_config.json)]
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ hpo_config              │ None
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ data_probe_algo         │ 'uniform'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ data_probe_ratio        │ 1.0
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ project_name            │ 'demo-process'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ executor_type           │ 'default'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ dataset_path            │ '/root/projects/data-juicer/demos/data/demo-da
taset.jsonl'                           │
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ export_path             │ '/root/projects/kdd_tutorial_notebooks/output
s/sandbox-process/demo-processed.jsonl' │
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ export_shard_size       │ 0
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ export_in_parallel      │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ keep_stats_in_res_ds    │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ keep_hashes_in_res_ds   │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ np                      │ 4
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ text_keys               │ 'text'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ image_key               │ 'images'
│
```

```
├──────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ image_special_token     │ '<__dj__image>'
│
├──────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ audio_key               │ 'audios'
│
├──────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ audio_special_token     │ '<__dj__audio>'
│
├──────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ video_key               │ 'videos'
│
├──────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ video_special_token     │ '<__dj__video>'
│
├──────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ eoc_special_token       │ '<|__dj__eoc|>'
│
├──────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ suffixes                │ []
│
├──────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ use_cache               │ True
│
├──────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ ds_cache_dir            │ '/root/.cache/huggingface/datasets'
│
├──────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ cache_compress          │ None
│
├──────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ use_checkpoint          │ False
│
├──────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ temp_dir                │ None
│
├──────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ open_tracer             │ False
│
├──────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ op_list_to_trace        │ []
│
├──────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ trace_num               │ 10
│
```

```
├─────────────────────────┼───────────────────────────────────────────────
──────────────────────────────────────┤
│ op_fusion               │ False
│
├─────────────────────────┼───────────────────────────────────────────────
──────────────────────────────────────┤
│ process                 │ [{'language_id_score_filter': {'accelerator':
None,                                  │
│                         │                                'audio_key': 'a
udios',                                │
│                         │                                'cpu_required':
1,                                    │
│                         │                                'image_key': 'i
mages',                                │
│                         │                                'lang': 'zh',
│
│                         │                                'mem_required':
0,                                    │
│                         │                                'min_score': 0.
8,                                     │
│                         │                                'num_proc': 4,
│
│                         │                                'stats_export_p
ath': None,                            │
│                         │                                'text_key': 'te
xt',                                   │
│                         │                                'video_key': 'v
ideos'}}]                              │
├─────────────────────────┼───────────────────────────────────────────────
──────────────────────────────────────┤
│ percentiles             │ []
│
├─────────────────────────┼───────────────────────────────────────────────
──────────────────────────────────────┤
│ export_original_dataset │ False
│
├─────────────────────────┼───────────────────────────────────────────────
──────────────────────────────────────┤
│ save_stats_in_one_file  │ False
│
├─────────────────────────┼───────────────────────────────────────────────
──────────────────────────────────────┤
│ ray_address             │ 'auto'
│
├─────────────────────────┼───────────────────────────────────────────────
──────────────────────────────────────┤
│ debug                   │ False
│
├─────────────────────────┼───────────────────────────────────────────────
──────────────────────────────────────┤
│ work_dir                │ '/root/projects/kdd_tutorial_notebooks/output
s/sandbox-process'                      │
├─────────────────────────┼───────────────────────────────────────────────
──────────────────────────────────────┤
│ timestamp               │ '20240809103112'
│
├─────────────────────────┼───────────────────────────────────────────────
──────────────────────────────────────┤
│ dataset_dir             │ '/root/projects/data-juicer/demos/data'
│
```

```
├─────────────────────────┼────────────────────────────────────────────────
───────────────────────────────────────┤
│ add_suffix              │ False
│
╘═════════════════════════╧════════════════════════════════════════════════
═══════════════════════════════════════╛
2024-08-09 10:31:13 | INFO     | data_juicer.core.sandbox.hooks:42 - Parsi
ng other configs in the job.
2024-08-09 10:31:13 | INFO     | data_juicer.utils.constant:61 - Begin to
track the usage of ops with a dummy data sample
2024-08-09 10:31:14 | WARNING  | data_juicer.config.config:405 - Cache man
agement of datasets is disabled.
2024-08-09 10:31:14 | WARNING  | data_juicer.config.config:416 - Set temp
directory to store temp files to [None].
2024-08-09 10:31:15 | INFO     | data_juicer.config.config:618 - Back up t
he input config file [/tmp/job_dj_config.json] into the work_dir [/root/pr
ojects/kdd_tutorial_notebooks/outputs/sandbox-process]
2024-08-09 10:31:15 | INFO     | data_juicer.config.config:640 - Configura
tion table:
╒═════════════════════════╤════════════════════════════════════════════════
═══════════════════════════════════════╕
│ key                     │ values
│
╞═════════════════════════╪════════════════════════════════════════════════
═══════════════════════════════════════╡
│ config                  │ [Path_fr(/tmp/job_dj_config.json)]
│
├─────────────────────────┼────────────────────────────────────────────────
───────────────────────────────────────┤
│ hpo_config              │ None
│
├─────────────────────────┼────────────────────────────────────────────────
───────────────────────────────────────┤
│ data_probe_algo         │ 'uniform'
│
├─────────────────────────┼────────────────────────────────────────────────
───────────────────────────────────────┤
│ data_probe_ratio        │ 1.0
│
├─────────────────────────┼────────────────────────────────────────────────
───────────────────────────────────────┤
│ project_name            │ 'demo-process'
│
├─────────────────────────┼────────────────────────────────────────────────
───────────────────────────────────────┤
│ executor_type           │ 'default'
│
├─────────────────────────┼────────────────────────────────────────────────
───────────────────────────────────────┤
│ dataset_path            │ '/root/projects/data-juicer/demos/data/demo-da
taset.tmp.jsonl'                       │
├─────────────────────────┼────────────────────────────────────────────────
───────────────────────────────────────┤
│ export_path             │ '/root/projects/kdd_tutorial_notebooks/output
s/sandbox-process/demo-processed.jsonl' │
├─────────────────────────┼────────────────────────────────────────────────
───────────────────────────────────────┤
│ export_shard_size       │ 0
│
├─────────────────────────┼────────────────────────────────────────────────
```

```
───────────────────────────────────────┤
│ export_in_parallel      │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ keep_stats_in_res_ds    │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ keep_hashes_in_res_ds   │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ np                      │ 4
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ text_keys               │ 'text'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ image_key               │ 'images'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ image_special_token     │ '<__dj__image>'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ audio_key               │ 'audios'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ audio_special_token     │ '<__dj__audio>'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ video_key               │ 'videos'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ video_special_token     │ '<__dj__video>'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ eoc_special_token       │ '<|__dj__eoc|>'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ suffixes                │ []
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ use_cache               │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ ds_cache_dir            │ '/root/.cache/huggingface/datasets'
│
├─────────────────────────┼───────────────────────────────────────────────
```

```
───────────────────────────────────────┤
│ cache_compress          │ None
│
├─────────────────────────┼────────────────────────────────────────────────
───────────────────────────────────────┤
│ use_checkpoint          │ False
│
├─────────────────────────┼────────────────────────────────────────────────
───────────────────────────────────────┤
│ temp_dir                │ None
│
├─────────────────────────┼────────────────────────────────────────────────
───────────────────────────────────────┤
│ open_tracer             │ False
│
├─────────────────────────┼────────────────────────────────────────────────
───────────────────────────────────────┤
│ op_list_to_trace        │ []
│
├─────────────────────────┼────────────────────────────────────────────────
───────────────────────────────────────┤
│ trace_num               │ 10
│
├─────────────────────────┼────────────────────────────────────────────────
───────────────────────────────────────┤
│ op_fusion               │ False
│
├─────────────────────────┼────────────────────────────────────────────────
───────────────────────────────────────┤
│ process                 │ [{'language_id_score_filter': {'accelerator':
None,                                  │
│                         │                                'audio_key': 'a
udios',                                │
│                         │                                'cpu_required':
1,                                    │
│                         │                                'image_key': 'i
mages',                                │
│                         │                                'lang': 'zh',
│
│                         │                                'mem_required':
0,                                    │
│                         │                                'min_score': 0.
8,                                     │
│                         │                                'num_proc': 4,
│
│                         │                                'stats_export_p
ath': None,                            │
│                         │                                'text_key': 'te
xt',                                   │
│                         │                                'video_key': 'v
ideos'}}]                              │
├─────────────────────────┼────────────────────────────────────────────────
───────────────────────────────────────┤
│ percentiles             │ []
│
├─────────────────────────┼────────────────────────────────────────────────
───────────────────────────────────────┤
│ export_original_dataset │ False
│
├─────────────────────────┼────────────────────────────────────────────────
```

```
───────────────────────────────────────┤
│ save_stats_in_one_file  │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ ray_address             │ 'auto'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ debug                   │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ work_dir                │ '/root/projects/kdd_tutorial_notebooks/output
s/sandbox-process'                      │
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ timestamp               │ '20240809103114'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ dataset_dir             │ '/root/projects/data-juicer/demos/data'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ add_suffix              │ False
│
╘═════════════════════════╧═══════════════════════════════════════════════
═══════════════════════════════════════╛
2024-08-09 10:31:15 | INFO     | data_juicer.core.analyzer:42 - Setting up
data formatter...
2024-08-09 10:31:15 | INFO     | data_juicer.core.analyzer:51 - Preparing
exporter...
2024-08-09 10:31:15 | INFO     | data_juicer.core.analyzer:75 - Loading da
taset from data formatter...
Setting num_proc from 4 back to 1 for the jsonl split to disable multiproc
essing as it only contains one shard.
Generating jsonl split: 1 examples [00:00, 188.23 examples/s]
2024-08-09 10:31:16 | INFO     | data_juicer.format.formatter:185 - Unifyi
ng the input dataset formats...
2024-08-09 10:31:16 | INFO     | data_juicer.format.formatter:200 - There
are 1 sample(s) in the original dataset.
num_proc must be <= 1. Reducing num_proc to 1 for dataset of size 1.
Filter: 100%|##########| 1/1 [00:00<00:00, 267.80 examples/s]
2024-08-09 10:31:16 | INFO     | data_juicer.format.formatter:214 - 1 samp
les left after filtering empty text.
2024-08-09 10:31:16 | INFO     | data_juicer.format.mixture_formatter:137
- sampled 1 from 1
2024-08-09 10:31:16 | INFO     | data_juicer.format.mixture_formatter:143
- There are 1 in final dataset
2024-08-09 10:31:16 | INFO     | data_juicer.core.analyzer:81 - Preparing
process operators...
2024-08-09 10:31:16 | INFO     | data_juicer.utils.model_utils:103 - Loadi
ng fasttext language identification model...
Warning : `load_model` does not return WordVectorModel or SupervisedModel
any more, but a `FastText` object which is very similar.
2024-08-09 10:31:16 | INFO     | data_juicer.core.analyzer:86 - Computing
the stats of dataset...
num_proc must be <= 1. Reducing num_proc to 1 for dataset of size 1.
Adding new column for stats: 100%|##########| 1/1 [00:00<00:00, 365.74 exa
```

```
mples/s]
num_proc must be <= 1. Reducing num_proc to 1 for dataset of size 1.
language_id_score_filter_compute_stats: 100%|##########| 1/1 [00:00<00:00,
285.35 examples/s]
num_proc must be <= 1. Reducing num_proc to 1 for dataset of size 1.
language_id_score_filter_process: 100%|##########| 1/1 [00:00<00:00, 541.9
0 examples/s]
2024-08-09 10:31:16 | INFO     | data_juicer.core.data:193 - OP [language_
id_score_filter] Done in 0.015s. Left 1 samples.
2024-08-09 10:31:16 | INFO     | data_juicer.core.analyzer:101 - Exporting
dataset to disk...
2024-08-09 10:31:16 | INFO     | data_juicer.core.exporter:111 - Exporting
computed stats into a single file...
Creating json from Arrow format: 100%|##########| 1/1 [00:00<00:00, 479.95
ba/s]
2024-08-09 10:31:16 | INFO     | data_juicer.core.analyzer:113 - Applying
overall analysis on stats...
100%|##########| 2/2 [00:00<00:00, 17119.61it/s]
2024-08-09 10:31:16 | INFO     | data_juicer.core.analyzer:120 - The overa
ll analysis results are:        lang lang_score
count   1.0        1.0
unique  1.0        NaN
top      en        NaN
freq    1.0        NaN
mean    NaN   0.995377
std     NaN        NaN
min     NaN   0.995377
25%     NaN   0.995377
50%     NaN   0.995377
75%     NaN   0.995377
max     NaN   0.995377
2024-08-09 10:31:16 | INFO     | data_juicer.core.analyzer:122 - Applying
column-wise analysis on stats...
Column: 100%|##########| 2/2 [00:00<00:00, 11.79it/s]
2024-08-09 10:31:16 | INFO     | tools.hpo.execute_hpo_3sigma:51 - Begin t
o modify the recipe with 3-sigma rule
2024-08-09 10:31:16 | INFO     | tools.hpo.execute_hpo_3sigma:73 - Using 3
-sigma rule, for op language_id_score_filter, changed its para min_score=
0.8 into min_score=0.9056360617985956
2024-08-09 10:31:16 | INFO     | data_juicer.core.sandbox.hooks:33 - Parsi
ng Data-Juicer configs in the job.
2024-08-09 10:31:19 | INFO     | data_juicer.config.config:618 - Back up t
he input config file [/tmp/job_dj_config.json] into the work_dir [/root/pr
ojects/kdd_tutorial_notebooks/outputs/sandbox-process]
2024-08-09 10:31:19 | INFO     | data_juicer.config.config:640 - Configura
tion table:
╒═══════════════════════╤═══════════════════════════════════════════════════
═══════════════════════════════════════╕
│ key                   │ values
│
╞═══════════════════════╪═══════════════════════════════════════════════════
═══════════════════════════════════════╡
│ config                │ [Path_fr(/tmp/job_dj_config.json)]
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────────────────────┤
│ hpo_config            │ None
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────────────────────┤
```

```
│ data_probe_algo         │ 'uniform'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ data_probe_ratio        │ 1.0
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ project_name            │ 'demo-process'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ executor_type           │ 'default'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ dataset_path            │ '/root/projects/data-juicer/demos/data/demo-da
taset.jsonl'                           │
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ export_path             │ '/root/projects/kdd_tutorial_notebooks/output
s/sandbox-process/demo-processed.jsonl' │
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ export_shard_size       │ 0
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ export_in_parallel      │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ keep_stats_in_res_ds    │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ keep_hashes_in_res_ds   │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ np                      │ 4
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ text_keys               │ 'text'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ image_key               │ 'images'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ image_special_token     │ '<__dj__image>'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ audio_key               │ 'audios'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
```

```
│ audio_special_token     │ '<__dj__audio>'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ video_key               │ 'videos'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ video_special_token     │ '<__dj__video>'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ eoc_special_token       │ '<|__dj__eoc|>'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ suffixes                │ []
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ use_cache               │ True
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ ds_cache_dir            │ '/root/.cache/huggingface/datasets'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ cache_compress          │ None
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ use_checkpoint          │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ temp_dir                │ None
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ open_tracer             │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ op_list_to_trace        │ []
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ trace_num               │ 10
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ op_fusion               │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ process                 │ [{'language_id_score_filter': {'accelerator':
None,                                  │
│                         │                               'audio_key': 'a
udios',                                │
```

```
│                          │                                 'cpu_required':
1,                                    │
│                          │                                 'image_key': 'i
mages',                               │
│                          │                                 'lang': 'zh',
│
│                          │                                 'mem_required':
0,                                    │
│                          │                                 'min_score': 0.
9056360617985956,                     │
│                          │                                 'num_proc': 4,
│
│                          │                                 'stats_export_p
ath': None,                           │
│                          │                                 'text_key': 'te
xt',                                  │
│                          │                                 'video_key': 'v
ideos'}}]                             │
├──────────────────────────┼──────────────────────────────────────────────────
──────────────────────────────────────┤
│ percentiles              │ []
│
├──────────────────────────┼──────────────────────────────────────────────────
──────────────────────────────────────┤
│ export_original_dataset │ False
│
├──────────────────────────┼──────────────────────────────────────────────────
──────────────────────────────────────┤
│ save_stats_in_one_file   │ False
│
├──────────────────────────┼──────────────────────────────────────────────────
──────────────────────────────────────┤
│ ray_address              │ 'auto'
│
├──────────────────────────┼──────────────────────────────────────────────────
──────────────────────────────────────┤
│ debug                    │ False
│
├──────────────────────────┼──────────────────────────────────────────────────
──────────────────────────────────────┤
│ work_dir                 │ '/root/projects/kdd_tutorial_notebooks/output
s/sandbox-process'                    │
├──────────────────────────┼──────────────────────────────────────────────────
──────────────────────────────────────┤
│ timestamp                │ '20240809103118'
│
├──────────────────────────┼──────────────────────────────────────────────────
──────────────────────────────────────┤
│ dataset_dir              │ '/root/projects/data-juicer/demos/data'
│
├──────────────────────────┼──────────────────────────────────────────────────
──────────────────────────────────────┤
│ add_suffix               │ False
│
╘══════════════════════════╪══════════════════════════════════════════════════
══════════════════════════════════════╛
2024-08-09 10:31:19 | INFO     | data_juicer.core.executor:47 - Using cach
e compression method: [None]
2024-08-09 10:31:19 | INFO     | data_juicer.core.executor:52 - Setting up
data formatter...
```

```
2024-08-09 10:31:19 | INFO     | data_juicer.core.executor:74 - Preparing
exporter...
2024-08-09 10:31:19 | INFO     | data_juicer.core.sandbox.hooks:169 - Begi
n to process the data with given dj recipe
2024-08-09 10:31:19 | INFO     | data_juicer.core.executor:151 - Loading d
ataset from data formatter...
2024-08-09 10:31:19 | INFO     | data_juicer.format.formatter:185 - Unifyi
ng the input dataset formats...
2024-08-09 10:31:19 | INFO     | data_juicer.format.formatter:200 - There
are 6 sample(s) in the original dataset.
Filter (num_proc=4): 100%|##########| 6/6 [00:00<00:00, 64.86 examples/s]
2024-08-09 10:31:20 | INFO     | data_juicer.format.formatter:214 - 6 samp
les left after filtering empty text.
2024-08-09 10:31:20 | INFO     | data_juicer.format.mixture_formatter:137
- sampled 6 from 6
2024-08-09 10:31:20 | INFO     | data_juicer.format.mixture_formatter:143
- There are 6 in final dataset
2024-08-09 10:31:20 | INFO     | data_juicer.core.executor:157 - Preparing
process operators...
2024-08-09 10:31:20 | INFO     | data_juicer.utils.model_utils:103 - Loadi
ng fasttext language identification model...
Warning : `load_model` does not return WordVectorModel or SupervisedModel
any more, but a `FastText` object which is very similar.
2024-08-09 10:31:20 | INFO     | data_juicer.core.executor:164 - Processin
g data...
Adding new column for stats (num_proc=4): 100%|##########| 6/6 [00:00<00:0
0, 69.22 examples/s]
language_id_score_filter_compute_stats (num_proc=4):   0%|          | 0/6
[00:00<?, ? examples/s]2024-08-09 10:31:20 | INFO     | data_juicer.utils.
model_utils:103 - Loading fasttext language identification model...
Warning : `load_model` does not return WordVectorModel or SupervisedModel
any more, but a `FastText` object which is very similar.
2024-08-09 10:31:20 | INFO     | data_juicer.utils.model_utils:103 - Loadi
ng fasttext language identification model...
Warning : `load_model` does not return WordVectorModel or SupervisedModel
any more, but a `FastText` object which is very similar.
2024-08-09 10:31:20 | INFO     | data_juicer.utils.model_utils:103 - Loadi
ng fasttext language identification model...
Warning : `load_model` does not return WordVectorModel or SupervisedModel
any more, but a `FastText` object which is very similar.
2024-08-09 10:31:20 | INFO     | data_juicer.utils.model_utils:103 - Loadi
ng fasttext language identification model...
Warning : `load_model` does not return WordVectorModel or SupervisedModel
any more, but a `FastText` object which is very similar.
language_id_score_filter_compute_stats (num_proc=4): 100%|##########| 6/6
[00:00<00:00, 36.78 examples/s]
language_id_score_filter_process (num_proc=4): 100%|##########| 6/6 [00:00
<00:00, 68.73 examples/s]
2024-08-09 10:31:20 | INFO     | data_juicer.core.data:193 - OP [language_
id_score_filter] Done in 0.479s. Left 2 samples.
2024-08-09 10:31:20 | INFO     | data_juicer.core.executor:171 - All OPs a
re done in 0.479s.
2024-08-09 10:31:20 | INFO     | data_juicer.core.executor:174 - Exporting
dataset to disk...
2024-08-09 10:31:20 | INFO     | data_juicer.core.exporter:111 - Exporting
computed stats into a single file...
Creating json from Arrow format: 100%|##########| 1/1 [00:00<00:00, 405.60
ba/s]
2024-08-09 10:31:20 | INFO     | data_juicer.core.exporter:140 - Export da
taset into a single file...
```

```
Creating json from Arrow format: 100%|##########| 1/1 [00:00<00:00, 882.64
ba/s]
2024-08-09 10:31:21 | WARNING  | data_juicer.config.config:381 - dataset_p
ath is empty by default.
2024-08-09 10:31:22 | INFO     | data_juicer.config.config:618 - Back up t
he input config file [/tmp/job_dj_config.json] into the work_dir [/root/pr
ojects/kdd_tutorial_notebooks/outputs]
2024-08-09 10:31:22 | INFO     | data_juicer.config.config:640 - Configura
tion table:
╒═══════════════════════╤═══════════════════════════════════════════════════
═══════════════════════╕
│ key                   │ values
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────┤
│ config                │ [Path_fr(/tmp/job_dj_config.json)]
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────┤
│ hpo_config            │ None
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────┤
│ data_probe_algo       │ 'uniform'
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────┤
│ data_probe_ratio      │ 1.0
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────┤
│ project_name          │ 'hello_world'
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────┤
│ executor_type         │ 'default'
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────┤
│ dataset_path          │ ''
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────┤
│ export_path           │ '/root/projects/kdd_tutorial_notebooks/output
s/hello_world.jsonl' │
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────┤
│ export_shard_size     │ 0
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────┤
│ export_in_parallel    │ False
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────┤
│ keep_stats_in_res_ds  │ False
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────┤
│ keep_hashes_in_res_ds │ False
```

```
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ np                      │ 4
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ text_keys               │ 'text'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ image_key               │ 'images'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ image_special_token     │ '<__dj__image>'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ audio_key               │ 'audios'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ audio_special_token     │ '<__dj__audio>'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ video_key               │ 'videos'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ video_special_token     │ '<__dj__video>'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ eoc_special_token       │ '<|__dj__eoc|>'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ suffixes                │ []
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ use_cache               │ True
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ ds_cache_dir            │ '/root/.cache/huggingface/datasets'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ cache_compress          │ None
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ use_checkpoint          │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ temp_dir                │ None
```

```
│
├─────────────────────────┼─────────────────────────────────────────────────
────────────────────┤
│ open_tracer             │ False
│
├─────────────────────────┼─────────────────────────────────────────────────
────────────────────┤
│ op_list_to_trace        │ []
│
├─────────────────────────┼─────────────────────────────────────────────────
────────────────────┤
│ trace_num               │ 10
│
├─────────────────────────┼─────────────────────────────────────────────────
────────────────────┤
│ op_fusion               │ False
│
├─────────────────────────┼─────────────────────────────────────────────────
────────────────────┤
│ process                 │ []
│
├─────────────────────────┼─────────────────────────────────────────────────
────────────────────┤
│ percentiles             │ []
│
├─────────────────────────┼─────────────────────────────────────────────────
────────────────────┤
│ export_original_dataset │ False
│
├─────────────────────────┼─────────────────────────────────────────────────
────────────────────┤
│ save_stats_in_one_file  │ False
│
├─────────────────────────┼─────────────────────────────────────────────────
────────────────────┤
│ ray_address             │ 'auto'
│
├─────────────────────────┼─────────────────────────────────────────────────
────────────────────┤
│ debug                   │ False
│
├─────────────────────────┼─────────────────────────────────────────────────
────────────────────┤
│ work_dir                │ '/root/projects/kdd_tutorial_notebooks/output
s'                    │
├─────────────────────────┼─────────────────────────────────────────────────
────────────────────┤
│ timestamp               │ '20240809103121'
│
├─────────────────────────┼─────────────────────────────────────────────────
────────────────────┤
│ dataset_dir             │ ''
│
├─────────────────────────┼─────────────────────────────────────────────────
────────────────────┤
│ add_suffix              │ False
│
╘═════════════════════════╧═════════════════════════════════════════════════
════════════════════╛
2024-08-09 10:31:22 | INFO     | data_juicer.core.sandbox.hooks:42 - Parsi
```

```
ng other configs in the job.
2024-08-09 10:31:22 | INFO     | data_juicer.core.sandbox.hooks:191 - Begi
n to train the model with given model config
2024-08-09 10:31:41 | INFO     | data_juicer.core.sandbox.hooks:33 - Parsi
ng Data-Juicer configs in the job.
2024-08-09 10:31:43 | INFO     | data_juicer.config.config:618 - Back up t
he input config file [/tmp/job_dj_config.json] into the work_dir [/root/pr
ojects/kdd_tutorial_notebooks/outputs/sandbox-process]
2024-08-09 10:31:43 | INFO     | data_juicer.config.config:640 - Configura
tion table:
╒═══════════════════════╤═══════════════════════════════════════════════════
═══════════════════════════════════════╕
│ key                   │ values
│
╞═══════════════════════╪═══════════════════════════════════════════════════
═══════════════════════════════════════╡
│ config                │ [Path_fr(/tmp/job_dj_config.json)]
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────────────────────┤
│ hpo_config            │ None
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────────────────────┤
│ data_probe_algo       │ 'uniform'
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────────────────────┤
│ data_probe_ratio      │ 1.0
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────────────────────┤
│ project_name          │ 'demo-process'
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────────────────────┤
│ executor_type         │ 'default'
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────────────────────┤
│ dataset_path          │ '/root/projects/data-juicer/demos/data/demo-da
taset.jsonl'                          │
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────────────────────┤
│ export_path           │ '/root/projects/kdd_tutorial_notebooks/output
s/sandbox-process/demo-processed.jsonl' │
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────────────────────┤
│ export_shard_size     │ 0
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────────────────────┤
│ export_in_parallel    │ False
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────────────────────┤
│ keep_stats_in_res_ds  │ False
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────────────────────┤
```

```
│ keep_hashes_in_res_ds   │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ np                      │ 4
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ text_keys               │ 'text'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ image_key               │ 'images'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ image_special_token     │ '<__dj__image>'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ audio_key               │ 'audios'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ audio_special_token     │ '<__dj__audio>'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ video_key               │ 'videos'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ video_special_token     │ '<__dj__video>'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ eoc_special_token       │ '<|__dj__eoc|>'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ suffixes                │ []
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ use_cache               │ True
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ ds_cache_dir            │ '/root/.cache/huggingface/datasets'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ cache_compress          │ None
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ use_checkpoint          │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
```

```
│ temp_dir                │ None
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ open_tracer             │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ op_list_to_trace        │ []
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ trace_num               │ 10
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ op_fusion               │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ process                 │ [{'language_id_score_filter': {'accelerator':
None,                                  │
│                         │                                'audio_key': 'a
udios',                                │
│                         │                                'cpu_required':
1,                                    │
│                         │                                'image_key': 'i
mages',                                │
│                         │                                'lang': 'zh',
│
│                         │                                'mem_required':
0,                                    │
│                         │                                'min_score': 0.
8,                                     │
│                         │                                'num_proc': 4,
│
│                         │                                'stats_export_p
ath': None,                            │
│                         │                                'text_key': 'te
xt',                                   │
│                         │                                'video_key': 'v
ideos'}}]                              │
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ percentiles             │ []
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ export_original_dataset │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ save_stats_in_one_file  │ False
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
│ ray_address             │ 'auto'
│
├─────────────────────────┼───────────────────────────────────────────────
───────────────────────────────────────┤
```

```
│ debug                    │ False
│
├──────────────────────────┼──────────────────────────────────────────────────────────────────────────────────────────────────────────┤
│ work_dir                 │ '/root/projects/kdd_tutorial_notebooks/outputs/sandbox-process'                                          │
├──────────────────────────┼──────────────────────────────────────────────────────────────────────────────────────────────────────────┤
│ timestamp                │ '20240809103142'
│
├──────────────────────────┼──────────────────────────────────────────────────────────────────────────────────────────────────────────┤
│ dataset_dir              │ '/root/projects/data-juicer/demos/data'
│
├──────────────────────────┼──────────────────────────────────────────────────────────────────────────────────────────────────────────┤
│ add_suffix               │ False
│
╞══════════════════════════╪══════════════════════════════════════════════════════════════════════════════════════════════════════════╡
2024-08-09 10:31:43 | INFO     | data_juicer.core.analyzer:37 - Using cache compression method: [None]
2024-08-09 10:31:43 | INFO     | data_juicer.core.analyzer:42 - Setting up data formatter...
2024-08-09 10:31:43 | INFO     | data_juicer.core.analyzer:51 - Preparing exporter...
2024-08-09 10:31:43 | INFO     | data_juicer.core.sandbox.hooks:63 - Begin to analyze data
2024-08-09 10:31:43 | INFO     | data_juicer.core.analyzer:75 - Loading dataset from data formatter...
2024-08-09 10:31:44 | INFO     | data_juicer.format.formatter:185 - Unifying the input dataset formats...
2024-08-09 10:31:44 | INFO     | data_juicer.format.formatter:200 - There are 6 sample(s) in the original dataset.
Filter (num_proc=4): 100%|##########| 6/6 [00:00<00:00, 60.25 examples/s]
2024-08-09 10:31:44 | INFO     | data_juicer.format.formatter:214 - 6 samples left after filtering empty text.
2024-08-09 10:31:44 | INFO     | data_juicer.format.mixture_formatter:137 - sampled 6 from 6
2024-08-09 10:31:44 | INFO     | data_juicer.format.mixture_formatter:143 - There are 6 in final dataset
2024-08-09 10:31:44 | INFO     | data_juicer.core.analyzer:81 - Preparing process operators...
2024-08-09 10:31:44 | INFO     | data_juicer.utils.model_utils:103 - Loading fasttext language identification model...
Warning : `load_model` does not return WordVectorModel or SupervisedModel any more, but a `FastText` object which is very similar.
2024-08-09 10:31:44 | INFO     | data_juicer.core.analyzer:86 - Computing the stats of dataset...
Adding new column for stats (num_proc=4): 100%|##########| 6/6 [00:00<00:00, 65.15 examples/s]
language_id_score_filter_compute_stats (num_proc=4):   0%|          | 0/6 [00:00<?, ? examples/s]2024-08-09 10:31:44 | INFO     | data_juicer.utils.model_utils:103 - Loading fasttext language identification model...
Warning : `load_model` does not return WordVectorModel or SupervisedModel any more, but a `FastText` object which is very similar.
2024-08-09 10:31:44 | INFO     | data_juicer.utils.model_utils:103 - Loading fasttext language identification model...
Warning : `load_model` does not return WordVectorModel or SupervisedModel any more, but a `FastText` object which is very similar.
```

```
2024-08-09 10:31:44 | INFO     | data_juicer.utils.model_utils:103 - Loadi
ng fasttext language identification model...
Warning : `load_model` does not return WordVectorModel or SupervisedModel
any more, but a `FastText` object which is very similar.
2024-08-09 10:31:44 | INFO     | data_juicer.utils.model_utils:103 - Loadi
ng fasttext language identification model...
Warning : `load_model` does not return WordVectorModel or SupervisedModel
any more, but a `FastText` object which is very similar.
language_id_score_filter_compute_stats (num_proc=4): 100%|##########| 6/6
[00:00<00:00, 35.43 examples/s]
language_id_score_filter_process (num_proc=4): 100%|##########| 6/6 [00:00
<00:00, 64.98 examples/s]
2024-08-09 10:31:45 | INFO     | data_juicer.core.data:193 - OP [language_
id_score_filter] Done in 0.536s. Left 6 samples.
2024-08-09 10:31:45 | INFO     | data_juicer.core.analyzer:101 - Exporting
dataset to disk...
2024-08-09 10:31:45 | INFO     | data_juicer.core.exporter:111 - Exporting
computed stats into a single file...
Creating json from Arrow format: 100%|##########| 1/1 [00:00<00:00, 393.43
ba/s]
2024-08-09 10:31:45 | INFO     | data_juicer.core.analyzer:113 - Applying
overall analysis on stats...
100%|##########| 2/2 [00:00<00:00, 16878.49it/s]
2024-08-09 10:31:45 | INFO     | data_juicer.core.analyzer:120 - The overa
ll analysis results are:        lang lang_score
count   6.0        6.0
unique  3.0        NaN
top      en        NaN
freq    3.0        NaN
mean    NaN    0.97077
std     NaN   0.021711
min     NaN   0.945098
25%     NaN   0.958538
50%     NaN   0.964044
75%     NaN   0.987662
max     NaN   0.999194
2024-08-09 10:31:45 | INFO     | data_juicer.core.analyzer:122 - Applying
column-wise analysis on stats...
Column: 100%|##########| 2/2 [00:00<00:00,  7.44it/s]
2024-08-09 10:31:47 | WARNING  | data_juicer.config.config:381 - dataset_p
ath is empty by default.
2024-08-09 10:31:47 | INFO     | data_juicer.config.config:618 - Back up t
he input config file [/tmp/job_dj_config.json] into the work_dir [/root/pr
ojects/kdd_tutorial_notebooks/outputs]
2024-08-09 10:31:47 | INFO     | data_juicer.config.config:640 - Configura
tion table:
╒═══════════════════════╤═══════════════════════════════════════════════════
═══════════════════════╕
│ key                   │ values
│
╞═══════════════════════╪═══════════════════════════════════════════════════
═══════════════════════╡
│ config                │ [Path_fr(/tmp/job_dj_config.json)]
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────┤
│ hpo_config            │ None
│
├───────────────────────┼───────────────────────────────────────────────────
───────────────────────┤
```

```
│ data_probe_algo         │ 'uniform'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ data_probe_ratio        │ 1.0
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ project_name            │ 'hello_world'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ executor_type           │ 'default'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ dataset_path            │ ''
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ export_path             │ '/root/projects/kdd_tutorial_notebooks/output
s/hello_world.jsonl' │
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ export_shard_size       │ 0
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ export_in_parallel      │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ keep_stats_in_res_ds    │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ keep_hashes_in_res_ds   │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ np                      │ 4
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ text_keys               │ 'text'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ image_key               │ 'images'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ image_special_token     │ '<__dj__image>'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ audio_key               │ 'audios'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
```

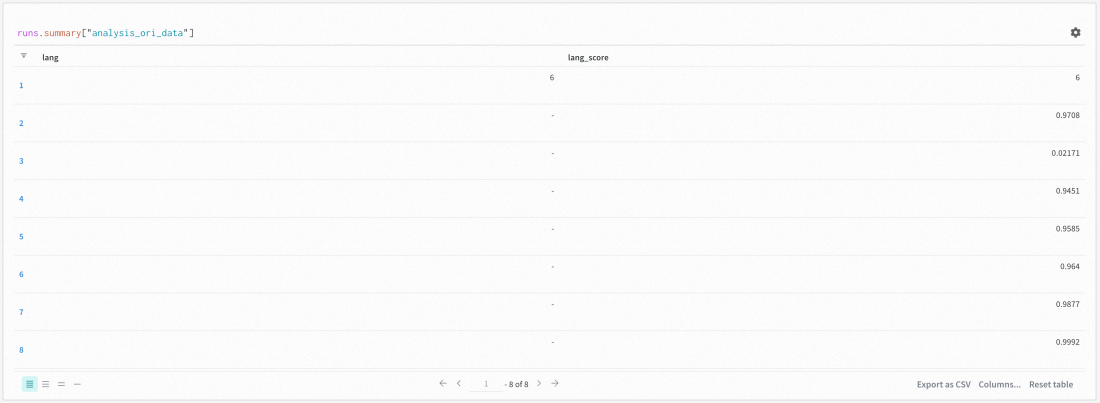
```
│ audio_special_token     │ '<__dj__audio>'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ video_key               │ 'videos'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ video_special_token     │ '<__dj__video>'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ eoc_special_token       │ '<|__dj__eoc|>'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ suffixes                │ []
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ use_cache               │ True
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ ds_cache_dir            │ '/root/.cache/huggingface/datasets'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ cache_compress          │ None
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ use_checkpoint          │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ temp_dir                │ None
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ open_tracer             │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ op_list_to_trace        │ []
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ trace_num               │ 10
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ op_fusion               │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ process                 │ []
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
```

```
│ percentiles             │ []
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ export_original_dataset │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ save_stats_in_one_file  │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ ray_address             │ 'auto'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ debug                   │ False
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ work_dir                │ '/root/projects/kdd_tutorial_notebooks/output
s'                  │
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ timestamp               │ '20240809103147'
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ dataset_dir             │ ''
│
├─────────────────────────┼───────────────────────────────────────────────
────────────────────┤
│ add_suffix              │ False
│
╘═════════════════════════╧═══════════════════════════════════════════════
════════════════════╛
2024-08-09 10:31:47 | INFO     | data_juicer.core.sandbox.hooks:42 - Parsi
ng other configs in the job.
2024-08-09 10:31:47 | INFO     | data_juicer.core.sandbox.hooks:237 - Begi
n to evaluate the data with given evaluator config
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setL
ogLevel(newLevel).
24/08/09 10:31:49 WARN NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
2024-08-09 10:31:49 | INFO     | tools.quality_classifier.qc_utils:44 - Sp
ark initialization done.
2024-08-09 10:31:49 | INFO     | tools.quality_classifier.qc_utils:64 - Pr
eparing scorer model in [/root/.cache/data_juicer/models/gpt3_quality_mode
l]...
2024-08-09 10:31:53 | INFO     | tools.quality_classifier.qc_utils:92 - Lo
ading dataset from [./outputs/sandbox-process/demo-processed.jsonl]...
2024-08-09 10:31:54 | INFO     | tools.quality_classifier.qc_utils:284 - S
tart scoring dataset...
2024-08-09 10:31:54 | INFO     | tools.quality_classifier.qc_utils:160 - E
xporting predicted result to [./outputs/sandbox-process/demo-processed.jso
nl.tmp_res.jsonl]
24/08/09 10:31:54 WARN DAGScheduler: Broadcasting large task binary with s
ize 2.2 MiB
24/08/09 10:31:55 WARN DAGScheduler: Broadcasting large task binary with s
```

```
ize 2.1 MiB
/root/projects/data-juicer/data_juicer/core/sandbox/evaluators.py:55: Futu
reWarning: Calling float on a single element Series is deprecated and will
raise a TypeError in the future. Use float(ser.iloc[0]) instead
  return float(overall_quality_stats.loc['mean'])
wandb: | 0.180 MB of 0.180 MB uploaded
wandb: Run history:
wandb: eval_data _
wandb:      loss ▇▁
wandb:
wandb: Run summary:
wandb: eval_data 0.90691
wandb:      loss 0.3416
wandb:
wandb: 🚀 View run sandbox-recipe-refinement-run0 at: https://wandb.ai/hyl
1024/sandbox-recipe-refinement/runs/3l1fauog
wandb: ⭐️ View project at: https://wandb.ai/hyl1024/sandbox-recipe-refinem
ent
wandb: Synced 5 W&B file(s), 2 media file(s), 2 artifact file(s) and 0 oth
er file(s)
wandb: Find logs at: ./wandb/run-20240809_103105-3l1fauog/logs
wandb: WARNING The new W&B backend becomes opt-out in version 0.18.0; try
it out with `wandb.require("core")`! See https://wandb.me/wandb-core for m
ore information.
```

After it's started, sandbox would run each group of jobs successively: probing the dataset, processing the dataset, training the model, analyze the processed dataset, and so on. These tasks are automatically run in the one-trial loop.

Sandbox integrates the WandB framework to "watch" the running states in the pipeline. For example, it would records analysis results:

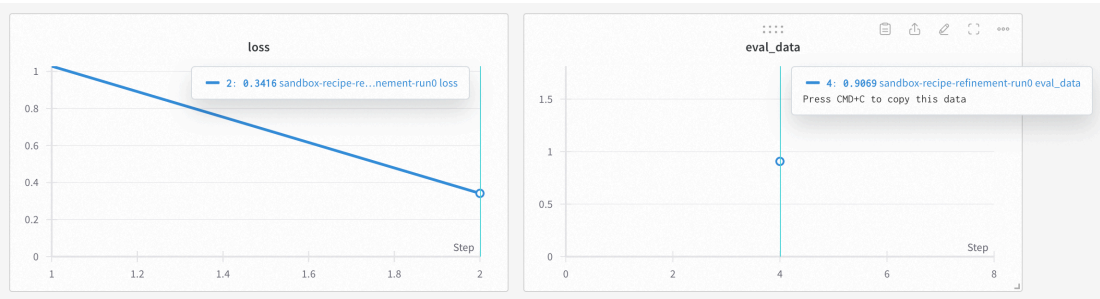


runs.summary["analysis_ori_data"]

	lang	lang_score
1	6	6
2	-	0.9708
3	-	0.02171
4	-	0.9451
5	-	0.9585
6	-	0.964
7	-	0.9877
8	-	0.9992

1 - 8 of 8 Export as CSV Columns... Reset table

training procedures & evaluation results:



and other detailed runtime information. Users could trace the whole pipeline according to these records.

Finally, we can clean up these demo config files.

In [6]:
```
!rm sandbox_pipeline.yaml
!rm dj_process.yaml
!rm gpt3_extra_train_config.yaml
!rm gpt3_data_quality_eval_config.yaml
```

Conclusion

In this notebook, we learn the basic usage of Data-Juicer sandbox, the default one-trial pipeline in sandbox and how to start to run a typical sandbox pipeline with config files of different components.